Ludwig-Maximilians Universität München
Fakultät für Mathematik, Informatik und Statistik

Klassifikation funktionaler Daten

Autor:
Thomas Maierhofer

Betreuerin:
Prof. Dr. Sonja Greven

Masterseminar:
Analyse funktionaler Daten
Wintersemester: 2016/2017

9. Februar 2017

Inhaltsverzeichnis

Zusammenfassung

In dieser Seminararbeit werden zwei Methoden zur Klassifikation funktionaler Daten vorgestellt. Nach eine Einführung der Begriffe Klassifikation und funktionale Daten werden der Nonparametrische Funktionale Kernschätzer von Ferraty and Vieu (2003) und das Nächste Nachbarn Ensemble von Fuchs et al. (2015) motiviert und eingeführt. Beide Methoden beruhen darauf, neue Beobachtungen aufgrund ihrer Distanz zu den Beobachtungen im Trainingsdatensatz zu klassifizieren, wobei als Distanzmaß Semimetriken verwendet werden. Semimetriken sind ein flexibles Werkzeug um maßgeschneiderte Distanzmaße zu definieren. Sie bieten einen sehr weiten Rahmen der für eine optimale Klassifikation des jeweiligen Problems genutzt werden kann. Der Nonparametrische Kernschätzer besteht aus einem Kernschätzer bezüglich einer Semimetrik auf einer funktionalen Kovariable. Das Nächste Nachbarn Ensemble erlaubt es mehrere Nächste Nachbarn Schätzer bezüglich mehrerer Semimetriken auf mehreren Transformationen (multivariater) funktionaler Kovariablen zu kombinieren. Die beiden Methoden werden auf einem Beispieldatensatz angewendet und ihre Prognosegüte verglichen. Häufig ist jedoch die kreuzvalidierte Prädiktionsgüte des Nächste Nachbarn Ensembles geringer als die des prädiktionsstärksten enthaltenen Einzelmodells. Durch die Selektion und Gewichtung der Semimetriken und Kovariablen im Nächsten Nachbarn Ensemble können wichtige Rückschlüsse auf das Klassifizierungsproblem gezogen werden.

1 Einleitung

Funktionale Daten sind ein Gebiet der Statistik das zunehmend an Bedeutung gewinnt und speziell von der Automatisierung der Erhebung von Messungen profitiert. So können Messungen in kurzen Abständen wiederholt erhoben werden, ohne dass menschliche Arbeit notwendig ist. Dies ermöglicht es zunehmend, Messungen als gesamte Funktion (z.B. über die Zeit) anstelle von einzelnen Messwerten zu erheben. Vorstellbar ist hier statt der Messung eines finalen Wertes die Messung des gesamten zeitlichen Verlaufs, zum Beispiel die Konzentration eines Biomarkers über Zeit. Automatisierung kann helfen, dass Prozesse die bisher nur binär aufgezeichnet wurden (Ereignis hat stattgefunden vs. Ereignis hat nicht stattgefunden) jetzt als Funktion aufgenommen werden können. So kann es interessant sein, beim Verbau eines Niets den gesamten Kraftverlauf aufzuzeichnen um sicherzustellen, dass der Niet einwandfrei verbaut wurde. Der gesamte Kraft-Zeit-Verlauf enthält zusätzliche Information über den Verbau des Niets, die nicht in der binären Information, ob die minimal für einen Verbau benötigte Kraft erreicht wurde, enthalten ist. Die Zuordnung eines solchen Kraft-Zeit-Verlaufs ist ein praxisrelevantes Beispiel für die Klassifikation funktionaler Daten. Funktionale Daten sind daher ein stark wachsendes Forschungsgebiet der Statistik mit hoher Anwendungsrelevanz.

Diese Seminararbeit behandelt zwei Ansätze zur Klassifikation funktionaler Daten, den Nonparametrischen Funktionalen Kernschätzer aus Ferraty and Vieu (2003) und das Nächste Nachbarn Ensemble aus Fuchs et al. (2015). Der Aufbau der Arbeit ist wie folgt: In Kapitel 2 werden die zugrundeliegenden Konzepte der funktionalen Daten und der Klassifikation motiviert und eingeführt, sowie eine Einführung in Semimetriken gegeben, die beiden Klassifikationsmethoden als zentraler Baustein zugrunde liegen. In Kapitel 3 werden der Nonparametrische Funktionale Kernschätzer und das Nächste Nachbarn Ensemble vorgestellt, kontrastiert und in bestehende Literatur eingeordnet. Anschließend werden die beiden Methoden in Kapitel 4 auf einem Beispieldatensatz verglichen. Im Anschluss an eine Zusammenfassung und Diskussion der zentralen Ergebnisse in Kapitel 5 endet die Arbeit in einem Ausblick auf mögliche Erweiterungen. Alle Analysen dieser Arbeit wurden mithilfe des Softwarepakets **R** (R Core Team, 2016) durchgeführt. Die Grafiken wurden mithilfe des Pakets **R**-Pakets *ggplot2* (Wickham, 2009) erstellt. Die hier verwendete Notation orientiert sich an Cuevas (2014).

2 Konzepte und Motivation

In diesem Kapitel werden die zentralen Begriffe funktionale Daten (Kapitel 2.1) und Klassifikation (Kapitel 2.2) eingeführt und mit anwendungsorientierten Beispielen motiviert. Außerdem werden Semimetriken definiert und anhand einiger Beispiele erläutert.

2.1 Funktionale Daten

Die Analyse funktionaler Daten (engl. Functional Data Analysis, FDA) ist ein Gebiet der Statistik, welches sich mit Kurven, Oberflächen und sämtlichen Daten befasst, die auf einem kontinuierlichen Träger erhoben werden (Ramsay, 2006). Im Allgemeinen wird bei FDA jede Beobachtung als Funktion dieses Trägers betrachtet, welches oft die Zeit ist, aber auch Raum, Wellenlänge, Winkel oder ein anderer kontinuierlicher Träger sein kann. Die Anforderung die gestellt wird ist, dass die Funktion theoretisch an beliebigen Stellen gemessen werden kann. In einem realen Datensatz wird diese zugrundeliegende Funktion nur an einer diskreten Menge von Punkten ausgewertet.

Als motivierendes Beispiel für Klassifikation funktionaler Daten dient in dieser Arbeit der Datensatz der Berkeley Growth Study (Tuddenham and Snyder, 1954). In dieser Studie wurde die Körpergröße von insgesamt 93 Kindern (39 Mädchen und 54 Jungen) über den Zeitraum von 0 - 18 Jahren erfasst, siehe Abbildung 1. Dieser Datensatz ist im **R**-Paket *fda* (Ramsay et al., 2014) enthalten.

