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. 3

Der Abgabetermin ist Dienstag der 15.05.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Bayes-Regeln) (5 Punkte)

In der Sigmatalsperre leben 2 Arten von Fischen, von denen eine äußerst schmackhaft (A), die andere jedoch ungenießbar (B) ist. Als begeisterter Angler haben Sie von diesen Fischen gehört und wollen für Ihren Freundeskreis ein Fisch-BBQ organsieren. Zunächst müssen Sie dazu eine ausreichende Anzahl der schmackhaften Fische fangen. Die Schwierigkeit liegt darin, dass Fische beider Arten nahezu identisch aussehen: gelbe Streifen auf schwarzem Schuppenkleid. Die Anzahl gelber Streifen variiert dabei von Fisch zu Fisch. Weiterhin sind aus der Literatur folgende Eigenschaften über die Fischpopulation bekannt:

(i) Die Anzahl der Streifen k der schmackhaften Fische ist binomial verteilt mit den Parametern $n = 4$ und $p = 0.5$.

(ii) Die Anzahl der Streifen k der ungenießbaren Fische ist geometrisch verteilt mit Parameter $p = 0.3$.

(iii) In der Sigmatalsperre leben 3 mal mehr ungenießbare als schmackhafte Fische.

Nachdem Sie einen Fisch gefangen haben, müssen Sie anhand der Anzahl seiner Streifen entscheiden, ob Sie den Fisch a) behalten und für Ihre Freunde zubereiten oder b) wieder freilassen. Ihre Freunde haben sich dazu entschieden, Sie für diesen Abend finanziell zu entlohnen: Für jeden zubereiteten schmackhaften Fisch erhalten Sie 40 Euro. Da Ihre Freunde jedoch Feinschmecker sind, müssen Sie Ihnen für jeden zubereiteten ungenießbaren Fisch 20 Euro zahlen. Weiterhin kostet Sie das Angeln eines jeden Fisches 5 Euro, das Zubereiten weitere 5 Euro.

a) (1 Punkt) Geben sie die a-priori-Wahrscheinlichkeiten π_i sowie die Kostenfunktion $c(i, j)$ für jeweils beide Fischarten, d.h. für $(i, j) \in \{A, B\}^2$ an.

b) (1 Punkt) Geben Sie die kostenminimale datenunabhängige Bayes-Regel an.

c) (1 Punkt) Wie lautet die kostenminimale datenabhängige Bayes-Regel?

d) (1 Punkt) Wie groß sind die Fehlklassifikationswahrscheinlichkeiten der beiden Regeln? D.h. bestimmen Sie $P(WahreKlasse \neq Entscheidung)$.

e) (1 Punkt) Wie hoch ist der erwartete Verlust der beiden Regeln? Welche Regel ist besser? Überrascht Sie das Ergebnis?

Aufgabe 2 (Empirische Bayes-Methode) (5 Punkte)

Sie erfahren nun, dass die beiden Fischarten auch in der benachbarten Omegatalsperre anzutreffen sind. Da diese für Sie besser zu erreichen ist, entscheiden Sie lieber hier zu angeln. Leider finden sich in der Literatur keine Angaben über die Fischpopulation in dieser Talsperre. In einem Feldversuch haben Sie (aufopferungsvoll, wie Sie sind) 1000 Fische gefangen, probiert und die Anzahl Ihrer Streifen notiert, die Ergebnisse haben Sie in dem Datensatz fish.RData archiviert. Schätzen Sie die

zugehörigen Bayes-Regeln auf folgende 3 Arten (je 0.5 Punkte). Welche Regel würden Sie verwenden? Warum?

1. Nehmen Sie eine Normalverteilung für die Streifenzahlen der beiden Fischarten ab. Schätzen Sie die Parameter der Normalverteilung durch das arithmetische Mittel $\bar{x}$ und die Standardabweichung $s(x)$.
2. Gehen Sie davon aus, dass die Streifenanzahlen weiterhin binomial bzw. geometrisch sind. Verwenden Sie bei der binomial-Verteilung das Maximum sowie $\frac{1}{\bar{x}}$, bei der geometrischen $\frac{1}{1+\bar{x}}$ als Schätzer für die Parameter.
3. Verwenden Sie die empirische Bayes-Methode.

