Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Prof. Dr. Claus Weihs
Dr. Thomas Liebig
Sebastian Buschjäger
Malte Jastrow
Jennifer Neuhaus-Stern

Übung zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2018

Übungsblatt Nr. 10

Der Abgabetermin ist Dienstag der 10.07.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Neuronale Netze) (6 Punkte)

Sie haben bereits Neuronale Netze als Klassifikationsmethode kennen gelernt. In dieser Aufgabe vertiefen wir das Verständnis dieser Methodik. Da Neuronale Netze sehr rechenaufwendig sind, verwenden wir in dieser Übung den `iris` Datensatz.

a) (3 Punkt) Implementieren Sie ein Perceptron (Folie 10 `WiD 180605 Künstliche Neuronale Netzwerke`) mit Sigmoid-Aktivierung und benutzen Sie dieses um die Klasse `virginica` von den übrigen beiden Klassen `setosa` und `versicolor` zu trennen (binäres Klassifikationsproblem). Berechnen Sie den Trainingsfehler und Testfehler ihres Modells. In der Beispieldatei finden Sie bereits eine geeignte Vorverarbeitung der Daten, sowie einen passenden Train/Test Split.
Hinweis: Sie dürfen als Ausgangspunkt ihre Lösung zu Blatt 5, Aufgabe 1 oder die entsprechende Musterlösung nutzen. Beachten Sie jedoch, dass ein Perceptron mittels *Stochastic* Gradient-Descent und einer anderen Modellfunktion trainiert wird.

b) (1 Punkt) Diskutieren Sie die Unterschiede und die Gemeinsamkeiten zwischen dem Perceptron, der SVM und der logistischen Regression!

c) (2 Punkt) Trainieren Sie ein neuronales Netz mit mindestens drei Layern auf dem `iris` Datensatz für das ursprüngliche Drei-Klassenproblem unter Verwendung des `neuralnet` Packages. Variieren Sie die Anzahl der Layer, sowie die Anzahl der Neuronen und testen Sie ihr Modell auf den Testdaten. Was fällt Ihnen auf? Welche Accuracy erreichen Sie?
Hinweis: Im Kontext von Neuronalen Netzen bietet es sich an, Mehrklassenprobleme in ein 1-Hot-Encoding zu überführen. In dieser Darstellungsform wird jedes Label in einen eindeutigen Vektor überführt, welcher einen 1-Eintrag für jede Klasse enthält. Zum Beispiel wird aus der Klasse `setosa` der Vektor $(1, 0, 0)^T$, für die Klasse `versicolor` ergibt sich der Vektor $(0, 1, 0)^T$ und für die Klasse `virginica` der Vektor $(0, 0, 1)^T$. Damit ist die Ausgabe y des Neuronalen Netzes ein drei-dimensionaler Vektor. Um diesen Vektor anschließend in eine Vorhersage zu überführen, können Sie das Maximum $\widehat{y} = \max\{y_1, y_2, y_3\}$aller Einträge benutzen.

Aufgabe 2 (Clustering angewendet) (5 Punkte)

In der Vorlesung haben Sie zwei Verfahren zum Clustern von Daten kennen gelernt, nämlich `KMEANS` und `DBScan`. In dieser Aufgabe sollen Sie diese Verfahren auf den `circles.csv`, `blobs.csv`, `moons.csv` und `random.csv` Daten praktisch anwenden und vergleichen.

a) (1 Punkt) Laden Sie die vier genannten Datensätze und plotten Sie diese jeweils in einem 2D Plot.

b) (2 Punkt) Wenden Sie `KMEANS` und `DBScan` an. `KMEANS` ist bereits in `R` vorhanden. Für `DBScan` verwenden Sie bitte das `dbscan` Paket. Variieren Sie drei verschiedene K bzw. ε Parameter. Benutzen Sie die Plots aus a) als Hilfestellung!

c) (2 Punkt) Wie bewerten Sie die zwei Verfahren? Welches Verfahren eignet sich für welchen Datensatz und wieso? Begründen Sie ihre Aussage.

Aufgabe 3 (Hierarchisches Clustern) **(6 Punkte)**

In Tabelle 1 sind als Wohlstandsindikatoren die durchschnittliche Lebenserwartung für Männer und Frauen (bei Geburt) sowie die Kleinkindersterblichkeitsrate aus dem Jahre 2016 für fünf verschiedene europäische Länder enthalten. Die Quelle dieser Daten ist die *IDB* des *U.S. Census Bureau*.

Table 1: Lebenserwartung und Kleinkindersterblichkeitsrate aus dem Jahre 2016 für fünf europäische Länder.

