Betrachtungen über ein distanzbasiertes Klassifikationsverfahren

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

im Studiengang Informatik

Friedrich-Schiller-Universität Jena
Fakultät für Mathematik und Informatik

eingereicht von Martin Friedrich Schirneck

geboren am 21. Februar 1990 in Gera

Betreuer: Prof. Dr. Joachim Giesen

Jena, im Sommersemester 2012

Abstract

In dieser Arbeit wird ein neuartiges distanzbasiertes Klassifikationsverfahren zur Kategorisierung numerischer Datenvektoren bei binärer Klassenteilung entworfen. Die dazu verwendeten Metriken basieren auf dem Volumen endlicher Vereinigungen d-dimensionaler Kugeln im Euklidischen Raum. Die angegebene Berechnungsvorschrift behandelt dabei konkret den zweidimensionalen Fall. Es werden außerdem einige topologische Eigenschaften der neuen Methode aufgezeigt. Von besonderem Interesse ist hier das Verhalten der medialen Achse der auftretenden Distanzfunktionen. Die gewonnen Erkenntnisse legen nahe, dass das entwickelte Verfahren fähig ist, zwischen der k-Nächste-Nachbarn-Methode und einem linearen Klassifikator zu interpolieren.

Inhaltsverzeichnis

Symbolverzeichnis

$\emptyset$	leere Menge
$\wp$	Potenzmenge
$i; k; l; n$	natürliche Zahlen
$\mathbb{N} = \{0; 1; 2; \cdots\}$	Menge der natürlichen Zahlen
$m; p; r$	reelle Zahlen
$\mathbb{R}$	Menge der reellen Zahlen
$\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$	Menge der erweiterten reellen Zahlen
$x; y; z$	Punkte;Vektoren
$\mathbb{R}^d$	d-dimensionaler Euklidischer Raum
$s; t$	Simplizes
$K; L; S; T$	Simplizialkomplexe
$\mathcal{K}$	Dualer Komplex
$\mathcal{P}$	Power-Diagramm
$\mathcal{Q}$	Schnittdiagramm
$\mathcal{R}$	reguläre Triangulation
b	d-dimensionale Kugel
B	Kugelmenge
$B(x; r)$	Kugel um x mit Radius r
B_r	Kugel um den Ursprung mit Radius r
$\leq$	kleiner als; Seite; Unterkomplex
$\subseteq$	Teilmenge
$\subset_{\text{fin}}$	endliche Teilmenge
$\sim$	verteilt wie
$=$	Gleichheit (Zahlen)
$\equiv$	Gleichheit (Abbildungen)
$+$	Addition; Minkowski-Summe
$\sum$	Summe
$\rightarrow$	Abbildung (Mengen); Konvergenz
$\mapsto$	Abbildung (Objekte)
conv	konvexe Hülle
dim	Dimension
inf	Infimum
lim	Limes
sgn	Signum
sup	Supremum
supp	Träger

δ	Dirac-Maß
λ	Lebesgue-Maß
$\mu; \nu$	Maße; Verteilungen
μ_n	empirisches Maß
$\mu_{n;r_0}$	Kugelmaß
$\lvert\cdot\rvert$	absoluter Betrag; Mächtigkeit; Polyeder
$\lVert\cdot\rVert$	Euklidische Norm
$\lVert\cdot\rVert_\infty$	Supremumsnorm
$c(\cdot)$	Klassifikator
$c_{n;m_0}(\cdot)$	empirischer Klassifikator
$c_{n;m_0;r_0}(\cdot)$	Kugelmaß-Klassifikator
$d(\cdot)$	Distanzfunktion
$d_{\mu_n;m_0}(\cdot)$	empirische Distanzfunktion
$d_{\mu_{n;r_0};m_0}(\cdot)$	Distanzfunktion des Kugelmaßes
$d(\cdot;\cdot)$	Euklidische Metrik
$B^\pm$	Bisektor des Trägers
$B_c^\pm$	Bisektor eines Klassifikators c
$B_{n;m_0}^\pm$	Bisektor des empirischen Klassifikators
$B_{n;r_0;m_0}^\pm$	Bisektor des Kugelmaß-Klassifikators
C	Punktwolke
C^+	Menge der Positivbeispiele
C^-	Menge der Negativbeispiele
n^+	Anzahl der Positivbeispiele
n^-	Anzahl der Negativbeispiele
$\bar{x}^+$	Mittelwert der Positivbeispiele
$\bar{x}^-$	Mittelwert der Negativbeispiele
X, Y	Zufallsvariablen
$\forall$	für alle
$\exists$	es gibt ein
$\mathcal{B}$	Borel-σ-Algebra
$\mathcal{O}$	Landau-O
$\square$	Definitionsschluss
$\blacksquare$	ohne Beweis
qed	Beweisschluss

Vorwort

Ich sage voraus, dass sich das Internet bald zu einer Supernova aufbläht und 1996 katastrophal kollabieren wird.

Robert M. Metcalfe, Erfinder des Ethernets, 1995

Nun, das ist es nicht. Im Gegenteil. Die Expansion des Internets hält bis heute an, statt abzuebben beschleunigt sie sich sogar immer weiter. Schätzungen zufolge verdoppeln sich die Informationen der Welt alle 20 Monate [16, S. V]. Die Bewältigung dieser schieren Masse an Daten ist im letzten Jahrzehnt zu *der* Herausforderung für die Informatik geworden. Will sie ihrem Selbstanspruch als Zukunftswissenschaft gerecht werden, so muss sie fähig sein Antworten zu liefern auf die Fragen der modernen Informationsgesellschaft und Lösungen anbieten, wie die allgegenwärtige Informationsflut sinnvoll verarbeitet, gefiltert und strukturiert werden kann. Sie muss Information in Wissen verwandeln.

Dieses Bestreben spiegelt sich in der theoretischen Informatik im Aufblühen des Data-Mining in der letzten Dekade. Data-Mining bezeichnet das Wissensfeld der Anwendung automatisierter Methoden auf sehr umfangreiche Datenbestände mit dem Ziel neue Erkenntnisse aus diesen zu generieren. In der vorliegenden Arbeit beschäftige ich mich mit einem neuartigen Ansatz zur Bearbeitung eines altbekannten Problems dieser Disziplin, der Klassifikation. Das ist die Aufgabenstellung mit Hilfe von bekanntem, bereits gegliederten Wissen Aussagen zu treffen, welcher Gattung ein neues, noch unklassifiziertes Objekt angehört. Hier handelt es sich konkret um Datenobjekte mit numerisch skalierten Merkmalen. Für den Entwurf des Klassifikators werden Distanzen verwendet, da diese sich meiner Meinung nach ganz natürlich aus der vorgefundenen Struktur des Objektraumes ergeben.

Ein häufig völlig verkanntes Anwendungsbeispiel für Klassifikation im Alltag ist die Arbeit von Internet-Suchmaschinen. Ohne *Google*, *Yahoo!*, *Bing*, *Ask* und Co. könnten wir uns heute nicht mehr im Internet zurecht finden, wir wären erschlagen von der unübersichtlichen Fülle an Webseiten und unfähig relevante Informationen herauszufiltern. Suchmaschinen verwenden nun Klassifikatoren an zwei ganz entscheidenden Punkten, nämlich bei der Datensortierung und der Kommunikation mit dem Nutzer. Auf der einen Seite müssen eine halbe Milliarde Webseiten [14] durchsucht und katalogisiert werden. Dabei werden die Dokumente nach ihrem Inhalt in verschiedene Themenbereiche eingeordnet. Auf der anderen müssen aber auch die Suchanfragen selbst klassifiziert werden. Nutzer formulieren ihre

Suche nicht in exakten Kategorien, sondern meist in kurzen natürlichsprachlichen Wortgruppen, die grob das umreißen, was gefunden werden soll. Es ist also in einem Vorverarbeitungsschritt nötig die Anfragen einem oder mehreren Themen zuzuordnen, um den Suchraum zumindest einzuschränken. Zur Anwendung von Klassifikation im Internet seien hier beispielsweise die Doktorarbeiten von Beitzel und Shen [2, 18] genannt.

Soweit in die Praxis will der Inhalt dieser Arbeit gar nicht gehen. Zunächst werden im ersten Kapitel einige Begriffe für die Darstellung der späteren Betrachtungen eingeführt, anschließend im zweiten das Konzept distanzbasierter Klassifikatoren anhand eines Beispiels erläutert und die mathematischen Grundlagen des neuen Verfahrens aufgeklärt. Daraus wird im dritten Teil eine Vorschrift zur Berechnung des Klassifikators abgeleitet. Dort findet sich auch die Gelegenheit Probleme und offene Fragen kritisch anzusprechen. Den Abschluss bilden dann die Ausführungen zu zwei Vermutungen über wichtige topologische Eigenschaften der entstehen Funktionen und eine Zusammenfassung der gewonnenen Erkenntnisse.

Ich hoffe mit dieser Arbeit einen bescheidenen Beitrag zur besseren Bewältigung der allgegenwärtigen Datenflut leisten zu können und mit meinen eigenen Ideen den Stand der informatischen Forschung zu erweitern, und sei es nur um ein kleines bisschen.

An dieser Stelle möchte ich mich bei denjenigen Menschen bedanken, die diese Arbeit ermöglicht haben. Meine Familie, allen voran meine Eltern Viola Schirneck und Michael Härig, mein Betreuer Dr. Joachim Giesen, mein Kommilitone und guter Freund Mario Biberhofer und nicht zuletzt die Frau an meiner Seite Natalie Siebelt haben mir mit fachlichem Rat, ihrer wertvollen Zeit, ihrem Zuspruch, vor allem aber mit menschlicher Wärme beigestanden, als dieses Projekt - mehr als einmal - vor dem Scheitern stand.

Kapitel 1

Vorbetrachtung und Bezeichnungen

1.1 Numerische Klassifikation

Das Klassifikationsproblem numerischer Datenvektoren lässt sich im Wesentlichen aus zwei verschiedenen Blickwinkeln betrachten. Bei dem einen wird eine feste Partition des Objektraumes - der hier stets der *d-dimensionale Euklidische Raum* $\mathbb{R}^d$ sein soll - in endlich viele Klassen unterstellt, die eigentliche *Klassifikation*. Die mit ihrer Klassenzugehörigkeit etikettierten Trainingsdaten werden als typische Repräsentanten ihrer jeweiligen Zerlegungsmengen aufgefasst. Ziel ist es, die genauen Klassengrenzen aus diesen Beispielen zu lernen und so ein Verfahren zu erhalten, das für einen *neuen*, noch unbekannten Punkt entscheidet, in welcher der Klassen er liegt. Also ein Verfahren, das ihn *klassifiziert*. Diese Interpretation findet sich exemplarisch bei Runkler [16, Kapitel 8]. Oft werden außerdem für eine effizientere Berechnung oder aus Modellierungsgründen zusätzliche geometrische Annahmen über die Gestalt der einzelnen Klassen getroffen. Dies geschieht zum Beispiel bei der *Support Vector Machine*, die explizit Halbräume als Klassenformen vorschreibt. Im Falle von genau zwei Klassen lässt sich die Klassifikation auch durch das Problem ausdrücken, für einen gegebenen Datenvektor algorithmisch zu entscheiden, ob er in einer zuvor fest gewählten Klasse angehört oder nicht, die jeweils andere Klasse ergibt sich dann implizit als Komplement der ersten. Hier sind sogar noch stärkere Forderungen an die feste Klasse möglich; beispielsweise, dass sie eine konvexes Polytop sei oder sich als eine glatte Mannigfaltigkeit beschreiben lasse. Ein *Klassifikator* ist dann eine berechenbare Funktion von den Punkten des Raumes in die Menge der möglichen Klassenmarkierungen.

Die andere Sicht ist eine stochastische, das heißt, die erwähnten Markierungen werden als Zufallsvariablen auf dem Objektraum aufgefasst, deren Verteilungen zunächst unbekannt sind. Allerdings werden die Lerndaten als konkrete Realisierungen des zugrundeliegenden Zufallsexperimentes angesehen. Mit Hilfe der Trainingsbeispiele und gegebenenfalls einigen zusätzlichen Annahmen ist daher die Klassenverteilung nun zu schätzen. Es wird angenommen, dass die Wahrscheinlichkeit für eine gegebene Markierung außerhalb einer einer bestimmten Menge verschwindet, dies stellt das stochastische Äquivalent zu den Klassengrenzen

der ersten Sichtweise dar. Mit der Aufgabe starrer Klassenstrukturen erreicht mensch hier eine erhöhte Flexibilität bei der geometrischen Modellierung und kann dadurch besser verrauschtes, durch Ausreißer verzerrtes oder gar inkonsistentes Datenmaterial verarbeiten. Notwendigerweise geht dabei allerdings die Eindeutigkeit der Zuordnung verloren. Ein Klassifikator hat nun eher den Charakter eines *Schätzers* für die unter den vorliegenden Voraussetzungen wahrscheinlichste Klassenmarkierung eines Punktes. Eine Einführung in das Data-Mining, die diesem Ansatz folgt, findet sich unter anderem bei Schukat-Talamazzini [17].

Selbstverständlich sind diese beiden Sichtweisen auf ein und denselben Sachverhalt ineinander überführbar und können parallel nebeneinander verwendet werden. In dieser Arbeit werden die Vorteile beider Standpunkte eingesetzt, um das Klassifikationsverfahren zu entwickeln. Da die Definitionen von Wahrscheinlichkeitsverteilungen über Maße einen natürlicheren Zugang zu Distanzfunktionen ermöglicht, wird dabei die Betonung auf der stochastischen Methode liegen. Gleichzeitig werden aber zur Berechnung ebendieser Distanzen geometrische Objekte verwendet, die ausgiebig die Interpretation der Lerndaten als Punkte im Raum benutzen.

