Mathematische Mustererkennung und Hidden Markov Modelle

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität
zu Kiel

vorgelegt von
Lars Willert
Kiel 2003

Referent: ..... Prof. Dr. A. Irle
Korreferent: ..... Prof. Dr. U. Rösler
Tag der mündlichen Prüfung: 11.02.2004
Zum Druck genehmigt: 11.02.2004

Der Dekan, gez. Depmeier

Für
Yvonne, Mette und Göran

Danksagung

Ich möchte meinem Betreuer, Herrn Prof. Dr. Albrecht Irle, für die Themenstellung, die zahlreichen Anregungen und die Freiheit in der Themenausarbeitung danken.

Den Mitgliedern der Fachgruppe Stochastik danke ich für die Möglichkeit, in freundlicher Atmosphäre fachliche Probleme zu diskutieren.

Bei Herrn Markus Wendt möchte ich mich für das freundschaftliche gemeinsame Verweilen in unserem Büro bedanken.

Ganz besonders danke ich meinen Kindern Mette und Göran, die viel zu oft einen am Schreibtisch sitzenden Vater mit ihrer kindlichen (Un-)Geduld ertragen mussten.

Einleitung

Zusammenfassung

Mit dem Begriff Mustererkennung wird das maschinelle Erkennen von verschiedenen Mustern bezeichnet. Das Zuordnen von Mustern in den sensorischen Signalen zu Wissen wird auch Klassifikation genannt. Lebewesen klassifizieren ständig ihre Sinneswahrnehmungen in für sie wichtige Ereignisse.

Die mathematische Mustererkennung befasst sich entsprechend mit der folgenden Situation: Es liegen eine durch den Zufall bestimmte Musterklasse vor, welche aus einer endlichen Menge stammt und nicht beobachtet werden kann, und eine Beobachtung, die von der Musterklasse abhängt. Das Paar bestehend aus der Musterklasse und der Beobachtung wird als Muster bezeichnet. Desweiteren existieren gegebenenfalls noch zusätzliche Beobachtungen, die zur Bewältigung der folgenden Aufgabe genutzt werden können. Anhand der gemachten Beobachtungen soll auf die Musterklasse zurückgeschlossen werden, d.h. es wird den Beobachtungen mittels einer Abbildung eine Musterklasse zugeordnet. Eine solche Abbildung wird Klassifikator genannt. Da jede Klassifikation, zumindest jede falsche, einen Verlust bedeuten kann, ergibt sich als eine Kenngröße für die Güte eine Klassifikators der zu erwartenden Verlust, den dieser Klassifikator liefert. Diese Kenngröße wird als Risiko des Klassifikators bezeichnet.

In dieser Arbeit wird das Musterklassifikationsmodell dahin gehend erweitert, dass eine zeitliche Komponente in das Modell eingebaut wird. Dies ermöglicht es, Situationen zu betrachten, in denen Prozesse von Mustern auftreten, bei denen die einzelnen Muster voneinander abhängig sind. Die Erweiterung des Musterklassifikationsmodell wird chronometrisches[1] Musterklassifikationsmodell genannt. Weiter wird der Spezialfall betrachtet, dass der Prozess der Muster ein Hidden-Markov-Modell ist. In dem allgemeinen Modell und in diesem Spezialfall werden die optimalen Klassifikationsrisiken untersucht.

[1] chronometrisch: auf genauer Zeitmessung beruhend

Im ersten Kapitel werden die Grundlagen der mathematischen Mustererkennung dargestellt, wobei bei der Einführung des Musterklassifikationsmodells eine Bezeichnungsweise gewählt wird, die es anschließend im zweiten Kapitel auf bequeme Weise ermöglicht, das chronometrische Musterklassifikationsmodell zu definieren. Im dritten Kapitel wird auf Hidden-Markov-Modelle und ihre Eigenschaften eingegangen.

Im zweiten Teil wird zunächst das optimale Klassifikationsrisiko im allgemeinen chronometrischen Musterklassifikationsmodell untersucht und anschließend diese Untersuchung auf den Spezialfall eines Hidden-Markov-Modells erweitert. Dabei werden die Fälle eines endlichen und eines kontinuierlichen Merkmalraumes einzeln betrachtet.

Im letzten Teil wird noch kurz auf die Berechenbarkeit von optimalen Klassifikationsrisiken eingegangen. Es werden zwei Rekursionsvorschriften hergeleitet, mit denen sich bedingte Wahrscheinlichkeiten berechnen lassen.

Abstract

Machine recognition of different patterns is termed pattern recognition. The assignation of patterns in sensory signals to knowledge is also known as classification. Organisms constantly classify their sensory perceptions to highlight this.

Mathematical pattern recognition can be demonstrated with the following situation: There are one of the coincidentally determined patterns, from a finite but unobservable set, and an observation dependent upon the pattern. The pair consisting of the pattern and the observation is called a sample. There may be further observations necessary which could be used to accomplish the following task. On the basis of these observations, the pattern should be recognized, that is the observations are assigned to a pattern with a function. Such a function is termed a classificator. Since each false classification can stand for a loss, the mean loss may be seen as a characteristic of the quality supplied by the classificator. This characteristic is termed the risk of the classificator.

In this text, the model of pattern classification is further extended, so that a temporal component may be built in the model. This makes it possible to regard processes of samples in which patterns arise as mutually dependent. The extension of the model ist termed Chronometric Model. In special cases, the sample process is a Hidden-Markov-Model. In the Chronometric Model the risk of optimal classificators are examined.

In the first and the second chapter, the bases of the mathematical pattern recognition and the definition of the Chronometric Model are represented. The third chapter concerns the characteristics of Hidden-Markov-Models.

In the second part the risk of optimal classificators in the Chronometric Model is examined, followed by further investigation of the special case of a Hidden-Markov-Model. The cases of finite an continuous sets of observation will regarded individually.

In the last part, two algorithms are given to compute the risk of an optimal classificator.

Bezeichnungen

Im Folgenden wird ein Reihe von Bezeichnungen zusammengestellt, die im laufenden Text ohne weitere Erklärungen verwendet werden.

- Ist M eine Menge und sind $m, n \in \mathbb{N}, m \leq n$, so bezeichne $x_m^n \in M^{n-m+1}$ den Vektor $(x_m, \ldots, x_n)$. Die Schreibweise

$$y_m^n = x_m^n \tag{1}$$

  wird als Abkürzung für $y_i = x_i$ für alle $i \in \{m, \ldots, n\}$ dienen.

- Ist für $m, n \in \mathbb{N}$ mit $m < n$ $(M_i)_{i=m}^n$ eine Folge von Mengen, so wird in Analogie zu (1)

$$x_m^n \in M_m^n$$

  anstelle von $x_i \in M_i$ für alle $i \in \{m, \ldots, n\}$ geschrieben.

- Die von dem Prozess $(X_t, Y_t)_{t \in \mathbb{Z}}$ erzeugten σ-Algebren werden wie folgt bezeichnet:

$$\mathcal{F}_s^t := \sigma(X_s^t, Y_s^t) \text{ für alle } s, t \in \mathbb{Z}, s \leq t\,.$$

$$\mathcal{F}_s^\infty := \sigma(X_s^t, Y_s^t, t \in \mathbb{Z}_{\geq s}) \text{ für alle } s \in \mathbb{Z}\,.$$

$$\mathcal{F}_{-\infty}^t := \sigma(X_s^t, Y_s^t, s \in \mathbb{Z}_{\leq t}) \text{ für alle } t \in \mathbb{Z}\,.$$

- Seien X und Y Zufallsvariablen. Dann wird die bedingte Verteilung von X gegeben $Y = y$ mit $P^{X|Y=y}$ bezeichnet.

- Liegen entsprechende Zufallsvariablen vor, so wird kurz $P(x|y, z)$ anstelle von $P(X = x|Y = y, Z = z)$ geschrieben.

Inhaltsverzeichnis

Teil I

Kapitel 1

Mathematische Mustererkennung

In der mathematischen Mustererkennung wird die folgende Problemstellung behandelt.

Es liegen beobachtete Daten vor, mittels derer eine diskrete Entscheidung getroffen werden soll. Dabei werden die Daten als **Merkmal**, der Vorgang des Entscheidens als **Musterklassifikation** (bzw. kurz als **Klassifikation**) und die möglichen Resultate der Entscheidung als **Musterklassen** bezeichnet. Ein Paar, bestehend aus einem Merkmal und der zugehörigen Musterklasse, wird **Muster** genannt. Als Beispiel kann die automatische Schrifterkennung genannt werden, bei der der Computer die Aufgabe hat, digitalisierte, per Hand geschriebene Buchstaben zu erkennen. Ein weiteres Beispiel findet sich in der Wettervorhersage, bei der aus den gemessenen Wetterdaten wie Luftdruck, Temperatur usw. eine Vorhersage z.B. in die Kategorien *schön, wechselhaft* und *regnerisch* denkbar wäre.

Die Abbildung 1.1 stellt den Vorgang einer Klassifikation dar. Es wird dabei angenommen, dass es möglich ist, einem Merkmal seine wahre Musterklasse durch eine **ideale Klassifikation** zuzuordnen.

Im Beispiel des handschriftlich arbeitenden Autoren ist das Schreiben an sich als realer Vorgang anzusehen, aus dem ein Schriftstück als weiterverarbeitbare Darstellung resultiert. Aus dieser Darstellung muss z.B. mit einem Scanner und einem Computerprogramm das Merkmal, welches für den Klassifikationsvorgang benutzt wird, extrahiert werden. Die ideale Klassifikation bleibt dem Autoren überlassen, der hoffentlich seine eigene Handschrift lesen kann.

Wird das Auftreten eines Musters als ein zufälliger Vorgang modelliert, so kann dieser Klassifikationsmechanismus durch das folgende Modell beschrieben werden.

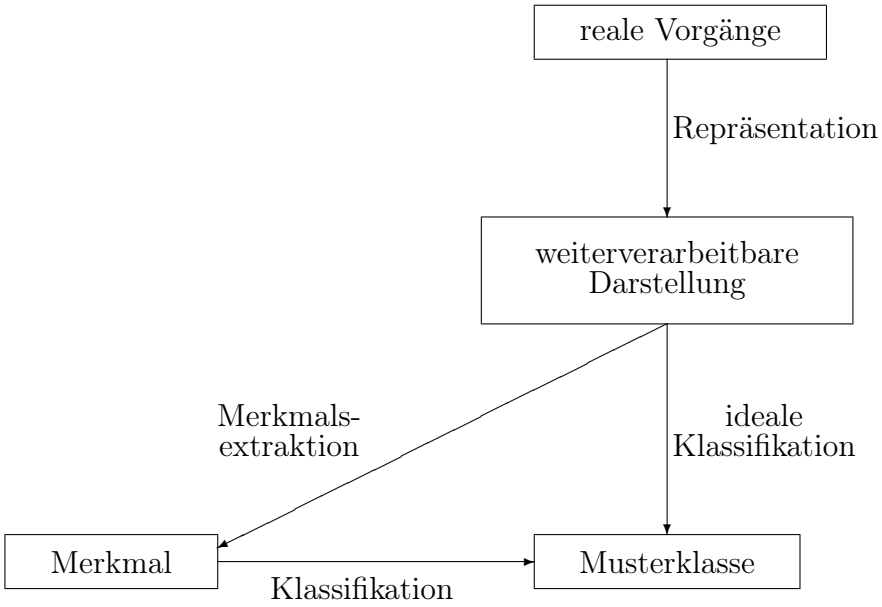

Abbildung 1.1: Vorgang einer Klassifikation in der mathematischen Mustererkennung

1.1 Das Musterklassifikationsmodell

1.1.1 Definition

Gegeben seien ein nicht spezifizierter Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und eine Zufallsvariable

$$(X, Y) : \Omega \to M \times \mathcal{Y} \ ,$$

die ein Muster modelliert, d.h., dem beobachteten Merkmal Y soll die zugehörige Musterklasse X zugeordnet werden. Dabei repräsentiert $M = \{1, \ldots, m\}$ die Menge der möglichen Musterklassen und $\mathcal{Y}$ die Menge der möglichen Merkmale.

Weiter sei eine Zufallsvariable $Z : \Omega \to \mathcal{Z}$ gegeben, die einen Informationsanteil beschreibt, den man aus zusätzlichen Beobachtungen heranziehen will. Häufig wird hier eine **Lernfolge** benutzt, d.h. weitere Muster

$$(X_1, Y_1), \ldots, (X_n, Y_n) : \Omega \to M \times \mathcal{Y},$$

welche stochastisch unabhängig von (X, Y) sind, die gleiche Verteilung wie (X, Y) besitzen und von denen die Merkmale und auch die Musterklassen bekannt sind.

Für den Klassifikationsvorgang wird eine messbare Abbildung

$$\delta : \mathcal{Y} \times \mathcal{Z} \to M$$

benutzt, die als **Klassifikator** bezeichnet wird. Geschieht die Klassifikation ohne die zusätzlichen Information Z, so liegt eine messbare Abbildung

$$\delta : \mathcal{Y} \to M$$

vor. Ein solcher Klassifikator wird **kontextfrei** genannt.

Um die Güte eines Klassifikators beschreiben zu können, wird zunächst eine Verlustfunktion eingeführt. Mit dieser kann dann die Güte in Form des Risikos eines Klassifikators, d.h. in Form des zu erwartenden Verlustes, angegeben werden.

1.1.2 Definition
Eine Abbildung

$$L : M \times M \to [0, \infty)$$

mit

$$L(i,j) = \begin{cases} 0 & , i = j \\ c_{i,j} & , i \neq j \end{cases},$$

$c_{i,j} > 0$, heißt **Verlustfunktion**. $L(i,j)$ gibt den Verlust an, der bei Vorliegen der wahren Musterklasse i und Klassifikationsentscheidung j auftritt.

Das **Risiko** eines Klassifikators δ wird dann definiert durch

$$R(\delta, P) = E_P \, L(X, \delta(Y, Z)) \,. \tag{1.1}$$

Einem kontextfreien Klassifikator wird natürlich das Risiko

$$R(\delta, P) = E_P \, L(X, \delta(Y))$$

zugeordnet.

1.1.3 Bemerkung
Ein naheliegendes Ziel ist es, einen Klassifikator zu finden, der ein möglichst geringes Klassifikationsrisiko besitzt.

Dabei ist zu beachten, dass das gesamte Problem der Musterklassifikation offensichtlich stark von der Verteilung von (X, Y, Z) bzw. (X, Y) bezüglich des Wahrscheinlichkeitsmaßes P geprägt wird. Folglich ist das Risiko $R(\delta, P)$ eines Klassifikators δ von dieser Verteilung abhängig.

Im Allgemeinen ist diese Verteilung jedoch unbekannt, und es werden Informationen gesucht, die eine direkte oder indirekte Schätzung dieser Verteilung zulassen. Wie oben schon erwähnt, wird hierzu häufig eine Lernfolge benutzt, mit deren Hilfe sich Klassifikatoren konstruieren lassen. Eine ausführliche Beschreibung solcher Verfahren findet man in [Dev, Devroye]. In [Hol, Holst] wird der k-NN Klassifikator betrachtet bei Vorliegen einer stochastisch abhängigen Lernfolge.

In diesem Text wird auf den Fall der unbekannten Verteilung nicht weiter eingegangen, sondern die Idealsituation, in der $P^{(X,Y,Z)}$ bekannt ist, betrachtet. Die Suche nach einem optimalen Klassifikator mündet dann in ein Minimierungsproblem für die reellwertige Größe $R(\delta, P)$, das mit den Methoden der Bayes'schen Statistik behandelt werden kann.

1.2 Bayes'sche Statistik

In der Bayes'schen Statistik wird davon ausgegangen, dass sämtliche unbekannten Größen durch Zufallsvariablen beschrieben werden können. Die Information, die über diese Zufallsvariablen verfügbar sind, wird durch deren Wahrscheinlichkeitsverteilung ausgedrückt.

Der Name *Bayes'sche Statistik* rührt vom englischen Priester Thomas Bayes her, der im 18. Jahrhundert unter anderem eine Abhandlung über die Wahrscheinlichkeiten von Hypothesen schrieb.

Für parametrische stochastische Modelle wird in konsequenter Weise der Parameter ϑ durch eine Zufallsvariable Θ und die Kenntnis über diesen Parameter durch seine Verteilung P^{Θ} beschrieben, die auch **a-priori-Verteilung des Parameters** genannt wird.

1.2.1 Definition
Für eine Zufallsvariable X mit der Verteilung W und dem Wertebereich $\mathcal{X}$ wird eine endliche Folge $X_1, \ldots, X_n$ von stochastisch unabhängigen Zufallsvariablen, die wie X verteilt sind, eine **Stichprobe von** X genannt, wobei eine Realisierung einer solchen Stichprobe ebenfalls als Stichprobe bezeichnet wird. Ein stochastisches Modell, bestehend aus X und W, wird im Folgenden kurz mit

$$X \sim W$$

notiert.

Im Fall parametrischer stochastischer Modelle mit unbekanntem Parameter ϑ aus einem bekannten Parameterraum Ξ wird die Notation

$$X \sim W_\vartheta, \vartheta \in \Xi$$

benutzt.

Liegen Dichten $f(\cdot|\vartheta)$ der Verteilung W_ϑ von X bezüglich eines Maßes μ für den Parameterwert ϑ vor, so wird

$$X \sim f(\cdot|\vartheta),\ \vartheta \in \Xi$$

geschrieben.

In der Behandlung eines statistischen Problems liegt also folgende Situation vor: Betrachtet werde ein stochastisches Modell $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$, eine a-priori-Verteilung von Θ und eine Stichprobe $X_1, \ldots, X_n$. Gemäß des Bayes'schen Ansatzes liefert das Zufallsexperiment in Abhängigkeit der a-priori-Verteilung einen unbekannten Parameter ϑ und eine Stichprobe. Anhand dieser Stichprobe wird mittels der Formel von Bayes eine Einschätzung des Parameters ϑ gewonnen. Dabei wird die durch die Stichprobe bedingte Verteilung von Θ **a-posteriori-Verteilung** genannt und mit $P^{\Theta|X_1^n=x_1^n}$ bezeichnet.

Das Bayes'sche Theorem

Liegt ein stochastisches Modell $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ vor, so wird eine Stichprobe $D = (x_1, \ldots, x_n)$ in folgender Art in Zusammenhang mit der Verteilung des Parameters ϑ gebracht: Aus der a-priori-Verteilung P^Θ wird die durch die Stichprobe bedingte a-posteriori-Verteilung $P^{\Theta|D}$, die mittels bedingter Wahrscheinlichkeiten bzw. mittels bedingter Dichten berechnet werden kann. Für diese Vorgehensweise ist das bekannte Bayes'sche Theorem von tragender Bedeutung. Es wird auch als Bayes'sche Formel bezeichnet.

1.2.2 Satz
In einem parametrischen stochastischen Modell $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ sei π die Dichte von P^Θ bezüglich eines Maßes ν. Dann gilt für alle $\vartheta \in \Xi$ und $x_1, \ldots, x_n \in \mathcal{X}$ mit $\varphi_X(x_1^n) = \int_\Xi \varphi(x_1^n|\theta) \cdot \pi(\theta)\, \nu(d\theta)$

$$\pi(\vartheta|x_1^n) = \frac{\varphi(x_1^n|\vartheta) \cdot \pi(\vartheta)}{\varphi_X(x_1^n)} \cdot 1_{\{\varphi_X>0\}} + \pi(\theta) \cdot 1_{\{\varphi_X=0\}}\,, \tag{1.2}$$

wobei $\pi(\cdot|x_1^n)$ die Dichte von $P^{\Theta|X_1^n=x_1^n}$ bezüglich des Maßes ν und $\varphi(\cdot|\vartheta)$ die Dichte von $P^{X_1^n|\Theta=\vartheta}$ ist.
*Dabei heißt $\pi(\cdot|x_1^n)$ **a-posteriori-Dichte** von Θ.*

1.2.3 Bemerkung
In einem diskreten parametrischen Modell liefert (1.2) gerade die wohlbekannte Formel von Bayes:

$$P(\Theta = \vartheta | X_1^n = x_1^n) = \frac{P(X_1^n = x_1^n | \Theta = \vartheta) \cdot P(\Theta = \vartheta)}{\sum_{\theta \in \Xi} P(X_1^n = x_1^n | \Theta = \theta) \cdot P(\Theta = \theta)}$$

für alle $\vartheta \in \Xi$ und $x_1, \ldots, x_n \in \mathcal{X}$ mit $P(\Theta = \vartheta) > 0$ und $P(X_1^n = x_1^n) > 0$.

Anwendungen der a-posteriori-Verteilung

Die Information, die durch eine Stichprobe in der a-posteriori-Verteilung vorhanden ist, kann auf verschiedene Arten verwendet werden, wobei für genauere Ausführungen auf die gängige Literatur verwiesen wird. Siehe auch [Con, Congdon],[Robert],[Ber-Smi,Bernado-Smith]. Hier seien nur zwei typische Anwendungen kurz skizziert.

(i) Prädiktivverteilungen:

In einem parametrischen stochastischen Modell $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ kann die a-posteriori-Dichte $\pi(\vartheta|x_1^n)$ zur Ermittlung einer Prognoseverteilung - auch *Prädiktivverteilung* genannt - verwendet werden. Dazu berechnet man zuerst die durch die Stichprobe bedingte gemeinsame Dichtefunktion g von (X, Θ) und daraus die Randdichte $f(x|x_1^n)$ von X, d.h.

$$f(x|x_1^n) = \int_\Xi g(x, \vartheta|x_1^n)\, \nu(d\vartheta) = \int_\Xi f(x|\vartheta) \cdot \pi(\vartheta|x_1^n)\, \nu(d\vartheta)\,.$$

(ii) A-posteriori-Bayes-Schätzer:

Für einen eindimensionalen Parameter ϑ eines stochastischen Modells $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ und beobachteter Stichprobe x_1^n bezeichnet man bei Existenz des Erwartungswertes bezüglich der a-posteriori-Verteilung von Θ diesen Erwartungswert als den **a-posteriori-Bayes-Schätzer** $\hat{\vartheta}$ für $E_\pi \Theta$, d.h.

$$\hat{\vartheta} := E_{x_1^n}\, \Theta = \int_\Xi \vartheta \cdot \pi(\vartheta|x_1^n)\, \nu(d\vartheta)\,.$$

Bayes'sche Entscheidungsregeln

Liegt eine Stichprobe vor, aus der der Parameter $\vartheta \in \Xi$ geschätzt werden soll, muss eine Entscheidung für ein Element aus Ξ getroffen werden. Da Entscheidungen oftmals mit Verlust verbunden sind, müssen bei statistischen Entscheidungen Verlustbetrachtungen mit einbezogen werden. Im Bayes'schen Modell können Entscheidungen unter Verwendung der a-priori-Information optimiert werden.

1.2.4 Definition
Sei $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ ein stochastisches Modell. Eine Abbildung

$$L : \Xi \times \Xi \to \mathbb{R}_{\geq 0}$$

mit $L(i,i) = 0$ für alle $i \in \Xi$ und $L(i,j) > 0$ für alle $i, j \in \Xi$ mit $i \neq j$ heißt **Verlustfunktion**.