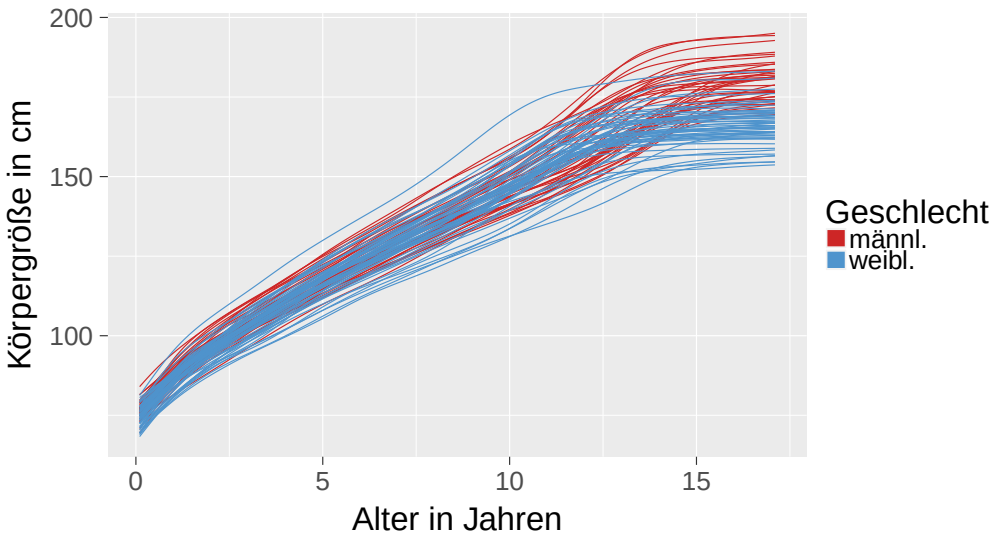

Abbildung 1: Darstellung der Daten der Berkely Growth Study. Abgebildet sind die Wachstumskurven der Jungen (rot) und Mädchen (blau) über den Beobachtungszeitraum von 0 bis 18 Jahren.

Für funktionale Daten sind neue theoretische Konzepte notwendig. Zunächst eine Abgrenzung zu verwandten Teilgebieten der Statistik, der multivariaten Statistik und der Zeitreihenanalyse. Dies geschieht hier nur informell und mithilfe einfacher Beispiele, um eine Idee der wichtigsten Konzepte zu erhalten. Für eine formale Definition siehe Ramsay (2006). Eine Wachstumskurve, die ausschließlich an fünf Punkten von Interesse ist und ausgewertet wird, fällt in den Bereich konventioneller multivariater Statistik, da die Messung als fünf-dimensionale Zufallsvariable aufgefasst werden kann. Wird die Wachstumskurve jedoch als stetige Funktion betrachtet, die an beliebig vielen Punkten ausgewertet werden kann, so handelt es sich um eine ∞-dimensionale Zufallsvariable. Diese theoretische ∞-Dimensionalität

ist die charakteristische Eigenschaft, die FDA zugrunde liegt und sie von multivariater Statistik unterscheidet. Ein anderer Teilbereich der Statistik, der sich mit einem ähnlichen Thema befasst, ist die Zeitreihenanalyse. Oft werden funktionale Daten als (wiederholbare) Realisation einer funktionalen Zufallsvariablen gesehen, im Gegensatz zu Zeitreihen, die als einmaliger nicht-wiederholbarer Verlauf gesehen werden. Eine Beobachtung ist in der FDA eine ganze beobachtete Kurve, in der Zeitreihenanalyse jedoch lediglich ein einzelner Beobachtungs- oder Messzeitpunkt. Oft wird in der Zeitreihenanalysen versucht, aus dem bisher beobachteten Verlauf Prognosen für den zukünftigen Verlauf zu gewinnen, wobei in der FDA generell die gesamte Kurve als bereits beobachtet und bekannt betrachtet wird. Als Beispiel dienen die Wachstumskurve von Kindern: Wenn es darum geht, die Unterschiede in den Wachstumskurven von Jungen und Mädchen zu finden, handelt es sich um ein klassisches FDA Problem. Die Wachstumskurven können an theoretisch beliebig vielen Mädchen und Jungen gemessen werden, die als Realisation der funktionalen Zufallsvariable „Wachstumskurve eines Kindes“ betrachtet werden können. Der Verlauf eines Aktienkurses ist jedoch grundsätzlich anders, insofern als dass der Aktienkurs einer Firma für ein festes Jahr nicht beliebig oft wiederholt werden kann. Der Aktienkurs steigt und fällt im Verlauf des Jahres, kann aber nicht wiederholt werden. Es handelt sich also um ein einmaliges „Experiment“. Auf Basis des Wachstums der vergangenen Jahre, können aber Prognosen für das Wachstum des nächsten Jahres gewonnen werden. Dies ist ein klassisches Problem für die Zeitreihenanalyse.

2.2 Klassifikation

Klassifikationsanalyse befasst sich mit der Zuordnung neuer Beobachtungen mit unbekannter Klasse zu bekannten Klassen. Dafür werden Beobachtungen mit bekannter Klasse verwendet, der sogenannte Trainingsdatensatz, um Regeln aufzustellen nach denen neue Beobachtungen den Klassen zugeordnet werden können. Beispiele sind die Klassifikation verschiedener Spezies von Iris-Blumen aufgrund festgelegter morphologischer Charakteristika wie der Länge und Breite ihrer Kelch- bzw. Kronblätter (Anderson (1935) und Fisher (1936)), siehe Abbildung 2. Ein weiteres Beispiel wäre die Klassifikation von Patienten in „Gesund“ und „Erkrankt“ oder die Art der Krankheit aufgrund ihrer Gesundheitsmerkmale.

Eine Übertragung des Konzepts der Klassifikation auf funktionale Daten ist konzeptionell einfach. Anstelle eines oder mehrerer Merkmale sollen jetzt ganze Funktionen klassifiziert, also einer Klasse zugeordnet werden. Anwendungsbeispiele sind die Klassifikation zellbasierter Sensor-Chips (Fuchs et al., 2015), der (nicht) korrekte Verbau von Schrauben auf Basis ihres Winkel-Drehmoment Verlaufs oder die Zuordnung von Wachstumskurven zum Geschlecht des Kindes.

2.3 Semimetriken

Dieser Abschnitt gibt eine Definition von Semimetriken und nennt einige Beispiele die zur Klassifikation funktionaler Daten verwendet werden können. Im Allgemeinen sind (Semi-) Metriken Distanzmaße, die sich eignen um die Ähnlichkeit oder Nähe

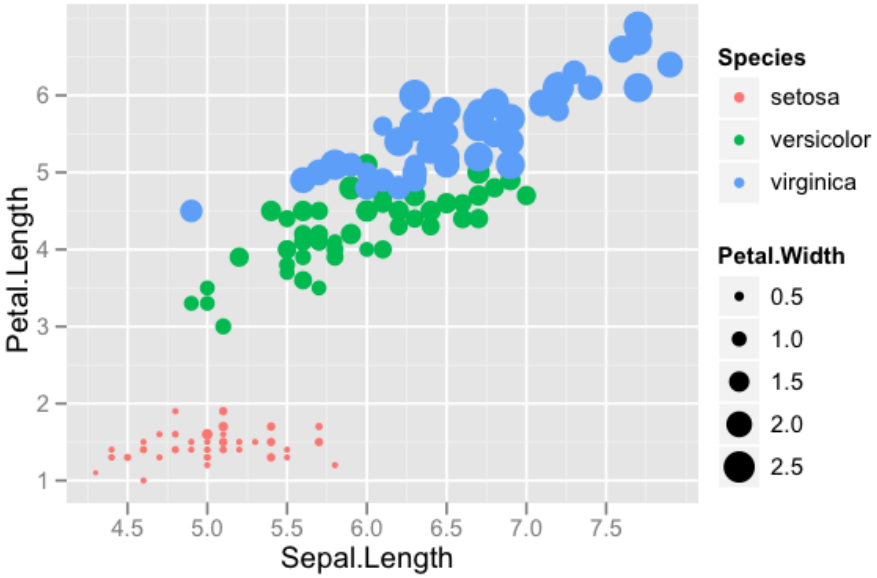

Abbildung 2: Klassifikation der Iris Datensatzes. Abgebildet sind die Länge des Kelchblattes auf der x-Achse und die Länge des Kronblattes auf der y-Achse. Der Durchmesser der Punkte gibt die Breite des Kronblattes. Die Punkte sind nach der Klassenzugehörigkeit (Spezies) der Blüte gefärbt. Die Grafik von https://www.r-bloggers.com/quick-introduction-to-ggplot2/.

zweier Beobachtungen zu definieren. Eine Funktion d ist eine Semimetrik auf einem Raum $\mathcal{X}$ wenn:

- $\forall x \in \mathcal{X} : d(x, x) = 0$,
- $\forall x_1, x_2, x_3 \in \mathcal{X} : d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$

Für eine Metrik wäre die zusätzliche Forderung notwendig, dass $\forall x_1, x_2 \in \mathcal{X} : d(x_1, x_2) = 0 => x_1 = x_2$.