1.2 Punktwolken und Distanzfunktionen

Es soll eine Familie von Klassifikatoren untersucht werden. Dazu ist es notwendig sich ein Begriffssystem zurecht zu legen, in dem es möglich ist, diese Funktionen zu definieren, Aussagen über sie zu formulieren und zu beweisen. Dazu wird der $\mathbb{R}^d$ als mit der *Euklidischen Norm* $\|\cdot\|$ versehen aufgefasst, es ergibt sich implizit die *Euklidische Metrik* $d(x;y) = \|x-y\|$, bezüglich der der Raum vollständig ist. Die Metrik wiederum induziert auf dem $\mathbb{R}^d$ die übliche Topologie. Außerdem wird der Objektraum stets als mit einer affinen - statt einer bloß linearen - Unterraumstruktur ausgestattet betrachtet. Eine Teilmenge des $\mathbb{R}^d$ heiße *Punktwolke*, wenn sie endlich, aber nicht leer ist. Punkte $\{x_0; x_1; x_2 \cdots ; x_n\} \subset_{\text{fin}} \mathbb{R}^d$ sollen weiter *affin unabhängig* heißen, wenn das System ihrer Richtungsvektoren $\{x_1 - x_0; x_2 - x_0; \cdots ; x_n - x_0\}$ linear unabhängig ist. Eine Punktwolke sei schließlich in *allgemeiner Lage*, wenn jede ihrer Teilmengen mit höchstens $d+1$ Elementen affin unabhängig ist. In den weiteren Ausführungen werden außerdem häufig endliche Systeme von abgeschlossenen d-dimensionalen Kugeln zur Darstellung von Berechnungen herangezogen. Für ein nicht-negatives $r \geq 0$ und einen Punkt $x \in \mathbb{R}^d$ bezeichne dann $B(x;r) = \{y \in \mathbb{R}^d \mid \|x-y\| \leq r\}$ die angeschlossene Kugel mit *Zentrum* x und *Radius* r, B_r stehe in diesem Zusammenhang abkürzend für $B(\underline{0};r)$ also eine Kugel um den Ursprung. Eine Menge von Kugeln sei in allgemeiner Lage, wenn es ihre Zentren sind und alle Elemente ein nicht-leeres Inneres besitzen, mit anderen Worten, wenn alle Radii positiv sind.

Die zu entwickelnden Klassifikatoren sollen auf Distanzen basieren, diese werden dafür als ein inverser Indikator für die Ähnlichkeit von Datenvektoren interpretiert. Nah beieinander liegende Punkte werden also als kategoriell gleichartig angesehen und sollen deshalb auch möglichst gleich klassifiziert werden. Ein Punkt gehöre außerdem genau dann zu einer bestimmten Klasse, wenn sein Abstand zu den Trainingsdaten aus dieser Klasse geringer ist als zu denen aller anderen Klassen.

Es ist also unbedingt notwendig zunächst diesen Distanzbegriff zu präzisieren. Da der Objektraum bereits mit einer natürlichen Metrik versehen ist, dient diese hier als Ausgangspunkt. Distanzfunktionen übertragen nun deren charakteristische Eigenschaften auf einstellige Abbildungen. Nach Chazal, Cohen-Steiner und Mérigot [6] soll eine nicht-negative Abbildung $d\colon \mathbb{R}^d \to \mathbb{R}$ *Distanzfunktion* heißen, wenn sie selbst 1-Lipschitz-stetig ist, also stets $|d(x) - d(y)| \leq \|x - y\|$ erfüllt, das Quadrat der Abbildung 1-semikonkav ist - das heißt $x \mapsto d^2(x) - \|x\|^2$ ist eine konkave Funktion - und schließlich $d(x)$ gegen Unendlich geht, wann immer $\|x\|$ dies tut. Dabei ist zu beachten, dass eine semikonkave Funktion nach einem Theorem von Alexandrov auch automatisch fast überall zweimal differenzierbar ist, für einen Beweis siehe Howard [13]. Der Prototyp einer Distanzfunktion ist der Abstand eines Punktes zu einer kompakten Menge $A \subseteq \mathbb{R}^d$, der durch $d_A(x) = \inf_{a \in A} \|x - a\|$ erklärt ist, wobei das Infimum wegen der Kompaktheit stets angenommen wird.

Streng genommen genügt es für Distanzfunktionen die letzten beiden Eigenschaften zu fordern, weil aus der 1-Semikonkavität des Quadrates bereits die 1-Lipschitz-Stetigkeit einer Abbildung folgt, vergleiche [6, Proposition 3.1]. Allerdings ist diese Stetigkeit genau die Eigenschaft, die eine Funktion anschaulich distanzartig macht: Es lässt sich leicht nachrechnen, dass eine Abbildung $d(\cdot)$ genau dann 1-Lipschitz-stetig ist, wenn sie mit der Dreiecksungleichung der Euklidischen Metrik verträglich ist, das bedeutet, dass für beliebige Punkte $x; y \in \mathbb{R}^d$ stets $d(x) \leq d(x; y) + d(y)$ gilt. In abschwächender Sprechweise heiße eine einstellige reellwertige Funktion, die keine negativen Werte annimmt und 1-Lipschitz-stetig ist, daher eine *Pseudo-Distanzfunktion*.

1.3 Stochastische Maßtheorie

Wie erwähnt werden einige Konzepte aus der Wahrscheinlichkeits- und der Maßtheorie für die weiteren Betrachtungen benötigt. Es bezeichne daher $\mathcal{B}(\mathbb{R}^d)$ die *Borel-σ-Algebra* auf dem d-dimensionalen Euklidischen Raum, das heißt die kleinste Σ-Algebra, die alle in der üblichen Topologie offenen Mengen enthält. Es sei bemerkt, dass das System dann natürlich auch alle abgeschlossenen Mengen enthalten muss. Ein *Maß* auf dem *Messraum* $(\mathbb{R}^d; \mathcal{B}(\mathbb{R}^d))$ ist eine nicht-negative Abbildung $\mu\colon \mathcal{B}(\mathbb{R}^d) \to \bar{\mathbb{R}}$ in die erweiterten reellen Zahlen, die zum einen der leeren Menge den Inhalt 0 zuweist und zum anderen σ-additiv ist. Letzteres bedeutet, dass für jede abzählbare Familie paarweise disjunkter Borel-Mengen $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{B}(\mathbb{R}^d)$ $\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i)$ gilt. Das Tripel $(\mathbb{R}^d; \mathcal{B}(\mathbb{R}^d); \mu)$ ist dann ein *Maßraum*.

Ist μ außerdem normiert, gilt also $\mu(\mathbb{R}^d) = 1$, so spricht mensch auch von einem *Wahrscheinlichkeitsraum* und nennt μ ein *Wahrscheinlichkeitsmaß* oder einfach *stochastisch*. Jede *messbare* Funktion von einem Wahrscheinlichkeits- in einen Messraum heißt nun *Zufallsvariable*. Hier wird sich auf den reellen Fall beschränkt, also nur Abbildungen $X\colon (\mathbb{R}^d; \mathcal{B}(\mathbb{R}^d); \mu) \to (\mathbb{R}^k; \mathcal{B}(\mathbb{R}^k))$ betrachtet, wobei auch in den meisten Fällen der Zielraum bloß eindimensional sein wird. Durch X überträgt sich das Wahrscheinlichkeitsmaß auf $\mathbb{R}^k$, denn durch $\mu_X(A) = \mu(X^{-1}(A))$, für jede Borel-Menge $A \in \mathcal{B}(\mathbb{R}^k)$, ist ein normiertes Maß erklärt. Die Variable X wird dann *gemäß μ_X verteilt* genannt, Schreibweise $X \sim \mu_X$. Häufig wird eine Zufallsvariable der Einfachheit halber mit ihrem Bild identifiziert und wird so zu einem *Punkt* im

$\mathbb{R}^k$, der gemäß μ_X verteilt ist. In dieser Interpretation tritt die Struktur der Quelle - vor allem das Maß μ selbst - zugunsten des Bildmaßes μ_X in den Hintergrund. Aus der Sicht des Klassifikatorentwurfs sind beide Deutungen gleichwertig, wenngleich sich auch die mathematische Behandlung im Einzelfall unterscheiden kann. Sind Verwechslungen zu befürchten, wird in dieser Arbeit explizit angegeben, welche Sichtweise gemeint ist.

Ein (allgemeines) Maß heiße *endlich*, falls der Wert $\mu(\mathbb{R}^d) < \infty$ endlich ist und *σ-endlich*, wenn sich der Raum durch abzählbar viele Mengen mit endlichem Inhalt überdecken lässt. Natürlich sind alle stochastischen Maße endlich und alle endlichen auch σ-endlich. Das d-dimensionale *Lebesgue-Maß* λ ist ein wichtiges Beispiel für ein Maß, das *nicht* endlich, aber σ-endlich ist. Betrachte dazu das System $\{[k_1; k_1 + 1] \times [k_2; k_2 + 1] \times \cdots \times [k_d; k_d + 1] \,|\, k_1; k_2; \cdots ; k_d \in \mathbb{Z}\}$. Bezüglich λ ist dies eine Überdeckung des $\mathbb{R}^d$ durch Hyperwürfel des Maßes 1.

Wie bereits angedeutet, sollen die hier untersuchten Maße außerhalb einer gewissen Menge verschwinden, diese Menge wird Träger des Maßes genannt, denn anschaulich befindet sich die gesamte Masse des Raumes in diesem Bereich. Für eine präzise Fassung des Begriffes sei nun $x \in \mathbb{R}^d$ ein Punkt und $U(x)$ bezeichne eine offene Umgebung von x. Der *Träger* eines Maßes μ sei dann durch $\operatorname{supp}(\mu) = \{x \in \mathbb{R}^d \,|\, \forall\, U(x) \colon \mu(U(x)) > 0\}$ definiert. Er muss dann notwendig abgeschlossen sein: Sei dazu $y \in \mathbb{R}^d$ ein Häufungspunkt von $\operatorname{supp}(\mu)$ und $U(y)$ eine offene Umgebung. Es gibt also einen weiteren Punkt $x \in U(y) \cap \operatorname{supp}(\mu)$; $y \neq x$. Allerdings ist $U(y)$ auch offene Umgebung von x und daher $\mu(U(y)) > 0$, es folgt $y \in \operatorname{supp}(\mu)$, wie gewünscht. Der Träger ist die (bezüglich der Inklusion als Halbordnung) kleinste abgeschlossene Teilmenge, die das gleiche Maß wie der gesamte Raum besitzt oder umgekehrt formuliert ist $\operatorname{supp}(\mu)$ das Komplement der größten offenen Nullmenge bezüglich μ. Die letzte Charakterisierung ist auch im Falle eines nicht-endlichen Maßes eindeutig.

Da bei der Klassifikation problemimmanent jeweils nur endlich viele Lernbeispiele für eine Klasse betrachtet werden können und die natürlich alle auch im Träger liegen sollen, ist es naheliegend diesen - als direkte Verallgemeinerung von endlichen Mengen in der Topologie - als kompakt anzunehmen. Dabei sollte in Erinnerung bleiben, dass, in Verbindung mit der gezeigten Abgeschlossenheit, der Träger nach dem Satz von Heine-Borel genau dann kompakt ist, wenn er beschränkt ist.

1.4 Simpliziale Komplexe

Für die eigentliche Berechnung der betrachteten Maße werden noch weitere Klassen von Objekten im Euklidischen Raum benötigt, die sogenannten Simplizes und simplizialen Komplexe. Es lassen sich dabei anschauliche geometrische und rein mengentheoretisch abstrakte Simplizes unterschieden. Sei $C \subset_{\text{fin}} \mathbb{R}^d$ eine affin unabhängige Punktwolke oder leer, so heiße ihre *konvexe Hülle* $s = \operatorname{conv}(C)$ *geometrischer Simplex*. Er habe *Dimension* $\dim s = |C| - 1$, ist $\dim s = k$, so wird s auch kurz ein *k-Simplex* genannt. Die konvexe Hülle t einer Teilmenge von C - die dann ebenfalls affin unabhängig - heiße *Teilsimplex* oder *Seite* $t \leq s$ des Simplex s. Handelt es sich bei t um einen 0-, 1- oder 2-Simplex, so sind auch die Bezeichnungen *Ecke*, *Kante* respektive *Fläche* üblich. Durch diese Festlegung wird

C zur Menge aller Ecken von s. Schließlich werde nun ein endliches, nicht-leeres System K von Simplizes *geometrischer Simplizialkomplex* genannt, wenn es zum einen mit jedem Simplex auch dessen sämtliche Seiten enthält und zum anderen der Durchschnitt je zweier Simplizes aus K eine Seite von beiden ist. Die *Dimension* eines solchen Komplexes sei die größte Dimension der enthaltenen Simplizes. Es ist klar, dass sie die des umgebenden Raumes $\mathbb{R}^d$ nicht überschreiten kann. Jedes nicht-leere Teilsystem $\emptyset \neq L \subseteq K$, das selbst wieder ein Simplizialkomplex ist, heiße *Unterkomplex* von K, Schreibweise $L \leq K$. Das ist offenbar genau dann der Fall, wenn L jede Seite jedes seiner Simplizes umfasst, denn die zweite definierende Eigenschaft wird direkt von K ererbt. Die Menge aller Punkte, die von allen Simplizes eines Komplexes K - aufgefasst als Teilmengen des Euklidischen Raumes - überdeckt werden, bilden dessen *Polyeder* $|K|$.

All diese Begriffe lassen sich allerdings auch auf viel allgemeinere Mengensysteme als bloß Sammlungen von konvexen Polytopen übertragen: Unter einem *abstrakten Simplex* soll daher nun einfach eine beliebige endliche Menge verstanden werden. Analog zum geometrischen Fall sei ihre Dimension um 1 geringer als die Anzahl ihrer Elemente. Eine Seite ist dann einfach eine beliebige Teilmenge dieses Simplex und ein *abstrakter Simplizialkomplex* ein endliches, nicht-leeres System von endlichen Mengen, das mit jeder Menge alle ihre Teilmengen umfasst. Auch die Definitionen der Dimension und des Unterkomplexes können im abstrakten Fall einfach vom geometrischen Gegenstück übernommen werden. Die Betrachtung des Polyeder allerdings ist nur dann sinnvoll, wenn die Ecken des Komplexes selbst Teilmengen eines gemeinsamen (zum Beispiel des Euklidischen) Raumes sind.