Der Verlust, der entsteht, falls ϑ durch $\hat{\vartheta}$ geschätzt wird, wird mittels einer solchen Verlustfunktion durch

$$L(\vartheta, \hat{\vartheta})$$

beschrieben.

1.2.5 Definition
Sei $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ ein stochastisches Modell und $(X_1, \ldots, X_n)$ eine Stichprobe, wobei der Wertebreich von X mit $\mathcal{X}$ bezeichnet sei. Dann heißt eine messbare Abbildung

$$\delta : \mathcal{X}^n \to \mathcal{D}$$

Entscheidungsregel, wobei $\mathcal{D}$ die Menge aller möglichen Entscheidungen bezeichnet.

Zur Ermittlung *guter* Entscheidungsregeln zieht man den zu erwartenden Verlust

$$r(\vartheta, \delta) := E_\vartheta \, L(\vartheta, \delta(X_1^n))$$

heran. Dieser ist abhängig von ϑ und daher eine Funktion von ϑ, welche **Risikofunktion der Entscheidungsregel** δ genannt wird.

Der Vergleich zweier Entscheidungsregeln mittels ihrer Risikofunktion ist nicht unmittelbar möglich, weil für verschiedene Werte des ϑ eine Entscheidungsregel δ_1 im Vergleich mit einer Entscheidungsregel δ_2 einmal günstiger, ein anderes Mal ungünstiger erscheinen kann. In der klassischen statistischen Entscheidungsanalyse zieht man bisweilen das sogenannte *Minimax-Prinzip*

heran, um möglichst gute Entscheidungsregeln zu finden, d.h., man sucht jene Entscheidungsregel δ^*, so dass das $\sup_{\vartheta\in\Xi} r(\vartheta,\delta^*)$ minimal ist. Bei dieser Vorgehensweise wird keinerlei Information über den Parameter verwendet. Eine umfassendere Verwertung aller vorhandenen Informationen, insbesondere der Verteilung von Θ, erfolgt durch Verwendung von Bayes'schen Entscheidungsregeln.

1.2.6 Definition
Sei $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ ein stochastisches Modell und sei π die a-priori-Dichte von Θ. Dann heißt das mittlere Risiko

$$R(\pi,\delta) := \int_\Xi r(\vartheta,\delta)\cdot\pi(\vartheta)\,\nu(d\vartheta)$$

Bayes'sches Risiko des Klassifikators δ.

Eine **Bayes'sche Entscheidungsregel** δ^* ist eine solche, deren Bayes'sches Risiko unter allen in Betracht kommenden Entscheidungsregeln minimal ist, d.h.

$$R(\pi,\delta^*) = \min_\delta R(\pi,\delta)\,,$$

wobei π wieder die a-priori-Dichte von Θ ist.

Wie anfangs angeführt, scheint es natürlich zu sein, die a-posteriori-Dichte $\pi(\vartheta|x_1^n)$ für eine Stichprobe x_1^n von X heranzuziehen. Dass dieses genau zu Bayes'schen Entscheidungsregeln führt, zeigt das Korollar des folgenden Satzes.

1.2.7 Satz
In einem kontinuierlichen stochastischen Modell $X \sim f(\cdot|\vartheta), \vartheta \in \Xi$ besitze die Entscheidungsregel δ die Eigenschaft

$$\int_\Xi L(\vartheta,\delta(x_1^n))\cdot\varphi(x_1^n|\vartheta)\cdot\pi(\vartheta)\,d\vartheta = \min_{\theta\in\Xi}\int_\Xi L(\vartheta,\theta)\cdot\varphi(x_1^n|\vartheta)\cdot\pi(\vartheta)\,\nu(d\vartheta) \quad (1.3)$$

für alle Stichproben x_1^n, wobei $\varphi(\cdot|\vartheta)$ die Dichte von $P^{X_1^n|\Theta=\vartheta}$ ist.

Dann ist δ eine Bayes'sche Entscheidungsregel.

Eine Bayes'sche Entscheidungsregel δ bezüglich einer a-priori-Verteilung $\pi(\vartheta)$ lässt sich also finden, indem für jede konkrete Stichprobe x_1^n eine Entscheidung $d = \delta(x_1^n)$ gewählt wird, welche

$$\int_\Xi L(\vartheta,\hat{d})\cdot\varphi(x_1^n|\vartheta)\cdot\pi(\vartheta)\,\nu(d\vartheta),\ \hat{d}\in\Xi$$

minimiert.

1.2.8 Korollar
Unter den Voraussetzungen des Satzes entstehen Bayes'sche Entscheidungen durch Minimierung des a-posteriori zu erwartenden Verlustes

$$E_{\vartheta|x_1^n} L(\Theta, \delta(x_1^n)) = \int_\Xi L(\vartheta, \delta(x_1^n)) \cdot \pi(\vartheta|x_1^n)\, \nu(d\vartheta)\,. \tag{1.4}$$

1.2.9 Bemerkung
Gemäß des Ansatzes der Bayes'schen Statistik kann zur Ermittlung einer Bayes'schen Entscheidungsregel die Minimierungsaufgabe (1.4) gelöst werden, wobei hierbei der zu erwartende Verlust mittels der a-posteriori Verteilung bestimmt wird.

Bevor nun die Methoden der Bayes'schen Statistik auf die Situation in der Mustererkennung übertragen werden, sei noch erwähnt, dass weder für die Herleitung des Satzes 1.2.7 noch für die des Korollars 1.2.8 die stochastische Unabhängigkeit der $X_1, \ldots, X_n$ benötigt wird.

1.3 Optimale Klassifikatoren

Bezieht man die Ideen und Vorgehensweisen der Bayes'schen Statistik in die Überlegungen zur Musterklassifikation mit ein, so liegt eine ideale Situation aus der Sicht des Klassifikators vor, wenn die Verteilung $P^{(X,Y,Z)}$ bekannt ist. Die Suche nach einem optimalen Klassifikator ergibt dann ein Minimierungsproblem für die reellwertige Größe $R(\delta, P)$, welches mit den Methoden der Bayes'schen Statistik behandelt werden kann.

1.3.1 Satz
In einem Musterklassifikationsmodell möge δ^* die Eigenschaft

$$\sum_{i\in M} L(i, \delta^*(y,z)) \cdot P(i|y,z) = \min_{j\in M} \sum_{i\in M} L(i,j) \cdot P(i|y,z) \;\; \text{für } P^{(Y,Z)}\text{-fast alle } (y,z) \tag{1.5}$$

besitzen.

Dann gilt:

$$R(\delta^*, P) \le R(\delta, P) \text{ für alle Klassifikatoren } \delta.$$

Beweis:

Es gilt für alle Klassifikatoren δ

$$R(\delta, P) \;=\; E\, L(X, \delta(Y,Z))$$

$$
\begin{aligned}
&= \int_{M\times\mathcal{Y}\times\mathcal{Z}} L(x,\delta(y,z))P^{(X,Y,Z)}(dx,dy,dz)\\
&= \int_{\mathcal{Y}\times\mathcal{Z}} \sum_{i\in M} L(i,\delta(y,z))\cdot P(X=i|y,z)\,P^{(Y,Z)}(dy,dz)\\
&\geq \int_{\mathcal{Y}\times\mathcal{Z}} \min_{j\in M}\sum_{i\in M} L(i,j)\cdot P(X=i|y,z)\,P^{(Y,Z)}(dy,dz)\\
&= \int_{M\times\mathcal{Y}\times\mathcal{Z}} L(x,\delta^*(y,z))P^{(X,Y,Z)}(dx,dy,dz)\\
&= R(\delta^*,P)\,.
\end{aligned}
$$

□

1.3.2 Definition
Gemäß 1.2.6 heißt ein Klassifikator δ^* **Bayes-Klassifikator** oder **optimaler Klassifikator**, falls für alle messbaren Abbildungen $\delta : \mathcal{Y}\times\mathcal{Z}\to M$ die Ungleichung

$$R(\delta^*,P)\leq R(\delta,P)$$

erfüllt ist.

Der Term

$$R(P) := R(\delta^*,P)$$

wird **minimales Klassifikationsrisiko zu** P genannt.

1.3.3 Bemerkung
Ein Klassifikator, der die Minimierungsaufgabe (1.5) löst, ist also schon ein Bayes-Klassifikator.

Wird der Spezialfall, dass stochastische Unabhängigkeit zwischen dem zu klassifizierenden Muster (X,Y) und der Information Z vorliegt, betrachtet, so ergibt sich folgende Variante des Satzes 1.3.1.

1.3.4 Satz
In einem Musterklassifikationsmodell seien (X,Y) und Z stochastisch unabhängig. Ferner sei δ^* ein kontextfreier Klassifikator mit der Eigenschaft

$$\sum_{i\in M} L(i,\delta^*(y))\cdot P(i|y) = \min_{j\in M}\sum_{i\in M} L(i,j)\cdot P(i|y) \quad \textit{für } P^Y\textit{- fast alle } y\,.$$

Dann gilt für alle Klassifikatoren δ

$$R(\delta^*,P)\leq R(\delta,P)\,.$$

Beweis:

Mit dem Korollar A.1.2 folgt wegen der stochastischen Unabhängigkeit von (X, Y) und Z

$$P(X = i|Y, Z) = P(X = i|Y)\,,$$

womit man

$$\begin{aligned}\sum_{i\in M} L(i, \delta^*(y)) \cdot P(i|y) &= \min_{j\in M} \sum_{i\in M} L(i, j) \cdot P(i|y) \\ &= \min_{j\in M} \sum_{i\in M} L(i, j) \cdot P(i|y, z)\end{aligned}$$

für $P^{(Y,Z)}$-fast alle (y, z) erhält. Damit erfüllt δ^* die Voraussetzungen des Satzes 1.3.1, was die Behauptung beweist.

□

1.3.5 Bemerkung
In einem Musterklassifikationsmodell, bei dem das Muster (X, Y) und die Information Z stochastisch unabhängig sind, erhält man somit zu gegebenem P einen kontextfreien Bayes-Klassifikator δ^* mit dem minimalen Klassifikationsrisiko

$$R(P) = E\, L(X, \delta^*(Y))\,.$$

Im Folgenden soll das Musterklassifikationsmodell etwas eingeschränkt werden, indem der Spezialfall betrachtet wird, dass eine Fehlklassifikation, also eine Entscheidung für eine falsche Musterklasse, mit dem Zahlenwert 1 bestraft wird.

1.3.6 Definition
Eine Verlustfunktion der Form

$$L : M \times M \to \{0, 1\}$$

$$(i, j) \mapsto \begin{cases} 0 & , i = j \\ c & , i \neq j \end{cases}$$

mit $c > 0$ heißt **symmetrische Verlustfunktion**, wobei im weiteren Verlauf ohne Einschränkung $c = 1$ gesetzt wird.

Durch die Einführung der symmetrischen Verlustfunktion erhält man aus dem Satz 1.3.1 folgendes Korollar.

1.3.7 Korollar

In einem Musterklassifikationsmodell mit symmetrischer Verlustfunktion ist ein Klassifikator δ^* optimal, falls er folgende Eigenschaft besitzt:

$$P(X \neq \delta^*(y,z)|Y=y,Z=z) = \min_{i\in M} P(X \neq i|Y=y,Z=z) \;\; \textit{für} \;\; P^{(Y,Z)}\textit{-fast alle } (y,z)\,. \tag{1.6}$$

Ein optimaler Klassifikator wählt also eine Musterklasse aus, bei der die Wahrscheinlichkeit einer Fehlklassifikation bei gegebenen y und z minimal wird.

Beweis:

Sei $y \in \mathcal{Y}$ und $z \in \mathcal{Z}$. Dann gilt für einen Klassifikator δ bei Anwendung der symmetrischen Verlustfunktion

$$\begin{aligned}\sum_{i\in M} L(i,\delta(y,z))\cdot P(X=i|Y=y,Z=z) &= \sum_{i\in M\setminus\{\delta(y,z)\}} P(X=i|Y=y,Z=z)\\ &= P(X\neq\delta(y,z)|Y=y,Z=z)\end{aligned}$$

und

$$\begin{aligned}\min_{j\in M}\sum_{i\in M} L(i,j)\cdot P(X=i|Y=y,Z=z) &= \min_{j\in M}\sum_{i\in M\setminus\{j\}} P(X=i|Y=y,Z=z)\\ &= \min_{j\in M} P(X\neq j|Y=y,Z=z)\,.\end{aligned}$$

Mit dem Satz 1.3.1 folgt dann die Behauptung.

□

In (3.2) wurde das Klassifikationsrisiko eines Klassifikators δ eingeführt durch

$$R(\delta,P) = E_P\, L(X,\delta(Y,Z))\,.$$

Dieser Term vereinfacht sich durch Benutzung der symmetrischen Verlustfunktion zu

$$R(\delta,P) = P(X\neq\delta(Y,Z))\,,$$

denn es gilt

$$\begin{aligned} E_P L(X, \delta(Y,Z)) &= E_P(1_{\{X \neq \delta(Y,Z)\}}) \\ &= P(X \neq \delta(Y,Z)) \,. \end{aligned}$$

Die folgende Darstellung des minimalen Klassifikationsrisikos wird in späteren Überlegungen benutzt werden.

1.3.8 Satz
Das minimale Klassifikationsrisiko in einem Musterklassifikationsmodell mit symmetrischer Verlustfunktion ist

$$R(P) = \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M} P(X \neq i | Y = y, Z = z)\, P^{(Y,Z)}(dy, dz) \,. \tag{1.7}$$

Beweis:

Es gilt

$$\begin{aligned} R(P) &= R(\delta^*, P) \\ &= P(X \neq \delta^*(Y,Z)) \\ &= \int_{\mathcal{Y}\times\mathcal{Z}} P(X \neq \delta^*(y,z) | y, z)\, P^{(Y,Z)}(dy, dz) \\ &= \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M} P(X \neq i | Y = y, Z = z)\, P^{(Y,Z)}(dy, dz) \,, \end{aligned}$$

wobei die letzte Gleichheit wegen (1.6) gilt.

□

Zum Abschluss dieses Kapitels wird noch kurz auf das Musterklassifikationsmodell mit Dichten eingegangen.

1.4 Modell mit Dichten

Es sei nun zu der bedingten Verteilung von (Y, Z) gegeben $X = i$, $i \in M$ jeweils eine Dichte

$$f_i : \mathcal{Y} \times \mathcal{Z} \to [0, \infty)$$

bezüglich eines Maßes μ gegeben, d.h., für alle $B \in \sigma(\mathcal{Y} \times \mathcal{Z})$ gilt

$$P((Y,Z) \in B | X = i) = \int_B f_i(y,z)\, \mu(dy, dz) \,.$$

1.4.1 Lemma
Die Dichte von (Y, Z) bezüglich des Maßes μ ist bestimmt durch

$$f = \sum_{i \in M} P(X = i) \cdot f_i \,.$$

Beweis:

Es gilt für alle messbaren $B \in \sigma(\mathcal{Y} \times \mathcal{Z})$

$$\begin{aligned} P((Y,Z) \in B) &= \sum_{i \in M} P((Y,Z) \in B | X = i) \cdot P(X = i) \\ &= \sum_{i \in M} \int_B f_i(y,z)\, \mu(dy, dz) \cdot P(X = i) \\ &= \int_B \sum_{i \in M} P(X = i) \cdot f_i(y,z)\, \mu(dy, dz) \\ &= \int_B f(y,z)\, \mu(dy, dz) \,. \end{aligned}$$

□

1.4.2 Lemma
Es gilt für alle $i \in M$ und $y, z \in \mathcal{Y} \times \mathcal{Z}$ mit $f(y,z) > 0$:

$$P(X = i | Y = y, Z = z) = \frac{P(X = i) \cdot f_i(y,z)}{f(y,z)} \,.$$

Beweis:

Die Behauptung folgt direkt durch Anwendung des Bayes'schen Theorems (1.2.2), denn damit gilt für $y \in \mathcal{Y}$ und $z \in \mathcal{Z}$ mit $f(x,z) > 0$

$$\begin{aligned} P(X = i | Y = y, Z = z) &= \frac{f_i(y,z) \cdot P(X = i)}{\sum_{j \in M} f_j(y,z) \cdot P(X = j)} \\ &= \frac{f_i(y,z) \cdot P(X = i)}{f(y,z)} \,. \end{aligned}$$

□

Im Satz 1.3.8 wurde eine Darstellung des optimalen Klassifikationsrisikos unter Anwendung der symmetrischen Verlustfunktion angegeben. Es wird nun eine aus diesem Satz folgende weitere Darstellung mit Einbeziehung der bedingten Dichten gegeben.

1.4.3 Korollar

Das minimale Klassifikationsrisiko in einem Musterklassifikationsmodell mit symmetrischer Verlustfunktion ist gegeben durch

$$R(P) = 1 - \int_{\mathcal{Y}\times\mathcal{Z}} \max_{i\in M}\{P(X=i)\cdot f_i(y,z)\}\,\mu(dy,dz)\,.$$

Beweis:

Mit dem Satz 1.3.8 und Lemma 1.4.2 gilt

$$\begin{aligned}
R(P) &= \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M} P(X\neq i|Y=y,Z=z)\,P^{(Y,Z)}(dy,dz)\\
&= \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M} P(X\neq i|Y=y,Z=z)\cdot f(y,z)\,\mu(dy,dz)\\
&= \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M}\{1-P(X=i|Y=y,Z=z)\}\cdot f(y,z)\,\mu(dy,dz)\\
&= \int_{\mathcal{Y}\times\mathcal{Z}} \min_{i\in M}\{1-\frac{P(X=i)\cdot f_i(y,z)}{f(y,z)}\}\cdot f(y,z)\,\mu(dy,dz)\\
&= \int_{\mathcal{Y}\times\mathcal{Z}} f(y,z)\,\mu(dy,dz) - \int_{\mathcal{Y}\times\mathcal{Z}} \max_{i\in M}\{P(X=i)\cdot f_i(y,z)\}\,\mu(dy,dz)\\
&= 1 - \int_{\mathcal{Y}\times\mathcal{Z}} \max_{i\in M}\{P(X=i)\cdot f_i(y,z)\}\,\mu(dy,dz)\,.
\end{aligned}$$

□

Kapitel 2

Chronometrische Mustererkennung

Das in Kapitel 1 beschriebene Modell zur Musterklassifikation beinhaltet keine Komponente, die eine zeitliche Einordnung des Auftretens eines Musters angibt. Es liegen lediglich eine Musterklasse X, die Beobachtung Y und die Information Z zu einem abstrakten, unspezifizierten Zeitpunkt vor.

Dieses Modell soll nun dahin gehend erweitert werden, dass das Auftreten von Mustern und Informationen mittels Prozessen $(X_t, Y_t)_{t\in\mathbb{Z}}$ und $(Z_t)_{t\in\mathbb{Z}}$ beschrieben werden kann, wobei als Zeitparametermenge die Menge $\mathbb{Z}$ gewählt wird, eine zeitliche Einordnung der einzelnen Muster und Informationen also möglich ist. Das aus dieser Erweiterung entstehende Modell wird im Folgenden **chronometrisches**[1] **Musterklassifikationsmodell** genannt.

2.1 Das chronometrische Musterklassifikationsmodell

2.1.1 Definition

Es sei ein unspezifizierter Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ zugrunde gelegt. Ferner sei ein stochastischer Prozess

$$(X_t, Y_t)_{t\in\mathbb{Z}}$$

gegeben, wobei die Zufallsvariablen $X_t : \Omega \to M$ und $Y_t : \Omega \to \mathcal{Y}$ für alle $t \in \mathbb{Z}$ als Paar (X_t, Y_t) im Sinne des Kapitels 1 das Muster modellieren, welches

[1] chronometrisch: auf genauer Zeitmessung beruhend

zum Zeitpunkt t vorliegt. Dabei repräsentiert $M = \{1, \ldots, m\}$ die Menge der möglichen Musterklassen und $\mathcal{Y}$ die Menge der möglichen Merkmale.

Weiter sei ein stochastischer Prozess $(Z_t)_{t\in\mathbb{Z}}$ gegeben, wobei die Zufallsvariable $Z_t : \Omega \to \mathcal{Z}_t$ die Information beschreibt, die aus zusätzlichen Beobachtungen zum Zeitpunkt t zur Klassifikation herangezogen werden soll.

2.1.2 Bemerkung
In diesem allgemein gehaltenen Modell sind dem Prozess $(Z_t)_{t\in\mathbb{Z}}$ keinerlei Restriktionen auferlegt worden, so dass es die folgenden, häufig in der Literatur auftretenden Probleme umfasst:

(i) Problem der Prognose:
Bei dem Problem der Prognose soll anhand von Messungen aus der Vergangenheit und eventuell der Gegenwart auf zukünftiges Verhalten des Prozesses der Muster geschlossen werden. So kann zum Beispiel X_0 (dabei ist $t = 0$ der gegenwärtige Zeitpunkt) die zu prognostizierende Größe sein, die anhand der Messungen $Y_0, \ldots, Y_{-l}$ für ein $l \in \mathbb{N}$ ermittelt werden soll. Hierbei ist dann offensichtlich $Z_0 = (Y_{-1}, \cdots, Y_{-l})$ zu wählen.

(ii) Problem der Filterung:
Beim Problem der Filterung soll mittels einer Zeitreihe auf ein Muster zu einem bestimmten Zeitpunkt geschlossen werden. So kann zum Beispiel das Ziel sein, aus der Zeitreihe $Y_0, \ldots, Y_l, l \in \mathbb{N}$ Aussagen über den Zustand $X_k, 0 < k < l$ herauszufiltern, wobei dazu $Z_0 = (Y_0, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_l)$ gewählt werden muss.

2.1.3 Bemerkung
Die im chronometrischen Musterklassifikationsmodell vorliegende Problemstellung lässt sich nun folgendermaßen zusammenfassen:

Befindet man sich im Zeitpunkt $t \in \mathbb{Z}$, so kann X_t nicht direkt beobachtet werden, sondern nur Y_t, und man besitzt die Information Z_t. Ziel ist es nun, sich anhand des Merkmals Y_t und der Information Z_t für eine Musterklasse zu entscheiden, also die wahre Musterklasse X_t möglichst gut zu schätzen.

Für diesen Klassifikationsvorgang benötigt man zu jedem Zeitpunkt $t \in \mathbb{Z}$ eine Abbildung, die mittels des in diesem Zeipunkt beobachteten Merkmals und der vorliegenden Information eine Entscheidung für eine Musterklasse fällt.

2.1.4 Definition
Sei $t \in \mathbb{Z}$. Eine messbare Abbildung δ_t, die ein Merkmal und die Information in die Menge der Musterklassen M abbildet, heißt **Klassifikator in** t oder

kurz **Klassifikator**, d.h.

$$\delta_t : \mathcal{Y} \times \mathcal{Z}_t \to M\,.$$

Um der zeitlichen Komplexität des Modells gerecht zu werden, lassen sich natürlich Folgen $(\delta_t)_{t\in\mathbb{Z}}$ von Klassifikatoren betrachten, was allerdings im weiteren Verlauf des Textes nicht geschehen wird.

Analog zum Musterklassifikationsmodell heißt für ein $t \in \mathbb{Z}$ eine messbare Abbildung

$$\delta_t : \mathcal{Y} \to M\,,$$

kontextfreier Klassifikatore in t.