Beispiel 2.1. *Euklidische Distanz der Ableitungen:* Die euklidische Distanz der Ableitung a-ten Grades $d_a^{\text{Eukl.}}(\cdot, \cdot)$ zweier Beobachtungen $x_1(t)$ und $x_2(t) \in \mathcal{X}$ ist definiert als

$$d_a^{\text{Eukl.}}\big(x_1(t), x_2(t)\big) = \sqrt{\int_{\mathcal{T}} \left(x_1^{(a)}(t) - x_2^{(a)}(t)\right)^2 dt},$$

wobei $x^{(a)}$ die a-te Ableitung von x bezeichnet. Für den Spezialfall $a = 0$ handelt es sich hierbei um eine Metrik. Für $a > 1$ handelt es sich jedoch um eine Semimetrik, da die Distanz zweier ungleicher Beobachtungen $x_1(t), x_2(t)$ mit $x_2(t) = x_1(t) + c$ gleich 0 ist. Eine Einschränkung bzw. Gewichtung des Trägers $\mathcal{T}$ ist in diesem Konzept ebenfalls möglich.

Für den Berkely Growth Study Datensatz sind in Abbildung 3 die Mittelwertskurven für die Körpergröße sowie deren erste und zweite Ableitung von Jungen und Mädchen abgebildet. Zunächst wurden die Daten mithilfe des *fda*-Pakets unter Verwendung von P-Spline Basen auf einem regelmäßigen Gitter dargestellt und anschließend die Ableitungen berechnet.

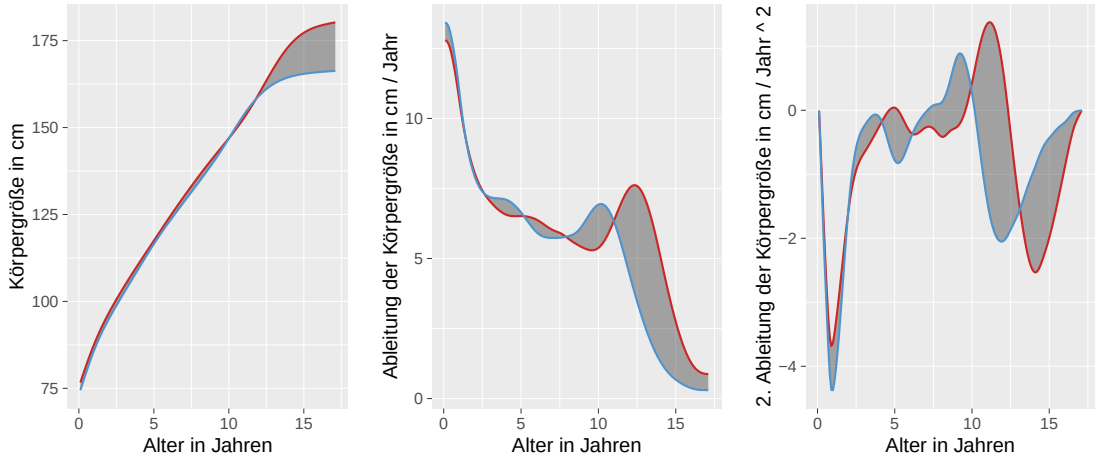

Abbildung 3: Darstellung der Mittelwertskurven der Jungen (rot) und Mädchen (blau) über den Beobachtungszeitraum. Relevant für die euklidische Distanz sind die Abstände in y-Richtung zwischen den Kurven, die grau hinterlegt sind.
Links: Alter in Jahren auf der x-Achse gegen die Körpergröße in cm auf der y-Achse; Mitte: Alter in Jahren auf der x-Achse gegen die erste Ableitung der Körpergröße auf der y-Achse (entspricht der Geschwindigkeit des Wachstums in cm / Jahr); Rechts: Alter in Jahren auf der x-Achse gegen die 2. Ableitung der Körpergröße auf der y-Achse (entspricht der Beschleunigung des Wachstums in cm / Jahr ^{2}).

Beispiel 2.2. *Differenz der globalen Maxima/Minima der Ableitungen:* Die Differenzen der globalen Minima/Maxima der Ableitung vom Grad a bzw. der Originalkurve ($a = 0$) mit

$$\begin{aligned} d_a^{\text{Max.}}\big(x_1(t), x_2(t)\big) &= |\max_t\big(x_1^{(a)}(t)\big) - \max_t\big(x_2^{(a)}(t)\big)| \\ d_a^{\text{Min.}}\big(x_1(t), x_2(t)\big) &= |\min_t\big(x_1^{(a)}(t)\big) - \min_t\big(x_2^{(a)}(t)\big)| \end{aligned}$$

sind Semimetriken. Für den Berkely Growth Study Datensatz zeigt Abbildung 4 die Maxima und Minima der Mittelwertskurven für die Körpergröße und deren Ableitungen von Jungen und Mädchen über den Erhebungszeitraum. Zusätzlich sind die Abstände der Maxima und Minima eingetragen.

3 Klassifikationsmethoden für funktionale Daten

In diesem Kapitel werden zwei Ansätze zur Klassifikation funktionaler Daten beschrieben. Der Nonparametrische Funktionale Kernschätzer von Ferraty and Vieu (2003) wird in Abschnitt 3.1 vorgestellt, das Nächste Nachbarn Ensemble von Fuchs et al. (2015) im Abschnitt 3.2. Beide Vorgehensweisen werden anschließend im Abschnitt 3.3 in bestehende Literatur eingeordnet.

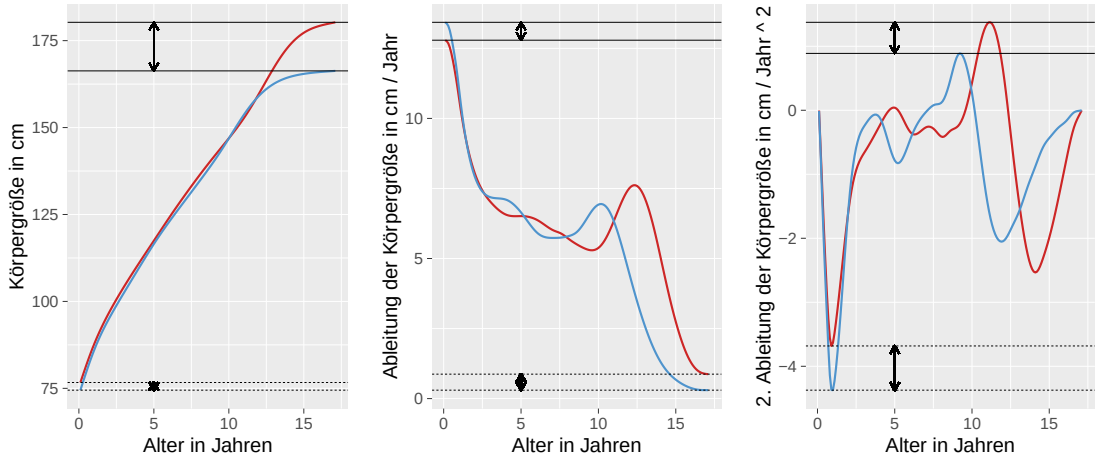

Abbildung 4: Darstellung der Mittelwertskurven der Jungen (rot) und Mädchen (blau) über den Beobachtungszeitraum analog zu Abbildung 3. Relevant für die Minimums- bzw. Maximumsmetrik sind jeweils die Minima (gestrichelte Linie) bzw. Maxima (durchgezogene Linie) der Kurven. Die Minimums- bzw. Maximumsmetrik misst den vertikalen Abstand (also in y-Richtung) der horizontalen Linien, der jeweils mit einem Pfeil eingezeichnet ist.