In diesem Text wird in der Sprache nicht zwischen geometrischen und abstrakten simplizialen Komplexen unterschieden, wenn aus dem Zusammenhang ersichtlich ist, von welcher Art das gemeinte System ist. In ihrer Verwendung jedoch spielt der Aufbau eine entscheidende Rolle. Trotz aller Unterschiedlichkeit existiert ein wichtiger Zusammenhang zwischen den beiden Ausprägungen: Jeder abstrakte Simplizialkomplex lässt sich *geometrisch realisieren*. Das heißt, es gibt für jeden Komplex K ein hinreichend großes k und eine Abbildung $K \to \wp(\mathbb{R}^k)$ derart, dass jedem abstrakten Simplex ein geometrischer gleicher Dimension zugeordnet werden kann. Diese Zuordnung soll weiterhin mit der Durchschnittsbildung verträglich sein. Das heißt, dass der Schnitt der Bilder je zweier Simplizes genau das Bild des Schnittes von diesen ist und damit eine Seite von beiden. Dadurch wird das entstehende Objekt tatsächlich zu einem geometrischen Simplizialkomplex nach obiger Festlegung. Ein Beweis der Existenz einer solchen *geometrischen Realisierung* findet sich zum Beispiel bei Green [12, S. 11].

Die hier verwendeten abstrakten Simplizes werden in der Regel aus d-dimensionalen Kugeln als Ecken bestehen. Den Gedanken von Attali und Edelsbrunner aus [1, 7] folgend, soll für eine endliche Menge B solcher Kugeln und einem Komplex $K \subseteq \wp(B)$ eine bestimmte Einbettung in den $\mathbb{R}^d$ festgehalten werden. Nämlich diejenige, die jedem Simplex sein *kanonisches Bild*, die konvexe Hülle der Zentren der enthaltenen Kugeln, zuordnet. Es sei darauf hingewiesen, dass diese Abbildung nicht notwendig auch eine Realisierung des betreffenden Komplexes zu sein braucht.

Kapitel 2

Distanzbasierte Klassifikationsverfahren

2.1 Der empirische Klassifikator

Basierend auf einer Idee von Giesen und Kühne [11] wird in diesem einleitenden Abschnitt modellhaft ein distanzbasierter Klassifikator betrachtet. Dieser dient als Richtschnur für die weiteren Untersuchungen. Es wird dafür zunächst auf dem Objektraum ein Maß definiert und aus diesem eine Distanzfunktion im Sinne der Theorie von Chazal et al. gewonnen. Das bedeutet, dass der Inhalt einer d-dimensionalen Kugel, die einen zuvor festgelegten Anteil der Trainingsbeispiele überdeckt, auf eine bestimmte Weise als der Abstand des Zentrums dieser Kugel zu den Daten aufgefasst wird. Die gewonnene Abbildung dient dann ihrerseits zur eigentlichen Klassifizierung, denn ein Punkt wird derjenigen Klasse zugewiesen zu der er den kleinsten Abstand hat. Es ist zu beachten, dass im Weiteren ein Punkt im $\mathbb{R}^d$ auf zwei verschiedene Arten betrachtet wird, einmal als geometrische Repräsentation eines Datenvektors mit d numerisch skalierten Merkmalen und einmal als Realisierung einer Zufallsvariable mit Bild im d-dimensionalem Euklidischen Raum. Dies korrespondiert mit den beiden eingangs erwähnten Sichtweisen auf das Klassifikationsproblem.

Definition 2.1 (Dirac-Maß)**:**
Sei $x \in \mathbb{R}^d$ ein Punkt, dann heiße die Abbildung

$$\delta_x \colon \mathcal{B}\left(\mathbb{R}^d\right) \to \mathbb{R}\ ;\ A \mapsto \begin{cases} 1 & x \in A \\ 0 & sonst \end{cases}$$

das Dirac-Maß *in x.*

□

Bemerkung. Es lässt sich leicht nachrechnen, dass es sich bei der definierten Zuordnung tatsächlich um ein normiertes Maß handelt. Sein Träger ist der einzelne Punkt x, dort konzentriert es seine ganze Masse. Dies charakterisiert das Dirac-Maß als Wahrscheinlichkeitsverteilung vollständig.

Definition 2.2 (Empirisches Maß)**:**
Seien $x_1; x_2; \cdots ; x_n \in \mathbb{R}^d$ *Realisierungen einer Zufallsvariable* $X \sim \mu$*, die Funktion*

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

heiße das empirische Maß *bezüglich der Menge* $\{x_1; \cdots ; x_n\}$*.*

□

Bemerkung. Anschaulich wird einer Borel-Menge der Anteil der von ihr überdeckten Punkte x_i zugewiesen Es gilt also stets $\mu_n(A) = |A \cap \{x_1; \cdots ; x_n\}|/n$, auch dies beschreibt ein Wahrscheinlichkeitsmaß. Sein Träger ist gerade die Punktwolke $\{x_1; \cdots ; x_n\}$. Da endliche Teilmengen des Euklidischen Raumes stets kompakt sind, ist es auch der Träger des empirischen Maßes.

Der Hauptsatz der Statistik - der Satz von Gliwenko-Cantelli - postuliert nun eine fast sichere Konvergenz der μ_n gegen μ für $n \to \infty$. Genauer konvergieren die gemäß μ_n verteilten Zufallsvariablen X_n (als Funktionen) punktweise λ-fast überall gegen X. Diese Konvergenz ist sogar gleichmäßig, vergleiche dazu Pollard [15, S. 6f.]. Für hinreichend viele Realisierungen lässt sich die ursprüngliche Verteilung also durch das empirische Maß annähern. Daher dient es hier auch als Schätzer für ein unbekannte Wahrscheinlichkeitsverteilung, beispielsweise die der Klassenmarkierungen.

Die bereits erwähnte Distanzfunktion d_A zu einer kompakten Menge $A \subseteq \mathbb{R}^d$, lässt sich auch als diejenige Abbildung interpretieren, die jedem Punkt $x \in \mathbb{R}^d$ des Raumes den Radius der kleinsten Kugel um x zuordnet, die noch Punkte von A enthält. Diese Deutung soll die weiteren Festlegungen inspirieren. Denn wird jetzt ein nicht-negatives reelles $m \geq 0$ fixiert, so ergibt sich ein Ansatz für eine analoge Übertragung dieser Ideen auf den vorliegenden maßtheoretischen Fall.

Definition 2.3:
Bezeichne μ *ein endliches Maß und* $0 \leq m < \mu(\mathbb{R}^d) < \infty$ *eine reelle Zahl. Im Weiteren sei mit*

$$\delta_{\mu;m} \colon \mathbb{R}^d \to \mathbb{R} \; ; \; x \mapsto \inf\{r > 0 \,|\, \mu(B(x;r)) > m\}$$

eine nicht-negative reelle Abbildung festgehalten.

□

Chazal et al. haben diese Funktion in [6] eingeführt und untersucht. Sie stellten dabei fest, dass $\delta_{\mu;m}$ selbst als eine Verallgemeinerung von Distanzfunktionen auf Maße noch nicht geeignet ist. Das liegt insbesondere daran, dass ihr Quadrat nicht notwendig für jedes m 1-semikonkav ist. Außerdem ergeben sich eklatante Stetigkeitsprobleme der Zuordnung $\mu \mapsto \delta_{\mu;m}$ zwischen den Maßen und den durch sie induzierten Abbildung für jede naheliegende Topologie auf dem Raum der endlichen Maße des $\mathbb{R}^d$. Dennoch spielt die oben definierte Hilfsfunktion beim Aufbau des eigentlichen Distanzbegriffes eine entscheidende Rolle, denn der Abstand eines Punktes zu einem Maß wird sich als ein normiertes L_2-Mittel von Werten von $\delta_{\mu;m}$ ergeben. Zunächst sollen allerdings einige nützliche Eigenschaften notiert werden,

die die gegebene Abbildung zumindest mit dem Begriff der Pseudo-Distanzfunktion in Beziehung setzt.

Lemma 2.4:
Seien die Bezeichnungen wie oben, so gilt:

(*i*) *Die Funktion $\delta_{\mu;m}$ ist eine Pseudo-Distanzfunktion.*

(*ii*) *Hat μ kompakten Träger, so divergiert $\delta_{\mu;m}(x)$ für $\|x\| \to \infty$.*

(*iii*) *$\delta_{\mu;0} \equiv d_{\mathrm{supp}(\mu)}$, bei kompaktem Träger ist also insbesondere $\delta^2_{\mu;0}$ 1-semikonkav.*

Beweis. Seien $x; y \in \mathbb{R}^d$ zwei Punkte.

(*i*) Wegen $B(y; \delta_{\mu;m}(y)) \subseteq B(x; d(x; y) + \delta_{\mu;m}(y))$ ist letztere *eine* Kugel um x mit Maß größer als m und daher gilt $\delta_{\mu;m}(x) \leq d(x; y) + \delta_{\mu;m}(y)$.

(*ii*) Sei $\mathrm{supp}(\mu)$ kompakt, das heißt beschränkt, es gibt daher ein r' mit $\mathrm{supp}(\mu) \subseteq B_{r'}$. Für hinreichend großes $\|x\|$ gilt nun $\delta_{\mu;m}(x) \geq d_{\mathrm{supp}(\mu)}(x) \geq d_{B_{r'}}(x) = \|x\| - r'$, woraus die Behauptung folgt.

(*iii*) Für alle $r > d_{\mathrm{supp}(\mu)}(x)$ ist $\mathring{B}(x; r) = \{z \in \mathbb{R}^d \mid \|x - z\| < r\}$ eine offene Umgebung eines Punktes aus dem Träger und hat damit positives Maß, umgekehrt liegt für jedes $r < d_{\mathrm{supp}(\mu)}(x)$ die Kugel $B(x; r)$ ganz in $\mathbb{R}^d \setminus \mathrm{supp}(\mu)$ und ist daher eine μ-Nullmenge.

qed

Auf diesen Erkenntnissen haben Chazal, et al. wie folgt eine tatsächliche Distanzfunktion aufgebaut.

Definition 2.5 (Distanzfunktion zu einem Maß)**:**
Sei μ ein endliches Maß mit zusätzlich endlichem zweiten Moment, das heißt, es gelte $\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x) < \infty$ und sei $0 < m_0 < \mu(\mathbb{R}^d) < \infty$ ein reeller Parameter. Die Distanzfunktion $d_{\mu;m_0}$ zum Maß μ *sei dann durch*

$$d^2_{\mu;mo} \colon \mathbb{R}^d \to \mathbb{R} \; ; \; x \mapsto \frac{1}{m_0} \int_0^{m_0} \delta^2_{\mu;m}(x) \, \mathrm{d}m$$

erklärt.

□

Theorem 2.6 (Chazal, Cohen-Steiner, Mérigot)**:**
Die Abbildung $d_{\mu;m_0}$ ist tatsächlich eine Distanzfunktion.

■

Bemerkung. Durch die eingeführe Regularisierung werden nun auch die erwähnten Unstetigkeitsprobleme der Zuordnung zwischen Maß und Distanzfunktion beseitigt. Vergleiche hierzu die Arbeiten von Bolley, Guillin und Villani, insbesondere [5, Theorem 1.1].

Für das empirische Maß lässt sich die Berechnung der Distanzfunktion durch eine Auswertung einer k-Nächste-Nachbar-Anfrage darstellen. Dieser Zusammenhang wird im folgenden Lemma präzisiert, es stammt in in seiner ursprünglichen Fassung aus [6].

Lemma 2.7:
Sei μ_n das empirische Maß bezüglich einer Punktwolke $C = \{x_1; \cdots ; x_n\}$ und $x \in \mathbb{R}^d$ ein weiterer Punkt. Außerdem bezeichne $X_i(x)$ jeweils den i-ten nächsten Nachbarn von x in C, wobei Punkte mit gleichem Abstand in einer beliebigen aber festen Reihenfolge auftreten sollen. Sei schließlich $1 \leq k < n$ eine ganze Zahl, so gilt:

$$d^2_{\mu_n;k/n}(x) = \frac{1}{k} \sum_{i=1}^{k} \|X_i(x) - x\|^2$$

Beweis. Es bezeichne k für einen Parameterwert $0 \leq m < 1$ die kleinste ganze Zahl echt größer als mn, dann ist $\delta_{\mu_n;m}(x)$ offenbar der Radius der kleinsten Kugel um x, die noch k Punkte aus C enthält. Mit anderen Worten, es gilt $\delta_{\mu_n;m}(x) = \|X_k(x) - x\|$. Insbesondere ist also $m \mapsto \delta_{\mu_n;m}(x)$ jeweils auf den Abschnitten $\left[\frac{i-1}{n}; \frac{i}{n}\right)$ konstant, diese Intervalle haben alle die Länge $\frac{1}{n}$. Zusammen ergibt sich:

$$d^2_{\mu_n;k/n}(x) = \frac{n}{k} \int_0^{k/n} \delta^2_{\mu_n;m}(x)\, \mathrm{d}m = \frac{n}{k} \sum_{i=1}^{k} \frac{1}{n} \|X_i(x) - x\|^2 = \frac{1}{k} \sum_{i=1}^{k} \|X_i(x) - x\|^2$$

qed

Korollar 2.8:
Es seien die Bezeichnungen wie oben und zusätzlich für einen reellen Parameter $0 < m_0 < 1$ sei k_0 die kleinste ganze Zahl echt größer als $m_0 n$, dann gilt:

$$d^2_{\mu_n;m_0}(x) = \|X_{k_0}(x) - x\|^2 + \frac{k_0 - 1}{m_0 n}(d^2_{\mu_n;(k_0-1)/n}(x) - \|X_{k_0}(x) - x\|^2)$$

Insbesondere ist also $m_0 \mapsto d_{\mu_n;m_0}(x)$ auf ganz $(0;1)$ stetig und fast überall - außer in ganzen Vielfachen von $\frac{1}{n}$ - differenzierbar.