Von der Einführung einer zeitabhängigen Verlustfunktion wird in diesem Modell abgesehen, denn diese würde zu zeitabhängigen Bewertungen mittels der Klassifikationsrisiken führen, was unerwünscht ist, da später zeitinvariante Prozesse (X_t, Y_t) und (Z_t) betrachtet werden.

Es wird also analog zum Musterklassifikationsmodell definiert:

2.1.5 Definition
Eine Abbildung

$$L : M \times M \to [0, \infty)$$

mit

$$L(i,j) = \begin{cases} 0 & ,i = j \\ c_{i,j} > 0 & ,i \neq j \end{cases}$$

heißt **Verlustfunktion**. Dabei gibt $L(i, j)$ den Verlust an, der bei Vorliegen der wahren Musterklasse i und der Klassifikationsentscheidung j auftritt.

In einem Zeitpunkt t wird zu einem Klassifikator $\delta_t : \mathcal{Y} \times \mathcal{Z}_t \to M$ und dem zugrunde liegenden Wahrscheinlichkeitsmaß P das **Klassifikationsrisiko von** δ_t, also der zu erwartende Verlust, den δ_t liefert, durch

$$R_t(\delta_t, P) = E_P\, L(X_t, \delta_t(Y_t, Z_t)) \tag{2.1}$$

definiert.

2.1.6 Bemerkung
Aus Gründen der Bequemlichkeit und der besseren Lesbarkeit wird im Folgenden der gegenwärtige Zeitpunkt mit $t = 0$ beschrieben und stets der Fall betrachtet, dass Aussagen über die Musterklasse X_0 anhand der Kenntnis von Y_0 und Z_0 getroffen werden sollen. Da im Folgenden also keine Folgen von Klassifikatoren betrachtet werden, sondern lediglich Klassifikatoren δ_0,

stellt dieses offenbar keine Einschränkung dar, ermöglicht es aber, den Index t durch 0 zu ersetzen bzw. einfach wegzulassen. Ebenfalls wird im weiteren Verlauf nicht mehr darauf hingewiesen, dass der gegenwärtige Zeitpunkt $t = 0$ betrachtet wird.

2.1.7 Bemerkung

Die Aussagen über **optimale Klassifikatoren** aus dem Abschnitt 1.3 lassen sich auf das chronometrische Musterklassifikationsmodell analog übertragen, da die Zufallsgröße Z sehr allgemein eingeführt wurde, sie somit durch Z_0 ersetzt werden kann.

Kapitel 3

Hidden-Markov-Modelle

In diesem Kapitel wird das Hidden-Markov-Modell eingeführt und anschließend werden die Eigenschaften und Rechenregeln hergeleitet, die benötigt werden, um im folgenden Kapitel ein chronometrische Musterklassifikationsmodell zu untersuchen, bei dem der Prozess der Muster durch ein Hidden-Markov-Modell modelliert wird.

3.1 Das Hidden-Markov-Modell

Hier und im Folgenden sei stets ein Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ zugrunde gelegt.

3.1.1 Definition
Es seien $X = (X_t)_{t\in\mathbb{Z}}$ und $Y = (Y_t)_{t\in\mathbb{Z}}$ zwei stochastische Prozesse. Dann heißt das Paar (X, Y) **Hidden-Markov-Modell**, wenn Folgendes erfüllt ist:

(i) X ist eine zeitlich homogene Markov-Kette mit endlichem Zustandsraum $M = \{1, \ldots, m\}, m \in \mathbb{N}$.
$P = (p_{ij})_{i,j\in M}$ sei die dazugehörige Übergangsmatrix und π die als existierend angenommene invariante Verteilung.

(ii) Y ist ein stochastischer Prozess mit dem Zustandsraum $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$.
Für alle $t \in \mathbb{Z}$ und $i \in M$ bezeichne f_i die Dichte von $P^{Y_t|X_t=i}$ bezüglich eines Maßes μ, d.h., für alle $t \in \mathbb{Z}$ und $B \in \mathcal{Y}$ gilt

$$P(Y_t \in B|X_t = i) = \int_B f_i(y)\mu(dy)\,.$$

(iii) Die Y_t sind bedingt stochastisch unabhängig gegeben die Markov-Kette X, d.h., für Indexmengen $I \subseteq \mathbb{Z}$ gilt mit den Notationen $\hat{Y} = (Y_i)_{i\in I}$, $\hat{X} = (X_i)_{i\in I}$ und $x = (x_i)_{i\in I} \in M^{|I|}$

$$P^{\hat{Y}|\hat{X}=x} = \bigotimes_{i\in I} P^{Y_i|X_i=x_i} \ . \tag{3.1}$$

Punkt (iii) heißt definitionsgemäß, dass für alle endlichen Indexmengen $I \subseteq \mathbb{Z}$, $x = (x_i)_{i\in I} \in M^{|I|}$ mit $P(\hat{X} = x) > 0$ und $B = (B_i)_{i\in I} \in \mathcal{Y}^{|I|}$ gilt

$$P(\hat{Y} \in B|\hat{X} = x) = \prod_{i\in I} \int_{B_i} f_{x_i}(y)\mu(dy) \, .$$

3.1.2 Bemerkung
Bei einem Hidden-Markov-Modell (X, Y) kann X als ein zugrunde liegender Prozess interpretiert werden, welcher nicht direkt, sondern nur über den Prozess Y beobachtet werden kann. Dabei kann Y als ein Prozess angesehen werden, der durch Überlagerung des Prozesses X mit Rauschen entsteht.

3.2 Eigenschaften von Hidden-Markov-Modellen

Im Folgenden werden elementare Eigenschaften von Hidden-Markov-Modellen angegeben, die in späteren Kapitel benötigt werden.

So impliziert die bedingte stochastische Unabhängigkeit der Merkmale gegeben die Musterklassen, dass eine Zufallsgröße Y_t lediglich nur von der Zufallsgröße X_t abhängig ist.

3.2.1 Lemma
Sei (X, Y) ein Hidden-Markov-Modell.
Dann gilt für alle $I \subseteq \mathbb{Z}$ und $B_i \in \mathcal{A}_y, i \in I$:

$$P(\bigcap_{i\in I}\{Y_i \in B_i\}|X, Y_s, s \in \mathbb{Z} \setminus I) = P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_i, i \in I) \ .$$

Insbesondere ist für $t \in \mathbb{Z}$ die Zufallsgröße Y_t somit nur von X_t abhängig, d.h.

$$P(Y_t \in B|X, Y_s, s \in \mathbb{Z} \setminus \{t\}) = P(Y_t \in B|X_t) \ .$$

Beweis:

Sei $I \subseteq \mathbb{Z}$ eine endliche Teilmenge.

Ferner seien $J \subseteq I^c$ endlich, $x_i \in M$ und $B_i \in \mathcal{A}_\mathcal{Y}$ für $i \in I + J$.

Dann gilt

$$\int_{\{X_i=x_i,i\in I\}\cap\{X_j=x_j,Y_j\in B_j,j\in J\}} P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_i, i \in I)dP$$

$$\begin{aligned}
&= P(\bigcap_{i\in I}\{Y_i \in B_i\}|\bigcap_{i\in I}\{X_i = x_i\}) \cdot P(\bigcap_{i\in I}\{X_i = x_i\} \cap \bigcap_{j\in J}\{X_j = x_j, Y_j \in B_j\}) \\
&= \prod_{i\in I} P(Y_i \in B_i|X_i = x_i) \cdot P(\bigcap_{i\in I}\{Y_i \in \mathbb{R}\} \cap \bigcap_{j\in J}\{Y_j \in B_j\}| \bigcap_{i\in I+J}\{X_i = x_i\}) \\
&\qquad \cdot P(\bigcap_{i\in I+J}\{X_i = x_i\}) \\
&= \prod_{i\in I} P(Y_i \in B_i|X_i = x_i) \cdot \prod_{j\in J} P(Y_j \in B_j|X_j = x_j) \cdot P(\bigcap_{i\in I+J}\{X_i = x_i\}) \\
&= P(\bigcap_{i\in I+J}\{X_i = x_i, Y_i \in B_i\}) \\
&= \int_{\{X_i=x_i,i\in I\}\cap\{X_j=x_j,Y_j\in B_j,j\in J\}} P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_i, i \in I, X_j, Y_j \in J)dP
\end{aligned}$$

Damit gilt

$$\begin{aligned}
&\int_C P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_t, X_s, Y_s, s \in \mathbb{Z}\backslash I)dP \\
&\qquad = \int_C P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_i, i \in I)dP \qquad (3.2)
\end{aligned}$$

für alle C der Menge $\mathcal{E}_I := \{\{X_i = x_i, i \in I\} \cap \{Y_j \in B_j, X_j = x_j, j \in J\}|x_i, x_j \in M, B_i \in \mathcal{A}_\mathcal{Y}, i \in I, J \subseteq I^c \text{ endlich}\}$. $\mathcal{E}_I$ ist aber ein $\cap$-stabiles Erzeugendensystem von $\mathcal{F}_I := \sigma(X, (Y_j)_{j\in\mathbb{Z}\backslash I})$, also gilt mit $\mathcal{D}_I := \{C \in \mathcal{F}_I|C \text{ erfüllt } (3.2)\}$

$$\mathcal{F}_I = \sigma(\mathcal{E}_I) = \delta(\mathcal{E}_I),\ \mathcal{D}_I \subseteq \mathcal{F}_I \text{ und } \mathcal{E}_I \subseteq \mathcal{D}_I\,.$$

Offensichtlich ist $\mathcal{D}_I$ ein Dynkin-System, womit die Behauptung für endliches I bewiesen ist.

Für abzählbar unendliches I existiert eine Bijektion $\varphi : \mathbb{N} \to I$ und es gilt mit $A_n := \bigcap_{k\le n}\{Y_{\varphi(k)} \in B_{\varphi(k)}\}$ für alle $n \in \mathbb{N}$.

$$
\begin{aligned}
&P(\bigcap_{i\in I}\{Y_i \in B_i\}|X, Y_s, s \in \mathbb{Z}\backslash I) \\
&\qquad = P(\bigcap_{n\in\mathbb{N}}\{Y_{\varphi(n)} \in B_{\varphi(n)}\}|X, Y_s, s \in \mathbb{Z} \setminus I) \\
&\qquad = P(\bigcap_{n\in\mathbb{N}} A_n|X, Y_s, s \in \mathbb{Z} \setminus I) \\
&\qquad = \lim_{n\to\infty} P(A_n|X, Y_s, s \in \mathbb{Z} \setminus I) \\
&\qquad = \lim_{n\to\infty} P(\bigcap_{k\le n}\{Y_{\varphi(k)} \in B_{\varphi(k)}\}|X, Y_s, s \in \mathbb{Z} \setminus I) \\
&\qquad = \lim_{n\to\infty} P(\bigcap_{k\le n}\{Y_{\varphi(k)} \in B_{\varphi(k)}\}|X_i, i \in I) \\
&\qquad = \lim_{n\to\infty} P(A_n|X_i, i \in I) \\
&\qquad = P(\bigcap_{n\in\mathbb{N}} A_n|X_i, i \in I) \\
&\qquad = P(\bigcap_{n\in\mathbb{N}}\{Y_{\varphi(n)} \in B_{\varphi(n)}\}|X_i, i \in I) \\
&\qquad = P(\bigcap_{i\in I}\{Y_i \in B_i\}|X_i, i \in I)\,.
\end{aligned}
$$

□

Das folgende Lemma zeigt, dass der gekoppelte Prozess $(X_t, Y_t)_{t\in\mathbb{Z}}$ schon eine Markov-Kette ist.

3.2.2 Lemma
Sei (X, Y) ein Hidden-Markov-Modell. Definiere für alle $t \in \mathbb{Z}$ die Zufallsvariable $M_t = (X_t, Y_t)$ und damit den stochastischen Prozess $\mathcal{M} = (M_t)_{t\in\mathbb{Z}}$. Dann gilt für alle $t \in \mathbb{Z}$ und $A \in \mathcal{P}(M) \otimes \mathcal{A}_\mathcal{Y}$

$$P(M_{t+1} \in A|M_{-\infty}^t) = P(M_{t+1} \in A|X_t)\,.$$

Insbesondere ist dann $\mathcal{M}$ schon eine Markov-Kette.

Beweis:

Definiere zum Einen das $\cap$-stabile Erzeugendensystem

$$\mathcal{E} := \{A \times B : A \in \mathcal{P}(M), B \in \mathcal{A}_\mathcal{Y}\}$$

von $\sigma := \mathcal{P}(M) \otimes \mathcal{A}_\mathcal{Y}$ und zum Anderen das $\cap$-stabile Erzeugendensystem

$$\mathcal{E}_{<t} := \{C \times D : C \in \bigotimes_{i<t} \mathcal{P}(M), D \in \bigotimes_{i<t} \mathcal{A}_\mathcal{Y}\}$$

von $\sigma_{<t} := \bigotimes_{i<t} \mathcal{P}(M) \otimes \bigotimes_{i<t} \mathcal{A}_{\mathcal{Y}}$.
Dann gilt für alle $A \times B \in \mathcal{E}$ und $C \times D \in \mathcal{E}_{<t}$ mit der Notation $a_{<t} = (a_{t-1}, a_{t-2}, \ldots)$

$$\begin{aligned}
&\int_{\{X_{<t}\in C, Y_{<t}\in D\}} P(X_t \in A, Y_t \in B | X_{<t}, Y_{<t}) dP \\
&= P(X_t \in A, Y_t \in B, X_{<t} \in C, Y_{<t} \in D) \\
&= \int_{\{X_t \in A, X_{<t}\in C\}} P(Y_t \in B, Y_{<t} \in D | X_t, X_{<t}) dP \\
&= \int E(1_{\{X_t\in A, X_{<t}\in C\}} \cdot P(Y_t \in B, Y_{<t} \in D | X_t, X_{<t}) | X_{<t}) dP \\
&= \int E(1_{\{X_t\in A, X_{<t}\in C\}} \cdot P(Y_t \in B | X_t) \cdot P(Y_{<t} \in D | X_{<t}) | X_{<t}) dP \\
&= \int_{\{X_{<t}\in C\}} P(Y_{<t} \in D | X_{<t}) \cdot E(1_{\{X_t\in A\}} \cdot P(Y_t \in B | X_t) | X_{t-1}) dP \\
&= \int_{\{X_{<t}\in C\}} P(Y_{<t} \in D | X_{<t}) \cdot E(1_{\{X_t\in A\}} \cdot E(1_{\{Y_t\in B\}} | X_{t-1}^t) | X_{t-1}) dP \\
&= \int_{\{X_{<t}\in C\}} P(Y_{<t} \in D | X_{<t}) \cdot E(E(1_{\{X_t\in A, Y_t\in B\}} | X_{t-1}^t) | X_{t-1}) dP \\
&= \int_{\{X_{<t}\in C\}} E(1_{\{Y_{<t}\in D\}} | X_{<t}) \cdot E(1_{\{X_t\in A, Y_t\in B\}} | X_{t-1}) dP \\
&= \int_{\{X_{<t}\in C\}} E(1_{\{Y_{<t}\in D\}} \cdot P(X_t \in A, Y_t \in B | X_{t-1}) | X_{<t}) dP \\
&= \int_{\{X_{<t}\in C, Y_{<t}\in D\}} P(X_t \in A, Y_t \in B | X_{t-1}) dP
\end{aligned}$$

Damit gilt

$$\int_{\{(X_{<t}, Y_{<t})\in C\}} P((X_t, Y_t) \in A | (X_{<t}, Y_{<t})) dP$$

$$= \int_{\{(X_{<t},Y_{<t})\in C\}} P((X_t, Y_t) \in A | X_{t-1}) dP \tag{3.3}$$

für alle $A \in \mathcal{E}$ und $C \in \mathcal{E}_{<t}$.

Zu $A \times B \in \sigma$ definiere die Menge

$$\mathcal{D}_{A\times B} := \{C \in \sigma_{<t} : \text{ (3.3) gilt für } A \times B \text{ und } C\}.$$

Dann ist $\mathcal{E}_{<t} \subseteq \mathcal{D}_{A\times B}$, und außerdem ist $\mathcal{D}_{A\times B}$ ein Dynkinsystem, womit $\mathcal{D}_{A\times B} = \sigma_{<t}$ folgt.

Die Behauptung gilt dann für alle Elemente von $\mathcal{E} \times \sigma_{<t}$.

Definiere weiter zu $D \in \sigma_{<t}$ die Menge

$$\mathcal{D}_D := \{A \in \sigma_t : \text{(3.3) gilt für } A \text{ und } D\}.$$

Dann ist $\mathcal{D}_D$ wieder ein Dynkinsystem und es folgt $\mathcal{D}_D = \sigma_t$, woraus folgt, dass die Behauptung für alle Elemente von $\sigma \otimes \sigma_{<t}$ gilt.

□

Im Folgenden sei mit $\mathcal{M}$ stets die Markov-Kette $(M_t)_{t\in\mathbb{Z}} = (X_t, Y_t)_{t\in\mathbb{Z}}$ gemeint.

Die Aussage des Lemmas führt dazu, dass Zukunft bzw. die Vergangenheit nur vom gegenwärtigen Zeitpunkt der Markov-Kette abhängig ist.

3.2.3 Korollar

Sei (X, Y) ein Hidden-Markov-Modell.

Dann gilt für alle $t \in \mathbb{Z}$

(i) $$P(A|X_s, Y_s, s \in \mathbb{Z}_{\leq t}) = P(A|X_t) \text{ für alle } A \in \mathcal{F}_{t+1}^{\infty}.$$

(ii) $$P(B|X_s, Y_s, s \in \mathbb{Z}_{\geq t}) = P(B|X_t) \text{ für alle } B \in \mathcal{F}_{-\infty}^{t-1}.$$

(iii) $$P(A|X_s, s \in \mathbb{Z}_{\leq t}) = P(A|X_t) \text{ für alle } A \in \mathcal{F}_{t}^{\infty}.$$

(iv) $$P(B|X_s, s \in \mathbb{Z}_{\geq t}) = P(B|X_t) \text{ für alle } B \in \mathcal{F}_{-\infty}^{t}.$$

(v) $$P(A|X_s, Y_s, s \in \mathbb{Z}_{\leq t}) = P(A|X_t, Y_t) \text{ für alle } A \in \mathcal{F}_{t}^{\infty}.$$

(vi) $$P(B|X_s, Y_s, s \in \mathbb{Z}_{\geq t}) = P(B|X_t, Y_t) \text{ für alle } B \in \mathcal{F}_{-\infty}^{t}.$$

Beweis:

Exemplarisch werde hier nur der Beweis zu (i) geführt und darauf hingewiesen, dass die Aussagen (ii) bis (vi) mit dem gleichen Vorgehen bewiesen werden können.

Zu (i):

Sei $s \in \mathbb{N}$. Für jede Zylindermenge $Z = A \times (M \times \mathcal{Y})^\infty$ mit $A \in \mathcal{F}_{t+1}^{t+s}$ erhält man unter Berücksichtigung der Markov-Eigenschaft des Prozesses $\mathcal{M}$

$$\begin{aligned} P((M_n)_{n>t} \in Z | X_t) & \\ &= P(M_{t+1}^{t+s} \in A | X_t) \\ &= P(M_{t+1}^{t+s} \in A | M_{-\infty}^{t}) \\ &= P((M_n)_{n>t} \in Z | M_{-\infty}^{t}) \,. \end{aligned}$$

Ein Dynkin-Schluss liefert dann die Behauptung.

□

Unabhängigkeit von Zukunft und Vergangenheit

Die beiden nächsten Lemmata geben die stochastische Unabhängigkeit von Zukunft und Vergangenheit an, sofern die Gegenwart der Markov-Kette und das gegenwärtige Merkmal gegeben sind bzw. nur die Gegenwart der Markov-Kette gegeben ist.

3.2.4 Lemma
Seien $t \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^{t}$ und $B \in \mathcal{F}_{t}^{\infty}$.

Dann gilt

$$P(A \cap B | X_t, Y_t) = P(A | X_t, Y_t) \cdot P(B | X_t, Y_t) \ .$$

Beweis:

Für alle $A, C \in (\mathcal{P}(M) \otimes \mathcal{A}_\mathcal{Y})^\infty$ und $B \in \mathcal{P}(M) \otimes \mathcal{A}_\mathcal{Y}$ gilt mit der Markov-Eigenschaft von $\mathcal{M}$

$$\int_{\{M_t \in B\}} P(M_t^\infty \in A, M_{-\infty}^{t} \in C | M_t) dP$$

$$
\begin{aligned}
&= P(M_t^\infty \in A, M_{-\infty}^t \in C, M_t \in B) \\
&= \int_{\{M_t^\infty \in A, M_t \in B\}} P(M_{-\infty}^t \in C | M_t^\infty) dP \\
&= \int_{\{M_t^\infty \in A, M_t \in B\}} P(M_{-\infty}^t \in C | M_t) dP \\
&= \int_{\{M_t \in B\}} E(1_{\{M_t^\infty \in A\}} P(M_{-\infty}^t \in C | M_t) | M_t) dP \\
&= \int_{\{M_t \in B\}} P(M_t^\infty \in A | M_t) \cdot P(M_{-\infty}^t \in C | M_t) dP \,,
\end{aligned}
$$

also ist $P^{(M_{-\infty}^t, M_t^\infty)|M_t} = P^{M_{-\infty}^t|M_t} \otimes P^{M_t^\infty|M_t}$ P-fast sicher.

□

3.2.5 Lemma

(i) Seien $t \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^t$ und $B \in \mathcal{F}_{t+1}^\infty$.
Dann gilt

$$P(A \cap B | X_t) = P(A|X_t) \cdot P(B|X_t) \,.$$

(ii) Seien $t \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^{t-1}$ und $B \in \mathcal{F}_t^\infty$.
Dann gilt

$$P(A \cap B | X_t) = P(A|X_t) \cdot P(B|X_t) \,.$$

Beweis:

Zu(i):

Seien $t \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^t, B \in \mathcal{F}_{t+1}^\infty$ und $i \in M$.

Dann gilt mit dem Lemma 3.2.4 und dem Korollar 3.2.3

$$
\begin{aligned}
P(A \cap B | X_t) &= E(E(1_{A \cap B} | X_t, Y_t) | X_t) \\
&= E(E(1_A | X_t, Y_t) \cdot E(1_B | X_t, Y_t) | X_t) \\
&= E(E(1_A | X_t, Y_t) \cdot E(1_B | X_t) | X_t) \\
&= E(E(1_A | X_t, Y_t) | X_t) \cdot E(1_B | X_t) \\
&= E(1_A | X_t) \cdot E(1_B | X_t) \\
&= P(A|X_t) \cdot P(B|X_t) \,.
\end{aligned}
$$

Zu (ii):

Seien $t \in \mathbb{Z}, A \in \mathcal{F}_{-\infty}^{t-1}, B \in \mathcal{F}_t^\infty$ und $i \in M$.

Dann gilt mit dem Lemma 3.2.4 und dem Korollar 3.2.3

$$\begin{aligned} P(A \cap B|X_t) &= E(E(1_{A\cap B}|X_t, Y_t)|X_t) \\ &= E(E(1_A|X_t, Y_t) \cdot E(1_B|X_t, Y_t)|X_t) \\ &= E(E(1_A|X_t) \cdot E(1_B|X_t, Y_t)|X_t) \\ &= E(1_A|X_t) \cdot E(E(1_B|X_t, Y_t)|X_t) \\ &= E(1_A|X_t) \cdot E(1_B|X_t) \\ &= P(A|X_t) \cdot P(B|X_t)\,. \end{aligned}$$

□

3.2.6 Bemerkung
Im Folgenden soll für Hidden-Markov-Modelle zusätzlich Folgendes gelten:

(i) Für alle $i, j \in M$ gelte $p_{ij} > 0$.