3.1 Nonparametrischer Funktionaler Kernschätzer

In Ferraty and Vieu (2003) wird ein nonparametrischer Ansatz zur Klassifikation funktionaler Daten vorgeschlagen, der auf einem Kerndichteschätzer basiert. Seien $(y_i, x_i(t)), i = 1, \ldots, N$, die Beobachtungen im Trainingsdatensatz mit bekannter Klassenzugehörigkeit $y_i \in 1, \ldots, G$, und $(y^*, x^*(t))$ eine neue zu klassifizierende Beobachtung. Die Distanzen zwischen allen Kurven werden mit einer im Voraus festgelegten Semimetrik $d(\cdot, \cdot)$ bestimmt. Die geschätzten Wahrscheinlichkeiten $\hat{\pi}_{g,h}\big(x(t)\big)$ der Zugehörigkeit einer Beobachtung $x(t)$ zu einer Klasse $g \in 1, \ldots, G$, gegeben einer Bandweite h und einer festen positiven Kernfunktion $K(\cdot)$ ist

$$\hat{\pi}_{g,h}\big(x(t)\big) = \frac{\sum_{i=1}^{N} I(y_i = g)\ K\Big(d\big(x(t), x_i(t)\big)/h\Big)}{\sum_{i=1}^{N} K\Big(d\big(x(t), x_i(t)\big)/h\Big)}.$$

mit $x_i(t), i = 1, \ldots, N$ den Beobachtungen im Trainingsdatensatz und der Indikatorfunktion $I(g_1 = g_2) = 1 \Leftrightarrow g_1 = g_2$ und 0 sonst. Die Kernfunktion $K(\cdot)$ bestimmt die Form des Kerns, seine Breite wird durch die Bandweite h bestimmt. So ist eine Kernfunktion mit Bandweite $h = 2$ doppelt so breit wie eine Kernfunktion mit Bandweite $h = 1$. Für die Prädiktion der Klassenzugehörigkeit einer Beobachtung wird die Klasse mit der höchsten Wahrscheinlichkeit $\underset{g \in 1,\ldots,G}{\operatorname{argmax}}\ \hat{\pi}_{g,h}(x(t))$ gewählt.

Beispiel 3.1. *Anwendung des Nonparametrischen Funktionalen Kernschätzers:* Für die beispielhafte Anwendung des Nonparametrischen Funktionalen Kernschätzers ziehen wir drei Beobachtungen aus der Berkely Growth Study, von denen eine männlich (x_m), eine weiblich (x_w) und eine unbekannten Geschlechts (x_{un}) ist, siehe Abbildung 5. Zur Klassifikation wird als Kernfunktion der Dreieckskern, $K(u) =$

$(1-|u|)\, I(|u| \leq 1)$, und eine Bandweite von $h = 20$ festgesetzt. Als Distanzmaß wird die Maximumsdistanz $d^{\text{Max.}}\big(x_1^{(0)}(t), x_2^{(0)}(t)\big)$ verwendet, vergleiche Beispiel 2.2. Die Distanz der zu klassifizierenden Beobachtung zu den beiden Beobachtungen im Trainingsdatensatz, $d^{\text{Max.}}\big(x_m(t), x_{un}(t)\big) = |182 - 179| = 3$ und $d^{\text{Max.}}\big(x_w(t), x_{un}(t)\big) = |168 - 179| = 11$, ist in Abbildung 6 abgebildet. Damit ergeben sich die geschätzten Wahrscheinlichkeiten für die Klassenzugehörigkeit der unbekannten Beobachtung zur Klasse Junge als

$$\hat{\pi}_{m,20}\big(x_{un}(t)\big) = \frac{K(3/20)}{K(3/20) + K(11/20)} = 0.65$$

und zur Klasse Mädchen als

$$\hat{\pi}_{w,20}\big(x_{un}(t)\big) = \frac{K(11/20)}{K(3/20) + K(11/20)} = 0.35.$$

Als Klasse mit maximaler Wahrscheinlichkeit wird damit die Klasse Junge vorhergesagt.

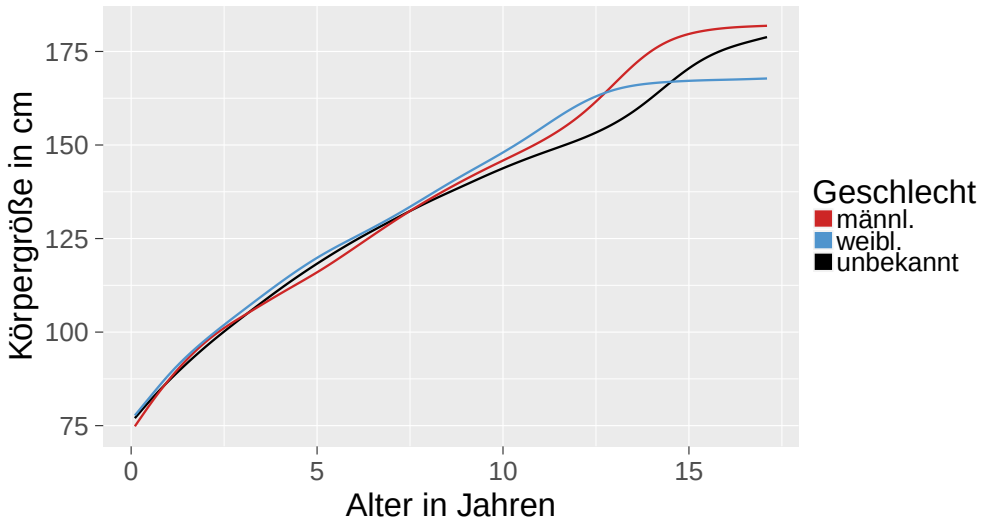

Abbildung 5: Wachstumskurven eines Jungen (rot), eines Mädchens (blau) und eines Kindes unbekannten Geschlechts (schwarz) aus der Berkely Growth Study.