Beweis.

$$\begin{aligned} d^2_{\mu_n;m_0}(x) &= \frac{1}{m_0} \left(\int_0^{(k_0-1)/n} \delta^2_{\mu;m}(x)\, \mathrm{d}m + \int_{(k_0-1)/n}^{m_0} \delta^2_{\mu;m}(x)\, \mathrm{d}m \right) \\ &= \frac{1}{m_0} \left(\frac{k_0 - 1}{n} d^2_{\mu_n;(k_0-1)/n}(x) + \left(m_0 - \frac{k_0 - 1}{n} \right) \|X_{k_0}(x) - x\|^2 \right) \end{aligned}$$

qed

Jetzt lässt sich einsehen, dass die $d_{\mu_n;k/n}(x)$ - aufgefasst als Werte des Regulationspfades $m_0 \mapsto d_{\mu_n;m_0}(x)$ - abschnittsweise durch Wurzeln streng monoton steigender, hyperbolischer Kurvenstücke mit niedriger Krümmung verbunden sind. Giesen et al. haben alternativ eine einfacher zu berechnende lineare Interpolation

vorgeschlagen, bei der die Krümmung natürlicherweise verschwindet, die aber dennoch die Eigenschaften der Funktion erhält, vergleiche [11, Distance functions.]. Des Weiteren ergibt sich aus dem obigen Korollar der wichtige Grenzwert $\lim_{m_0 \searrow 0} \|d_{\mu_n;m_0} - d_{\mathrm{supp}(\mu_n)}\|_\infty = 0$ in der*Supremumsnorm* reellwertiger Funktionen vom $\mathbb{R}^d$. Das bedeutet, dass sich die Distanzfunktion zum empirischen Maß für hinreichend kleine Parameterwerte m_0 im Wesentlichen so verhält wie die Abstandsfunktion zum seinem Träger. Dies ist auch bei anderen Maßen mit kompakten Träger zu erwarten, vergleiche dazu den Abschnitt 4.1 dieser Arbeit.

Theorem 2.6 und Korollar 2.8 zusammen sagen aus, dass der Übergang von $\delta_{\mu_n;m}$ zu $d_{\mu_n;m_0}$ als die Distanzfunktion des empirischen Maßes die Stetigkeit überall und die Differenzierbarkeit λ-fast überall sowohl im Parameter m_0 als auch im Argument x sichert. Allerdings sind diejenigen Stellen, in denen der Pfad immer noch nicht differenzierbar ist, mit den Unstetigkeitsstellen von $m \mapsto \delta_{\mu_n;m}(x)$ identisch. Auch der Einfluss des Parameters wird durch diese Aussagen aufgeklärt, je größer m_0 ist, von desto mehr Punkten aus C hängt die Distanzfunktion ab und desto größer ist ihr numerischer Wert.

Es kann nun wie angekündigt aus der definierten Distanzfunktion ein erster Klassifikator gewonnen werden. Er markiert einen Punkt genau dann als der fixierten Klasse zugehörig, wenn der gemittelte kleinste Radius um ihn, der einen m_0 Anteil der Positivbeispiele überdeckt, echt kleiner ist als derjenige, der zur Überdeckung des gleichen Anteils der Negativbeispiele benötigt wird. Bei der Modellierung wird unterstellt, dass die Klassenmarkierungen gemäß zweier Wahrscheinlichkeitsmaße verteilt seien, diese müssen zunächst aus den Lernbeispielen geschätzt werden.

Definition 2.9:
Seien $\{(x_1; y_1); \cdots; (x_n; y_n)\} \in \mathbb{R}^d \times \{\pm 1\}$ *Realisierungen einer Zufallsvariable* $(X; Y) \sim \mu$ *und* n^+ *beziehungsweise* n^- *die Anzahl der Positiv- respektive Negativbeispiele, so bezeichnen*

$$\mu_n^+ = \frac{1}{n^+} \sum_{i;y_i=+1} \delta_{x_i} \quad \textit{und} \quad \mu_n^- = \frac{1}{n^-} \sum_{i;y_i=-1} \delta_{x_i}$$

die empirischen Schätzer für die bedingten Verteilungen $\mu^+ = \mu(\,\cdot\,|\,Y = +1)$ *und* $\mu^- = \mu(\,\cdot\,|\,Y = -1)$.

□

Definition 2.10 (Empirischer Klassifikator)**:**
Seien die Bezeichnung wie oben, dann ist die durch

$$c_{n;m_0} \colon \mathbb{R}^d \to \{+1; 0; -1\}\ ;\ x \mapsto \begin{cases} +1 & \textit{falls } d_{\mu_n^+;m_0}(x) < d_{\mu_n^-;m_0}(x) \\ 0 & \textit{falls } d_{\mu_n^+;m_0}(x) = d_{\mu_n^-;m_0}(x) \\ -1 & \textit{sonst} \end{cases}$$

definierte Abbildung der empirische Klassifikator.

□

Bemerkung. Der empirische Klassifikator lässt sich auch kürzer als Signum einer reellwertigen Funktion beschreiben: $c_{n;m_0} \equiv \mathrm{sgn}(d_{\mu_n^-;m_0} - d_{\mu_n^+;m_0})$

Die Einführung der Null als einen möglichen Wert der Klassifizierung dient hier nur Untersuchungszwecken, denn bei der Betrachtung distanzbasierter Verfahren sind diejenigen Punkte von besonderem Interesse, die zu beiden Mengen von Lernbeispielen gleichen Abstand haben. In der Praxis mögen sie - in Abhängigkeit von der jeweiligen Problemstellung - wahlweise einer der beiden Klassen zugerechnet werden, für die hier angestellten Überlegungen ist aber die Entwicklung in eine eigene Kategorie unbedingt notwendig.

Definition 2.11 (Mediale Achse)**:**
Sei $c\colon \mathbb{R}^d \to \{+1; 0; -1\}$ *ein distanzbasierter Klassifikator, so heiße das Urbild* $B_c^{\pm} = c^{-1}(0)$ *der Null die* mediale Achse *oder* Bisektor*s von c.*

□

Analog zu den Betrachtungen der Grenzwerte empirischer Maße und der Distanzfunktionen lassen sich auch Konvergenzaussagen für die Bisektoren formulieren. So geht $B_{n;m_0}^{\pm} = B_{c_{n;m_0}}^{\pm}$ mit wachsendem n fast sicher gegen $B_{m_0}^{\pm} = \{x \in \mathbb{R}^d \,|\, d_{\mu^+;m_0}(x) = d_{\mu^-;m_0}(x)\}$, letzterer wiederum hat für $m_0 \searrow 0$ den Limes $B^{\pm} = \{x \in \mathbb{R}^d \,|\, d_{\mathrm{supp}(\mu^+)}(x) = d_{\mathrm{supp}(\mu^-)}(x)\}$, falls beide Maße jeweils kompakten Träger besitzen. Dabei werden die Grenzwerte in der *Hausdorff-Distanz* betrachtet, die anschaulich misst, wie gut sich zwei Teilmengen des $\mathbb{R}^d$ gegenseitig überdecken. Die Beweise dieser Aussagen bedienen sich der entsprechenden oben angesprochenen Konvergenz der μ_n beziehungsweise $d_{\mu_n;m_0}$. Abschätzungen zur Konvergenzgeschwindigkeit finden sich in [5, 6, 11, 15].

2.2 Motivation eines neuen Verfahrens

Im vorangegangenen Abschnitt wurde der Prototyp eines distanzbasierten Klassifikators entwickelt, das verwendete Konstruktionsverfahren wird auch die Basis für die anderen Entwürfe in dieser Arbeit bilden. Außerdem wurden erste Überlegungen zur medialen Achse der entstehenden Abbildungen angestellt. Der maßtheoretische Ansatz bringt dabei einen entscheidenden Vorteil, denn, um charakteristische Eigenschaften des Klassifikators aufzudecken, genügt es nun häufig schon das zugrundeliegende Maß zu untersuchen.

Experimentelle Berechnungen von Giesen und Kühne [11] legen den Schluss nahe, dass sich der empirische Klassifikator ähnlich dem klassischen k-Nächste-Nachbarn-Verfahren verhält. Das betrifft sowohl den Aufwand für das Training und die Auswertungen als auch die Topologie des Klassifikators, genauer seines Bisektors. Wie kNN muss das distanzbasierte Verfahren praktisch nicht trainiert werden, da das unterliegende Maß nicht explizit berechnet wird, dies kommt im Lemma 2.7 zum Ausdruck. Stattdessen werden zum Zeitpunkt der Anfrage alle nötigen Distanzen bestimmt und der Klassifikator aus dieses berechnet, das macht die Auswertung allerdings auch erheblich aufwendiger als bei trainingsorientierten Ansätzen. Das empirische Verfahren kann sich gut einer komplexen Geometrie des Datenmateriales anpassen, insbesondere in niedrigen Dimensionen, was auf eine flexible Topologie der medialen Achse hindeutet. Defizite ergeben sich bei einfacheren Datensätzen, zum Beispiel linear trennbaren. Der Generalisierungsfehler

von Klassifikatoren mit starreren geometrischen Modellen - hier idealerweise natürlich die lineare Support Vector Machine - ist in diesem Fall deutlich geringer, da der distanzbasierte Ansatz Überanpassung begünstigt. Das ist zunächst kontraintuitiv, übereinstimmend vermutet wird, dass das empirische Verfahren bei wachsendem Stichprobenumfang gegen den idealen also den *Bayes-Klassifikator* strebt, dies aber bekanntlich für die SVM nicht gilt.

Es wäre nun wünschenswert ein Verfahren zur Hand zu haben, dass sich den verschiedenen Anwendungssituationen anpassen kann. Denkbar ist beispielsweise eine nutzerseitige Regulierung der Kapazität - also der geometrischen Flexibilität - des zugrundeliegenden Modells je nach Herkunft des Datenmaterials oder entsprechend eventueller Vorkenntnisse über die Verteilung der Lernbeispiele. Unter dem Gesichtspunkt der informatischen Forschung wäre außerdem hilfreich, wenn auch die neu entstehende Abbildung - aufgefasst als Funktion ihrer Parameter wieder differenzierbar wäre, da dies die Anwendung globaler Methoden der Analysis erleichtert.

2.3 Das Kugelmaß

Eine Möglichkeit ein anpassungsfähigeres Verfahren zu erreichen besteht darin, den Einfluss der Masse eines einzelnen Punktes auf das stochastische Maß über einen größeren Bereich zu verwischen. Statt eines singulären Sprunges wie beim empirischen Fall soll ein Datenvektor nun einen möglichst gleichmäßigen Anstieg von $\delta_{\mu;m}$ erzeugen. Es ist allerdings dennoch zunächst nötig den Bereich dieses Einflusses zu beschränken, damit in der aus dem Maß konstruierten Distanzfunktion tatsächlich noch Informationen über die relative Lage der Lernbeispiele zueinander erhalten bleibt. Hätte die Wirkung eines jeden Punktes unendliche Reichweite so bestünde die Gefahr, dass der entworfene Klassifikator seine Bindung an die Datenmaterial verliert, was die Idee von Abstandsmaßen schließlich kontakarieren würde.

Der hier gewählte Ansatz besteht darin, das Gewicht eines Datenvektors uniform auf eine abgeschlossene Kugel um den Punkt zu verteilen. Diese Kugel soll ausserdem in ihrem Radius skalierbar sein, was die gewünschte Regulierung umsetzt. Das entspricht dem Übergang von einer Punktwolke C als Bezugsmenge zu einer Vereinigung von endlich vielen d-dimensionalen Kugeln $C + B_{r_0} = \{x \in \mathbb{R}^d \,|\, d_C(x) \leq r_0\}$, wobei $+$ hier die Minkowski-Summe notiert. Das neue Maß soll nun einer Borel-Menge den von ihr überdeckten Anteil an dem Gesamtvolumen der Kugelmenge zuordnen.

Definition 2.12 (Kugelmaß)**:**

Sei λ das d-dimensionale Lebesgue-Maß, C eine Punktwolke mit Mächtigkeit n und $r_0 > 0$ reell, so bezeichne

$$\mu_{n;r_0} \colon \mathcal{B}\left(\mathbb{R}^d\right) \to \bar{\mathbb{R}} \; ; \; A \mapsto \frac{\lambda(A \cap (C + B_{r_0}))}{\lambda(C + B_{r_0})}$$

das Kugelmaß *bezüglich C mit Parameter r_0.*

□

Lemma 2.13:

Das Kugelmaß ist tatsächlich ein normiertes Maß mit kompaktem Träger $C + B_{r_0}$.

Beweis. Zunächst überzeugt sich mensch davon, dass $C + B_{r_0}$ positives, endliches Lebesgue-Maß besitzt, also $\mu_{n;r_0}$ als Abbildung existiert. Sei dazu $C = \{x_1; \cdots ; x_n\}$, dann lässt sich die Bezugsmenge schreiben als $C + B_{r_0} = \bigcup_{i=1}^{n} B(x_i; r_0)$ und dessen Volumen ist wegen der Translationsinvarianz von λ von unten durch $\lambda(B_{r_0}) = r_0^d \pi^{d/2} / \Gamma(\frac{d}{2} + 1) > 0$ und von oben durch $n \cdot \lambda(B_{r_0}) < \infty$ beschränkt, wobei Γ die Gamma-Funktion bezeichne. Die Monotonie des Lebesgue-Maßes sichert außerdem $0 = \mu_{n;r_0}(\emptyset) \leq \mu_{n;r_0}(A) \leq \mu_{n;r_0}(\mathbb{R}^d) = 1$ für jede messbare Menge A. Sei nun $\{A_i\}_{i \in \mathbb{N}}$ eine abzählbare Familie paarweise disjunkter Borel-Mengen, so gilt:

$$\left(\bigcup_{i \in \mathbb{N}} A_i \right) \cap (C + B_{r_0}) = \bigcup_{i \in \mathbb{N}} (A_i \cap (C + B_{r_0}))$$

Letzteres ist selbst wieder eine Vereinigung von abzählbar vielen disjunkten messbaren Mengen. Die Σ-Additivität von λ ausnutzend ergibt sich nun:

$$\mu_{n;r_0} \left(\bigcup_{i \in \mathbb{N}} A_i \right) = \frac{\sum_{i \in \mathbb{N}} \lambda(A_i \cap (C + B_{r_0}))}{\lambda(C + B_{r_0})} = \sum_{i \in \mathbb{N}} \frac{\lambda(A_i \cap (C + B_{r_0}))}{\lambda(C + B_{r_0})} = \sum_{i \in \mathbb{N}} \mu_{n;r_0}(A_i)$$

Es bleibt zu zeigen, dass $\operatorname{supp}(\mu_{n;r_0}) = C + B_{r_0}$ gilt. Die Kompaktheit ergibt sich dann unmittelbar durch die obige Darstellung als endliche Vereinigung von abgeschlossenen, beschränkten Mengen.