(ii) Für alle $i \in M$ und $y \in \mathcal{Y}$ gelte $f_i(y) > 0$.

Außerdem sei mit (X, Y) immer ein Hidden-Markov-Modell bezeichnet.

3.2.7 Bemerkung
Aus der Definition des Hidden-Markov-Modells folgt unmittelbar, dass für alle $s, t \in \mathbb{Z}, s < t$ der Zufallsvektor $(X_s, \ldots, X_t, Y_s, \ldots, Y_t)$ folgende Dichte f bezüglich des Zählmaßes und des Maßes μ^{t-s+1} besitzt:

Für $x_s^t \in M^{t-s+1}$ und $y_s^t \in \mathcal{Y}^{t-s+1}$ gilt:

$$f(x_s, \cdots x_t, y_s, \ldots, y_t) = \pi(x_s) \cdot \prod_{j=s}^{t-1} p_{x_j x_{j+1}} f_{x_j}(y_j) \cdot f_{x_t}(y_t)\ .$$

Teil II

Kapitel 4

Klassifikationsrisiken

Auch in diesem Teil werde weiter wie im Kapitel 2 angenommen, dass man Aussagen über die wahre Musterklasse X_0 treffen möchte. Dazu liegen wieder das beobachtete Merkmal Y_0 und die Information Z_0 vor.
Es werden im Folgenden zunächst die Zufallsvariable Z_0 auf verschiedene Arten spezifiziert und die dadurch resultierenden optimalen Klassifikationsrisiken verglichen. Anschließend wird das Hidden-Markov-Modell als Grundlage des Prozesses (X, Y) der Muster gewählt. In dem so entstehenden Modell werden ebenfalls die optimalen Klassifikationsrisiken für verschiedene Besetzungen der Zufallsgröße Z_0 miteinander verglichen.

4.1 Verschiedene Informationen

4.1.1 Bemerkung
Da im weiteren Verlauf die Klassifikationsrisiken zweier chronometrischer Musterklassifikationsmodelle miteinander verglichen werden sollen, ist eine zusätzliche Schreibweise für Klassifikationsrisiken in den einzelnen Modellen notwenig. Der einzige Unterschied zwischen den zu betrachtenden Modellen soll in der Information Z_0 liegen, d.h., die Verteilungen der Muster seien identisch und lediglich die Informationen $Z_0^{Modell\,1}$ und $Z_0^{Modell\,2}$ seien verschieden.

Da also derselbe Wahrscheinlichkeitsraum zugrundeliegt, reicht die Schreibweise

$$R(Z_0^{Modell\,1})$$

anstelle von $R(P, Z_0^{Modell\,1})$ für das optimale Klassifikationsrisiko aus.

4.1.2 Bemerkung
Mit der Notation

$$Cr(\Omega, \mathcal{A}, P, X, Y, Z)$$

sei im Folgenden stets ein chronometrisches Musterklassifikationsmodell gemeint, wobei $(\Omega, \mathcal{A}, P)$ der unspezifizierte Wahrscheinlichkeitsraum, (X, Y) der Prozess der Muster und Z der Prozess der Informationen ist.

4.1.3 Definition
Zwei chronometrische Musterklassifikationsmodelle $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ und $Cr(\hat{\Omega}, \hat{\mathcal{A}}, \hat{P}, \hat{X}, \hat{Y}, \hat{Z})$ heißen **vergleichbar**, falls beiden Modellen derselbe Wahrscheinlichkeitsraum zugrundeliegt und beide Modelle denselben Prozess (X, Y) der Muster beinhalten, d.h. falls gilt

$$Cr(\hat{\Omega}, \hat{\mathcal{A}}, \hat{P}, \hat{X}, \hat{Y}, \hat{Z}) = Cr(\Omega, \mathcal{A}, P, X, Y, \hat{Z})\,.$$

Abkürzend wird in vergleichenden Situationen die Schreibweise $Cr(Z)$ bzw. $Cr(\hat{Z})$ für $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ bzw. $Cr(\Omega, \mathcal{A}, P, X, Y, \hat{Z})$ benutzt.

Im ersten Satz wird die nahe liegende Vermutung bestätigt, dass mit wachsender Information das optimale Klassifikationsrisiko nicht größer werden kann.

4.1.4 Satz
Seien $Cr(Z)$ und $Cr(\hat{Z})$ vergleichbare chronometrische Musterklassifikationsmodelle.

Ist dann die Zufallsvariable Z_0 $\sigma(\hat{Z}_0)$-messbar, so gilt

$$R(\hat{Z}_0) \leq R(Z_0)\,.$$

Beweis:

Es gilt

$$\begin{aligned}
R(\hat{Z}_0) &= E \min_{s\in M} P(X_0 \neq s | Y_0, \hat{Z}_0) \\
&= E\left(E(\min_{s\in M} P(X_0 \neq s | Y_0, \hat{Z}_0) \ | Y_0, Z_0 \ ) \right) \\
&\leq E\left(\min_{s\in M} E(\ P(X_0 \neq s | Y_0, \hat{Z}_0) \ | Y_0, Z_0 \ ) \right) \\
&= E\left(\min_{s\in M} E(\ E(1_{\{X_0 \neq s\}} | Y_0, \hat{Z}_0) \ | Y_0, Z_0 \ ) \right)
\end{aligned}$$

$$\begin{aligned} &= E \min_{s\in M} E(1_{\{X_0\neq s\}}|Y_0, Z_0) \\ &= R(Z_0)\,. \end{aligned}$$

□

Es ist ebenfalls nahe liegend, dass die Information Z_0 gänzlich überflüssig ist, falls diese stochastisch unabhängig vom zu klassifizierenden Muster (X_0, Y_0) ist.

4.1.5 Satz
Sei $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ ein chronometrisches Musterklassifikationsmodell. Sind Z_0 und (X_0, Y_0) stochastisch unabhängige Zufallsvariablen, so ergibt sich als optimales Klassifikationsrisiko

$$R(P, Z_0) = \int_{\mathcal{Y}} \min_{s\in M} P(X_0 \neq s|Y_0 = y) P^{Y_0}(dy)\,,$$

wobei zu beachten ist, dass die Zufallsgröße Z_0 im rechten Term nicht auftaucht.

Beweis:

Mit dem Korollar A.1.2 folgt aus der stochastischen Unabhängigkeit

$$\begin{aligned} R(P, Z_0) &= E \min_{s\in M} P(X_0 \neq s|Y_0, Z_0) \\ &= E \min_{s\in M} P(X_0 \neq s|Y_0)\,. \end{aligned}$$

□

4.1.6 Bemerkung
Betrachtet man den einfachen Fall, dass der Prozess (X, Y) der Muster eine Folge von stochastisch unabhängigen Zufallsvariablen ist und die Information Z_0 lediglich die Form

$$Z_0 = (X_{i_1}, \ldots, X_{i_n}, Y_{j_1}, \ldots, Y_{j_m})$$

mit $n, m \in \mathbb{N}$, $i_1, \ldots, i_n, j_1, \ldots, j_m \in \mathbb{Z} \setminus \{0\}$ besitzt, so sind die Voraussetzungen des Satzes 4.1.5 erfüllt, womit dann diese Information überflüssig wird.

Werden keinerlei Eigenschaften den stochastischen Prozessen zugesprochen, lassen sich auch keine weiteren Aussagen über Klassifikationsrisiken erzielen, so dass der Übergang zu einem weiter spezifizierten Modell sinnvoll erscheint. Um dem nachzukommen wird im nächsten Kapitel der Prozess (X, Y) der Muster durch ein Hidden-Markov-Modell modelliert.

Kapitel 5

Ein HMM als Prozess der Muster

5.1 Vergleich optimaler Klassifikationsrisiken

Im weiteren Verlauf sei der Prozess (X, Y) der Muster ein Hidden-Markov-Prozess, d.h., der Prozess $X = (X_t)_{t\in\mathbb{Z}}$ ist eine homogene Markov-Kette mit endlichem Zustandsraum $M = \{1, \ldots, m\}, m \in \mathbb{N}$, mit der Übergangsmatrix $P = (p_{ij})_{i,j\in M}$, $p_{ij} > 0, i, j \in M$ und mit der invarianten Verteilung π. Der Prozess $Y = (Y_t)_{t\in\mathbb{Z}}$ besitzt den Wertebereich $\mathcal{Y}$, und für alle $i \in M$ bezeichne f_i die Dichte von $P^{Y_t|X_t=i}$ für sämtliche $t \in \mathbb{Z}$ bezüglich eines Maßes μ.

Wie im Abschnitt 4.1 werden im Folgenden die optimalen Klassifikationsrisiken vergleichbarer chronometrischer Musterklassifikationsmodelle untersucht, wobei im Sinne eines Prognosesystems die unterschiedlichen Informationen stets aus der Vergangenheit stammen werden und nur aus dem Hidden-Markov-Modell resultieren.

Zunächst sei angenommen, dass die Muster, also die Merkmale und die wahren Musterklassen, der letzten $l, l \in \mathbb{N}$, Zeitpunkte bekannt sind, was ein Höchstmaß an Information darstellt. Es wird sich im ersten Satz zeigen, dass die Markov-Eigenschaft den Einfluss auf das optimale Klassifikationsrisiko dieser Information auf die Kenntnis der letzten Musterklasse reduziert.

5.1.1 Satz
Seien $Cr(Z)$ und $Cr(\hat{Z})$ zwei vergleichbare chronometrische Musterklassifikationsmodelle mit

$$Z_0 = X_{-1} \text{ und } \hat{Z}_0 = (X_{-1}, \ldots, X_{-l}, Y_{-1}, \ldots, Y_{-l}),\ l \in \mathbb{N}\,.$$

Dann gilt

$$R(\hat{Z}_0) = R(Z_0)\,.$$

Die Berücksichtigung von $X_{-1}, \ldots, X_{-l}, Y_{-1}, \ldots, Y_{-l}$ im Vergleich zu X_{-1} erbringt also keinen Vorteil bezüglich des optimalen Klassifikationsrisikos.

Zum Beweis dieses Satzes wird ein Lemma gebraucht, welches aus den Eigenschaften eines Hidden-Markov-Modells folgt.

5.1.2 Lemma
Für alle $A \subseteq M$ und $l \in \mathbb{N}$ gilt

$$P(X_0 \in A | Y_0, X_{-\infty}^{-1}, Y_{-\infty}^{-1}) = P(X_0 \in A | Y_0, X_{-1}) \,.$$

Beweis:

Seien $A \subseteq M$, $B \subseteq \mathcal{Y}$ und $C \in \mathcal{F}_{-\infty}^{-1}$, wobei B messbar ist.

Dann gilt

$$\begin{aligned}
&\int_{\{Y_0 \in B, (X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} P(X_0 \in A | Y_0, X_{-\infty}^{-1}, Y_{-\infty}^{-1}) dP \\
&= P(X_0 \in A, Y_0 \in B, (X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C) \\
&= \int_{\{(X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} E(1_{\{X_0 \in A\}} \cdot 1_{\{Y_0 \in B\}} | X_{-\infty}^{-1}, Y_{-\infty}^{-1}) dP \\
&= \int_{\{(X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} E(1_{\{X_0 \in A\}} \cdot 1_{\{Y_0 \in B\}} | X_{-1}) dP \\
&= \int_{\{(X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} E(E(1_{\{X_0 \in A\}} | Y_0, X_{-1}) \cdot 1_{\{Y_0 \in B\}} | X_{-1}) dP \\
&= \int_{\{(X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} E(P(X_0 \in A | Y_0, X_{-1}) \cdot 1_{\{Y_0 \in B\}} | X_{-\infty}^{-1}, Y_{-\infty}^{-1}) dP \\
&= \int_{\{Y_0 \in B, (X_{-\infty}^{-1}, Y_{-\infty}^{-1}) \in C\}} P(X_0 \in A | Y_0, X_{-1}) dP \,.
\end{aligned}$$

Mittels eines Dynkin-Schlusses folgt die Behauptung.

□

Mit diesem Lemma lässt sich der Satz einfach beweisen.

Beweis:

Es gilt mit dem obigen Lemma

$$\begin{aligned} R(\hat{Z}_0) &= E \min_{s\in M} P(X_0 \neq s|Y_0, \hat{Z}_0) \\ &= E \min_{s\in M} P(X_0 \neq s|Y_0, X_{-1}, \ldots, X_{-l}, Y_{-1}, \ldots, Y_{-l}) \\ &= E \min_{s\in M} P(X_0 \neq s|Y_0, X_{-1}) \\ &= R(Z_0)\,. \end{aligned}$$

□

Als Korollar ergibt sich aus diesem Satz, dass keine Information aus der Vergangenheit so wertvoll ist wie die Kenntnis der wahren Musterklasse des letzten Zeitpunktes.

5.1.3 Korollar
Seien $Cr(Z)$ und $Cr(\hat{Z})$ zwei vergleichbare chronometrische Musterklassifikationsmodelle mit

$$Z_0 = X_{-1} \;\; und \;\; \hat{Z}_0 = (X_{i_1}, \ldots, X_{i_l}, Y_{j_1}, \ldots, Y_{j_k})\,,$$

wobei $l, k \in \mathbb{N}$ und $i_1, \ldots, i_l, j_1, \ldots j_k \in \mathbb{Z}_{\leq -1}$ sind.

Dann gilt

$$R(Z_0) \leq R(\hat{Z}_0)\,,$$

die Berücksichtigung von Informationen aus der Vergangenheit erbringt also keinen Vorteil bezüglich des optimalen Klassifikationsrisikos gegenüber der Kenntnis der wahren Musterklasse im letzten Zeitpunkt.

Beweis:

Mit dem Satz 5.1.1 und $\rho := \max\{l, k\}$ gilt

$$R(Z_0) = R(X_{-1}, \ldots, X_{-\rho}, Y_{-1}, \ldots, Y_{-\rho})\,.$$

Der Satz 4.1.4 liefert dann die Behauptung.

□

Es wird nun der interessantere Fall behandelt, in dem keine Musterklasse aus der Vergangenheit bekannt ist und somit die Markov-Eigenschaft keinen direkten Einfluss auf weiter zurückliegende Musterklassen und Merkmale haben kann. In den folgenden beiden Abschnitten wird das optimale Klassifikationsrisiko für den Fall untersucht, dass lediglich Merkmale der letzten l Zeitpunkte bekannt sind. Die natürliche Zahl l gibt also an, wie viele Schritte

in die Vergangenheit zurückgegangen wird, um die Information Z_0 zu bilden. Im weiteren Verlauf wird diese Schrittweite l **Erinnerungstiefe** genannt.

Es erfolgt zunächst die leichtere Betrachtung mit der Annahme eines endlichen Merkmalsraumes, weil in ihr einerseits die Vorgehensweise und wesentlichen Ideen klarer zum Vorschein kommen und andererseits die in beiden Fällen auftauchenden Abschätzungen voneinander abweichen. Dabei wird in beiden Betrachtungen die Differenz der optimalen Klassifikationsrisiken zweier chronometrischer Musterklassifikationsmodelle, die sich lediglich darin unterscheiden, dass in einem dieser Modelle die Erinnerungstiefe l und im anderen Modell die Erinnerungstiefe $l + k$ vorliegen, abgeschätzt. Es wird also die Frage beantwortet, welchen Gewinn die Hinzunahme der Merkmale $Y_{-l-1}, \ldots, Y_{-l-k}$ zur Information $Z_0 = (Y_0, \ldots, Y_{-l})$ erbringen. Anschließend wird sich zeigen, dass das optimale Klassifikationsrisiko exponentiell mit der Erinnerungstiefe l fällt.

5.2 Endlicher Merkmalsraum

In diesem Abschnitt wird ein Hidden-Markov-Modell (X, Y) betrachtet, wobei der Wertebereich der Merkmale endlich ist. Damit ergeben sich folgende Notationen:

(i) $X = (X_t)_{t\in\mathbb{Z}}$ ist eine homogene Markovkette mit endlichem Zustandsraum M, der Übergangsmatrix $P = (p_{ij})_{i,j\in M}$, wobei $p_{ij} > 0$ für alle $i, j \in M$ angenommen werde, und invarianter Verteilung π.

(ii) $Y = (Y_t)_{t\in\mathbb{Z}}$ ist ein stochastischer Prozess mit endlichem Zustandsraum $\mathcal{Y}$, der bezüglich X bedingt stochastisch unabhängig ist.

(iii) Es sei $Q = (q_{ij})_{i\in M, j\in\mathcal{Y}}$ die Matrix der bedingten Wahrscheinlichkeiten, die unabhängig vom Zeitparameter sein sollen, d.h. $q_{ij} = P(Y_t = j | X_t = i) > 0$ für alle $t \in \mathbb{Z}$.

Im ersten Lemma dieses Abschnittes wird eine Ungleichung bewiesen, die im späteren Verlauf nützlich sein wird.

5.2.1 Lemma
Sei $n \in \mathbb{N}$ und seien $x_1, \ldots, x_n \in \mathbb{R}$ und , $y_1, \ldots, y_n > 0$.

Dann gilt

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \leq \max_{j\in\{1,\ldots,n\}} \frac{x_j}{y_j} \ .$$

Beweis:

Sei $j = \underset{i\in\{1,\ldots,n\}}{\operatorname{argmax}} \frac{x_i}{y_i}$. Dann gilt

$$\begin{aligned}
\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} &= \frac{\sum_{i=1}^n \frac{x_i}{y_i} y_i}{\sum_{i=1}^n y_i} \\
&\leq \frac{\sum_{i=1}^n \frac{x_j}{y_j} y_i}{\sum_{i=1}^n y_i} \\
&= \frac{\frac{x_j}{y_j} \cdot \sum_{i=1}^n y_i}{\sum_{i=1}^n y_i} \\
&= \frac{x_j}{y_j} = \max_{i\in\{1,\ldots,n\}} \frac{x_i}{y_i}
\end{aligned}$$

□

Mit Hilfe dieses Lemmas lässt sich nun die folgende wichtige Abschätzung gewinnen.

5.2.2 Lemma
Für alle $l \in \mathbb{N}_0$, beschränkten $T \subset \mathbb{Z}$, $\{y_t \in \mathcal{Y} : t \in T\}$ und $i, r, s \in M$ gilt

$$\frac{P(X_{-l} = r | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = s | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \leq \begin{cases} \max\limits_{y,i,r,s,j} \frac{q_{ry} \cdot p_{ir} \cdot p_{rj}}{q_{sy} \cdot p_{is} \cdot p_{sj}} & , \textit{falls } -l \in T \\ \max\limits_{i,r,s,j} \frac{p_{ir} \cdot p_{rj}}{p_{is} \cdot p_{sj}} & , \textit{falls } -l \notin T \end{cases}.$$

Beweis:

1.Fall $-l, -l+1 \in T$:

Mit der Bezeichnung $T_{\geq x} := T \cap \{t \in \mathbb{Z} : t \geq x\}$ und $\mathcal{Y}_{\geq x} := \{Y_t = y_t : t \in T_{\geq x}\}$ gilt

$$\frac{P(X_{-l} = r | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = s | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}$$

$$
\begin{aligned}
&= \frac{P(X_{-l}=r|X_{-l-1}=i,\mathcal{Y}_{\geq -l})}{P(X_{-l}=r|X_{-l-1}=i,\mathcal{Y}_{\geq -l})} \\
&= \frac{P(X_{-l}=r,X_{-l-1}=i,\mathcal{Y}_{\geq -l})}{P(X_{-l}=s,X_{-l-1}=i,\mathcal{Y}_{\geq -l})} \\
&= \frac{P(X_{-l}=r,\mathcal{Y}_{\geq -l}|X_{-l-1}=i)}{P(X_{-l}=s,\mathcal{Y}_{\geq -l}|X_{-l-1}=i)} \\
&= \frac{\sum_j P(X_{-l+1}=j,X_{-l}=r,\mathcal{Y}_{\geq -l}|X_{-l-1}=i)}{\sum_j P(X_{-l+1}=j,X_{-l}=s,\mathcal{Y}_{\geq -l}|X_{-l-1}=i)} \\
&= \frac{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j,X_{-l}=r,X_{-l-1}=i,Y_{-l}^{-l+1}=y_{-l}^{-l+1})}{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j,X_{-l}=s,X_{-l-1}=i,Y_{-l}^{-l+1}=y_{-l}^{-l+1})} \\
&\quad \frac{\cdot P(X_{-l+1}=j,X_{-l}=r,Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l-1}=i)}{\cdot P(X_{-l+1}=j,X_{-l}=s,Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l-1}=i)} \\
&= \frac{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot P(Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l+1}=j,X_{-l}=r,X_{-l-1}=i)}{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot P(Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l+1}=j,X_{-l}=s,X_{-l-1}=i)} \\
&\quad \frac{\cdot P(X_{-l+1}=j,X_{-l}=r|X_{-l-1}=i)}{\cdot P(X_{-l+1}=j,X_{-l}=s|X_{-l-1}=i)} \\
&= \frac{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot P(Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l+1}=j,X_{-l}=r)}{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot P(Y_{-l}^{-l+1}=y_{-l}^{-l+1}|X_{-l+1}=j,X_{-l}=s)} \\
&\quad \frac{\cdot P(X_{-l+1}=j|X_{-l}=r,X_{-l-1}=i)\cdot P(X_{-l}=r|X_{-l-1}=i)}{\cdot P(X_{-l+1}=j|X_{-l}=s,X_{-l-1}=i)\cdot P(X_{-l}=s|X_{-l-1}=i)} \\
&= \frac{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot q_{ry_{-l}}\cdot q_{jy_{-l+1}}\cdot p_{rj}\cdot p_{ir}}{\sum_j P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot q_{sy_{-l}}\cdot q_{jy_{-l+1}}\cdot p_{sj}\cdot p_{is}} \\
&\leq \max_j \frac{P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot q_{ry_{-l}}\cdot q_{jy_{-l+1}}\cdot p_{rj}\cdot p_{ir}}{P(\mathcal{Y}_{\geq -l+2}|X_{-l+1}=j)\cdot q_{sy_{-l}}\cdot q_{jy_{-l+1}}\cdot p_{sj}\cdot p_{is}}
\end{aligned}
$$

$$= \max_j \frac{q_{ry_{-l}} \cdot p_{rj} \cdot p_{ir}}{q_{sy_{-l}} \cdot p_{sj} \cdot p_{is}}$$

$$\leq \max_{j,y,r,s,i} \frac{q_{ry} \cdot p_{rj} \cdot p_{ir}}{q_{sy} \cdot p_{sj} \cdot p_{is}} .$$

2.Fall $-l \in T, -l+1 \notin T$:

In dieser Situation fallen in obiger Rechnung sämtliche Terme $Y_{-l+1} = y_{-l+1}$ weg, die ab der sechsten Zeile auftauchen, um schließlich durch Kürzen wieder zu verschwinden.

Somit liefern diese beiden ersten Rechnungen den ersten Teil der Behauptung.