		Lebens-erwartung	Kleinkinder-sterblichkeitsrate
1	Deutschland	80.7	3.4
2	Frankreich	81.8	3.3
3	Irland	80.8	3.7
4	Island	83.0	2.1
5	Schweden	82.1	2.6

Bearbeiten Sie die folgenden Teilaufgaben a)–c) mit Stift und Papier oder schreiben Sie eigene `R`-Funktionen zur Berechnung. Sie sollen hier jedoch keine existierenden `R`-Funktionen verwenden, um die Distanzen zu berechnen und Single-Linkage und Complete-Linkage durchzuführen. Erst in Teilaufgabe d) sollen Sie diese Funktionen hinzuziehen.

a) (1 Punkt) Visualisieren Sie die Daten geeignet. Bestimmen Sie die Distanzmatrix bzgl. der Merkmalsträger auf Basis der euklidischen Abstände.

b) (2.5 Punkte) Führen Sie nun das Single-Linkage und das Complete-Linkage Verfahren durch. Bestimmen Sie jeweils die Vereinigung in jeder Stufe, berechnen Sie die nötigen Heterogenitätsmaße sowie die inverse mittlere Klassenheterogenität für jede Stufe als Gütemaß der Partition. Wie würden Sie die Länder nun gruppieren?

c) (0.5 Punkte) Zeichnen Sie das Dendrogramm für das Single-Linkage und das Complete-Linkage Verfahren. Wie würden Sie die Länder gruppieren?

d) (2 Punkte) Standardisieren Sie nun die Daten auf Mittelwert 0 und Varianz 1. Bestimmen Sie erneut die Distanzmatrix, führen Sie das Single-Linkage und das Compelete-Linkage Verfahren durch und zeichnen Sie erneut die jeweiligen Dendrogramme. In dieser Teilaufgabe dürfen Sie `R`-Funktionen für die Berechnungen verwenden. Wie viele Cluster würden Sie nun wählen? Vergleichen Sie mit Teilaufgabe a)–c) und erklären Sie eventuelle Unterschiede.

Aufgabe 4 (Fluch der hohen Dimensionen) **(3 Punkte)**

In der Vorlesung wurde bereits kurz der "Fluch der hohen Dimensionen" erwähnt. Dieser "Fluch" beschreibt verschiedene Phänomene die bei hochdimensionalen Daten auftreten und häufig die Intuition stören, welche bei niedrigdimensionalen Daten noch gut funktioniert hat. In dieser Aufgabe sollen verschiedene Manifestationen dieses "Fluches" untersuchen.

a) (0.5 Punkt) (Kombinatorik) Gegeben sei eine drei-dimensionale Beobachtung $\vec{x} \in \{0, 1, 2, 4\}^3$, wobei jede Dimension einen von vier möglichen Werten annehmen kann. Wie viele verschiedene Beobachtungen gibt es in diesem Fall?
Wie verändert sich die Anzahl der verschiedenen Beobachtungen, wenn man eine weitere Dimension mit gleichem Beobachtungsraum, d.h. $\vec{x} \in \{0, 1, 2, 4\}^4$ hinzufügt?

b) (1 Punkt) (Sampling) Als Sampling-Dichte bezeichnet man die Anzahl der gleichmäßig verteilten Beobachtungen pro Standardintervall (= Intervall der Länge 1). Werden z.B. 10 Punkte gleichmäßig aus dem Intervall $[0, 1]$ gesampled, so ist die Sampling-Dichte $\frac{1}{10}$.

Begründen Sie, dass die Sampling-Dichte proportional zu $\left(\frac{1}{N}\right)^{\frac{1}{d}}$ ist, wobei d die Anzahl der Dimensionen und N die Anzahl der Beobachtungen ist. Diskutieren Sie insbesondere, wieso die Sampling-Dichte nicht exakt $\left(\frac{1}{N}\right)^{\frac{1}{d}}$ für $d \geq 2$ entspricht.

c) (1.5 Punkt) (Distanzmaße) Angenommen Sie konstruieren einen Ball um einen d-dimensionalen Punkt $\vec{x} = (x_1, x_2, \ldots, x_d)$ mit Radius r. Dann hat der kleinste Würfel, welcher den Ball einschließt eine Kantenlänge von $2r$. Das Volumen dieses Würfels ist dann durch $(2r)^d$ gegeben, wohingegen das Volumen des Balls mit $\frac{2r^d\pi^{d/2}}{d\Gamma(d/2)}$ gegeben ist. Hier bezeichnet $\Gamma(\cdot)$ die Gamma-Funktion.

- Wie verhalten sich diese beiden Volumina zueinander für $d \to \infty$?
- Angenommen Sie verwenden ein distanzbasiertes Verfahren, wie z.B: KMEANS und legen ein euklidisches Distanzmaß $d(\vec{x}, \vec{x}') = \sqrt{\sum_{i=1}^{d}(\vec{x}_i - \vec{x}'_i)^2}$ zugrunde. Dann induziert die Distanz $d(\vec{x}, \vec{x}')$ einen Ball um $\vec{x}$ (bzw. $\vec{x}'$) mit Radius $r = d(\vec{x}, \vec{x}')$. Was sagt ihnen die vorherige Betrachtung der Volumina über die Qualität dieses Abstandsmaßes?

Hinweis: Sie dürfen annehmen, dass $\Gamma(x)$ schneller als π^x wächst für $x \to \infty$.