3.2 Nächste Nachbarn Ensemble

In Fuchs et al. (2015) wird ein Ansatz zur Klassifikation funktionaler Daten vorgestellt, der auf einem Ensemble von Wahrscheinlichkeiten basiert, die über Nächste Nachbarn Klassifikationen bestimmt wurden. Die grundsätzliche Idee ist, eine Vielzahl von Nächste Nachbarn Schätzern auf verschiedenen Semimetriken (und falls vorhanden Kovariablen) zu berechnen, und diese anschließend zu einem Ensemble-Schätzer zusammenzufassen. Fuchs et al. (2015) fordern von den Semimetriken

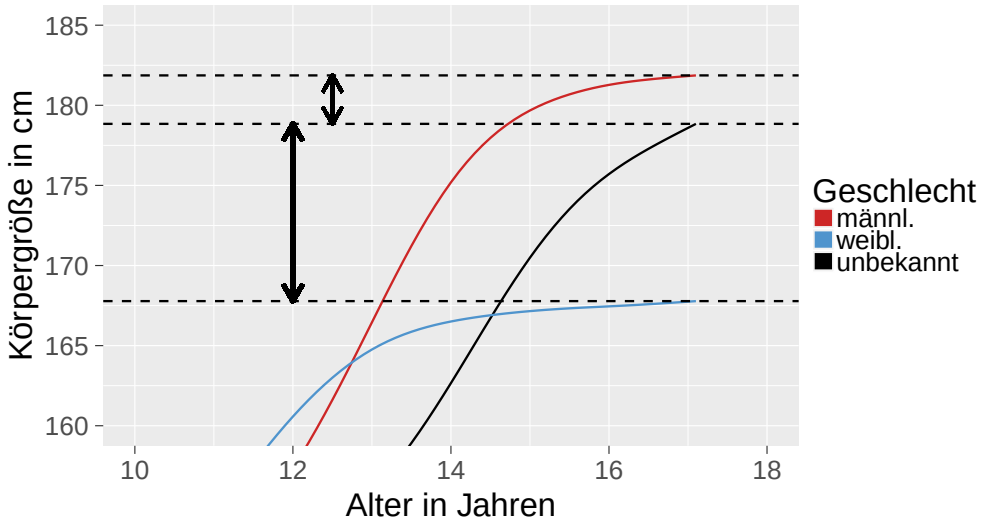

Abbildung 6: Ausschnitt aus Abbildung 5. Zusätzlich sind die Maxima der Wachstumskurven eingetragen und die jeweiligen Differenzen zum Maximum der neuen Beobachtung.

zusätzlich zu den in Abschnitt 2.3 genannten Eigenschaften Symmetrie, also dass $\forall x_1, x_2 \in \mathcal{X} : d(x_1, x_2) = d(x_2, x_1)$.

Betrachtet wird zunächst der Spezialfall, einen Nächste Nachbarn Schätzer für eine Semimetrik $d(\cdot, \cdot)$ und eine Kovariable $x(t)$ zu erstellen. Die Beobachtungen im Trainingsdatensatz $(y_i, x_i(t)), i = 1, \ldots, N$, werden nach ihrem Abstand zur neuen Beobachtung $(y^*, x^*(t))$ bezüglich $d(\cdot, \cdot)$ sortiert, d.h.

$$d(x^*(t), x_{(1)}(t)) \leq \ldots \leq d(x^*(t), x_{(k)}(t)) \leq \ldots \leq d(x^*(t), x_{(N)}(t)).$$

Die Nachbarschaft $\mathcal{N}(x^*(t))$ der k Nächsten Nachbarn von $x^*(t)$ ist definiert als

$$\mathcal{N}(x^*(t)) = \{x_i(t) : d(x^*(t), x_i(t)) \leq d(x^*(t), x_{(k)}(t))\}.$$

Daraus ergibt sich der Nächste Nachbarn Schätzer für die Wahrscheinlichkeit der Zugehörigkeit der Beobachtung $x^*(t)$ zur Klasse $g \in G$ als

$$\hat{\pi}_g = \frac{1}{k} \sum_{x_i(t) \in \mathcal{N}(x^*(t))} I(y_i = g). \tag{1}$$

Das unbekannte y^* wird durch die Klasse mit der höchsten Wahrscheinlichkeit geschätzt, also

$$\hat{y}^* = \underset{g \in G}{\operatorname{argmax}}(\hat{\pi}_g).$$

Für das Vorgehen der Nächsten Nachbarn Klassifikation gibt Abbildung 7 eine Intuition. Diese Abbildung veranschaulicht das Vorgehen bei der Nächsten Nachbarn Klassifikation zweidimensionaler Beobachtungen. Als Distanz wird hier die euklidische Distanz verwendet. Die Vorhersage der Klassenzugehörigkeit einer Beobachtung unbekannter Klasse basiert auf den relativen Häufigkeiten der Klassen der k Nächsten Nachbarn.

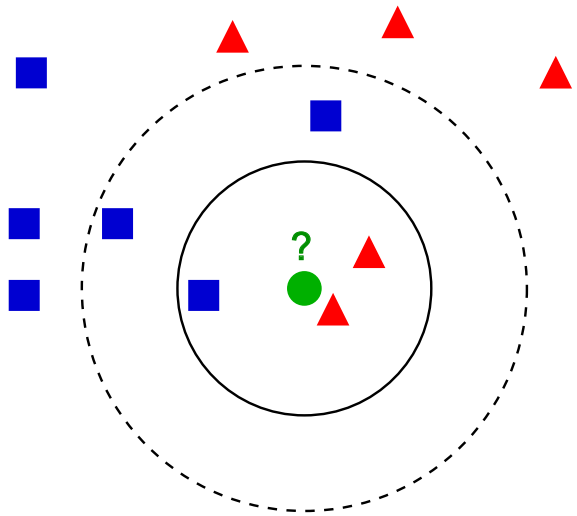

Abbildung 7: Schematische Darstellung der Nächste Nachbarn Klassifikation zweidimensionaler Beobachtungen. Für die Anzahl der verwendeten Nächsten Nachbarn $k = 3$ wird die Klasse „rote Dreiecke“ geschätzt (mit Wahrscheinlichkeit $\frac{2}{3}$), für $k = 5$ die Klasse „blaue Quadrate“ (mit Wahrscheinlichkeit $\frac{3}{5}$).
(Grafik von https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm)

Die Erweiterung zu einem Ensemble von Nächste Nachbarn Schätzern verläuft wie folgt. Es werden anstelle einer einzigen Semimetrik p Semimetriken $d_l, l = 1, \ldots, p$, verwendet. Für jede Semimetrik wird eine eigene Nachbarschaft $\mathcal{N}_l(x^*(t))$ bestimmt. Der Nächste Nachbarn Schätzer für jede Semimetrik ergibt sich analog zu (1) als

$$\hat{\pi}_{gl} = \frac{1}{k} \sum_{x_i(t) \in \mathcal{N}_l\left(x^*(t)\right)} I(y_i = g).$$