3.Fall $-l \notin T$:

Eine analoge Rechnung, wobei die Terme $Y_{-l} = y_{-l}$ und somit $q_{ry_{-l}}$ und $q_{sy_{-l}}$ jeweils nicht auftauchen können, liefert den zweiten Teil der Behauptung und schließt den Beweis damit ab.

□

5.2.3 Korollar
Für alle $l \in \mathbb{N}_0$, $T \subset \mathbb{Z}$, $\{Y_t = y_t : t \in T\}$, $i, j \in M$ *und*

$$\alpha = \max \left\{ \max_{j,y,\lambda,\mu,i} \frac{q_{\lambda y} \cdot p_{\lambda j} \cdot p_{i\lambda}}{q_{\mu y} \cdot p_{\mu j} \cdot p_{i\mu}}, \max_{j,\lambda,\mu,i} \frac{p_{\lambda j} \cdot p_{i\lambda}}{p_{\mu j} \cdot p_{i\mu}} \right\}$$

gilt

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq (1 + (|M| - 1)\alpha)^{-1} > 0 .$$

Beweis:

Mit dem Lemma gilt

$$\frac{1}{P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}$$

$$= \sum_k \frac{P(X_{-l} = k | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}$$

$$\leq 1 + (|M| - 1)\alpha$$

Also folgt

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq (1 + (|M| - 1)\alpha)^{-1} > 0. \quad (5.1)$$

□

Da die von α abhängige Konstante in der Ungleichung (5.1) später noch benutzt wird, sei $\eta_\alpha = (1 + (|M| - 1)\alpha)^{-1}$ definiert, also

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq \eta_\alpha > 0.$$

Mit diesem Korollar lässt sich der nun folgende Satz beweisen, der eine tragende Rolle im Vergleich zweier chronometrischer Modelle spielen wird.

5.2.4 Definition
Seien $l, n \in \mathbb{N}$, $A \subseteq M$ und $y_0, \ldots, y_{-n} \in \mathcal{Y}$.
Definiere

$$m_l^+ := m_l^+(A, y_0, \ldots, y_{-n}) = \max_{i \in M} P(X_0 \in A | X_{-l} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n})$$

und

$$m_l^- := m_l^-(A, y_0, \ldots, y_{-n}) = \min_{i \in M} P(X_0 \in A | X_{-l} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n}) .$$

5.2.5 Satz
Es existiert ein $\beta \in (0, 1)$, *so dass für alle* $l, n \in \mathbb{N}$, $A \subseteq M$ *und* $y_0, \ldots, y_{-n} \in \mathcal{Y}$ *gilt*

$$m_l^+ - m_l^- \leq \beta^l.$$

Beweis:

Für $A = M$ ist $m_l^+ - m_l^- = 1 - 1 = 0$. Also ist die Behauptung für $A \neq M$ zu zeigen.

Der Beweis wird mittels des Beweisprinzips der vollständigen Induktion über l geführt.

Induktionsverankerung $l = 1$:

Wegen $m_1^+ = \max_{i \in M} P(X_0 \in A | X_{-1} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n})$ gilt mit $k \in A^c$

$$\begin{aligned} 1 - m_1^+ &= \min_{i \in M} P(X_0 \in A^c | X_{-1} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n}) \\ &\geq \min_{i \in M} P(X_0 = k | X_{-1} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n}) \\ &\geq \eta_\alpha . \end{aligned}$$

Weiter gilt für ein $k \in A$ wegen $m_1^- = \min_{i \in M} P(X_0 \in A | X_{-1} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n})$

$$\begin{aligned} m_l^- &\geq \min_{i \in M} P(X_0 = k | X_{-1} = i, Y_0 = y_0, \ldots, Y_{-n} = y_{-n}) \\ &\geq \eta_\alpha \,. \end{aligned}$$

Und somit folgt:

$$m_1^+ - m_1^- \leq 1 - 2\eta_\alpha \,,$$

was die Induktionsverankerung mit $\beta := 1 - 2\eta_\alpha$ liefert.

Induktionsschritt:

Die Behauptung gelte für die Erinnerungstiefe l.

Dann gilt:

$$\begin{aligned} m_{l+1}^+ &= \max_{i \in M} P(X_0 \in A | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \\ &= \max_{i \in M} \sum_{k=1}^{|M|} P(X_0 \in A | X_{-l} = k, X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \\ &\quad \cdot P(X_{-l} = k | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \\ &= \max_{i \in M} \{ \sum_{k=1, k \neq arg(m_l^-)}^{|M|} P(X_0 \in A | X_{-l} = k_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \\ &\quad \cdot P(X_{-l} = k | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \\ &\quad + \underbrace{P(X_0 \in A | X_{-l} = arg(m_l^-), X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0)}_{m_l^-} \\ &\quad \cdot P(X_{-l} = arg(m_l^-) | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \} \\ &\leq \max_{i \in M} \{ \sum_{k=1, k \neq arg(m_l^-)}^{|M|} m_l^+ \cdot P(X_{-l} = k | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) + m_l^- \\ &\quad \cdot P(X_{-l} = arg(m_l^-) | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0) \} \\ &= \max_{i \in M} \{ m_l^+ \cdot (1 - P(X_{-l} = arg(m_l^-) | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0)) + m_l^- \\ &\quad \cdot \underbrace{P(X_{-l} = arg(m_l^-) | X_{-(l+1)} = i, Y_{-n}^0 = y_{-n}^0)}_{\geq \eta_\alpha} \} \end{aligned}$$

$$
\begin{aligned}
&\leq \max_{i\in M}\{m_l^+\cdot(1-\eta_\alpha)+m_l^-\cdot\eta_\alpha\}\\
&= m_l^+\cdot(1-\eta_\alpha)+m_l^-\cdot\eta_\alpha,
\end{aligned}
$$

wobei für die letzte Ungleichung ein Konvexitätsargument benutzt wurde. Weiter gilt

$$
\begin{aligned}
m_{l+1}^- &= \min_{i\in M} P(X_0\in A|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&= \min_{i\in M}\sum_{k=1}^{|M|} P(X_0\in A|X_{-l}=k,X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&\quad\cdot P(X_{-l}=k|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&= \min_{i\in M}\{\sum_{k=1,k\neq arg(m_l^+)}^{|M|} P(X_0\in A|X_{-l}=k,X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&\qquad\cdot P(X_{-l}=k|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&\qquad+\underbrace{P(X_0\in A|X_{-l}=arg(m_l^+),X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)}_{m_l^+}\\
&\qquad\cdot P(X_{-l}=arg(m_l^+)|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\}\\
&\geq \min_{i\in M}\{\sum_{k=1,k\neq arg(m_l^+)}^{|M|} m_l^-\cdot P(X_{-l}=k|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\\
&\qquad+m_l^+\cdot P(X_{-l}=arg(m_l^+)|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)\}\\
&= \min_{i\in M}\{m_l^-\cdot(1-P(X_{-l}=arg(m_l^+)|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0))\\
&\qquad+m_l^+\cdot\underbrace{P(X_{-l}=arg(m_l^+)|X_{-(l+1)}=i,Y_{-n}^0=y_{-n}^0)}_{\geq\eta_\alpha}\}\\
&\geq \min_{i\in M}\{m_l^-\cdot(1-\eta_\alpha)+m_l^+\cdot\eta_\alpha\}\\
&= m_l^-\cdot(1-\eta_\alpha)+m_l^+\cdot\eta_\alpha\,.
\end{aligned}
$$

Damit erhält man die beiden Bedingungen

$$
\begin{aligned}
I)\quad & m_{l+1}^+\leq(1-\eta_\alpha)m_l^++\eta_\alpha m_l^-\,,\\
II)\quad & m_{l+1}^-\geq(1-\eta_\alpha)m_l^-+\eta_\alpha m_l^+\,.
\end{aligned}
$$

Die Induktionsannahme, I) und II) liefern dann

$$\begin{aligned} m_{l+1}^{+} - m_{l+1}^{-} &\leq (1-2\eta_\alpha)(m_l^{+} - m_l^{-}) \\ &\leq (1-2\eta_\alpha)(1-2\eta_\alpha)^l \\ &= (1-2\eta_\alpha)^{l+1} \ . \end{aligned}$$

Mit $\beta = 1 - 2\eta_\alpha$ folgt die Behauptung.

□

Bei der direkten Abschätzung der Differenz der optimalen Klassifikationsrisiken wird folgendes Korollar benutzt.

5.2.6 Korollar
Es existiert ein $\beta \in (0,1)$, so dass für alle $l, k \in \mathbb{N}$, $A \subseteq M$ und $y_0, y_{-1}, \ldots \in \mathcal{Y}$ gilt

$$|P(X_0 \in A | Y_{-l}^0 = y_{-l}^0) - P(X_0 \in A | Y_{-l-k}^0 = y_{-l-k}^0)| \leq \beta^{l+1}.$$

Beweis:

Es gilt mit dem Satz **5.2.5**

$$P(X_0 \in A | Y_{-l}^0 = y_{-l}^0) - P(X_0 \in A | Y_{-l-k}^0 = y_{-l-k}^0)$$

$$\begin{aligned} &= \sum_{i \in M} P(X_0 \in A | Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \cdot P(X_{-l-1} = i | Y_{-l}^0 = y_{-l}^0) \\ &\quad - \sum_{i \in M} P(X_0 \in A | Y_{-l-k}^0 = y_{-l-k}^0, X_{-l-1} = i) \cdot P(X_{-l-1} = i | Y_{-l-k}^0 = y_{-l-k}^0) \\ &= \sum_{i \in M} P(X_0 \in A | Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \cdot P(X_{-l-1} = i | Y_{-l}^0 = y_{-l}^0) \\ &\quad - \sum_{i \in M} P(X_0 \in A | Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \cdot P(X_{-l-1} = i | Y_{-l-k}^0 = y_{-l-k}^0) \\ &\leq \sum_{i \in M} m_{l+1}^{+} \cdot P(X_{l-1} = i | Y_{-l}^0 = y_{-l}^0) - m_{l+1}^{-} \cdot P(X_{-l-1} = i | Y_{-l-k}^0 = y_{-l-k}^0) \\ &= m_{l+1}^{+} - m_{l+1}^{-} \\ &\leq \beta^{l+1} . \end{aligned}$$

□

5.2.7 Vergleich zweier optimaler Klassifikationsrisiken

Wie schon angedeutet wurde, sollen jetzt die optimalen Klassifikationsrisiken zweier chronometrischer Musterklassifikationsmodelle miteinander verglichen werden, wobei sich die Modelle nur darin unterscheiden, dass beim ersten die letzten l Merkmale und beim zweiten dagegen die letzten $l+k$ Merkmale bekannt sind, d.h. Modelle, die sich in der Erinnerungstiefe um k Schritte unterscheiden.

Da in diesem Abschnitt der Wertebereich der Merkmale endlich ist, ergibt sich aus (1.7) das optimale Klassifikationsrisiko für $Z_0^l = (Y_{-1}, \ldots, Y_{-l})$ als

$$R(Z_0^l) := \sum_{y_{-l}^0} \min_{i \in M} P(X_0 \neq i | Y_{-l}^0 = y_{-l}^0) \cdot P(Y_{-l}^0 = y_{-l}^0)\,.$$

Die anschließende Definition liefert eine Schreibweise für die folgende Situation: Es liege ein chronometrisches Musterklassifikationsmodell $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ vor, und von Interesse ist ein weiteres chronometrisches Musterklassifikationsmodell, das mit dem Vorliegenden vergleichbar ist aber nur eine Teilinformation $\hat{Z}$ von Z benutzt.

5.2.8 Definition
Sei $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ ein chronometrisches Musterklassifikationsmodell. Dann heißt ein weiteres chronometrisches Musterklassifikationsmodell $Cr(\hat{\Omega}, \hat{\mathcal{A}}, \hat{P}, \hat{X}, \hat{Y}, \hat{Z})$ **das auf $\hat{Z}$ eingeschränkte Modell zu** $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$, falls diese beiden Modelle vergleichbar sind, und $\hat{Z}$ $\sigma(Z)$-messbar ist.

5.2.9 Satz
Sei $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ ein chronometrisches Musterklassifikationsmodell mit $Z_t = (Y_t, Y_{t-1}, \ldots)$ für alle $t \in \mathbb{Z}$.

Dann existiert ein $\beta \in (0,1)$, so dass für alle $l, k \in \mathbb{N}$ und alle eingeschränkten Modelle $Cr(Z^l)$ und $Cr(Z^{l+k})$ zu $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ mit

$$Z_0^l = (Y_{-1}, \ldots, Y_{-l}) \text{ und } Z_0^{l+k} = (Y_{-1}, \ldots, Y_{-l-k})$$

gilt

$$R(Z_0^l) - R(Z_0^{l+k}) \leq \beta^{l+1}\,.$$

Beweis:

Es gilt

$$R(Z_0^l) - R(Z_0^{l+k})$$

$$
\begin{aligned}
&= \sum_{y^0_{-l}} \Bigg\{ \min_{i\in M} P(X_0 \neq i | Y^0_{-l} = y^0_{-l}) P(Y^0_{-l} = y^0_{-l}) \\
&\qquad - \sum_{y^{-l-1}_{-l-k}} \min_{j\in M} P(X_0 \neq j | Y^0_{-l-k} = y^0_{-l-k}) P(Y^0_{-l-k} = y^0_{-l-k}) \Bigg\} \\
&= \sum_{y^0_{-l}} \Bigg\{ \min_{i\in M} P(X_0 \neq i | Y^0_{-l} = y^0_{-l}) \sum_{y^{-l-1}_{-l-k}} P(Y^0_{-l-k} = y^0_{-l-k}) \\
&\qquad - \sum_{y^{-l-1}_{-l-k}} \min_{j\in M} P(X_0 \neq j | Y^0_{-l-k} = y^0_{-l-k}) P(Y^0_{-l-k} = y^0_{-l-k}) \Bigg\} \\
&= \sum_{y^0_{-l}} \Bigg\{ \sum_{y^{-l-1}_{-l-k}} \Big[\min_{i\in M} P(X_0 \neq i | Y^0_{-l} = y^0_{-l}) P(Y^0_{-l-k} = y^0_{-l-k}) \\
&\qquad - \min_{j\in M} P(X_0 \neq j | Y^0_{-l-k} = y^0_{-l-k}) P(Y^0_{-l-k} = y^0_{-l-k}) \Big] \Bigg\} \\
&= \sum_{y^0_{-l}} \Bigg\{ \sum_{y^{-l-1}_{-l-k}} \Big[\Big(\min_{i\in M} P(X_0 \neq i | Y^0_{-l} = y^0_{-l}) \\
&\qquad - \min_{j\in M} P(X_0 \neq j | Y^0_{-l-k} = y^0_{-l-k}) \Big) P(Y^0_{-l-k} = y^0_{-l-k}) \Big] \Bigg\} \\
&= \sum_{y^0_{-l}} \Bigg\{ \sum_{y^{-l-1}_{-l-k}} \beta^{l+1} \cdot P(Y^0_{-l-k} = y^0_{-l-k}) \Bigg\} \\
&= \beta^{l+1} \cdot \sum_{y^0_{-l-k}} P(Y^0_{-l-k} = y^0_{-l-k}) \\
&= \beta^{l+1} .
\end{aligned}
$$

□

Mit Hilfe dieses Satzes lässt sich offensichtlich leicht abschätzen, wie stark das optimale Klassifikationsrisiko bei einer Erinnerungstiefe l vom Idealzustand der Kenntnis sämtlicher Merkmale der Vergangenheit abweicht.

5.2.10 Korollar
Es existiert ein $\beta \in (0,1)$, so dass für alle $l \in \mathbb{N}$ gilt

$$R^* := \lim_{n\to\infty} R(Z_0^n) \leq R(Z_0^l) \leq R^* + \beta^{l+1}\,,$$

wobei $Z_0^k = (Y_{-1}, \ldots, Y_{-k})$ *für alle* $k \in \mathbb{N}$.

Dieses β *wird* ***Rate des chronometrischen Musterklassifikationsmodells*** *genannt.*

Sind die letzten l Merkmale bekannt, so weicht das optimale Klassifikationsrisiko also maximal um den Wert β^{l+1} vom minimal erreichbaren optimalen Klassifikationsrisiko ab.

5.2.11 Beispiel

Die einfachste Situation der vorliegenden Problematik ist gegeben, falls es nur zwei Musterklassen und auch nur zwei Merkmale gibt, d.h. wenn $M = \{0, 1\}$ und $\mathcal{Y} = \{0, 1\}$ vorliegen. Wegen

$$\alpha = \max\left\{ \max_{i,j,\mu,\lambda,y} \frac{q_{\lambda y} p_{i\lambda} p_{\lambda j}}{q_{\mu y} p_{i\mu} p_{\mu j}}, \max_{i,j,\mu,\lambda} \frac{p_{i\lambda} p_{\lambda j}}{p_{i\mu} p_{\mu j}} \right\},$$

$$\eta_\alpha = \frac{1}{1+\alpha} \text{ und } \beta = 1 - 2\eta_\alpha$$

ist es leicht einzusehen, dass die Rate β des vorliegenden Modells schon in einfachen Beispielen nahe bei 1 liegen kann, was daran liegt, dass die Abschätzungen, die zum Korollar 5.2.10 geführt haben, sehr grobe Abschätzungen waren.

Betrachtet man zum Beispiel die Übergangsmatrix P und die Matrix der bedingten Wahrscheinlichkeiten Q mit

$$P = \begin{pmatrix} 0,9 & 0,1 \\ 0,2 & 0,8 \end{pmatrix} \text{ und } Q = \begin{pmatrix} 0,7 & 0,3 \\ 0,4 & 0,6 \end{pmatrix},$$

dann ergibt sich eine Rate von $\beta \approx 0.9722$.

In den Abbildungen 5.1 und 5.2 sind die Graphen der Funktion $x \mapsto \beta^x$ für verschiedene Raten β abgebildet. In der Abbildung 5.2 kann man erkennen, dass für den Wert $\beta = 0.9722$ eine Erinnerungstiefe von circa $l = 80$ gewählt werden muss, wenn der Schätzer $R(Z_{-l}^0)$ um weniger als 10% von R^* abweichen soll.

Im Kapitel über Algorithmen werden die Werte der Folge $R(Z_0^l)$, $l \in \{0, \ldots, 20\}$ an einem Beispiel berechnet und mit der Abschätzung aus dem Korollar 5.2.10 verglichen.

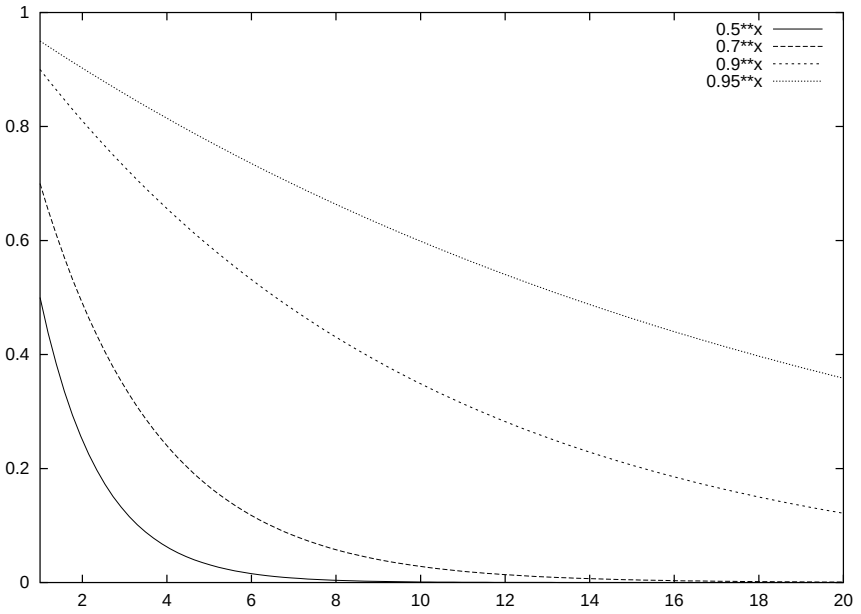

Abbildung 5.1: Graphen der Funktion $x \mapsto \beta^x$ für $\beta = 0.5, 0.7, 0.9, 0.95$

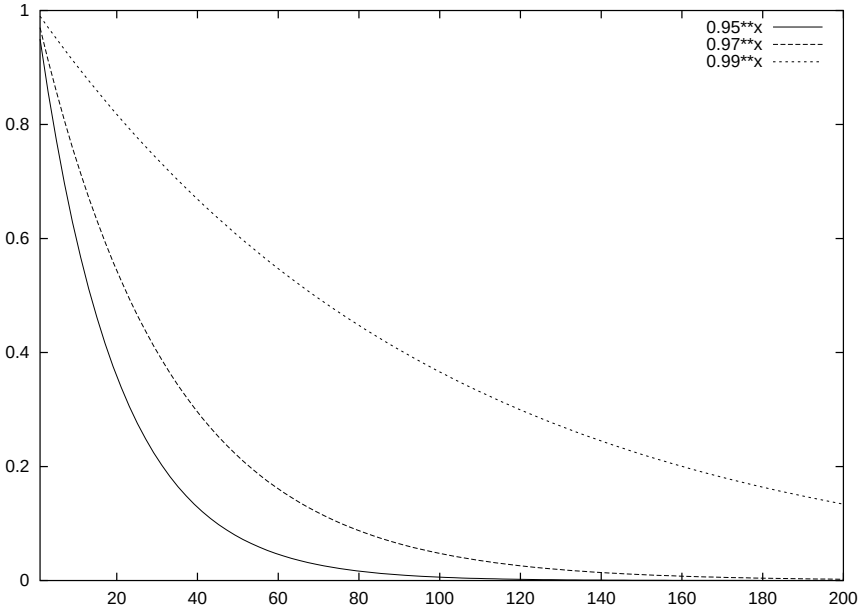

Abbildung 5.2: Graphen der Funktion $x \mapsto \beta^x$ für $\beta = 0.95, 0.97, 0.99$

5.3 Kontinuierlicher Merkmalsraum

In diesem Abschnitt wird auf ähnliche Art wie im vorherigen Abschnitt gezeigt, dass auch im Modell mit kontinuierlichem Wertebereich der Merkmale das Klassifikationsrisiko exponentiell mit der Erinnerungstiefe fällt.

Es wird ein Hidden-Markov-Modell betrachtet, wobei der Wertebereich der Merkmale kontinuierlich ist. Dabei werden folgende Notationen verwendet:

(i) $X = (X_t)_{t\in\mathbb{Z}}$ ist ein homogene Markov-Kette mit endlichem Zustandsraum M, $|M| = 2$, der Übergangsmatrix $P = (p_{ij})_{i,j\in M}$, wobei $p_{ij} > 0$ für alle $i, j \in M$ angenommen werde, und invarianter Verteilung π.

(ii) $Y = (Y_t)_{t\in\mathbb{Z}}$ ist ein stochastischer Prozess mit dem Zustandsraum $\mathcal{Y}$, der bezüglich X bedingt stochastisch unabhängig ist.