Allerdings ist $\mathbb{R}^d \setminus (C + B_{r_0})$ gerade die größte offene $\mu_{n;r_0}$-Nullmenge: Als Komplement einer kompakten Menge ist sie zunächst einmal offen und sie hat per definitionem Maß 0. Elementare Topologie ergibt außerdem, dass jede ihrer offenen echten Obermengen notwendig das Innere von $C + B_{r_0}$ treffen muss. Da nicht-leere offene Mengen aber stets positives Lebesgue-Maß besitzen, wird für diese Obermengen auch $\mu_{n;r_0}$ positiv.

qed

Aus den Ausführungen im letzten Abschnitt dieser Arbeit ergibt sich jetzt sofort die Pseudo-Distanzfunktion

$$\delta_{\mu_{n;r_0};m}(x) = \inf\{r > 0 \,|\, \mu_{n;r_0}(B(x;r)) > m\},$$

die außerdem bei wachsendem $\|x\|$ divergiert sowie die Distanzfunktion $d_{\mu_{n;r_0};m_0}$, definiert durch

$$d^2_{\mu_{n;r_0};m_0}(x) = \frac{1}{m_0} \int_0^{m_0} \delta^2_{\mu_{n;r_0};m}(x) \, \mathrm{d}m$$

und schließlich der entsprechende Klassifikator:

$$c_{n;r_0;m_0}(x) = \operatorname{sgn}(d_{\mu^-_{n;r_0};m_0}(x) - d_{\mu^+_{n;r_0};m_0}(x))$$

Auch die oben erwähnten Konvergenzeigenschaften der empirischen Schätzer $\mu_{n;r_0}$, der Distanzfunktion $d_{\mu_{n;r_0};m_0}$ sowie des Bisektors $B^{\pm}_{n;r_0;m_0} = B^{\pm}_{c_{n;r_0;m_0}}$ für

$n \to \infty$ übertragen sich unmittelbar auf das Kugelmaß. Wird der Umfang des Datenmaterials sowie die Kennzahl m_0 für den Anteil der zu berücksichtigenden Lernbeispiele fixiert, so ergibt sich eine Familie $\{c_{n;r_0;m_0}\}_{r_0}$ von Klassifikatoren, die in den positiven reellen Zahlen parametrisiert ist. Es ist außerdem zweckmäßig den empirischen Klassifikator für $r_0 = 0$ hinzuzunehmen. Dies geschieht nicht bloß, weil anschaulich mit verschwindendem Radius die Kugelmenge $C+B_{r_0}$ zur Punktwolke C entartet, sondern auch, weil $c_{n;m_0}$ in allen wesentlichen Eigenschaften als Grenzwert der Klassifikatoren $c_{n;r_0;m_0}$ für $r_0 \searrow 0$ aufgefasst werden kann.

Lemma 2.14:

Unter den gegebenen Bezeichnungen gilt folgende Konvergenz:

$$\lim_{r_0 \searrow 0} \|d_{\mu_{n;r_0};m_0} - d_{\mu_n;m_0}\|_\infty = 0$$

Beweis. Es genügt hier für hinreichend kleine r_0 nachzuweisen, dass sich $\|\delta_{\mu_{n;r_0};m} - \delta_{\mu_n;m}\|_\infty$ für beliebige $0 \leq m < 1$ durch r_0 nach oben beschränken lässt.

Sei dafür $x \in \mathbb{R}^d$ ein Punkt und k die kleinste ganze Zahl echt größer als mn, wieder bezeichne $X_i(x)$ den i-ten nächsten Nachbarn von x. Offenbar lässt sich r_0 so klein wählen, dass jede Kugel $B(x_i; r_0)$ stets nur noch einen einzigen Datenpunkt - nämlich gerade das Zentrum x_i - enthält. Es gilt dann:

$$\mu_{n;r_0}(B(x; \|X_k(x) - x\| - r_0)) = \frac{k-1}{n} \leq m$$
$$\mu_{n;r_0}(B(x; \|X_k(x) - x\| + r_0)) = \frac{k}{n} > m$$

Woraus zusammen $\|X_k(x) - x\| - r_0 < \delta_{\mu_{n;r_0};m}(x) \leq \|X_k(x) - x\| + r_0$ folgt, die bereits bekannte Aussage $\delta_{\mu_n;m}(x) = \|X_k(x) - x\|$ vollendet nun den Beweis.

qed

Der andere Grenzwertes - für $r_0 \to \infty$ - ist dagegen deutlich aufwendiger zu bestimmen und wird erst im 4. Kapitel der Arbeit untersucht. Zunächst einmal soll aber ein Algorithmus zur eigentlichen Auswertung des Klassifikators entwickelt werden.

Kapitel 3

Berechnung

3.1 Der Duale Komplex

Die Essenz der Auswertung des neuen Klassifikators besteht in der Bestimmung der im vorangegangenen Kapitel definierten Distanzen. Anders als beim empirischen Ansatz ist es dafür aber nun notwendig das zugrundeliegende Maß tatsächlich explizit zu berechnen. Liegt hierfür erst einmal ein Algorithmus vor, so ist auch klar, wie dieser zur Abstandsbestimmung eingesetzt werden kann und schließlich zur eigentlichen Klassifikation führt. Für die Berechnung des Kugelmaßes müssen Vorschriften entwickelt werden, die angeben wie sich zum einen das Volumen der Vereinigung von endlich vielen Kugeln zum anderen aber auch deren Durchschnitt bestimmen lässt. Es wird sich zeigen, dass die Lösungen für beide Aufgaben sehr eng ineinander greifen und es daher möglich ist, einmal gewonnene Methoden auf mehrere Teilprobleme im Laufe der Berechnung anzuwenden. Parallel zu dieser Arbeit befindet sich eine Referenzimplementierung in Entwicklung, vergleiche dazu Biberhofer [4]. Für eine bessere Visualisierung der Ergebnisse wird sich hier für die weiteren Betrachtungen auf den Fall $d = 2$, also die Euklidische Ebene, konzentriert. Es sei im Weiteren außerdem gefordert, dass die Lerndaten - aufgefasst als Punkte des affinen Raumes - in allgemeiner Lage vorliegen. Da es sich hierbei um recht starke Voraussetzungen handelt, sei an dieser Stelle ausdrücklich auf deren kritische Diskussion im Abschnitt 3.4 dieser Arbeit verwiesen.

Für die hier gewählte Darstellung der Berechnung werden noch einige zusätzlichen Begriffe benötigt. Das sind vor allem geometrische Diagramme, die zugrundeliegenden Ideen stammen dabei von Attali und Edelsbrunner und sind in [1, 7] dokumentiert. Sie wurden für die Verwendung in diesem Verfahren geringfügig angepasst. Im Weiteren notiere nun B stets eine Menge von n Vollkreisen der Form $b = B(x_b; r_b)$ in allgemeiner Lage. Es ist für die hier angestellten Betrachtungen nicht notwendig, streng zwischen dem System B und dem von ihm erzeugten abstrakten Simplizialkomplex $\wp(B)$ zu unterscheiden, so sei von der Eckenmenge B immer als mit allen ihren Teilmengen gedacht. Das kanonische Bild von B ist dann eine Punktwolke - natürlich ebenfalls in allgemeiner Lage - der Ebene, insbesondere ist die konvexe Hülle des Bildes einer höchstens dreielementigen Teilmenge von B nach Voraussetzung ein Simplex. Allerdings sei bemerkt, dass der gesamte Komplex nicht notwendig kanonisch realisierbar sein muss. Im vorliegenden

Fall lässt sich der Polyeder trotz der gebotenen Abstraktion sinnvoll definieren, so sei dann $|B| = \bigcup_{b \in B} b$ also eine endliche Vereinigung von Kreisen. Nun lassen sich mit Hilfe von B eine Reihe weiterer Komplexe definieren, die sich bei der Berechnung als hilfreich erweisen.

Definition 3.1 (Power-Diagramm)**:**

Sei $x \in \mathbb{R}^2$ ein Punkt und $b \in B$ ein Kreis, dann sei:

(*I*) *Die* Power-Distanz *von x zu b durch $\pi_b(x) = \|x - x_b\|^2 - r_b^2$ erklärt.*

(*II*) *Die* Power-Zelle *von b sei dann $p_b = \{x \in \mathbb{R}^2 \mid \forall b' \in B \colon \pi_b(x) \leq \pi_{b'}(x)\}$ und allgemeiner für eine Teilmenge $T \subseteq B$ sei $p_T = \bigcap_{b \in T} p_b$.*

(*III*) *Das* Power-Diagramm *von B sei schließlich $\mathcal{P} = \mathcal{P}(B) = \{p_T \mid \emptyset \neq T \subseteq B\}$*

□

Bemerkung. Die Power-Zellen sind auf Grund der allgemeinen Lage von B entweder leer oder (3-$|T|$)-dimensionale konvexe Polyeder. Da außerdem aus $T \subseteq T'$ auch $p_T \supseteq p_{T'}$ folgt, lässt sich $\mathcal{P}$ selbst als ein (abstrakter) Simplizialkomplex auffassen. Aus topologischer Sicht ist er genauer sogar eine Zellenzerlegung der gesamten Euklidischen Ebene. Ein wichtiger Spezialfall ergibt sich, wenn alle Radii r_b übereinstimmen, also alle Power-Distanzen gleich gewichtet sind. Dies deckt sich nämlich mit dem ungewichteten Fall und damit mit dem klassischen *Voronoi-Diagramm*.

Lemma und Definition 3.2:

Falls ein Kreis aus B die Power-Zelle eines anderen trifft, so trifft sie sogar diesen Kreis selbst. Dies motiviert die Definitionen $q_b = p_b \cap b$ und allgemeiner $q_T = \bigcap_{b \in T} q_b$. Dann gilt für für jedes $T \subseteq B$, $b \in T$ und $b' \in B \setminus T$ die Beziehung:

$$q_T = p_T \cap b \supseteq p_T \cap b'$$

Daraus
folgt schließlich, dass das System $\mathcal{Q} = \{q_T \mid \emptyset \neq T \subseteq B\} = \{p_T \cap |B| \mid p_T \in \mathcal{P}\}$ der Schnitt des Power-Diagramms mit dem Polyeder $|B|$ ist.

□

Beweis. Seien $b; b' \in B$ zwei Kreise. Zu zeigen: $p_b \cap b' \subseteq p_b \cap b$

Zum Beweis genügt offenbar schon $p_b \cap b' \subseteq b$. Beachtet mensch, dass sowohl Quadrieren als auch Wurzelziehen monotone Funktionen im Sinne der

Ordnungstheorie sind, ergibt sich zusammen:

$$\begin{aligned} & x \in p_b \cap b' \\ \Rightarrow\ & \pi_b(x) \leq \pi_{b'}(x) \wedge \|x - x_{b'}\| \leq r_{b'} \\ \Rightarrow\ & \|x - x_b\|^2 - r_b^2 \leq \|x - x_{b'}\|^2 - r_{b'}^2 \wedge \|x - x_{b'}\|^2 \leq r_{b'}^2 \\ \Rightarrow\ & \|x - x_b\|^2 - r_b^2 \leq \|x - x_{b'}\|^2 - r_{b'}^2 \leq 0 \\ \Rightarrow\ & \|x - x_b\|^2 \leq r_b^2 \\ \Rightarrow\ & \|x - x_b\| \leq r_b \\ \Rightarrow\ & x \in b \end{aligned}$$

Alle anderen Inklusionen ergeben sich dann sofort als einfache Folgerungen unter Zuhilfenahme der Definition 3.1 und den Eigenschaften der Durchschnittsbildung.

qed

Bemerkung. Analog zu $\mathcal{P}$ ist $\mathcal{Q}$ ebenfalls eine Zellenzerlegung - diesmal von $|B|$ - und ist damit insbesondere wieder ein simplizialer Komplex.

Im Folgenden sei noch eine weitere Eigenschaft von $\mathcal{Q}$ gesondert notiert.

Korollar 3.3:

Falls eine Zelle q_T nicht leer ist, so gilt dies auch für den Durchschnitt $\bigcap_{b \in T} b$.

Beweis. Nach den Überlegungen im Lemma 3.2 gilt $\forall b \in T \colon q_T \subseteq b$ und damit $q_T \subseteq \bigcap_{b \in T} b$.

qed

Zu diesen abstrakten Konstruktionen lassen sich die in einem gewissen Sinne *dualen* Komplexe definieren, diese sind nun tatsächlich auch geometrische Simplizialkomplexe. Als abkürzende Schreibweise bezeichne im Weiteren σ_T für eine Teilmenge $T \subseteq B$ das kanonische Bild von T und $b_T = \bigcap_{b \in T} b$ den Schnitt der entsprechenden Kreise.

Definition 3.4 (Reguläre Triangulation, Dualer Komplex)**:**

Seien die Bezeichnungen wie oben.

(*I*) $\mathcal{R} = \mathcal{R}(B) = \{\sigma_T \,|\, \emptyset \neq p_T \in \mathcal{P}\} \cup \{\emptyset\}$ *heiße die* reguläre Triangulation *von* B.

(*II*) $\mathcal{K} = \mathcal{K}(B) = \{\sigma_T \,|\, \emptyset \neq q_T \in \mathcal{Q}\} \cup \{\emptyset\}$ *heiße der* Duale Komplex *zu* $\mathcal{Q}$.