Der Ensemble-Schätzer wird als gewichtete Summe der einzelnen Schätzungen definiert, mit

$$\hat{\pi}_g = \sum_{l=1}^{p} c_l \hat{\pi}_{gl}, \tag{2}$$

wobei die Gewichten c_l folgenden Restriktionen unterliegen:

$$c_l \geq 0 \ \forall l \text{ und } \sum_{l=1}^{p} c_l = 1. \tag{3}$$

Die unbekannten Gewichte werden durch die Minimierung des Brier Scores ausgewählt. Der Brier Score ist eine propere Bewertungsregel, deren Score Q gegeben ist durch

$$Q = \sum_{i=1}^{N} \sum_{g=1}^{G} (z_{ig} - \hat{\pi}_{ig})^2 \tag{4}$$

mit $z_{ig} = 1$, falls $y_i = g$, und $z_{ig} = 0$ sonst. Er zeichnet sich dadurch aus, weder zu sensitiv noch zu insensitiv für Änderungen in sehr kleinen oder sehr großen Wahrscheinlichkeiten zu sein (Brier, 1950). Der globale Brier Score des Ensembles (4) kann in Matrixschreibweise als Funktion in Abhängigkeit der Gewichte $\mathbf{c} = (c_1, \ldots, c_p)^T$ geschrieben werden:

$$Q(\mathbf{c}) = \left(\underbrace{\mathbf{z}}_{N \cdot G \times 1} - \underbrace{\mathbf{P}}_{N \cdot G \times p} \underbrace{\mathbf{c}}_{p \times 1} \right)^T \left(\underbrace{\mathbf{z}}_{N \cdot G \times 1} - \underbrace{\mathbf{P}}_{N \cdot G \times p} \underbrace{\mathbf{c}}_{p \times 1} \right), \tag{5}$$

mit $\mathbf{z} = (\mathbf{z_1}| \ldots |\mathbf{z_N})^T$, $\mathbf{z_i} = (z_{i1}, \ldots, z_{iG})^T$, $i = 1, \ldots, N$, $g = 1, \ldots, G$, und $\mathbf{P} = (\mathbf{P_1^T}| \ldots |\mathbf{P_N^T})$, wobei

$$\mathbf{P_i} = \begin{bmatrix} \hat{\pi}_{11i} & \ldots & \hat{\pi}_{1pi} \\ \vdots & \ddots & \vdots \\ \hat{\pi}_{G1i} & \ldots & \hat{\pi}_{Gpi} \end{bmatrix}$$

die Schätzer für Wahrscheinlichkeiten π_{gli} für Person i enthält. Diese Matrizen haben die Klassen $g = 1, \ldots, G$, in den Zeilen und alle verwendeten Kombinationen aus Semimetriken, Ableitungsgraden und Anzahl an Nächsten Nachbarn in den $l = 1, \ldots, p$ Spalten. Diese π_{gli} werden für jede Person durch „leave-one-out“ Kreuzvalidierung bestimmt, da ansonsten jede Beobachtung im Trainingsdatensatz ihr eigener Nächster Nachbar wäre. Durch die Minimierung von Gleichung (5) bezüglich der Ensemblegewichte $\mathbf{c}$ unter den Restriktionen (3) wird eine Lasso-artige Bestrafung verwendet, die manche Gewichte c_l auf 0 setzt und dadurch eine Selektion der verwendeten Kombinationen aus Semimetriken, Ableitungsgraden und Anzahl an Nächsten Nachbarn ermöglicht. Durch eine Implementierung als quadratisches Programm ist die Lösung des Optimierungsproblems mithilfe des **R**-Pakets *limSolve* (Soetaert et al., 2009) sehr effizient implementiert. Die Ensemblegewichte können aufschlussreiche Informationen über das Modell und die Daten liefern. So lassen hohe Gewichte einer Semimetrik auf eine hohe Wichtigkeit des in dieser Semimetrik gemessenen Charakteristikums schließen. Eine Gewichtung von 0 bedeutet, dass diese Semimetrik nicht ausreichend zur Prädiktion der Klassenzugehörigkeit beiträgt und daher nicht in das Ensemble mit aufgenommen wird. Dies ermöglicht eine sparsame Modellierung.

3.3 Kontrastierung und Einordnung der Methoden

In diesem Abschnitt werden die wichtigsten Gemeinsamkeiten und Unterschiede der beiden hier vorgestellten Methoden, dem Nonparametrischen Funktionalen Kernschätzer von Ferraty and Vieu (2003) (Abschnitt 3.1) und dem Nächste Nachbarn Ensembles von Fuchs et al. (2015) (Abschnitt 3.2), herausgearbeitet. Anschließend folgt eine Einordnung der Methoden in etablierte Methoden zur Klassifikation funktionaler Daten.

Die wichtigste Gemeinsamkeit der beiden Ansätze liegt darin, dass sie darauf basieren, die Ähnlichkeit einer neuen Beobachtung zu den Beobachtungen im Trai-

ningsdatensatz mithilfe von Semimetriken bestimmen. Diese Ähnlichkeitsmaße werden genutzt, um eine neue Beobachtung mit unbekannter Klassenzugehörigkeit einer der Klassen zuzuordnen. Der zentrale Unterschied liegt in der Zuordnungsvorschrift einer neuen Beobachtung zu einer Klasse. Im Nonparametrischen Funktionalen Kernschätzer wird ein Kernschätzer zur Schätzung der Klassenzugehörigkeit verwendet, im Nächste Nachbarn Ensemble ein Nächste Nachbarn Schätzer. Die beiden Vorgehensweisen sind sich insofern ähnlich, als dass sie die Klassen der ähnlichsten Beobachtungen (gewichtet) auszählen und darauf basierend die neue Beobachtung zuordnen. Dadurch wird für jede Klasse eine Zugehörigkeitswahrscheinlichkeit bestimmt. Ein weiterer wichtiger Unterschied und großer Vorteil des Nächste Nachbarn Ensembles im Vergleich zum Nonparametrischen Funktionalen Kernschätzer ist, dass die gleichzeitige Verwendung mehrerer Semimetriken und Kovariablen möglich ist. Die Verwendung mehrerer Semimetriken erleichtert nicht nur die Wahl der Semimetriken, im Zweifelsfall werden alle in Betracht kommenden Semimetriken ausgewählt, sondern kann durch ihre gewichtete Kombination zusätzlich die Präzision der Vorhersage verbessern. Außerdem können mehrere funktionale Kovariablen gleichzeitig in das Modell aufgenommen werden und auch Transformationen der Kovariablen als eigene Kovariablen verwendet werden, beispielsweise deren Ableitungen oder gewarpte Funktionen. Eine Erweiterung des Nonparametrischen Funktionalen Kernschätzers auf mehrere Semimetriken oder Kovariablen von den Autoren nicht vorgesehen.

Nun zu einer Einordnung der Vorgehensweisen in etablierte Methoden zur Klassifikation funktionaler Daten. Beide Vorgehensweisen fallen in den Bereich der non- bzw. semi-parametrischen statistischen Modellierung. Dadurch grenzen sie sich vom logistischen funktionalen Regressionsmodell (siehe beispielsweise Gertheiss et al. (2013)) ab. Viele Methoden zur Klassifikation funktionaler Daten basieren auf einer Reduktion der Dimensionalität der Daten durch die Projektion in eine Basisdarstllung, beispielsweise funktionale Hauptkomponentenanalyse (Besse et al., 1997) oder die Verwendung von Splines. Eine andere Möglichkeit ist die Diskretisierung der Daten. Anschließendend können die Daten mit einem beliebigen multivariaten Klassifikationsverfahren ausgewertet werden, beispielsweise logistischen/multinomialen Regressionsmodellen (Fahrmeir et al., 1996) oder Machine Learning Verfahren wie Random Forests (Breiman et al., 1984) und Support Vector Machines (Suykens and Vandewalle, 1999). Bei den Machine Learning Verfahren geht jedoch oft die Interpretierbarkeit des Modells verloren. Es handelt sich um Black-Box Verfahren, bei denen ein Informationsgewinn über die reine Prädiktion hinaus schwer möglich ist. Das Nächste Nachbarn Ensemble bietet die Möglichkeit die Wichtigkeit einiger Charakteristika der Kurven über die Gewichtung im Ensemble zu interpretieren.

4 Empirischer Vergleich der Methoden

In diesem Kapitel werden die beiden Methoden zur Klassifikation funktionaler Daten die in Kapitel 3 vorgestellt wurden auf einem Beispieldatensatz verglichen. Hierzu wird der Datensatz Berkely Growth Study aus Tuddenham and Snyder (1954) verwendet, der in Ramsay et al. (2014) enthalten ist. Ein besonderes Augenmerk wird auf den Vergleich der Stärken und Schwächen der Methoden gelegt.