(iii) Für alle $i \in M$ und $t \in \mathbb{Z}$ sei $f_i(\cdot) > 0$ eine Dichte der Verteilung $P^{Y_t|X_t=i}$ bezüglich eines Maßes μ. Für alle messbaren $A \subseteq \mathcal{Y}$, $t \in \mathbb{Z}$ und $i \in M$ gilt also

$$P(Y_t \in A|X_t = i) = \int_A f_i(y)\,\mu(dy)\,. \tag{5.2}$$

Die Funktion $f_\cdot(\cdot)$ wird auch **bedingte Dichte** genannt. Die bedingte Dichte ist somit zeitinvariant.

Aus Gründen der Bequemlichkeit wird im Folgenden auf die Angabe des Maßes μ bei entsprechenden Integrationen verzichtet; so wird zum Beispiel anstelle der Gleichung (5.2) kurz

$$P(Y_t \in A|X_t = i) = \int_A f_i(y)\,dy$$

geschrieben. Weiter werden nun zwei Funktionen und zwei Konstanten eingeführt, die im weiteren Verlauf vielfach benutzt werden.

5.3.1 Definition
Definiere eine Abbildung $\alpha : \mathcal{Y} \to \mathbb{R}_{\geq 1}$ durch

$$\alpha(y) = \max_{1\leq\iota,\kappa,i,j\leq m} \frac{p_{i\iota}p_{\iota j}\cdot f_\iota(y)}{p_{i\kappa}p_{\kappa j}\cdot f_\kappa(y)}\ ,$$

eine Konstante

$$\alpha = \max_{1\leq\iota,\kappa,i,j\leq m} \frac{p_{i\iota}p_{\iota j}}{p_{i\kappa}p_{\kappa j}} \geq 1\ ,$$

eine Abbildung $\eta : \mathcal{Y} \to (0, \frac{1}{2}]$ durch

$$\eta(y) = (1 + (m-1)\alpha(y))^{-1}$$

und eine Konstante

$$\eta = (1 + (m-1)\alpha)^{-1} \in (0, \frac{1}{2}] \ .$$

Es wird nun das dem Lemma 6.1.1 entsprechende Lemma bewiesen.

5.3.2 Lemma
Seien $l \in \mathbb{N}_0$, $T \subset \mathbb{Z}$ beschränkt, $y_t \in \mathcal{Y}$, $t \in T$ und $i, \iota, \kappa \in M$.
Dann gilt

(1) im Fall $-l \in T$

$$\frac{P(X_{-l} = \iota | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = \kappa | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \leq \alpha(y_{-l}) \ ,$$

(2) im Fall $-l \notin T$

$$\frac{P(X_{-l} = \iota | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = \kappa | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \leq \alpha \ .$$

Beweis:

Im ersten Fall gilt mit der Bezeichnung $T_{\geq x} = T \cap \{t \in \mathbb{Z} : t \geq x\}$

$$\begin{aligned}
&\frac{P(X_{-l} = \iota | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = \kappa | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \\
&= \frac{P(X_{-l} = \iota | X_{-l-1} = i, \{Y_t = y_t : t \in T_{\geq -l}\})}{P(X_{-l} = \kappa | X_{-l-1} = i, \{Y_t = y_t : t \in T_{\geq -l}\})} \\
&= \frac{P(X_{-l-1} = i, X_{-l} = \iota | \{Y_t = y_t : t \in T_{\geq -l}\})}{P(X_{-l-1} = i, X_{-l} = \kappa | \{Y_t = y_t : t \in T_{\geq -l}\})} \\
&= \frac{\sum_j P(X_{-l-1} = i, X_{-l} = \iota, X_{-l+1} = j | \{Y_t = y_t : t \in T_{\geq -l}\})}{\sum_j P(X_{-l-1} = i, X_{-l} = \kappa, X_{-l+1} = j | \{Y_t = y_t : t \in T_{\geq -l}\})} \\
&= \frac{\sum_j P(X_{-l-1} = i, X_{-l} = \iota, X_{-l+1} = j) \cdot f_{i,\iota,j}(y_t : t \in T_{\geq -l}) / f(y_t : t \in T_{\geq -l})}{\sum_j P(X_{-l-1} = i, X_{-l} = \kappa, X_{-l+1} = j) \cdot f_{i,\kappa,j}(y_t : t \in T_{\geq -l}) / f(y_t : t \in T_{\geq -l})}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_j P(X_{-l-1}=i, X_{-l}=\iota, X_{-l+1}=j)\cdot f_{i,\iota,j}(y_t : t\in T_{\geq -l})}{\sum_j P(X_{-l-1}=i, X_{-l}=\kappa, X_{-l+1}=j)\cdot f_{i,\kappa,j}(y_t : t\in T_{\geq -l})}\\
&= \frac{\sum_j P(X_{-l-1}=i)\cdot p_{i\iota}\cdot p_{\iota j}\cdot f_{i,\iota,j}(y_t : t\in T_{\geq -l})}{\sum_j P(X_{-l-1}=i)\cdot p_{i\kappa}\cdot p_{\kappa j}\cdot f_{i,\kappa,j}(y_t : t\in T_{\geq -l})}\\
&= \frac{p_{i\iota}\cdot \sum_j p_{\iota j}\cdot f_{i,\iota,j}(y_t : t\in T_{\geq -l})}{p_{i\kappa}\cdot \sum_j p_{\kappa j}\cdot f_{i,\kappa,j}(y_t : t\in T_{\geq -l})}\\
&= \frac{p_{i\iota}\cdot \sum_j p_{\iota j}\cdot f_\iota(y_{-l})\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})}{p_{i\kappa}\cdot \sum_j p_{\kappa j}\cdot f_\kappa(y_{-l})\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})} \quad (*)\\
&= \frac{p_{i\iota}\cdot f_\iota(y_{-l})\cdot \sum_j p_{\iota j}\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})}{p_{i\kappa}\cdot f_\kappa(y_{-l})\cdot \sum_j p_{\kappa j}\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})}\\
&\leq \frac{p_{i\iota}\cdot f_\iota(y_{-l})}{p_{i\kappa}\cdot f_\kappa(y_{-l})}\cdot \max_j \frac{p_{\iota j}\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})}{p_{\kappa j}\cdot f_j(y_{-l+1})\cdot f_j(y_t : t\in T_{\geq -l+2})}\\
&= \frac{p_{i\iota}\cdot f_\iota(y_{-l})}{p_{i\kappa}\cdot f_\kappa(y_{-l})}\cdot \max_j \frac{p_{\iota j}}{p_{\kappa j}}\\
&\leq \max_{\iota,\kappa,i,j} \frac{p_{i\iota}\cdot p_{\iota j}\cdot f_\iota(y_{-l})}{p_{i\kappa}\cdot p_{\kappa j}\cdot f_\kappa(y_{-l})},
\end{aligned}
$$

wobei der sich später herauskürzende Term $f_j(y_{-l+1})$ in der Zeile $(*)$ im Nenner und Zähler nur auftritt, falls $-l+1\in T$ ist.

Im zweiten Fall ist $-l\notin T$, so dass die Terme $f_\iota(y_{-l})$ und $f_\kappa(y_{-l})$ in der Zeile $(*)$ nicht erscheinen. Dadurch vereinfachen sich die darauf folgenden Umformungen und es entfällt die Abhängigkeit von y_{-l}.

□

Das Korollar 5.2.3 verändert sich dahin gehend, dass die Konstante η_α durch einen von y_{-l} abhängenden Term $\eta(y_{-l})$ ersetzt werden.

5.3.3 Korollar
Es seien die Voraussetzungen wie im obigen Lemma gegeben.
Dann gilt

(1) im Fall $-l \in T$

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq \eta(y_{-l}) > 0 \ ,$$

(2) im Fall $-l \notin T$

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq \eta > 0 \ .$$

Beweis:

Es gilt mit $\hat{\alpha} = \begin{cases} \alpha(y_{-l}) & , -l \in T \\ \alpha & , -l \notin T \end{cases}$ und dem Lemma 5.3.2

$$\begin{aligned} & \frac{1}{P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \\ &= \sum_k \frac{P(X_{-l} = k | X_{-l-1} = i, \{Y_t = y_t : t \in T\})}{P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\})} \\ &\leq 1 + (m-1)\hat{\alpha} \ , \end{aligned}$$

also

$$P(X_{-l} = j | X_{-l-1} = i, \{Y_t = y_t : t \in T\}) \geq (1 + (m-1)\hat{\alpha})^{-1} \ .$$

□

Die Entsprechung des Satzes 5.2.5 erweist sich in der Abschätzung als etwas komplizierter, da die Abhängigkeit von der Realisierung der Folge der Merkmale erhalten bleibt, wobei der Beweisgang nahezu identisch ist.

5.3.4 Satz
Seien $n, l \in \mathbb{N}$, $A \subseteq M$ *und* $y_0, \cdots, y_{-n} \in \mathcal{Y}$.
Definiere die Folge $m_1^+, m_2^+, \ldots$ *durch*

$$m_l^+ = \max_{i \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l} = i), \ \ l \in \mathbb{N}$$

und die Folge $m_1^-, m_2^-, \ldots$ *durch*

$$m_l^- = \min_{i \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l} = i), \ \ l \in \mathbb{N} \, .$$

Dann gilt für alle $l \in \mathbb{N}$

$$m_l^+ - m_l^- \leq \prod_{j=-l+1}^{0} (1 - 2\hat{\eta}_j) \ ,$$

wobei $\hat{\eta}_j = \begin{cases} \eta(y_j) & , j \in \{-n, \cdots, 0\} \\ \eta & , j \notin \{-n, \cdots, 0\} \end{cases}$.

Beweis:

Im Fall $A = M$ gilt $m_l^+ - m_l^+ = 1 - 1 = 0$. Sei $A \neq M$.

Mit dem Hilfsmittel der vollständigen Induktion ergibt sich:

Induktionsverankerung $l = 1$:

Sei $j \in A^c$. Dann gilt

$$\begin{aligned} m_1^+ - m_1^- &= \max_{i \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-1} = i) - m_1^- \\ &= 1 - \min_{i \in M} P(X_0 \notin A | Y_{-n}^0 = y_{-n}^0, X_{-1} = i) - m_1^- \\ &\leq 1 - \min_{i \in M} P(X_0 = j | Y_{-n}^0 = y_{-n}^0, X_{-1} = i) - m_1^- \\ &\leq 1 - \hat{\eta}_0 - \hat{\eta}_0 \\ &\leq (1 - 2\hat{\eta}_0)\,. \end{aligned}$$

Induktionsschritt:

Es gelte die Behauptung für $l \in \mathbb{N}$. Es gilt einerseits

$$\begin{aligned} m_{l+1}^+ &= \max_{i \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\ &= \max_{i \in M} \{ \sum_{j \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l} = j, X_{-l-1} = i) \\ &\qquad \cdot P(X_{-l} = j | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \} \\ &= \max_{i \in M} \{ \sum_{j \in M} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l} = j) \\ &\qquad P(X_{-l} = j | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \} \\ &= \max_{i \in M} \{ \sum_{\substack{j \in M \\ j \neq arg(m_l^-)}} P(X_0 \in A | Y_{-n}^0 = y_{-n}^0, X_{-l} = j) \\ &\qquad \cdot P(X_{-l} = j | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\ &\qquad + m_l^- \cdot P(X_{-l} = arg(m_l^-) | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \} \\ &\leq \max_{i \in M} \{ \sum_{\substack{j \in M \\ j \neq arg(m_l^-)}}^{n} m_l^+ \cdot P(X_{-l} = j | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\ &\qquad + m_l^- \cdot P(X_{-l} = arg(m_l^-) | Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \} \end{aligned}$$

$$
\begin{aligned}
&= \max_{i\in M}\{m_l^+ \cdot (1 - P(X_{-l} = arg(m_l^-)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)) \\
&\qquad + m_l^- \cdot \underbrace{P(X_{-l} = arg(m_l^-)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)}_{\geq \hat{\eta}_{-l}}\} \\
&\leq \max_{i\in M}\{m_l^+(1-\hat{\eta}_{-l}) + m_l^-\hat{\eta}_{-l}\} \\
&= m_l^+(1-\hat{\eta}_{-l}) + m_l^-\hat{\eta}_{-l}\,,
\end{aligned}
$$

wobei die letzte Ungleichung das Vorliegen einer Konvexkombination ausnutzt.

Und andererseits gilt

$$
\begin{aligned}
m_{l+1}^- &= \min_{i\in M} P(X_0 \in A|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\
&= \min_{i\in M}\{\sum_{j\in M} P(X_0 \in A|Y_{-n}^0 = y_{-n}^0, X_{-l} = j, X_{-l-1} = i) \\
&\qquad \cdot P(X_{-l} = j|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)\} \\
&= \min_{i\in M}\{\sum_{j\in M} P(X_0 \in A|Y_{-n}^0 = y_{-n}^0, X_{-l} = j) \\
&\qquad P(X_{-l} = j|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)\} \\
&= \min_{i\in M}\{\sum_{\substack{j\in M \\ j\neq arg(m_l^+)}} P(X_0 \in A|Y_{-n}^0 = y_{-n}^0, X_{-l} = j) \\
&\qquad \cdot P(X_{-l} = j|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\
&\qquad + m_l^+ \cdot P(X_{-l} = arg(m_l^+)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)\} \\
&\geq \min_{i\in M}\{\sum_{\substack{j\in M \\ j\neq arg(m_l^+)}}^{n} m_l^- \cdot P(X_{-l} = j|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i) \\
&\qquad + m_l^+ \cdot P(X_{-l} = arg(m_l^+)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)\} \\
&= \min_{i\in M}\{m_l^- \cdot (1 - P(X_{-l} = arg(m_l^+)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i))
\end{aligned}
$$

$$+m_l^+ \cdot \underbrace{P(X_{-l} = arg(m_l^+)|Y_{-n}^0 = y_{-n}^0, X_{-l-1} = i)}_{\geq \hat{\eta}_{-l}}\}$$

$$\begin{aligned} &\geq \min_{i \in M}\{m_l^-(1-\hat{\eta}_{-l}) + m_l^+\hat{\eta}_{-l}\} \\ &= m_l^-(1-\hat{\eta}_{-l}) + m_l^+\hat{\eta}_{-l}\,. \end{aligned}$$

Damit folgt dann zusammenfassend:

$$I) \quad m_{l+1}^+ \leq (1-\hat{\eta}_{-l})m_l^+ + \hat{\eta}_{-l}m_l^- \ ,$$

$$II) \quad m_{l+1}^- \geq (1-\hat{\eta}_{-l})m_l^- + \hat{\eta}_{-l}m_l^+ \ ,$$

und schließlich

$$\begin{aligned} I)-II) \quad m_{l+1}^+ - m_{l+1}^- &\leq (1-2\hat{\eta}_{-l})m_l^+ - (1-2\hat{\eta}_{-l})m_l^- \\ &= (1-2\hat{\eta}_{-l})(m_l^+ - m_l^-) \\ &\leq (1-2\hat{\eta}_{-l}) \prod_{j=-l+1}^{0} (1-2\hat{\eta}_j) \\ &= \prod_{j=t-l}^{t} (1-2\hat{\eta}_j)\,. \end{aligned}$$

□

5.3.5 Bemerkung
Für die Abschätzung des später auftretenden Klassifikationsrisikos benötigen wir in Lemma 5.3.4 den Fall -$n \leq -l$. In diesem Fall muss man jedes $\hat{\eta}_j$ durch $\eta(y_j)$ ersetzen, so dass die Abhängigkeit von den Beobachtungen $y_0, \cdots y_{-n}$ erhalten bleibt. Das ist der wesentliche Unterschied zum Fall des diskreten Merkmalsraumes.

Das folgende Korollar beschreibt eine Ungleichung, die für die Abschätzung von Klassifikationsrisiken von Bedeutung ist.

5.3.6 Korollar
Für alle $l, k \in \mathbb{N}$, $A \subseteq M$ *und* $y_0, y_{-1}, \ldots \in \mathcal{Y}$ *gilt:*

$$|P(X_0 \in A|Y_{-l}^0 = y_{-l}^0) - P(X_0 \in A|Y_{-l-k}^0 = y_{-l-k}^0)| \leq \prod_{j=-l}^{0} (1-2\eta(y_j)) \ .$$

Beweis:

Es gilt:

$$
\begin{aligned}
& |P(X_0 \in A|Y_{-l}^0 = y_{-l}^0) - P(X_0 \in A|Y_{-l-k}^0 = y_{-l-k}^0)| \\
= \; & |\sum_{\iota \in M} P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = \iota) P(X_{-l-1} = \iota|Y_{-l}^0 = y_{-l}^0) \\
& \quad - \sum_{\kappa \in M} P(X_0 \in A|Y_{-l-k}^0 = y_{-l-k}^0, X_{-l-1} = \kappa) P(X_{-l-1} = \kappa|Y_{-l-k}^0 = y_{-l-k}^0)| \\
= \; & |\sum_{\iota \in M} P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = \iota) P(X_{-l-1} = \iota|Y_{-l}^0 = y_{-l}^0) \\
& \quad - \sum_{\kappa \in M} P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = \kappa) P(X_{-l-1} = \kappa|Y_{-l-k}^0 = y_{-l-k}^0)| \\
\leq \; & \max_{i,j \in M} |P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \cdot \sum_{\iota \in M} P(X_{-l-1} = \iota|Y_{-l}^0 = y_{-l}^0) \\
& \quad - P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = j) \cdot \sum_{\kappa \in M} P(X_{-l-1} = \kappa|Y_{-l-k}^0 = y_{-l-k}^0)| \\
= \; & \max_{i,j \in M} |P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \\
& \quad - P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = j)| \\
= \; & \max_{i \in M} P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \\
& \quad - \min_{i \in M} P(X_0 \in A|Y_{-l}^0 = y_{-l}^0, X_{-l-1} = i) \\
= \; & m_{l+1}^+ - m_{l+1}^- \\
\leq \; & \prod_{j=-l}^{0} (1 - 2\hat{\eta}_j) \; .
\end{aligned}
$$

□

5.3.7 Bemerkung

Man sieht also, dass im Gegensatz zum Fall mit diskretem Merkmalsraum die Differenz

$$|P(X_0 \in A|Y_{-l}^0 = y_l^0) - P(X_0 \in A|Y_{-l-k}^0 = y_{-l-k}^0)|$$

von der Realisierung von Y_{-l}^0 abhängt, wobei die $\eta(y_j)$ in ungünstigen Fällen sehr kleine Werte annehmen können.

5.3.8 Das optimale Klassifikationsrisiko

Im momentan betrachteten Fall des kontinuierlichen Merkmalraumes ist das optimale Klassifikationsrisiko für $Z_0^l = (Y_{-1}, \ldots, Y_{-l})$ nach (1.7) gegeben durch

$$R(Z_0^l) = \int_{\mathcal{Y}^l} \min_{i \in M} P(X_0 \neq i | y_{-l}^0) \cdot f(y_{-l}^0) dy_{-l}^0 \ ,$$

wobei l wieder die Erinnerungstiefe und f die Dichte von $P^{Y_{-l}^0}$ bzgl. μ ist.

Bevor nun eine Abschätzung zweier chronometrischer Musterklassifikationsmodelle wie im Satz (5.2.9) geschehen kann, werden noch zwei einfache Aussagen benötigt.

5.3.9 Lemma
Für alle $x \in M$ gilt mit den Abkürzungen

$$c = \max_{\iota,\kappa,i,k} \frac{p_{i\iota} \cdot p_{\iota k}}{p_{i\kappa} \cdot p_{\kappa k}}$$

und

$$\beta = \frac{1}{mc} \int_{\mathcal{Y}} \min_{\iota,\kappa} \frac{(f_\kappa(y))^2}{f_\iota(y)} dy \ :$$

$$\int_{\mathcal{Y}} \frac{1}{1 + (m-1)\alpha(y)} f_x(y) dy \geq \beta > 0 \, .$$

5.3.10 Bemerkung
Für die im Lemma auftretenden Terme gilt, dass $c \geq 1$ und $\beta \in (0, \frac{1}{2}]$ ist.

Beweis:

Wegen

$$\begin{aligned} \alpha(y) &= \max_{1 \leq \iota,\kappa,i,k \leq m} \frac{p_{i\iota} p_{\iota k} f_\iota(y)}{p_{i\kappa} p_{\kappa k} f_\kappa(y)} \\ &\leq \underbrace{\max_{1 \leq \iota,\kappa,i,k \leq m} \frac{p_{i\iota} p_{\iota k}}{p_{i\kappa} p_{\kappa k}}}_{=c} \cdot \max_{1 \leq \iota,\kappa \leq m} \frac{f_\iota(y)}{f_\kappa(y)} \\ &= c \cdot \max_{1 \leq \iota,\kappa \leq m} \frac{f_\iota(y)}{f_\kappa(y)} \end{aligned}$$

gilt

$$\begin{aligned}
\int_{\mathcal{Y}} \frac{1}{1+(m-1)\alpha(y)} f_x(y)dy &\geq \int_{\mathcal{Y}} \frac{1}{\alpha(y)+(m-1)\alpha(y)} f_x(y)dy \\
&= \frac{1}{m}\int_{\mathcal{Y}} \frac{1}{\alpha(y)} f_x(y)dy \\
&\geq \frac{1}{mc}\int_{\mathcal{Y}} \frac{1}{\max_{1\leq\iota,\kappa\leq m} \frac{f_\iota(y)}{f_\kappa(y)}} f_x(y)dy \\
&= \frac{1}{mc}\int_{\mathcal{Y}} \frac{\min_\kappa f_\kappa(y)}{\max_\iota f_\iota(y)} f_x(y)dy \\
&\geq \frac{1}{mc}\int_{\mathcal{Y}} \frac{\min_\kappa f_\kappa(y)}{\max_\iota f_\iota(y)} \min_\kappa f_\kappa(y)dy \\
&= \frac{1}{mc}\int_{\mathcal{Y}} \min_{\iota,\kappa} \frac{(f_\kappa(y))^2}{f_\iota(y)} dy \\
&> 0\,.
\end{aligned}$$

□

5.3.11 Korollar
Es gilt für alle $x \in M$ mit den Abkürzungen

$$\beta = \frac{1}{mc}\int_{\mathcal{Y}} \min_{\iota,\kappa} \frac{(f_\kappa(y))^2}{f_\iota(y)} dy$$

und

$$\gamma = 1 - 2\beta \ :$$

$$\int_{\mathcal{Y}} (1-2\eta(y)) \cdot f_x(y)dy \leq \gamma \ ,$$

wobei $\gamma \in [0,1)$.

Beweis:

Sei $y \in \mathcal{Y}$. Dann ist

$$1 - 2\eta(y) = 1 - \frac{2}{1+(m-1)\alpha(y)} \ .$$

Also folgt

$$\begin{aligned}\int_{\mathcal{Y}}(1-2\eta(y))\cdot f_x(y)dy &= \int_{\mathcal{Y}}1\cdot f_x(y)dy - 2\int_{\mathcal{Y}}\frac{1}{1+(m+1)\alpha(y)}\cdot f_x(y)dy \\ &\leq 1-2\beta\ ,\end{aligned}$$

für ein $\beta \in (0, \frac{1}{2}]$.

□

Mit den beiden letzten Aussagen lässt sich nun eine zum diskreten Fall analoge Behandlung der Differenz zweier optimaler Klassifikationsrisiken gewinnen.

5.3.12 Satz
Sei $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ ein chronometrisches Musterklassifikationsmodell mit $Z_t = (Y_t, Y_{t-1}, \ldots)$ für alle $t \in \mathbb{Z}$.