□

Bemerkung. Die angesprochende Dualität wird durch den Umstand vermittelt, dass $\sigma_T \neq \emptyset$ genau dann in $\mathcal{R}$ liegt, wenn $p_T \neq \emptyset$ in $\mathcal{P}$ liegt, analoges gilt für $\mathcal{K}$ und $\mathcal{Q}$.

Lemma 3.5:

Seien die Bezeichnungen wie oben.

(*i*) *$\mathcal{R}$ und $\mathcal{K}$ sind geometrische simpliziale Komplexe, insbesondere ist $\mathcal{K} \leq \mathcal{R}$ ein Unterkomplex.*

(*ii*) *Für ein nicht-leeres $\emptyset \neq T \subseteq B$ mit $\sigma_T \in \mathcal{K}$ ist b_T nicht-leer.*

Beweis.

(*i*) Für Teilmengen T mit $|T| > 3$ sind die p_T und q_T nach Voraussetzung der allgemeinen Lage allesamt leer. Daher sind die σ_T tatsächlich stets Simplizes im $\mathbb{R}^2$. Ist weiter $s \leq \sigma_T$ eine Seite, so muss es eine Teilmenge $\emptyset \neq S \subseteq T$ geben mit $s = \sigma_S$ oder aber $s = \emptyset$. Im ersten Fall gilt - wie bereits bei den jeweiligen Definitionen angesprochen - $p_S \supseteq p_T$ beziehungsweise $q_S \supseteq q_T$ und damit $\emptyset \neq p_S \in \mathcal{P}$ sowie $\emptyset \neq q_S \in \mathcal{Q}$ wie gewünscht, der zweite Fall ist explizit behandelt. Tatsächlich dient das Hinzufügen der leeren Menge einzig und allein der Vervollständigung der Mengensysteme als Simplizialkomplexe. Wegen der Eigenschaft von $\mathcal{P}$ als Zellenzerlegung der Ebene können sich außerdem je zwei entstehende Simplizes ausschließlich in der konvexen Hülle der ihnen gemeinsamen Ecken schneiden, dies ist dann jeweils eine Seite von beiden. Schließlich liegt jeder Simplex aus $\mathcal{K}$ wegen $q_T \subseteq p_T$ (vergleiche Definition 3.2) auch in $\mathcal{R}$.

(*ii*) Dies ist einfach die Übertragung von Korollar 3.3 auf den dualen Komplex.

qed

Bemerkung. Analog zum Power-Diagramm geht der duale Komplex im gleich- beziehungsweise ungewichteten Fall in die wohlbekannte *Delaunay-Triangulation* über.

3.2 Endliche Vereinigungen von Kreisscheiben

Der erste Schritt bei der Berechnung des Kugelmaßes ist die Bestimmung der Gesamtfläche der Bezugsmenge $C + B_{r_0}$. Diese ergibt sich in der Theorie durch eine Formel nach dem Prinzip von Inklusion und Exklusion, die die Fläche des Schnittes jeder Menge mit einer geraden Anzahl von Kreisen aufsummiert und im ungeraden Fall abzieht. Diese Berechnungsmethode ist aber praktisch unbrauchbar, weil bei ihr die Zahl der Terme exponentiell in der Anzahl n der Kreise wächst. Allerdings lässt sich für eine effizientere Lösung ein fundamentaler Zusammenhang zwischen dem dualen Komplex einer Menge von Kreisscheiben und der von ihr überdeckten Fläche ausnutzen, diese Aussage stammt aus [7].

Theorem 3.6 (Edelsbrunner)**:**

Fasst mensch $C + B_{r_0}$ wieder als endliches System von Kreisen auf und überträgt die Bezeichnungen von oben, insbesondere sei $\mathcal{K} = \mathcal{K}(C + B_{r_0})$ der duale Komplex, so gilt:

$$\lambda(C + B_{r_0}) = \sum_{\sigma_T \in \mathcal{K}} (-1)^{|T|-1} \lambda(b_T)$$

■

Die Bedeutung dieses Lemmas ergibt sich durch die beiden Umstände, dass für den dualen Komplex - hier nämlich der Delaunay-Triangulation - effiziente Algorithmen zur Verfügung stehen deren Rechenzeit im Bereich von $\mathcal{O}(n \log n)$ liegt und die Summe nur noch über linear viele Terme läuft, vergleiche dazu DE BERG et al. [3, S. 205f.].

Offen ist noch die Berechnung der Flächen $\lambda(b_T)$, dabei ist zu beachten, dass auf Grund der allgemeinen Lage nur die Fälle $1 \leq |T| \leq 3$ vorkommen können. Im einfachste Fall eines einzelnen Kreises ergibt sich schlicht die Kreisformel $\lambda(b) = \pi r_0^2$. $|T| = 2$ führt auf einen Spezialfall des Kreisschnittproblems der planaren Geometrie, wobei beide Kreise gleiche Radii haben. Dessen Lösung [19] für Kreise $b_1 = B(x_1; r_0); b_2 = B(x_2; r_0)$ ist bekanntlich

$$\lambda(b_1 \cap b_2) = 2r_0^2 \arccos\left(\frac{\|x_1 - x_2\|}{2r_0}\right) - \frac{1}{2}\|x_1 - x_2\|\sqrt{4r_0^2 - \|x_1 - x_2\|^2}$$

Deutlich komplizierter ist die Bestimmung des gemeinsamen Schnittes dreier Kreise. Zu diesem Problem existiert eine umfangreiche Arbeit [10] von FEWELL. Aus dieser werden die folgenden Formeln und Berechnungsvorschriften bloß übernommen, zusätzlich wird deren Herleitung kurz umrissen. Der Autor beleuchtet dort ausschließlich den Fall, dass sich drei Kreisscheiben in der Ebene in einem konvexen Kreisbogendreieck mit nicht-leerem Inneren schneiden. Im Weiteren bezeichne der einfache Begriff *Kreisbogendreieck* nur noch ein ebensolches, wenn nicht anders angegeben sei es außerdem *nicht* zu einem Vollkreis entartet. Es muss sich nun kurz davon überzeugt werden, dass der genannte Fall für diese Betrachtung ausreichend ist. Dazu wird noch ein zusätzlicher Begriff gewissermaßen als Verbindungsstück benötig, er stammt aus [1].

Definition 3.7 (Unabhängiger Simplex)**:**

Ein abstrakter Simplex s, dessen Ecken aus Kreisscheiben bestehen, heiße unabhängig*, falls es für jede - auch leere - Teilmenge $t \subseteq s$ einen Punkt des $\mathbb{R}^2$ gibt, der in* jedem *Kreis aus t aber in* keinem *aus $s \setminus t$ liegt.*

□

Bemerkung. Es gibt tatsächlich unabhängige Simplizes in der Ebene, so ist das klassische *Venn-Diagramm* dreier Kreise ein Beispiel.

Lemma 3.8:

Bei einem unabhängigen 2-Simplex schneiden sich die beteiligten Kreise in einem Kreisbogendreieck.

Beweis. Auf Grund der allgemeinen Lage besitzen alle betrachteten Durchschnitte genau dann ein nicht-leeres Inneres, wenn sie selbst nicht leer sind. Auch die Konvexität ergibt sich sofort per definitionem.

Nach [10] schneiden sich nun drei Kreise der Ebene entweder in einem Kreisbogendreieck, einem Vollkreis, einer Linse - das heißt dem Schnitt genau zweier Kreise - oder überhaupt nicht. Alle anderen Fälle außer dem Kreisbogendreieck

widersprechen aber der Unabhängigkeit. So sind leere Schnitte unmöglich, die Linse müsste ganz im dritten Kreis enthalten sein und der Vollkreis wäre mit einem der drei betrachteten Kreise identisch.

qed

Bemerkung. Es gilt auch die Umkehrung des obigen Satzes, diese wird aber hier nicht weiter betrachtet.

Die Feststellung von Attali und Edelsbrunner in [1], dass alle Teilmengen $T \subseteq B$ derart, dass $\sigma_T \in \mathcal{K}$ ist, unabhängig sind, in Verbindung mit Lemma 3.8 rechtfertigt nun die Konzentration auf die Formeln aus [10].

Fewell gibt in seiner Arbeit zunächst eine geschlossene Flächeninhaltsformel für Kreisbogendreiecke an, die die Radii und die Abstände zwischen je zwei Eckpunkten benutzt. Letztere liegen hier aber nicht vor und müssen daher gesondert berechnet werden. Er verwendet dafür drei verschiedene Koordinatensysteme, die jeweils auf bestimmte Weiße durch Rotation und Translation auseinander hervorgehen. Die Koordinatenursprünge liegen dabei jeweils in den Zentren der drei Kreisscheiben und auch die Achsenorientierung hängt von deren relativen Lage ab. In jedem einzelnen System berechnet er die Koordinaten desjenigen Schnittpunktes der gerade betrachteten Kreise, der schließlich einen Eckpunkt des Dreiecks bildet. Zuletzt überführt er die lokalen Koordinaten durch Rücktransport in das erste System und bestimmt in diesem die Länge der Verbindungsstrecken. Außerdem muss noch der Spezialfall behandelt werden, dass einer der Kreise mehr als die Hälfte seiner Fläche zu dem Dreieck beiträgt. Im Weiteren haben zunächst alle Kreise identische Radii. Bei der Notation ist zu beachten, dass die verwendeten Doppelindizes nur der Benennung dienen, ihre Reihenfolge ist irrelevant.

Theorem 3.9 (Fewell)**:**

Angenommen die Kreise $B(z_1; r_0)$, $B(z_2; r_0)$ und $B(z_3; r_0)$ bilden einen unabhängigen Simplex, dann lässt sich die Fläche A des Schnittes dieser Kreise wie folgt berechnen:

Seien $d_{ij} = \|z_j - z_i\|$ die paarweisen Abstände der Zentren, die lokalen Koordinaten der Eckpunkte gleich

$$\begin{array}{ccc} x_{12} = d_{12}/2 & x'_{13} = d_{13}/2 & x''_{23} = d_{23}/2 \\ y_{12} = \sqrt{r_0^2 - d_{12}^2/4} & y'_{13} = \sqrt{r_0^2 - d_{13}^2/4} & y''_{23} = \sqrt{r_0^2 - d_{23}^2/4} \end{array}$$

und weiter

$$\begin{array}{cc} \cos\theta' = \frac{d_{12}^2 + d_{13}^2 - d_{23}^2}{2 d_{12} d_{13}} & \sin\theta' = \sqrt{1 - \cos\theta'} \\ \cos\theta'' = \frac{d_{12}^2 + d_{23}^2 - d_{13}^2}{2 d_{12} d_{23}} & \sin\theta' = \sqrt{1 - \cos\theta''} \end{array}$$

die Rotatonswinkel. Dann sind

$$\begin{array}{cc} x_{13} = x'_{13} cos\theta' - y'_{13} \sin\theta' & x_{23} = x''_{23} \cos\theta'' - y''_{23} \sin\theta'' + d_{12} \\ y_{13} = x'_{13} \sin\theta' + y'_{13} \cos\theta' & y_{23} = x''_{23} \sin\theta'' + y''_{23} \cos\theta'' \end{array}$$

die Koordinaten im ersten System und daher beträgt die Länge der

Verbindungsstrecken jeweils $c_k = \sqrt{(x_{ik} - x_{jk})^2 + (y_{ik} - y_{jk})^2}$.

Die Bedingung, ob ein Kreis zu mehr als die Häfte im Dreieck liegt lautet:

$$(*) \quad d_{13} \sin\theta' < y_{13} + \frac{y_{23} - y_{13}}{x_{23} - x_{13}}(d_{13}\cos\theta' - x_{13})$$

Schließlich ergibt sich die gesuchte Fläche dann zu

$$A = \frac{1}{4}\sqrt{(c_1 + c_2 + c_3)(c_2 + c_3 - c_1)(c_1 + c_3 - c_2)(c_1 + c_2 - c_3)} + \sum_{k=1}^{3} r_0^2 \arcsin\frac{c_k}{2r_0}$$
$$- \sum_{k=1}^{2} \frac{c_k}{4}\sqrt{4r_0^2 - c_k^2} + \begin{cases} \frac{c_3}{4}\sqrt{4r_0^2 - c_3^2} & \textit{falls } (*) \textit{ wahr ist} \\ \frac{-c_3}{4}\sqrt{4r_0^2 - c_3^2} & \textit{sonst} \end{cases}$$

■

Der wohlbekannte Algorithmus, der die Delaunay-Triangulation $\mathcal{K}$ der Punktwolke C bestimmt, vervollständigt schließlich die Berechnungsvorschrift für das Lebesgue-Maß $\lambda(C+B_{r_0})$. Bisher wurde der eigentliche zu klassifizierende Punkt bei der Berechnung noch überhaupt nicht verwendet, der hier dargestellte Teil muss daher auch nur ein einziges Mal zur Laufzeit ausgeführt werden. Von diesem Moment an steht der Inhalt der Bezugsmenge des Klassifikators für den gesamten folgenden Rechenweg zur Verfügung.