Die Methoden werden für den Datensatzes systematisch verglichen. Für beide Methoden werden sowohl auf den Originaldaten als auch auf den ersten beiden Ableitungen jeweils drei verschiedenen Semimetriken berechnet, die Euklidische Distanz (siehe Beispiel 2.1), die Maximums- und die Minimumsdistanz (siehe Beispiel 2.2). Als Parameter k des Nächste Nachbarn Ensembles wird $k = \{1, 5, 11\}$ in einem Ensemble verwendet. Als Kernfunktion des Nonparametrischen Funktionalen Kernschätzers wird der Dreieckskern verwendet mit automatisiert über Kreuzvalidierung geschätzter Bandweite h. Für jede Parameterkombination wird die mittlere Missklassifikationsrate (MMCR), definiert als $\frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i)$, mithilfe einer zehn mal wiederholten zehnfachen Kreuzvalidierung geschätzt. Dies ist analog zum Vorgehen in Fuchs et al. (2015). Die Ergebnisse für jede Parameterkombination sind in Tabelle 1 zusammengefasst. Die geringsten MMCR sind fett markiert und wurden sowohl für den Kernschätzer als auch für das Nächste Nachbarn Ensemble für die Semimetrik der Euklidische Distanz auf den Originaldaten erreicht. Eine gleich gute Prädiktion wird vom Nächsten Nachbarn Ensemble für das Ensemble aus allen Semimetriken auf den Originaldaten erreicht. Dies ist damit zu begründen, dass die Schätzer die auf Maximums- und Minimumsdistanz basieren das Gewicht 0 im Ensemble erhalten haben.

Da das Nächste Nachbarn Ensemble im Gegensatz zum Nonparametrischen Funktionalen Kernschätzer mit mehreren Semimetriken und Ableitungen zugleich umgehen kann, sind in den Randfeldern nur für dieses Vorgehen die Missklassifikationsrate eingetragen. Erkennbar ist, dass das Nächste Nachbarn Ensemble für die einzelnen Parameterkombinationen besser in der Prognose ist, was möglicherweise durch eine ungünstig gewählte Kernfunktion zu erklären ist. Das Nächste Nachbarn Ensemble bietet den Vorteil, dass keine weiteren Hyperparameter übergeben werden müssen. Ein weiterer Vorteil ist, dass mehrere Semimetriken und Ableitungen gleichzeitig ins Modell aufgenommen werden können. Die Selektion und Gewichtung der wichtigen Semimetriken und Ableitungen funktioniert in diesem Fall gut, vergleiche dazu die Randfelder in Tabelle 1. Diese Ensembleschätzer erzielen generell nicht die optimale Prädiktionsgüte, sind jedoch nicht weit davon entfernt.

Ein weiterer Vorteil des Nächsten Nachbarn Ensembles ist die Möglichkeit, die Ensemblegewichte zu interpretieren. Für das Modell, welches alle verwendeten Kombinationen aus Semimetriken und Ableitungsgraden enthält (in Tabelle 1 unten rechts), sind die Ensemblegewichte in Abbildung 8 dargestellt. Die Hintergrundfarbe kodiert die Missklassifikationsrate (MMCR) in % der jeweiligen einzelnen Ableitungs/Semimetrik-Kombinationen (also dem jeweiligen inneren Feld von Tabelle 1), die Zahl entspricht der Gewichtung im Ensembleschätzer aus allen Kombinationen. Die in Abbildung 8 berichteten Gewichte sind die Summe über die Gewichte der Nächsten Nachbarn Schätzer mit dieser Ableitungs/Semimetrik-Kombination und verschiedenem k. Erkennbar ist, dass Ableitungs/Semimetrik-Kombinationen mit hoher Missklassifikationsrate tendenziell geringe Gewichte im Ensemble bekommen und Ableitungs/Semimetrik-Kombinationen mit geringer Missklassifikationsrate hohe Gewichte. Schlecht klassifizierende Ableitungs/Semimetrik-Kombinationen wurden zum Teil auch mit 0 gewichtet und damit ausselektiert. Diese Selektion ist im Allgemeinen wünschenswert, da sie ein sparsameres Modell erzeugt.

Tabelle 1: Mittlere Missklassifikationsrate (MMCR) in % für das Nächste Nachbarn Ensemble (oben, in schwarz) und den Nonparametrischen Funktionalen Kernschätzer (unten, in blau) auf den Berkeley Growth Study Daten. Als Anzahl Nächster Nachbarn wurde $k = 1, 5, 11$ verwendet. Es handelt sich damit für jede Ableitungs/Semimetrik-Kombination um Ensembles aus 3 Nächsten Nachbarn Schätzern. Die Bandbreite h des Nonparametrischen Funktionalen Kernschätzers wurde über Kreuzvalidierung optimal gewählt.

	Euklid.	**Maximum**	**Minimum**	**Alle Semimtr.**
Original	**3.2**	12.3	43.9	**3.2**
	8.7	19.9	46.6	-
1. Ableitung	7.5	55.8	28.9	7.5
	11.9	49.8	33.4	-
2. Ableitung	5.8	32.8	52.6	5.7
	25.1	34.7	48.6	-
Alle Ablt.	6.0	14.1	30.3	5.7
	-	-	-	-

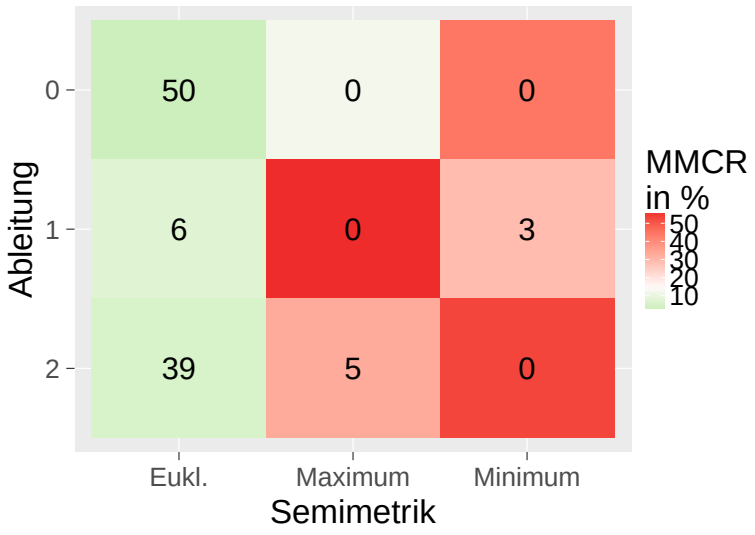

Abbildung 8: Ensemblegewichte in % des Nächste Nachbarn Ensembles für alle 3 verwendeten Ableitungsgrade und 3 Semimetriken. Die Hintergrundfarbe kodiert die Missklassifikationsrate (MMCR) in % der Ableitungs/Semimetrik-Kombination (aus Tabelle 1). Die Zahl entspricht der Gewichtung im Ensembleschätzer. Zu beachten ist, dass die Gewichte für jede Ableitungs/Semimetrik-Kombination über die verschiedenen $k = \{1, 5, 11\}$ summiert wurden.