Dann existiert ein $\gamma \in (0,1)$, so dass für alle $l, k \in \mathbb{N}$ und alle eingeschränkten Modelle $Cr(Z^l)$ und $Cr(Z^{l+k})$ zu $Cr(\Omega, \mathcal{A}, P, X, Y, Z)$ mit

$$Z_0^l = (Y_{-1}, \ldots, Y_{-l}) \text{ und } Z_0^{l+k} = (Y_{-1}, \ldots, Y_{-l-1})$$

gilt

$$R(Z_0^l) - R(Z_0^{l+k}) \leq \gamma^{l+1}\ .$$

Beweis:

Wegen

$$f(y_{-l-k}^0) = \sum_{x_{-l-k}^0} P(X_{-l-k}^0 = x_{-l-k}^0)\cdot f_{x_{-l-k}^0}(y_{-l-k}^0)$$

und

$$f_{x_{-l-k}^0}(y_{-l-k}^0) = \prod_{j=-l-k}^{0} f_{x_j}(y_j)$$

gilt

$R(Z_0^l) - R(Z_0^{l+k})$

$$= \int_{\mathcal{Y}^{l+1}} \min_{i\in M} P(X_0 \neq i | y_{-l}^0)\cdot f(y_{-l}^0)dy_{-l}^0$$

$$- \int_{\mathcal{Y}^{l+k+1}} \min_{i\in M} P(X_0 \neq i | y_{-l-k}^0)\cdot f(y_{-l-k}^0)dy_{-l-k}^0$$

$$
\begin{aligned}
&= \int_{\mathcal{Y}^{l+1}} \min_{i\in M} P(X_0 \neq i|y_{-l}^0) \cdot \int_{\mathcal{Y}^k} f(y_{-l-k}^0) dy_{-l-k}^{-l-1} dy_{-l}^0 \\
&\quad - \int_{\mathcal{Y}^{l+k+1}} \min_{i\in M} P(X_0 \neq i|y_{-l-k}^0) \cdot f(y_{-l-k}^0) dy_{-l-k}^0 \\
&= \int_{\mathcal{Y}^{l+k+1}} \min_{i\in M} P(X_0 \neq i|y_{-l}^0) \cdot f(y_{-l-k}^0) dy_{-l-k}^0 \\
&\quad - \int_{\mathcal{Y}^{l+k+1}} \min_{i\in M} P(X_0 \neq i|y_{-l-k}^0) \cdot f(y_{-l-k}^0) dy_{-l-k}^0 \\
&= \int_{\mathcal{Y}^{l+k+1}} (\min_{i\in M} P(X_0 \neq i|y_{-l}^0) - \min_{i\in M} P(X_0 \neq i|y_{-l-k}^0)) \cdot f(y_{-l-k}^0) dy_{-l-k}^0 \\
&\leq \int_{\mathcal{Y}^{l+k+1}} \prod_{j=-l}^{0} (1-2\eta(y_j)) \cdot f(y_{-l-1}^0) dy_{-l-1}^0 \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \int_{\mathcal{Y}^{l+k+1}} \prod_{j=-l}^{0} (1-2\eta(y_j)) \cdot f_{x_{-l-k}^0}(y_{-l-k}^0) dy_{-l-k}^0\} \qquad (*) \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \int_{\mathcal{Y}^{l+k+1}} \prod_{j=-l}^{0} (1-2\eta(y_j)) \cdot \prod_{j=-l-k}^{0} f_{x_j}(y_j) dy_{-l-k}^0\} \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \int_{\mathcal{Y}^{l+k+1}} \prod_{j=-l}^{0} (1-2\eta(y_j)) f_{x_j}(y_j) \cdot f_{x_{-l-k}^{-l-1}}(y_{-l-k}^{-l-1}) dy_{-l-k}^0\} \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \int_{\mathcal{Y}^{l+1}} \prod_{j=-l}^{0} (1-2\eta(y_j)) f_{x_j}(y_j) dy_{-l}^0\} \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \prod_{j=-l}^{0} \int_{\mathcal{Y}} (1-2\eta(y_j)) f_{x_j}(y_j) dy_j\} \\
&\leq \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0) \prod_{j=-l}^{0} \gamma\} \\
&= \sum_{x_{-l-k}^0} \{P(X_{-l-k}^0 = x_{-l-k}^0)\} \gamma^{l+1} \\
&= \gamma^{l+1} \ .
\end{aligned}
$$

Da der Integrand in der Zeile $(*)$ sich als Produkt darstellen lässt und η nur von y_j abhängt, kann das Integral über $\mathcal{Y}^{l+k+1}$ in eine Produkt von Integralen

über $\mathcal{Y}$ überführt werden.

□

Wie im vorherigen Abschnitt ergibt sich mit dem Satz folgendes Korollar.

5.3.13 Korollar
Es existiert ein $\gamma \in (0,1)$ *mit*

$$R^* := \lim_{n\to\infty} R(Z_0^n) \leq R(Z_0^l) \leq R^* + \gamma^{l+1} \ .$$

5.3.14 Bemerkung
Es liegt also im kontinuierlichen sowie im diskreten Fall ein exponentieller Abfall des Klassifikationsrisikos mit wachsender Erinnerungstiefe vor. Bei Anwendungen lässt sich, falls die Rate explizit berechnet werden kann, somit eine Erinnerungstiefe wählen, so dass die Abweichung vom optimalen Klassifikationsrisiko R^* einen vorgegeben Wert nicht überschreiten kann. Dieses kann eine erhebliche Ersparnis an Rechenaufwand und somit an Zeitaufwand bedeuten.

5.3.15 Beispiel

Seien $M = \{0,1\}$, $\mathcal{Y} = \mathbb{R}$. Weiter seien $P^{Y_t|X_t=0}$ $N(0,1)$-verteilt und $P^{Y_t|X_t=1}$ $N(\mu,1)$-verteilt. Dann ist zur Bestimmung der Rate $\gamma = 1-2\beta$ zunächst einmal

$$\beta = \frac{1}{mc}\int_{\mathbb{R}} \min_{\iota,\kappa} \frac{(f_\iota(y))^2}{f_\kappa(y)} dy \tag{5.3}$$

mit $c = \max_{i,j,a,b} \frac{p_{ia}\cdot p_{aj}}{p_{ib}\cdot p_{bj}}$ und den Dichten $f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$ zu berechnen. Die Graphen der Dichten und der Funktion $x \mapsto \min_{\iota,\kappa} \frac{(f_\iota(x))^2}{f_\kappa(x)}$ sind in den Abbildungen 5.3 und 5.4 mit $\mu = 1$ bzw. mit $\mu = 2$ dargestellt. Der Schnittpunkt der beiden Dichten liegt offensichtlich bei $\frac{\mu}{2}$, so dass sich für das Integral in (5.3) ergibt

$$\begin{aligned}
\int_{\mathbb{R}} \min_{\iota,\kappa} \frac{(f_\iota(y))^2}{f_\kappa(y)} dy &= \int_{-\infty}^{\frac{\mu}{2}} \frac{(f_1(y))^2}{f_0(y)} dy + \int_{\frac{\mu}{2}}^{\infty} \frac{(f_0(y))^2}{f_1(y)} dy \\
&= \frac{1}{\sqrt{2\pi}} e^{\mu^2} \int_{-\infty}^{\frac{\mu}{2}} e^{-\frac{1}{2}(y-2\mu)^2} dy + \frac{1}{\sqrt{2\pi}} e^{\mu^2} \int_{\frac{\mu}{2}}^{\infty} e^{-\frac{1}{2}(y+\mu)^2} dy \\
&= \frac{1}{\sqrt{2\pi}} e^{\mu^2} \int_{-\infty}^{-\frac{3\mu}{2}} e^{-\frac{x^2}{2}} dx + \frac{1}{\sqrt{2\pi}} e^{\mu^2} \int_{\frac{3\mu}{2}}^{\infty} e^{-\frac{x^2}{2}} dx \\
&= 2\cdot e^{\mu^2} \cdot \Phi(-\frac{3}{2}\mu) \,,
\end{aligned}$$

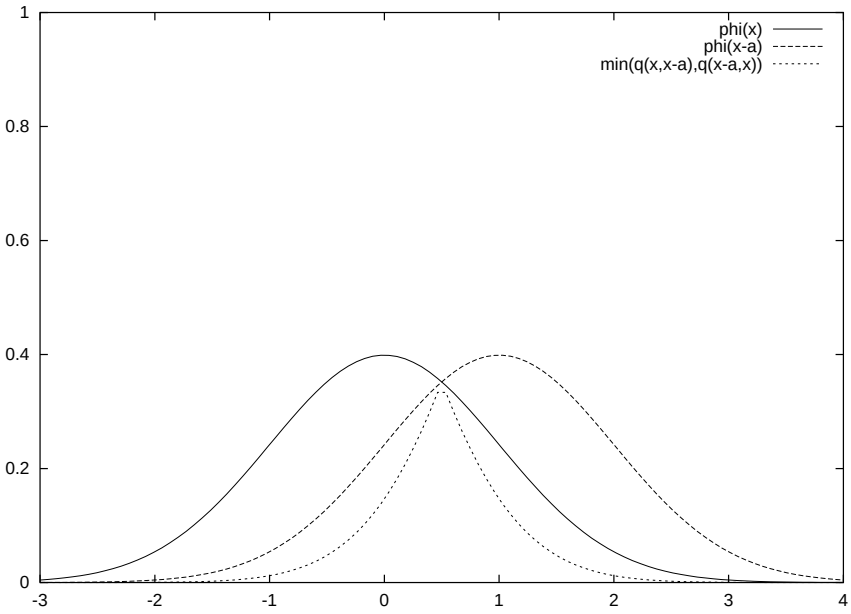

Abbildung 5.3: $\mu = 1$

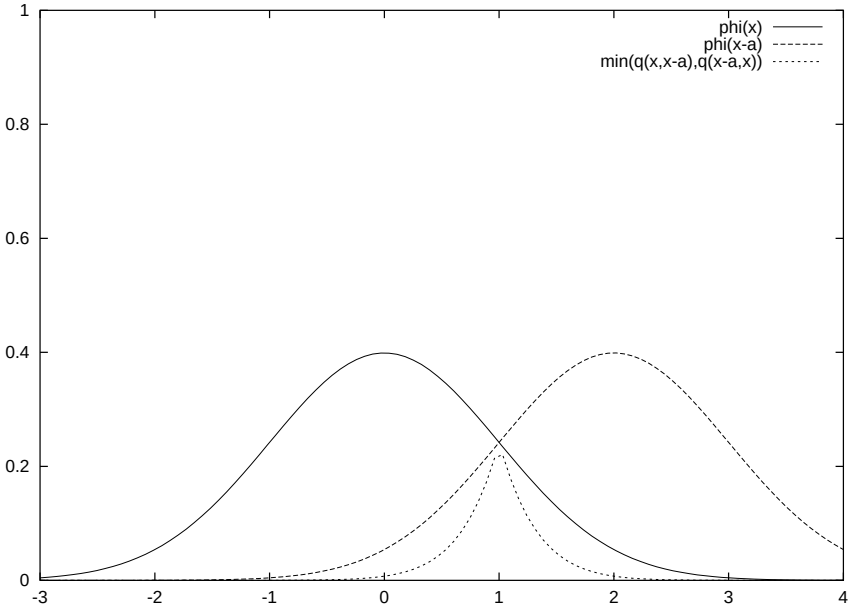

Abbildung 5.4: $\mu = 2$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Für $\mu = 1$ und $\mu = 2$ ergeben sich damit die Werte $0,3632$ und $0,01995$ für das Integral, woraus sich mit der Übergangsmatrix P schließlich das γ bestimmen lässt.

Als Beispiel sei hier die Matrix $P = \begin{pmatrix} 0,8 & 0,2 \\ 0,3 & 0,7 \end{pmatrix}$ betrachtet. Für $\mu = 1$ ergibt sich ein Rate $\gamma = 0.983$ und für $\mu = 2$ eine Rate $\gamma = 0.999$.

Teil III

Algorithmen

Kapitel 6

Algorithmen

Eine zentrale Größe in der Betrachtung von optimalen Klassifikationsrisiken ist bei gegebenen Werten $s \in M$ und $y_0, y_{-1}, y_{-2}, \ldots \in \mathcal{Y}$ die Folge

$$P(X_0 = s|Y_{-l}^0 = y_{-l}^0)\,,\ l \in \mathbb{N}\,. \tag{6.1}$$

Um in Anwendungen optimale Klassifikationsrisiken abschätzen zu können, ist es nötig zumindest für eine vorgegebene Erinnerungstiefe $\lambda \in \mathbb{N}$ die Werte der Folge

$$P(X_0 = s|Y_{-l}^0 = y_{-l}^0),\ l = 0, \ldots, \lambda$$

berechnen zu können.

In diesem Kapitel werden für den einfachen Fall eines diskreten Merkmalraumes $\mathcal{Y}$ zwei Rekursionsformeln für die Folge (6.1) hergeleitet. Dabei werden die Notationen aus Abschnitt 5.2 verwendet.

Die direkte Berechnung von $P(X_0 = s|Y_{-l}^0 = y_{-l}^0)$ für ein festes $l \in \mathbb{N}$ erfordert wegen

$$\begin{aligned} &P(X_0 = s|Y_{-l}^0 = y_{-l}^0) \\ &\quad = \frac{\displaystyle\sum_{x_{-l}^{-1} \in M^l} \prod_{i=-l}^{-1} \{q_{x_i y_i}\} \cdot q_{s y_0} \cdot \prod_{i=-l}^{-2} \{p_{i,i+1}\} p_{x_{-1} s} \cdot P(X_{-l} = x_{-l})}{\displaystyle\sum_{x_0 \in M} \sum_{x_{-l}^{-1} \in M^l} \prod_{i=-l}^{-1} \{q_{x_i y_i}\} \cdot q_{x_0 y_0} \cdot \prod_{i=-l}^{-2} \{p_{i,i+1}\} p_{x_{-1} x_0} \cdot P(X_{-l} = x_{-l})} \end{aligned}$$

alleine im Nenner schon $|M|^{l+1}$ Summationen, so dass der Rechenaufwand für große $l \in \mathbb{N}$ selbst mit modernster Technik nicht bewältigt werden kann.

Im nächsten Abschnitt wird eine Vorwärtsrekursion vorgestellt, die es ermöglicht die Terme $P(X_{m+2} = s|Y_1^{m+1} = y_1^{m+1})$ aus den Größen $P(X_{m+1} = v|Y_1^m = y_1^m)$, $v \in M$ zu bestimmen.

6.1 Vorwärtsrekursion

6.1.1 Lemma
Sei $s \in M$ und seien $y_0, y_1, y_2, \ldots \in \mathcal{Y}$.

Definiere

$$h_0(s) = P(X_1 = s|Y_0 = y_0)$$

und rekursiv für alle $m \in \mathbb{N}$

$$h_m(s) = P(X_{m+1} = s|Y_0^m = y_0^m)\,.$$

Dann gilt für alle $m \in \mathbb{N}$

$$h_{m+1}(s) = \frac{\sum\limits_{v\in M} p_{vs} \cdot q_{vy_{m+1}} \cdot h_m(v)}{\sum\limits_{v\in M} q_{vy_{m+1}} \cdot h_m(v)}\,. \tag{6.2}$$

Beweis:

Es gilt für alle $m \in M$

$h_{m+1}(s)$

$$\begin{aligned}
&= P(X_{m+2} = s|Y_0^{m+1} = y_0^{m+1}) \\
&= \sum_{v\in M} P(X_{m+2} = s|X_{m+1} = v) \cdot P(X_{m+1} = v|Y_0^{m+1} = y_0^{m+1}) \\
&= \frac{\sum_{v\in M} p_{vs} \cdot P(X_{m+1} = v, Y_0^{m+1} = y_0^{m+1})}{P(Y_0^{m+1} = y_0^{m+1})} \\
&= \frac{\sum_{v\in M} p_{vs} \cdot P(Y_0^{m+1} = y_0^{m+1}|X_{m+1} = v) \cdot P(X_{m+1} = v)}{\sum_{v\in M} P(Y_0^{m+1} = y_0^{m+1}|X_{m+1} = v) \cdot P(X_{m+1} = v)} \\
&= \frac{\sum_{v\in M} p_{vs} \cdot P(Y_{m+1} = y_{m+1}|X_{m+1} = v) \cdot P(X_{m+1} = v, Y_0^m = y_0^m)}{\sum_{v\in M} P(Y_{m+1} = y_{m+1}|X_{m+1} = v) \cdot P(X_{m+1} = v, Y_0^m = y_0^m)}
\end{aligned}$$

$$= \frac{\sum_{v\in M} p_{vs} \cdot q_{vy_{m+1}} \cdot P(X_{m+1} = v | Y_0^m = y_0^m)}{\sum_{v\in M} q_{vy_{m+1}} \cdot P(X_{m+1} = v | Y_0^m = y_0^m)}$$

$$= \frac{\sum_{v\in M} p_{vs} \cdot q_{vy_{m+1}} \cdot h_m(s)}{\sum_{v\in M} q_{vy_{m+1}} \cdot h_m(s)} .$$

□

6.1.2 Lemma
Sei $s \in M$ und seien $y_0, y_1, y_2, \ldots \in \mathcal{Y}$.

Dann gilt

$$P(X_{m+1} = s | Y_0^{m+1} = y_0^{m+1}) = \frac{h_m(s) \cdot q_{sy_{m+1}}}{\sum_{v\in M} h_m(v) \cdot q_{vy_{m+1}}} .$$

Beweis:

Es gilt

$$\frac{h_m(s) \cdot q_{sy_{m+1}}}{\sum_{v\in M} h_m(v) \cdot q_{vy_{m+1}}}$$

$$= \frac{P(X_{m+1} = s | Y_1^m = y_1^m) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = s)}{\sum_{v\in M} P(X_{m+1} = v | Y_1^m = y_1^m) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = v)}$$

$$= \frac{P(X_{m+1} = s, Y_1^m = y_1^m) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = s)}{\sum_{v\in M} P(X_{m+1} = v, Y_1^m = y_1^m) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = v)}$$

$$= \frac{P(Y_1^m = y_1^m | X_{m+1} = s) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = s) \cdot P(X_{m+1} = s)}{\sum_{v\in M} P(Y_1^m = y_1^m | X_{m+1} = v) \cdot P(Y_{m+1} = y_{m+1} | X_{m+1} = v) \cdot P(X_{m+1} = v)}$$

$$= \frac{P(Y_1^{m+1} = y_1^{m+1} | X_{m+1} = s) \cdot P(X_{m+1} = s)}{\sum_{v\in M} P(Y_1^{m+1} = y_1^{m+1} | X_{m+1} = v) \cdot P(X_{m+1} = v)}$$

$$= \frac{P(Y_1^{m+1} = y_1^{m+1}, X_{m+1} = s)}{P(Y_1^{m+1} = y_1^{m+1})}$$

$$= P(X_{m+1} = s | Y_1^{m+1} = y_1^{m+1}) .$$

□

6.1.3 Bemerkung
Mit Hilfe der Rekursionsformel (6.2) lassen sich für ein $\mu \in \mathbb{N}$ die Folgen

$$(P(X_m = s|Y_0^m = y_0^m))_{m=0}^{\mu}\,,\ s \in M$$

simultan berechnen.

Offensichtlich erhält man damit durch geeignete Indexverschiebungen eine umständliche Methode zur Bestimmung der Werte des Folge

$$\big(P(X_0 = s|Y_{-l}^0 = y_{-l}^0)\big)_{l=0}^{\lambda}\,.$$

Es ist nicht möglich eine direkte Rückwärtsrekursion anzugeben, die es ermöglicht aus den Termen $P(X_0 = s|Y_{-l}^0 = y_{-l}^0)$, $s \in M$ rekursiv die Terme $P(X_0 = s|Y_{-l-1}^0 = y_{-l-1}^0)$, $s \in M$ zu bestimmen. Um trotzdem zu einer Rekursionsvorschrift zu gelangen, kann mit Hilfe der Formel von Bayes ein Umweg über eine Ruckwärtsinduktion der Folgen

$$P(Y_{-l}^0 = y_{-l}^0, X_{-l} = j|X_0 = i),\ l \in \mathbb{N}_0$$

gemacht werden. Dieses ist der Inhalt des nächsten Abschnittes.

6.2 Rückwärtsrekursion

Im folgenden Lemma wird eine Rekursionsvorschrift beschrieben, mit der indirekt bei gegebener Folge $y_0, y_{-1}, y_{-2}, \ldots \in \mathcal{Y}$ die Folge

$$\big(P(X_0 = s|Y_{-l}^0 = y_{-l}^0)\big)_{l\in\mathbb{N}_0}$$

beliebig weit berechnen kann.

6.2.1 Lemma
Definiere für alle $i, j \in M$

$$a_0(j|i) = P(Y_0 = y_0|X_0 = i) \cdot 1_{\{i=j\}}$$

und für alle $i, j \in M$ und $l \in \mathbb{N}$

$$a_l(j|i) = P(Y_{-l}^0 = y_{-l}^0, X_{-l} = j|X_0 = i)\,.$$

Dann gilt für alle $l \in \mathbb{N}$ und $i, j \in M$

$$a_{l+1}(j|i) = q_{jy_{-l-1}} \cdot \pi_j \sum_{k\in M} \frac{p_{jk}}{p_k} \cdot a_l(k|i)\,. \tag{6.3}$$

Beweis:

Es gilt für alle $l \in \mathbb{N}$ und $i, j \in M$

$a_{l+1}(j|i)$

$$
\begin{aligned}
&= P(Y_{-l-1}^0 = y_{-l-1}^0, X_{-l-1} = j | X_0 = i) \\
&= \sum_{k \in M} P(Y_{-l-1}^0 = y_{-l-1}^0, X_{-l} = k, X_{-l-1} = j | X_0 = i) \\
&= \sum_{k \in M} P(Y_{-l-1} = y_{-l-1}, X_{-l-1} = j | Y_{-l}^0 = y_{-l}^0, X_{-l} = k, X_0 = i) \\
&\quad \cdot P(Y_{-l}^0 = y_{-l}^0, X_{-l} = k | X_0 = i) \\
&= \sum_{k \in M} P(Y_{-l-1} = y_{-l-1} | X_{-l-1} = j) \cdot P(X_{-l-1} = j | X_{-l} = k) \cdot a_l(k|i) \\
&= \sum_{k \in M} q_{j y_{-l-1}} \cdot \pi_j \frac{p_{jk}}{p_k} \cdot a_l(k|i) \,.
\end{aligned}
$$

□

6.2.2 Bemerkung
Für $i \in M$ und $y_0, y_{-1}, y_{-2}, \ldots$ lassen sich die Folgen

$$(a_l(j|i))_{l \in \mathbb{N}_0} \ , \ j \in M$$

simultan berechnen. Bei der Umsetzung dieser Rekursion mit einem Computer ist allerdings zu beachten, dass es sich um Nullfolgen handelt und somit der beschränkten Rechengenauigkeit eines Computers mit einer Normierung in jedem Rekursionsschritt entgegengewirkt werden muss. Diese Normierungen werden sich aber bei der Bestimmung der Größen $P(X_0 = s | Y_{-l}^0 = y_{-l}^0)$ wieder herauskürzen, so dass diese Rekursionsvorschrift in Verbindung mit der Normierung auch in Anwendungen benutzt werden kann.