3.3 Hinzunahme des zu klassifizierenden Punktes

Der nächste Schritt, nämlich die Berechnung von $\lambda(B(x;r) \cap C + B_{r_0})$ für einen zu klassifizierenden Punkt $x \in \mathbb{R}^2$ und positives r, verfolgt einen ganz ähnlichen Ansatz. Zunächst wird der duale Komplex $\mathcal{K}_r = \mathcal{K}(B(x;r) \cup C + B_{r_0})$ berechnet. Hierfür wurde von Edelsbrunner und Shah in [9] ein inkrementeller Algorithmus mit Laufzeit in $\mathcal{O}(n \log n)$ im zweidimensionalen Fall angegeben. Nun ergibt sich das Maß der betrachteten Fläche als

$$\lambda(B(x;r) \cap C + B_{r_0}) = \sum_{\substack{\sigma_T \in \mathcal{K}_r \\ x \in \sigma_T \wedge |T| > 1}} (-1)^{|T|} \lambda(b_T)$$

Die Korrektheit dieser Berechnung ergibt sich sofort aus den Eigenschaften des dualen Komplexes und dem Prinzip von Inklusion und Exklusion. Allerdings muss nun die Berechnung der $\lambda(b_T)$ für den Fall ungleicher Radii angepasst werden. Bei $|T| = 2$ ergibt sich - wieder nach [19] - für zwei Kreise $b_1 = B(x_1; r_0); b_2 = B(x; r)$ mit dem Mittelpunktsabstand $d = \|x_1 - x\|$ die elementargeometrische Flächeninhaltsformel

$$\lambda(b_1 \cap b_2) = r_0^2 \arccos\left(\frac{d^2 + r_0^2 - r^2}{2dr_0}\right) + r^2 \arccos\left(\frac{d^2 + r^2 - r_0^2}{2dr}\right)$$
$$- \frac{1}{2}\sqrt{(-d + r_0 + r)(d + r_0 - r)(d - r_0 + r)(d + r_0 + r)}$$

Die Formeln für den Fall $|T| = 3$ stammen erneut aus [10], auch die Korrektheit ihrer Anwendung ergibt sich wie im vorherigen Abschnitt. Es ist dabei zu beachten, dass sie den noch allgemeineren Fall *dreier* unterschiedlicher Radii behandeln, diese werden als absteigend nach ihrer Größe geordnet aufgefasst. Tatsächlich unterscheiden sich also die Berechnungen für $r \leq r_0$ und $r > r_0$, allerdings nur im Detail.

Theorem 3.10 (Fewell)**:**

Angenommen die Kreise $B(z_1; r_1)$, $B(z_2; r_2)$ und $B(z_3; r_3)$ mit $r_1 \geq r_2 \geq r_3$ schneiden sich in einem Kreisbogendreieck, so lässt sich dessen Fläche analog zum Theorem 3.9 bestimmen, nur die Berechnung der lokalen Koordinaten der Eckpunkte unterscheidet sich:

Alle Bezeichnungen seien wie in Theorem 3.9, dann gilt:

$$x_{12} = \frac{r_1^2 - r_2^2 + d_{12}^2}{2d_{12}} \qquad y_{12} = \frac{1}{2d_{12}} \sqrt{2d_{12}^2(r_1^2 + r_2^2) - (r_1^2 - r_2^2)^2 - d_{12}^4}$$

$$x'_{13} = \frac{r_1^2 - r_3^2 + d_{13}^2}{2d_{13}} \qquad y'_{13} = \frac{-1}{2d_{13}} \sqrt{2d_{13}^2(r_1^2 + r_3^2) - (r_1^2 - r_3^2)^2 - d_{13}^4}$$

$$x''_{23} = \frac{r_2^2 - r_3^2 + d_{23}^2}{2d_{23}} \qquad y''_{23} = \frac{1}{2d_{23}} \sqrt{2d_{23}^2(r_2^2 + r_3^2) - (r_2^2 - r_3^2)^2 - d_{23}^4}$$

■

Da nun die Berechnung des zweidimensionalen Kugelmaßes – zumindest für angeschlossene Vollkreise um einen Punkt $x \in \mathbb{R}^2$ - bekannt ist, lässt es sich für die auf diesem Maß basierende Klassifikation heranziehen. Zusammen mit der Infimumsbildung für $\delta_{\mu^-_{n;r_0};m}(x)$ beziehungsweise $\delta_{\mu^+_{n;r_0};m}(x)$ durch das fortlaufende Testen schrittweise ansteigender Radii r sowie der numerischen Integration zur Bestimmung von $d_{\mu^-_{n;r_0};m_0}(x)$ und $d_{\mu^+_{n;r_0};m_0}(x)$ ergibt sich daraus ein Algorithmus zum Beantworten der Anfragen $c_{n;r_0;m_0}(x)$ an den neuen Klassifikator.

3.4 Problembetrachtung

Die dargestellten Vorschriften liefern eine vollständige Anleitung zur Auswertung des neuentwickelten Klassifikators, allerdings ergeben sich für den entstehenden Algorithmus gleich mehrere Probleme. Es fällt deutlich auf, dass bei der Berechnung sehr ausgiebig und an entscheidenden Stellen die getroffenen Annahmen verwendet werden. Die meisten angegebenen Rechenwege sind in dieser Einfachheit nur durch das intensive Ausnutzen der geforderten allgemeinen Lage der Lerndaten aufrechtzuerhalten. Die Aufgabe dieser Annahme würde eine ausufernde Einzelfallbetrachtung zur Folge haben, so könnte beispielsweise nicht mehr davon ausgegangen werden, dass nur drei Spezialfälle von Kreisschnitten zu behandeln wären, selbst der gesamte Ansatz des dualen Komplexes müsste in der jetztigen Form auf den Prüfstand. Edelsbrunner und Mücke haben versucht in [8] zu zeigen, dass die Voraussetzung allgemeiner Lage keine deutliche Einschränkung

darstellt, wenn mensch die Eingangsdaten mit einem (simulierten) schwachen Rauschen versieht, um so degenerierte Fälle zu vermeiden. Je nach Herkunft des Datenmaterials kann sich diese Annahme dennoch als sehr problematisch erweisen, so zum Beispiel wenn die behandelten Punkte von der Abtastung eines flachen Werkstückes oder ähnlichen Anwendungen stammen. Hierfür müssten das hier vorgestellte Verfahren speziell angepasst werden.

Die Konzentration auf zwei Dimensionen verringert die praktische Anwendbarkeit des Algorithmus außerdem doch ganz erheblich. Während sich die Überlegungen zum Power-Diagramm und dem dualen Komplex noch wortgleich auf höhere Dimensionen übertragen lassen würden, betrachten die Formeln für die Kreischnitte ausschließlich den ebenen Fall. Sie skalieren praktisch überhaupt nicht mit wachsendem d. Außerdem müssen in höheren Dimensionen deutlich mehr als nur drei Spezialfälle betrachtet werden, so lässt sich zeigen, dass es im $\mathbb{R}^d$ ganz allgemein $d+1$ verschiedene Arten von unabhängigen Simplizes gibt und sie alle in der verwendeten Volumenformel auftreten können (vgl. [1]). Nicht zuletzt wächst auch der Aufwand zur Berechnung des dualen Komplex überdimensional in der Größe der Dimension, siehe dazu die Ausführungen in [9].

Dabei ist die Komplexität schon im Zweidimensionalen eine bedeutende Schwachstelle des Algorithmus. Eine einzige Auswertung des Maßes wird zwar durch die Berechnung des dualen Komplex dominiert und liegt daher ebenso in $\mathcal{O}(n \log n)$, doch dieser Teil muss im Laufe der Berechnung der Distanzfunktion immer wieder ausgeführt werden und stellt so den Hauptaufwand der gesamten Methode dar. Zum einen lässt sich die Bestimmung des gesuchten Infimum kaum anders realisieren, als den Radius r solange schrittweise zu erhöhen bis das Maß des betrachteten Kreises den aktuellen Anteil m überschreitet. Zum anderen muss auch bei der numerischen Integration eben dieser Anteil innerhalb der Intervallgrenzen 0 und m_0 immer wieder leicht erhöht werden und die Berechnung beginnt erneut. Es ist richtig, dass der genaue Aufwand des Algorithmus von den jeweils konkret gewählten Schrittweiten abhängt und es lässt sich ausnutzen, dass der Regulationspfad $m \mapsto \delta^2_{\mu_{n;r_0};m}$ nicht-fallend ist und daher bei steigendem m echt kleinere als der zuletzt bestimmte Radius r nicht mehr berücksichtigt zu werden brauchen. Aber mensch kommt nicht umhin zuzugeben, dass die Rechenzeit mit einer angemessenen Genauigkeit weit jenseits aller aktuell praxisrelevanter Methoden zu erwarten ist. Insbesondere ist die Laufzeit nicht mit der des empirischen Klassifikators zu vergleichen (für entsprechenden Werte einer Referenzimplementierung siehe [11]). Dies liegt in der einfachen Tatsache begründet, dass bei diesem das Maß nicht explizit berechnet werden muss, sondern mensch stattdessen für geeignete Parameterwerte auf eine Variante des k-Nächsten-Nachbarns ausweichen kann. Auch für den hier vorliegenden Fall wäre es wünschenswert ein möglichst einfaches Kriterium zur Hand zu haben, dass angibt, für welche Radii sich der duale Komplex der Bezugsmenge jeweils kombinatorisch verändert, denn dann könnte mensch in allen anderen Fällen von einer Neuberechnung absehen. Empirische Überprüfungen legen nahe, dass diese Veränderungen in der Regel nur bei einigen vergleichsweise wenigen Grenzradien eintritt. Im Ergebnis ließe sich der Gesamtaufwand der Methode erheblich nach unten korrigieren.

Erschwerend kommt bei der Berechnung der Distanzen hinzu, dass das

Integrieren schlicht ein numerisch instabiles Problem ist, dies wird für Parameterwerte m_0 nahe 0 durch die vorgenommene Normierung, also der Division durch eine betragskleine Zahl, noch verschärft. Es ist daher verfahrensbedingt bei der Verwendung gängiger Gleitkomma-Typen gemäß der Norm IEEE 754-2008 selbst bei doppelter Genauigkeit mit erheblichen Informationsverlusten zu rechnen.

Ein weiteres Problem liegt in der korrekten Wahl der beiden Parameter m_0 und r_0. Es ist in Ansätzen bekannt, wie diese sich einzeln auf die Berechnung auswirken - siehe dazu die Aussagen im Abschnitt 2.3 dieser Arbeit - doch über deren Korrelation und mögliche optimale Kombinationen in verschiedenen Anwendungsfällen liegen praktisch noch überhaupt keine Erkenntnisse vor.

Zusammenfassend wird klar, dass die Angabe einer Berechnungsvorschrift erst der allererste Schritt bei der Betrachtung einer neuen Klassifikationsmethode sein kann und es unbedingt nötig ist hier weitere Untersuchungen anzustellen, den Algorithmus tatsächlich lauffähig zu implementieren und stetig zu optimieren. Das nächste Kapitel beschäftigt sich daher mit weiteren Eigenschaften der definierten Familie von Klassifikatoren, nämlich ihrem Verhalten bei steigenden Bezugsradius r_0.

Kapitel 4

Zusammenfassung und Ausblick

4.1 Erste Grenzwertvermutung

Im Laufe der Untersuchungen zum Kugelmaß-Klassifikator liesen sich einige topologische Eigenschaften beobachten, für deren Beweis oder Widerlegung aber das Verständnis der Zusammenhänge des gerade erst definierten Klassifikationsverfahren noch nicht umfassend genug war. Dies betrifft vor allem ein bestimmtes Grenzverhalten der entstandenen Distanzfunktionen beziehungsweise der Bisektoren bezüglich einzelner Parameter sowie dem Zusammenspiel der Regulatoren untereinander. In diesem letzten Kapiteln sollen diese Sachverhalte als Motivation für weitere Forschungen in Form zweier zentraler Vermutungen formuliert werden. Es wird sich außerdem bemüht diese in den Kontext der bisherigen Betrachtungen einzubetten.

Die erste der beiden Vermutungen betrifft den Grenzwert der zum Kugelmaß gehörenden Distanzfunktion $d_{\mu_{n;r_0};m_0}$ für $m_0 \searrow 0$. Zunächst einmal soll aber eine schwächere Aussage zum gleichen Sachverhalt gezeigt werden. Dazu ist jedoch etwas Vorarbeit nötig.

Lemma 4.1:

Für jeden Parameter $r_0 > 0$ und jeden Punkt $x \in \mathbb{R}^d$ gibt es ein positives $m' > 0$ derart, dass die Abbildung $m \mapsto \delta_{\mu_{n;r_0};m}(x)$ auf dem Intervall $[0; m')$ stetig ist.

Beweis. In jedem $x \in \mathbb{R}^d$ lässt sich $\delta_{\mu_{n;r_0};m}(x)$ durch $d_{C+B_{r_0}}(x)$ nach unten und $d_{C+B_{r_0}}(x) + \max_{a \in C+B_{r_0}} \|x - a\|$ nach oben beschränken, die Kompaktheit der Bezugsmenge sichert die Endlichkeit dieser Werte. Als beschränkte, reelle, nichtfallende Funktion, die überall auf dem abgeschlossenen Einheitsintervall definiert ist, kommt für eine Unstetigkeitsstelle von $m \mapsto \delta_{\mu_{n;r_0};m}(x)$ nur ein positiver Sprung in Frage.

Ein solcher kann nur dann auftreten, wenn die Vergrößerung des Radius r auf einem Abschnitt positiver Länge vorübergehend keine Zunahme des Maßes $\mu_{n;r_0}(B(x;r))$ mehr zur Folge hat. Diese Länge ist im Übrigen dann auch genau die Höhe des Sprunges. Das kann aber nach Konstruktion des Kugelmaßes nur dann geschehen, wenn $B(x;r)$ gerade eine Kugel aus $C + B_{r_0}$ vollständig eingeschlossen hat und gleichzeitig keine weitere schneidet. Zum einen kann dies also nur an endlich

vielen - genauer höchstens n -Stellen der Fall sein. Zum anderen ist die betrachtete Abbildung in $m = 0$ stetig, denn $B(x; \delta_{\mu_{n;r_0};0}(x))$ hat per definitionem Maß 0 und enthält daher noch gar keine Kugel aus $C + B_{r_0}$.

Die Wahl von m' als die kleinste Sprungstelle - oder 1 falls es keine solche gibt - liefert nun die Behauptung.

qed

Diese Eigenschaft vereinfacht die Herleitung einer ersten Grenzwertaussage.