5 Zusammenfassung und Diskussion

In dieser Seminararbeit wurden zwei Methoden zur Klassifikation funktionaler Daten vorgestellt, der Nonparametrische Funktionale Kernschätzer von Ferraty and Vieu (2003) und das Nächste Nachbarn Ensemble von Fuchs et al. (2015). In beiden Methoden basiert die Klassifikation neuer Beobachtungen auf einem (gewichteten) Auszählen der Klassen der ähnlichsten Beobachtungen im Datensatz. Dabei ist Ähnlichkeit über die Distanz zwischen den jeweiligen Beobachtungen definiert. Als Distanzmaß werden Semimetriken verwendet, die einen sehr offenen Rahmen für eine optimale Klassifikation des jeweiligen Problems bieten. Der zentrale Vorteil des Nächsten Nachbarn Ensembles im Gegensatz zum Nonparametrischen Funktionalen Kernschätzer ist, dass es mit multivariaten funktionalen Daten, mehreren Transformationen der Daten und mehreren Distanzmaßen für die Beobachtungen im Datensatz zugleich umgehen kann. Diese höhere Flexibilität zeigt sich als vorteilhaft in der beispielhaften Analyse des Berkeley Growth Study Datensatzes. Hier zeigt sich die höhere Flexibilität und Anwendbarkeit des Nächsten Nachbarn Ensemble im Vergleich zum historisch älteren Nonparametrischen Funktionalen Kernschätzer. Ein weiterer Vorteil ist die Interpretierbarkeit des Nächsten Nachbarn Ensembles durch die Selektion und Gewichtung der Semimetriken und Kovariablen. Die Ensemblegewichte liefern ein intuitives Maß für die Wichtigkeit der einzelnen Semimetriken in Kombination mit dem Ableitungsgrad der Daten. Automatische Selektion ermöglicht die Definition eines sparsameren Modells.

6 Ausblick

Dieses Kapitel gibt einen Ausblick auf ausgewählte weitere Semimetriken und eine mögliche Erweiterung der Schätzung der Ensemblegewichtung. Beide vorgestellten Methoden können durch die Wahl der Semimetriken prinzipiell beliebig angepasst werden. Denkbar wären Semimetriken die auf (shape) Dynamic Time Warping (Zhao and Itti, 2016) basieren, wie die Distanz der gewarpten Funktionen und die Distanz der Warping Funktionen. Für die Daten der Berkeley Growth Study konnte der Nonparametrische Funktionale Kernschätzer jedoch für diese Distanzmaße keine zufriedenstellende Klassifikation erreichen, siehe Tabelle 2. Semimetriken die auf Hauptkomponentenscores basieren sind ein weiterer vielversprechender Ansatz.

Tabelle 2: Mittlere Missklassifikationsrate (MMCR) in % des Nonparametrischen Funktionalen Kernschätzers unter Verwendung der Distanz der gewarpten Funktionen und der Warping-Funktionen als Semimetrik.

	Gewarpte Funktion	Warping Funktion
Original	**15.9**	16.6
1. Ableitung	27.6	19.7
2. Ableitung	29.1	20.3

Ein möglicher Ansatzpunkt zur Verbesserung des Nächste Nachbarn Ensembles ist, anstelle einer gewichteten Summe der einzelnen Nächsten Nachbarn Schätzungen, die einzelnen Nächsten Nachbarn Schätzungen nichtlinear zu kombinieren. Dies wäre beispielsweise mithilfe eines Random Forests (Breiman et al., 1984) umsetzbar, der die $\hat{\pi}_{gl}$ als Einflussgrößen und die wahren Klassenzugehörigkeiten im Trainingsdatensatz als Zielgröße hat. Dies würde komplexere Interaktionen der Nächste Nachbarn Schätzungen und dadurch den Distanzen der Semimetriken erlauben. Anstelle der c_l könnte die Variablenwichtigkeit im Random Forest betrachtet werden. Als Anschauung dient Abbildung 9, welche die eingeschränkte Flexibilität der gewichteten Summe für zwei Nächste Nachbarn Schätzer zeigt. Durch die Verwendung eines Random Forests sind auch nichtlineare Trennungen möglich.

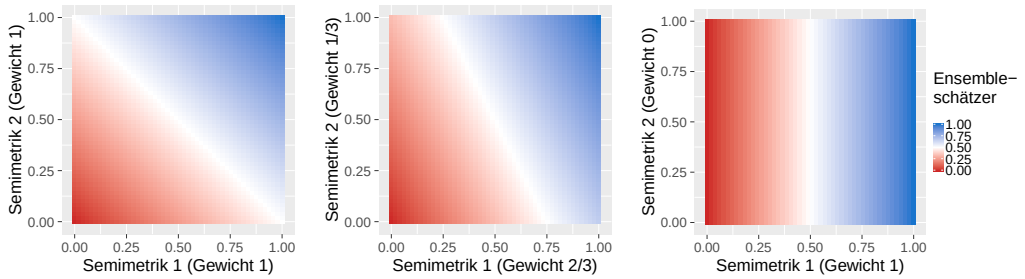

Abbildung 9: Auswirkung verschiedener Gewichte zweier Semimetriken auf den Ensemble Schätzer. Abgebildet ist die prognostizierte Wahrscheinlichkeit des Ensembles $\hat{\pi}_g$ in Abhängigkeit der prognostizierten Wahrscheinlichkeit der beiden im Ensemble enthaltenen Schätzer $\hat{\pi}_{g1}$ und $\hat{\pi}_{g2}$.
Links: Gleiche Gewichtung der Schätzer im Ensemble; Mitte: Stärkere Gewichtung des ersten Schätzers; Rechts: Volle Gewichtung des ersten Schätzers.

Literatur

Anderson, E. (1935). The irises of the Gaspe peninsula. *Bulletin of the American Iris Society 59*, 2–5.

Besse, P. C., H. Cardot, and F. Ferraty (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis 24*(3), 255–270.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. CRC press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference 147*, 1–23.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multivariate statistische Verfahren*. Walter de Gruyter GmbH & Co KG.

Ferraty, F. and P. Vieu (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis 44*(1), 161–173.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Fuchs, K., J. Gertheiss, and G. Tutz (2015). Nearest neighbor ensembles for functional data with interpretable feature selection. *Chemometrics and Intelligent Laboratory Systems 146*, 186–197.

Gertheiss, J., A. Maity, and A.-M. Staicu (2013). Variable selection in generalized functional linear models. *Stat 2*(1), 86–101.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. (2006). *Functional Data Analysis*. Wiley Online Library.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2014). *fda: Functional Data Analysis*. R package version 2.4.4.

Soetaert, K., K. Van den Meersche, and D. van Oevelen (2009). *limSolve: Solving Linear Inverse Models*. R package 1.5.1.

Suykens, J. A. and J. Vandewalle (1999). Least squares support vector machine classifiers. *Neural processing letters 9*(3), 293–300.

Tuddenham, R. D. and M. M. Snyder (1954). Physical growth of California boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley 1*(2), 183.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zhao, J. and L. Itti (2016). shapeDTW: shape Dynamic Time Warping. *arXiv preprint arXiv:1606.01601*.

Elektronischer Anhang

Teil dieser Seminararbeit ist ebenfalls ein elektronischer Anhang. Dieser besteht aus dem Ordner `Maierhofer FDA Seminar 2017 Klassifikation`, der Code enthält um alle Analysen und Grafiken in dieser Arbeit zu reproduzieren. Weitere Details sind der im Ordner enthaltenen Datei `README.txt` zu entnehmen.

Eidesstattliche Erklärung

Hiermit versichere ich, Thomas Maierhofer, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

(Ort, Datum)	(Unterschrift)