Mit der Formel von Bayes erhält man das folgende Lemma.

6.2.3 Lemma
Sei $s \in M$ und seien $y_0, y_{-1}, y_{-2}, \ldots \in \mathcal{Y}$.

Dann gilt für alle $l \in M$

$$P(X_0 = s | Y_{-l}^0 = y_{-l}^0) = \frac{\pi_s \cdot \sum_{j \in M} a_l(j|s)}{\sum_{t \in M} \pi_t \cdot \sum_{j \in M} a_l(j|t)} \,. \tag{6.4}$$

Beweis:

Es gilt

$$\frac{\pi_s \cdot \sum_{j\in M} a_l(j|s)}{\sum_{t\in M} \pi_t \cdot \sum_{j\in M} a_l(j|t)}$$

$$\begin{aligned} &= \frac{\pi_s \cdot \sum_{j\in M} P(Y_{-l}^0 = y_{-l}^0, X_{-l} = j|X_0 = s)}{\sum_{t\in M} \pi_t \cdot \sum_{j\in M} P(Y_{-l}^0 = y_{-l}^0, X_{-l} = j|X_0 = t)} \\ &= \frac{\pi_s \cdot P(Y_{-l}^0 = y_{-l}^0|X_0 = s)}{\sum_{t\in M} \pi_t \cdot P(Y_{-l}^0 = y_{-l}^0|X_0 = t)} \\ &= P(X_0 = s|Y_{-l}^0 = y_{-l}^0)\,. \end{aligned}$$

□

Zum Abschluss dieses Abschnittes wird noch der für die Anwendung erforderliche Algorithmus angegeben, wobei noch einmal angemerkt sei, dass sich die Normierungsfaktoren, die lediglich dazu dienen computerbedingte Rundungsfehler zu verringern, im Ausdruck (6.4) faktorisieren und sich somit kürzen lassen.

6.2.4 Algorithmus zur Berechnung von $P(X_0 = s|Y_{-\lambda}^0 = y_{-\lambda}^0)$

Sei $\lambda \in \mathbb{N}$.

1° Berechne für alle $i, j \in M$ die Startwerte $a_0(i|j)$, und setze $l = 0$.

2° Berechne für alle $i, j \in M$ mittels der Rekursionsvorschrift (6.3) die Werte $a_{l+1}(i|j)$.

3° Berechne den Normierungsfaktor

$$N = \sum_{i,j\in M} a_{l+1}(i|j),$$

und ersetze für sämtliche $i, j \in M$ die $a_{l+1}(i|j)$ durch $\frac{a_{l+1}(i|j)}{N}$.

4° Ist $l < \lambda$, so ersetze l durch $l + 1$ und gehe zu 2°.

5° Gib $P(X_0 = s|Y_{-l}^0 = y_{-l}^0) = \frac{p_s \cdot \sum_{j\in M} a_l(j|s)}{\sum_{t\in M} p_t \cdot \sum_{j\in M} a_l(j|t)}$ aus.

6.2.5 Bemerkung

In einem chronometrischem Musterklassifikationsmodell mit diskretem Merkmalsraum liegt zu einer Erinnerungstiefe $l \in \mathbb{N}$ das optimale Klassifikationsrisiko

$$R(Z_0^l) = \sum_{y_{-l}^0 \in \mathcal{Y}} \min_{s \in M} P(X_0 = s | Y_{-l}^0 = y_l^0) \cdot P(Y_{-l}^0 = y_l^0) \tag{6.5}$$

vor. In Abschnitt 6.2 wurde gezeigt, wie die kombinatorischen Explosion bei der Bestimmung der Summanden $\min_{s \in M} P(X_0 = s | Y_{-l}^0 = y_l^0)$ umgangen werden kann. Allerdings liefert die Bildung der Summen über die y_{-l}^0 in (6.5) wieder eine kombinatorische Explosion, so dass das optimale Klassifikationsrisiko nur für kleine $l \in \mathbb{N}$ exakt ausgerechnet werden kann.

Die Bildung des Minimums über die Musterklassen verhindert die Übertragung der Rekursionsvorschrift der bedingten Wahrscheinlichkeiten auf die optimalen Klassifikationsrisiken.

Als abschließendes Beispiel werden für den Fall $M = \{0, 1\}$ und $\mathcal{Y} = \{0, 1\}$ in einem konkreten Zahlenbeispiel die Rate des Modells und die optimalen Klassifikationsrisiken für die Erinnerungstiefen $l = 0, \ldots, 20$ berechnet.

6.2.6 Beispiel

Berechnet werden sollen die Werte der Folge

$$R(Z_0^l),\ l \in \{0, \ldots, 20\}\,,$$

wobei sich die optimalen Klassifikationsrisiken zu

$$R(Z_0^l) = \sum_{y_{-l}^0 \in \mathcal{Y}} \min_{s \in M} P(X_0 = s, Y_{-l}^0 = y_l^0)$$

ergeben. Dazu wird ähnlich zur Rückwärtsinduktion eine indirekte Rekursionsvorschrift für die Folge

$$\left(P(X_0 = s, Y_{-l}^0 = y_{-l}^0)\right)_{l \in \mathbb{N}}$$

bei gegebener Folge $y_0, y_{-1}, \ldots \in \mathbb{R}$ und gegebenem $s \in M$ hergeleitet, d.h. es wird eine Rekursionsvorschrift für eine Folge von Termen erstellt, mit denen man die gewünschten Wahrscheinlichkeiten simultan berechnen kann.

6.2.7 Lemma

Definiere für alle $l \in \mathbb{N}$ und $j \in M$

$$b_l(j, s) := P(X_0 = s, Y_{-l}^0 = y_{-l}^0 | X_{-l} = j)\,.$$

Dann gilt für alle $j \in M$

$$b_0(j,s) = q_{sy_0} \cdot 1_{\{j=s\}}$$

und für alle $l \in \mathbb{N}$ *und* $j \in M$

$$b_{l+1}(j,s) = \sum_{k \in M} b_l(k,s) \cdot q_{j,y_{-l-1}} \cdot p_{jk}\,.$$

Beweis:

Es gilt für alle $j \in M$

$$\begin{aligned} b_0(j,s) &= P(X_0 = s, Y_0 = y_0 | X_0 = j) \\ &= q_{sy_0} \cdot 1_{\{j=s\}}\,. \end{aligned}$$

Weiter gilt für alle $l \in \mathbb{N}$ und $j \in M$

$$\begin{aligned} b_{l+1}(j,s) &= P(X_0 = s, Y_{-l-1}^0 = y_{-l-1}^0 | X_{-l-1} = j) \\ &= \sum_{k \in M} P(X_0 = s, Y_{-l-1}^0 = y_{-l-1}^0, X_{-l} = k | X_{-l-1} = j) \\ &= \sum_{k \in M} P(X_0 = s, Y_{-l}^0 = y_{-l}^0 | X_{-l} = k) \cdot P(Y_{-l-1} = y_{-l-1} | X_{-l-j} = j) \\ &\qquad \cdot P(X_{-l} = k | X_{-l-1} = j) \\ &= \sum_{k \in M} b_l(k,s) \cdot q_{jy_{-l-1}} \cdot p_{jk}\,. \end{aligned}$$

□

6.2.8 Bemerkung

Mit dieser Rekursionsvorschrift lassen sich nun die Werte der Folge

$$P(X_0 = s, Y_{-l}^0 = y_{-l}^0),\ l \in \mathbb{N}$$

berechnen, denn es gilt

$$P(X_0 = s, Y_{-l}^0 = y_{-l}^0) = \sum_{j \in M} b_l(j,s) \cdot \pi_j\,.$$

Dabei ist wie oben anzumerken, dass die Folgen $(a_l(j,s))_{l \in \mathbb{N}}$ simultan für alle $j \in M$ berechnet werden müssen und dass diese Folgen Nullfolgen sind. Um der beschränkten Rechengenauigkeit eines Computers entgegen zu wirken, wird im unten aufgeführten Algorithmus eine Normierung auftauchen.

Das optimale Klassifikationsrisiko bestimmt sich für $l \in \mathbb{N}$ durch

$$R(Z_0^l) = \sum_{y_{-l}^0} \min_{s \in M} \sum_{j \in M} a_l(j,s) \cdot \pi_j \,.$$

Das bedeutet, dass weiterhin eine kombinatorische Explosion vorliegt, die nur für kleine $l \in \mathbb{N}$ ein exaktes Resultat zulässt.

6.2.9 Algorithmus zur Berechnung von $P(X_0 = s, Y_{-\lambda}^0 = y_{-\lambda}^0)$

Sei $\lambda \in \mathbb{N}$.

1° Berechne für alle $j \in M$ die Startwerte $a_0(j,s)$, und setze $l = 0$.

2° Berechne für alle $j \in M$ mittels der Rekursionsvorschrift die Werte $a_{l+1}(j,s)$.

3° Berechne den Normierungsfaktor

$$N_{l+1} = \sum_{j \in M} a_{l+1}(j,s),$$

und ersetze für sämtliche $j \in M$ die $a_{l+1}(j,s)$ durch $\frac{a_{l+1}(j,s)}{N_{l+1}}$.

4° Ist $l < \lambda$, so ersetze l durch $l+1$ und gehe zu 2°.

5° Gib $P(X_0 = s, Y_{-l}^0 = y_{-l}^0) = \sum_{j \in M} a_l(j,s) \cdot \pi_j \cdot \prod_{i=1}^{l} N_i$ aus.

6.2.10 Zahlenbeispiel 1

Die Übergangsmatrix

$$P = \begin{pmatrix} 0.12 & 0.88 \\ 0.89 & 0.11 \end{pmatrix}$$

und die Matrix der bedingten Wahrscheinlichkeiten

$$Q = \begin{pmatrix} 0,44 & 0,56 \\ 0,60 & 0,40 \end{pmatrix}$$

ergeben eine hohe Rate $\beta = 0.979$. Die daraus resultierenden optimalen Klassifikationsrisiken und der Graph der Abbildung $x \mapsto \beta^x$ sind in den Abbildungen 6.1 und 6.2 dargestellt. Die bei 1 liegende Rate β ist dafür verantwortlich, dass erst bei hohen Erinnerungstiefen die Werte β^l klein werden, so ist z.B. eine Erinnerungstiefe von $l = 109$ nötig bis der Wert von β^l kleiner als $0,1$ ist. Im Gegensatz dazu ist zu erkennen, dass schon bei einer Erinnerungstiefe ab $l = 10$ keine wesentlichen Verbesserungen mehr auftreten. Es ist natürlich nicht auszuschließen ist, dass eventuell bei größeren Erinnerungstiefen noch Sprünge auftreten können.

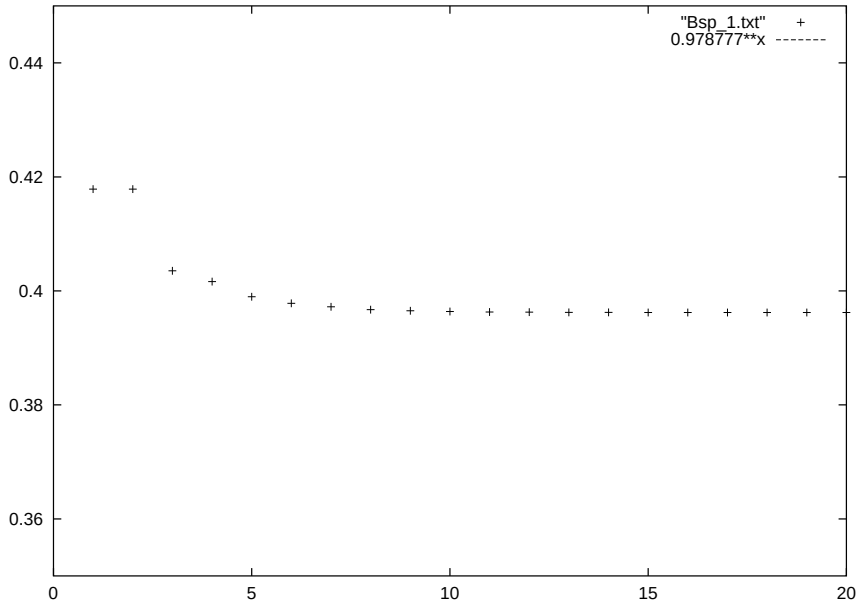

Abbildung 6.1: die optimalen Klassifikationsrisiken $R(Z_0^l)$, $l = 0, \ldots, 20$ im Intervall $[0,35, 0,45]$

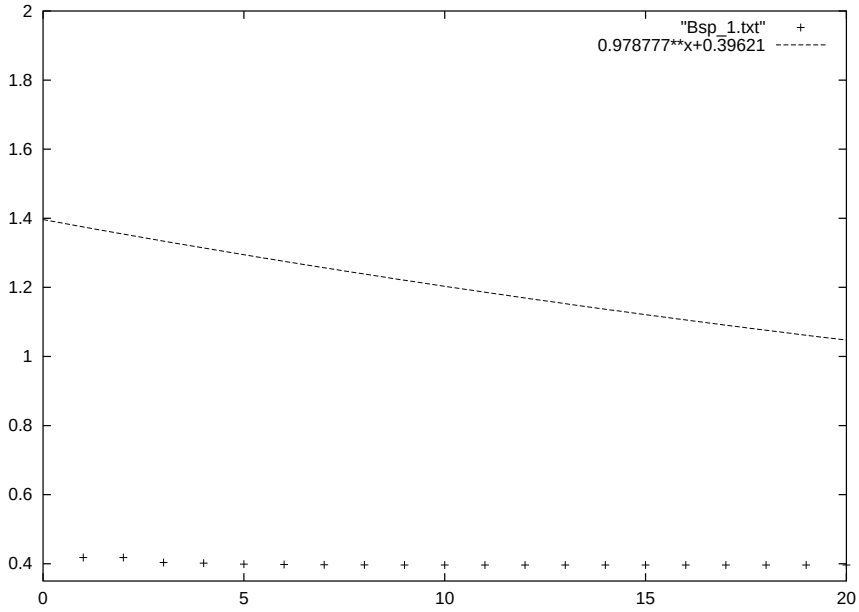

Abbildung 6.2: $x \mapsto \beta^x$ mit $\beta = 0.979$ und die optimalen Klassifikationsrisiken $R(Z_0^l)$, $l = 0, \ldots, 20$ im Intervall $[0,35, 2]$

6.2.11 Zahlenbeispiel 2

In einem zweiten Beispiel werden die Matrizen

$$P = \begin{pmatrix} 0,53 & 0,47 \\ 0,63 & 0,37 \end{pmatrix} \text{ und } Q = \begin{pmatrix} 0,47 & 0,53 \\ 0,56 & 0,44 \end{pmatrix}$$

betrachtet. Hieraus ergeben sich eine kleine Rate $\beta = 0,4307$ und die Abbildung 6.3.

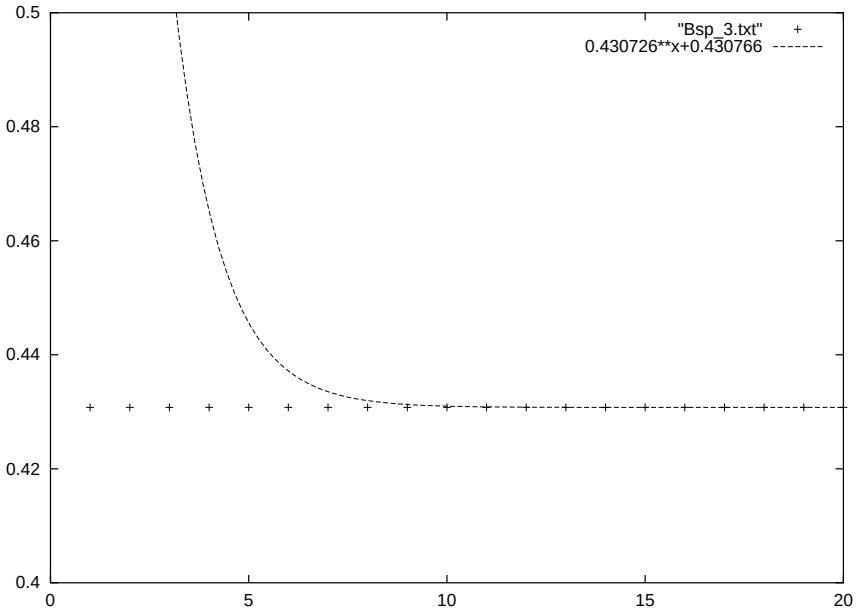

Abbildung 6.3: $x \mapsto \beta^x$ mit $\beta = 0,4307$ und die optimalen Klassifikationsrisiken $R(Z_0^l)$, $l = 0, \ldots, 20$ im Intervall $[0,4, 0,5]$

Anhang A

A.1 Resultate aus der Wahrscheinlichkeits- und Maßtheorie

A.1.1 Satz
Seien $\mathcal{C}_1$ und $\mathcal{C}_2$ Unter-σ-Algebren von $\mathcal{A}$, $\mathcal{C} := \sigma(\mathcal{C}_1, \mathcal{C}_2)$ die von $\mathcal{C}_1$ und $\mathcal{C}_2$ erzeugte σ-Algebra und X eine integrierbare Zufallsvariable. Ist dann die von X und $\mathcal{C}_1$ erzeugte σ-Algebra $\sigma(X, \mathcal{C}_1)$ unabhängig von $\mathcal{C}_2$, so gilt fast sicher

$$E(X|\mathcal{C}) = E(X|\mathcal{C}_1)\,.$$

A.1.2 Korollar
Seien X, Y, Z Zufallsvariablen mit den Wertebereichen $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ und sei $h : \mathcal{X} \to \mathcal{H}$ eine messbare Abbildung.

Ist Z stochastisch unabhängig von (X, Y), so gilt

$$E(h(X)|Y, Z) = E(h(X)|Y)\,.$$

A.1.3 Satz
Seien $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $\mathcal{F}$ eine Unter-σ-Algebra von $\mathcal{A}$ und $X, Y, X_1, X_2, \ldots$ nichtnegative bzw. integrierbare Zufallsgrößen.

Dann gilt

(i) $E(\alpha X + \beta Y|\mathcal{F}) = \alpha E(X|\mathcal{F}) + \beta E(Y|\mathcal{F})$ *P-f.s. für alle* $\alpha, \beta \geq 0$ *bzw.* $\alpha\,\beta \in \mathbb{R}$.

(ii) $X \leq Y$ *P-f.s. impliziert* $E(X|\mathcal{F}) \leq E(Y|\mathcal{F})$ *P-f.s.*

(iii) $E(E(X|\mathcal{F})) = EX$ *und* $|E(X|\mathcal{F})| \leq E(|X||\mathcal{F})$ P*-f.s.*

(iv) $X_n \uparrow X$ P*-f.s. impliziert* $E(X_n|\mathcal{F}) \uparrow E(X|\mathcal{F})$ P*-f.s.*

(v) $X_n \to X$ P*-f.s. und* $\sup_{n\geq 1} |X_n| \in Y \in \mathcal{L}^1$ *impliziert* $E(X_n|\mathcal{F}) \to E(X|\mathcal{F})$ P*-f.s.*

A.1.4 Satz
Seien $(\Omega, \mathcal{A}, P)$ *ein Wahrscheinlichkeitsraum,* $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$ *Unter-*σ*-Algebren von* $\mathcal{A}$ *und* X, Y *Zufallsgrößen auf* $(\Omega, \mathcal{A}, P)$*.*

Dann gilt

(i) Sind X *und* Y *nichtnegativ bzw.* $X, Y \in \mathcal{L}^2$ *so gilt*

$$Y \ \mathcal{F}\text{-messbar impliziert } E(XY|\mathcal{F}) = YE(X|\mathcal{F}) \ P\text{-f.s.}$$

(ii) Ist X *integrierbar, so gilt*

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \text{ impliziert } E(E(X|\mathcal{F}_1)|\mathcal{F}_2) = E(E(X|\mathcal{F}_2)|\mathcal{F}_1) = E(X|\mathcal{F}_1) \ P\text{-f.s.}$$

A.1.5 Satz
Ungleichung von Jensen für bedingte Erwartungswerte

Es seien X *eine Zufallsgröße mit Werten in* (a, b)*,* $-\infty \leq a < b \leq +\infty$ *und* $\varphi : (a, b) \to \mathbb{R}$ *eine konvexe Funktion mit* $E|X| < \infty$ *und* $E|\varphi(X)| < \infty$*. Ferner sei* $\mathcal{F}$ *eine Unter-*σ*-Algebra. Dann gilt*

$$\varphi(E(X|\mathcal{F})) \leq E(\varphi(X)|\mathcal{F})) \,.$$

A.1.6 Satz
Seien X *und* Y *Zufallsvariablen auf einem Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$ *und* $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ *eine messbare Abbildung. Ferner sei* Z *eine Abbildung auf* $\mathcal{X}$ *definiert durch*

$$Z(x) := \int_{\mathcal{Y}} h(x, y) P^{Y|X=x}(dy) \ .$$

Dann gilt:

$Z(X)$ *ist eine Version des bedingten Erwartungswertes* $E(h(X, Y)|X)$*.*

A.1.7 Korollar
Mit den Notationen des vorigen Satzes gilt

$$E\, h(X, Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} h(x, y) P^{Y|X=x}(dy) P^X(dx) \ .$$

Literaturverzeichnis

[Als] G. Alsmeyer; Stochastische Prozesse; 2002 Skripten zur Mathematischen Statistik des Fachbereichs Mathematik der Wilhelms-Universität Münster

[Bau] H.Bauer; Wahrscheinlichkeitstheorie. 1991 de Gruyter

[Ber-Smi] Bernado, Smith; Bayesian Theory. 2000 Wiley

[Bru] V. Bruhn; Ein Überblick über mathematische Grundlagen des Hidden-Markov-Modells. 1992 Diplomarbeit der mathematischen Fakultät der CAU

[Con] P. Congdon; Bayesian statistical modelling. 2001 Wiley

[Dev] L. Devroye, L.Györfi, G. Lugosi; A Probabilistic Theory of Pattern Recognition. 1996 Springer

[Hol] M. Holst; Der k-NN Klassifikator bei stochastisch abhängigen Lernfolgen. 1997 Dissertation an der CAU

[Irl] A. Irle; Statistische Methoden der Mustererkennung; (1992) Vorlesungsskript, CAU

[Keh] A. Kehagias, Bayesian classification of hidden markov models; mathematical and computer modelling, 23(5), 1996

[Kim] N.S. Kim, D.K. Kim; Filtering on Hidden Markov Models, IEEE signal processing, 7(9), p. 253, 2000

[Koc] K.R. Koch; Einführung in die Bayes-Statistik. 2000 Springer

[Ler] B.G. Leroux; Maximum-likelihood estimation for hidden Markov models. Stoch.Proc.Appl. 40, 127-43

[Mee] G. Meeden, S. Vardeman; A Simple Hidden Markov Model for Bayesian Modeling with Time Dependent Data. Communications in statistics: theory an methods, 29 (8), p. 1801, 2000

[Rob] C.P. Robert; The Bayesian choice. 2001 Springer

[Rob] C.P. Robert, T. Rydén, D.M. Titterington; Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. Journal of the royal statistical society. serie., 62(1), p. 57,2000

Index