Lemma 4.2:

Die Abstandsfunktion des Trägers des Kugelmaßes ist der punktweise Grenzwert der Distanzfunktion des Maßes selbst, wenn m_0 gegen 0 geht. Mit anderen Worten, für jeden Punkt $x \in \mathbb{R}^d$ und jeden Radius $r_0 > 0$ gilt

$$\lim_{m_0 \searrow 0} d_{\mu_{n;r_0};m_0}(x) = d_{C+B_{r_0}}(x)$$

Beweis. Die Grenzwertsätze reeller Funktionen sichern

$$\lim_{m_0 \searrow 0} d^2_{\mu_{n;r_0};m_0}(x) = \left(\lim_{m_0 \searrow 0} d_{\mu_{n;r_0};m_0}(x) \right)^2$$

Da außerdem sowohl $d_{\mu_{n;r_0};m_0}$ als auch $d_{C+B_{r_0}}$ nicht-negativ sind, genügt es folgende Beziehung zu zeigen:

$$\lim_{m_0 \searrow 0} d^2_{\mu_{n;r_0};m_0}(x) = d^2_{C+B_{r_0}}(x)$$

Das obige Lemma impliziert, dass die Zuordnung $m_0 \mapsto \int_0^{m_0} \delta^2_{\mu_{n;r_0};m}(x)\,\mathrm{d}m$ auf dem Intervall $[0; m')$ eine Stammfunktion von $m \mapsto \delta^2_{\mu_{n;r_0};m}(x)$ ist. Nach der Regel von l'Hôpital ergibt sich nun

$$\begin{aligned}\lim_{m_0 \searrow 0} d^2_{\mu_{n;r_0};m_0}(x) &= \lim_{m_0 \searrow 0} \frac{\int_0^{m_0} \delta^2_{\mu_{n;r_0};m}(x)\,\mathrm{d}m}{m_0} \\ &= \lim_{m_0 \searrow 0} \frac{\mathrm{d}(\int_0^{m_0} \delta^2_{\mu_{n;r_0};m}(x)\,\mathrm{d}m)/\mathrm{d}m_0}{\mathrm{d}m_0/\mathrm{d}m_0} = \lim_{m_0 \searrow 0} \delta^2_{\mu_{n;r_0};m}(x)\end{aligned}$$

Der letzte Grenzwert ist aber nach Lemma 2.4.(*iii*) in Verbindung mit Lemma 2.13 gerade $d^2_{C+B_{r_0}}(x)$, wie gewünscht.

qed

Die eigentliche Vermutung bezieht sich nun auf einen stärkeren Grenzwertbegriff. Falls sie wahr ist, so würde der zuletzt bewiesene Satz sofort als Korollar abfallen.

Vermutung 4.3:

Die Abstandsfunktion des Trägers ist sogar der Limes der Distanzfunktion des Kugelmaßes bezüglich der Supremumsnorm. Das heißt, für hinreichend kleine m_0 existiert eine kleinste obere Schranke zu $|d_{\mu_{n;r_0};m_0}(x) - d_{C+B_{r_0}}(x)|$ für alle *Punkte $x \in \mathbb{R}^d$ und diese unterschreitet jede positive reelle Zahl, wenn der Parameter gegen 0 geht.*

■

Diese Vermutung überträgt die analoge Aussage über die Distanzfunktion des empirischen Maßes (vergleiche die Bemerkungen zu Korollar 2.8) auch auf das Kugelmaß mit positivem Bezugsradius. Das passt gut in die Anschauung, dass der Parameter m_0 bestimmt, wie viele der Lernbeispiele bei der Klassifikation eines Punktes berücksichtigt werden. Für $m_0 \to 0$ würden nur die nächstgelegensten Datenvektoren für $m_0 \to 1$ dagegen alle. Außerdem würde sich als Folgerung aus der Vermutung ergeben, dass die mediale Achse des Kugelmaß-Klassifikators für kleine Parameterwerte gegen den Bisektor des Klassifikators $\operatorname{sgn}(d_{C^- + B_{r_0}} - d_{C^+ + B_{r_0}})$ geht. Dabei bezeichnen C^+ und C^- die Mengen der Positiv- beziehungsweise Negativbeispiele. Die im letzteren verwendeten Distanzen lassen sich allerdings ungleich einfacher als das eigentliche Maß berechnen. Sie ergeben sich erneut durch einfache Auswertungen von Nächsten-Nachbar-Anfragen mit einer kleinen Korrektur zur Berücksichtigung der Wirkung von r_0. Dies würde den Aufwand zur Auswertung des neuen Klassifikators gegebenenfalls erheblich reduzieren.

4.2 Zweite Grenzwertvermutung

Bei der zweiten Vermutung handelt es sich um eine Überlegung über das Verhalten der medialen Achse des Klassifikators, wenn die Reichweite eines jeden Datenpunktes gegen Unendlich geht. Dies wurde bei den Definition zunächst explizit vermieden, um nicht Gefahr zu laufen den Bezug zu den Lernbeispielen zu verlieren. Eine Motivation zur Untersuchung für dieses Verfahren war auch der Wunsch nach einer Kapazitätskontrolle des Modells durch Skalierung eines Parameters. Auch diese spiegelt sich in der Vermutung wieder, denn es wird angenommen, dass mit wachsendem r_0 der Bisektor des betreffenden Klassifikators gegen eine Hyperebene geht, im zweidimensionalen Fall also einer Gerade. Genauer bedeutet dies das Folgende:

Vermutung 4.4:

Seien $\bar{x}^+ = \frac{1}{n^+} \sum_{x \in C^+} x$ *und* $\bar{x}^- = \frac{1}{n^-} \sum_{x \in C^-} x$ *die arithmetischen Mittel der Positiv- beziehungswiese Negativbeispiele, diese seien von einander verschieden. Dann konvergieren die Bisektoren der Klassifikatoren* $c_{n;r_0;m_0}$ *für* $r_0 \to \infty$ *in der Hausdorff-Distanz gegen die Hyperebene im Punkt* $\frac{1}{2}(\bar{x}^+ + \bar{x}^-)$ *mit dem Normalenvektor* $\bar{x}^- - \bar{x}^+$.

■

Die beschriebene Hyperebene ist genau die Menge aller Punkte, die zu den beiden Mittelwerten den gleichen Abstand besitzen. Die Vermutung sagt also aus, dass mit steigendem Bezugsradius r_0 der Einfluss eines konkreten Datenpunktes aus einer Klasse zugunsten ihres Zentroides - beziehungsweise seiner besten empirischen Näherung dem Mittelwert der zur Klasse gehörigen Beispiele - zurücktritt. Dies wird durch die Anschauung unterstützt. Die Punktwolken C^+ und C^- haben bloß endlichen Durchmesser, für hinreichend großes r_0 sind also die Systeme $C^+ + B_{r_0}$ respektive $C^- + B_{r_0}$ (im topologischen Sinne) zusammenhängende Kugelhaufen mit weiter ansteigendem Radius wird die Krümmung an ihren Rändern beliebig klein

und auch die Struktur einzelner Kugeln verschwindet. Dann verhält sich eine Menge $C+B_{r_0}$ im Wesentlichen wie eine einzige Kugel um den Mittelwert von C mit Radius r_0. Für einen Punkt $x \in \mathbb{R}^d$ des Raumes ist also $d_{\mu^+_{n;r_0};m_0}(x)$ genau dann kleiner als $d_{\mu^-_{n;r_0};m_0}(x)$, wenn sein Abstand zu $\bar{x}^+$ kleiner ist als zu $\bar{x}^-$, denn dem Fall werden Kugeln um x stets mehr Masse aus $C^+ + B_{r_0}$ aufnehmen als aus $C^- + B_{r_0}$.

Allerdings ist nich zu erwarten, dass die entstehende Hyperebene einen besonders großen Abstand zu den einzelnen Datenpunkten besitzt, wie das beispielsweise bei der Support Vector Machine der Fall ist. Während sich die SVM auf Datenpunkte am Rand einer Klassenstichprobe stützt, hängt der Bisektor hier ausschließlich von den Mittelwerten der Klassen ab. Er ist daher unempfindlicher gegen eventuelle Ausreißer, das darauf beruhende Verfahren klassifiziert dafür aber auch Punkte die näher am Zentroid der anderen Klasse liegen konsequent falsch.

Falls die Vermutung zutrifft, ergäbe sich eine denkbar einfache Berechnung der Klassifikation für hinreichend große Parameterwerte r_0. Es müssten bloß die Daten klassenweise gemittelt, die Gleichung der betreffenenden Hyperebene aufgestellt und bei der Bearbeitung der Anfragen geprüft werden auf welcher Seite sich der zu klassifizierende Punkt befindet.

4.3 Fazit

In dieser Arbeit ist der Weg nachvollzogen worden, aus einer einfachen Idee, nämlich dem Übergang von Einzelpunkten zu einer Vereinigung von Kugeln, einen neuartigen distanzbasierten Klassifikator zu entwickeln. Es konnte eine Berechnungsvorschrift für diesen angegeben, einige seiner Eigenschaften bewiesen und Beziehungen zu anderen Klassifikationsverfahren aufgezeigt werden. Insgesamt ist eine Ahnung gewonnen worden, welche Möglichkeiten des nutzerseitigen Einflusses die neue Technik bietet. So kann zum einen der Anteil der zu berücksichtigenden Daten zum anderen aber auch die Reichweite ihrer Wirkung reguliert werden.

Sehr schnell wurden aber auch sowohl die theoretischen als auch die praktischen Schwierigkeiten deutlich. Viel zu viele Wirkungsweisen der neuen Methode sind noch unbekannt. Die Auswahl geeigneter Parameterwerte erfolgt noch immer anhand von Faustregeln und *best practice* statt auf wissenschaftlichen Erkenntnissen zu beruhen. Auch die massiven Effizienzprobleme einer maschinengestützten Berechnung scheinen zum gegenwärtigen Zeitpunkt noch fast unüberwindlich.

Dies sollte uns aber nicht davon abhalten diesen Klassifikator weiter zu untersuchen und die gewonnenen Erkenntnisse auszubauen. Denn selbst wenn es niemals eine praxisrelevante Implementierung dieses Verfahrens geben sollte, so ist von der weiteren Beschäftigung mit dem Thema dennoch die Vertiefung unseres Verständnisses über die mathematische Theorie distanzbasierter Klassifikation zu erwarten. Nur wenn wir immer neue Ideen wagen, neue Wege gehen und auch bereit sind, das Risiko von Fehlschlägen in Kauf zu nehmen, kann die Informatik ihren Charakter als innovative Zukunftswissenschaft erhalten.

Literaturverzeichnis

[1] Attali, Dominique ; Edelsbrunner, Herbert: Inclusion-Exclusion Formulas from Independent Complexes. In: *Discrete and Computational Geometry* 37 (2007), Januar, Nr. 1, S. 59–77

[2] Beitzel, Steven M.: *On Understanding and Classifiying Web Queries*, Illinois Institute of Technology, Diss., May 2006

[3] Berg, Mark de ; Cheong, Otfried ; Kreveld, Marc van ; Overmas, Mark: *Computational Geometry: Algorithms and Applications*. 3rd Edition. Heidelberg : Springer, 2008

[4] Biberhofer, Mario: *Proof of Concept Utilities for a Theoretical, Measurement-based Classificator.* – Preliminary work. Bachelor thesis, FSU Jena

[5] Bolley, François ; Guillin, Arnaud ; Villani, Cédric: Quantitative concentration inequalities for empirical measures on non-compact spaces. In: *Probability Theory and Related Fields* 137 (2007), Nr. 3, S. 541–593

[6] Chazal, Frédéric ; Cohen-Steiner, David ; Mérigot, Quentin: Geometric Inference for Probability Measures. In: *Foundations of Computational Mathematics* 11 (2011), Nr. 6, S. 733–751

[7] Edelsbrunner, Herbert: The Union of Balls and its Dual Shape. In: *Proceedings of the 9th annual symposium on Computational geometry*. New York, NY : American Mathematical Society, 1993 (SCG '93), 218–231

[8] Edelsbrunner, Herbert ; Mücke, Ernst P.: Simulation of Simplicity: A Technique to Cope with Degenerate Cases in Geometric Algorithms. In: *ACM Transaction on Graphics* 9 (1990), Nr. 1, S. 66–104

[9] Edelsbrunner, Herbert ; Shah, Nimish R.: Incremental topological flipping works for regular triangulations. In: *Proceedings of the eighth annual symposium on Computational geometry*. New York, NY, USA : ACM, 1992 (SCG '92), S. 43–52

[10] Fewell, Matthew P.: *Area of Common Overlap of Three Circles*. DSTO Defence Science and Technology Organisation (Australia), 2006. – Edinburgh, South Australia

[11] Giesen, Joachim ; Kühne, Lars: *Revisiting the Geometry of Large Margin Classification*. 2011. – Preliminary work. Under review by the International Conference on Machine Learning (ICML)

[12] Green, David J.: *Algebraische Topologie, Vorlesung im Sommersemester 2012. Kapitel 01 - Simplex*. Fakultät für Mathematik und Informatik der FSU Jena, 2012

[13] Howard, Ralph: *Alexandrov's theorem on the second derivates of convex functions*. Lecture notes of the functional analysis seminar at the University of South Carolina, 1998

[14] Patzwaldt, Klaus: *@-web Suchmaschinen Blog*. `http://www.at-web.de/blog/20111212/mehr-als-555-millionen-websites-im-internet.htm`. Version: September 2012. – Internetquelle

[15] Pollard, David: *Convergence of Stochastic Processes*. New York : Springer, 1984 (Springer Series in Statistics)

[16] Runkler, Thomas A.: *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. 2. Auflage. Wiesbaden : Vieweg+Teubner, 2012 (Computational Intelligence)

[17] Schukat-Talamazzini, Ernst G.: *Mustererkennung, Vorlesung im Sommersemester 2011.* Fakultät für Mathematik und Informatik der FSU Jena, 2011

[18] Shen, Dou: *Learning-based Web Query Understanding*, Hong Kong University of Science and Technology, Diss., June 2007

[19] Wolfram Math World: *Circle-Circle Intersection*. `http://www.mathworld.wolfram.com/Circle-CircleIntersection.html`. Version: Juli 2012. – Internetquelle

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Seitens des Verfassers bestehen keine Einwände die vorliegende Bachelorarbeit für die öffentliche Benutzung im Universitätsarchiv zur Verfügung zu stellen.

Martin Schirneck, Jena, den 28. September 2012