Aufgabe 3 (Nächste Nachbarn) **(5 Punkte)**

Schreiben Sie in R eine Funktion mknn mit den Eingabeparameter:

- features: Ein data.frame mit n Zeilen und p Spalten, der die Trainingsdaten enthält.
- y: Ein Vektor der Länge n, der die zugehörigenn Werte der Zielvariablen enthält.
- k: Ein Integer, gibt die Anzahl der zu verwendenden Nachbarn an
- new.data: Ein data.frame mit p Spalten. Enthält die neuen Beobachtungen.

Diese Funktion soll die in der Vorlesung angedeutete Methode der nächsten Nachbarn realisieren.

a) (3 Punkte) Betrachten Sie zunächst den Fall der Klassifikation (d.h. eine diskrete Zielvariable). Hier ist der Vorhersagewert für eine neue Beobachtung diejenige Klasse, die unter den k nächsten Nachbarn (nach euklidischer Distanz) des Trainingsdatensatzes am häufigsten vorkommt. Implementieren Sie diesen Fall. Zum Testen Ihrer Funktion können Sie die iris-Daten verwenden. Ziehen Sie dazu zufällig 120 Beobachtungen als Trainingsdatensatz und sagen Sie die Klassen der übrigen 30 Beobachtungen vorher.

b) (2 Punkte) Überlegen Sie sich, wie die Vorhersage von knn aussehen könnte, falls ein Regressionsproblem (d.h. eine numerische Zielvariable) vorliegt. Implementieren Sie diese! Ihre Funktion sollte dabei automatisch entscheiden können, ob es sich bei den eingegebenen Daten um ein Klassifikations oder um ein Regressionsproblem handelt. Hier können Sie zum Testen die trees-Daten verwenden.

Aufgabe 4 (Frequent-Item Sets) **(5 Punkte)**

Auf dem vergangenen Aufgabenblatt haben Sie bereits praktisch mit Frequent-Item-Set Mining gearbeitet. In dieser Aufgabe sollen Sie ihr Verständnis für Frequent-Item-Set Mining theoretisch vertiefen.

Betrachten Sie eine Transaktionsdatenbank $\mathcal{T} = \{T_1, \ldots, T_N\}$ mit $T_i \subseteq \mathcal{I}$,wobei $\mathcal{I}$ die Menge aller Items in dieser Datenbank sei. Rufen Sie sich folgende Definitionen ins Gedächnis (vgl. Foliensatz `"WiD 180426 Häufige Mengen und Assoziationsregeln"` und Foliensatz `"WiD 180503 Häufige Mengen 2"`):

$$
\begin{aligned}
\text{Support einer Menge: } supp(A) &= |\{T_i | A \subseteq T_i, T_i \in \mathcal{T}\}| \\
\text{Support einer Regel: } s(A \to B) &= \frac{supp(A \cup B)}{N} \\
\text{Confidence einer Regel: } c(A \to B) &= \frac{supp(A \cup B)}{supp(A)} \\
\text{Lift einer Regel: } lift(A \to B) &= \frac{\frac{supp(A \cap B)}{N}}{\frac{supp(A)}{N} \cdot \frac{supp(B)}{N}} = \frac{N \cdot supp(A \cap B)}{supp(A) \cdot supp(B)}
\end{aligned}
$$

Hinweis: In der Vorlesung wird für eine Regel der Implikationspfeil benutzt, d.h. eine Regel wird mit $A \Rightarrow B$ benannt. Um an dieser Stelle Verwirrung mit dem Implikationspfeil aus der Logik zu verhindern benutzen wir den einfachen Pfeil, d.h. eine Assoziationsregel ist mit $A \rightarrow B$ gekennzeichnet, wohingegen $A \Rightarrow B$ eine logische Aussage der Form "A impliziert B" meint.

a) (1 Punkt) In der Vorlesung haben Sie die Anti-Monotonie Eigenschaft des supports kennen gelernt. An dieser Stelle sollen Sie diese Eigenschaft formal beweisen. Zeigen Sie:

$$\forall B \subseteq C: \ supp(A \cup B) \geq supp(A \cup C)$$

Hinweis: Ein informeller Beweis bzw. eine sinnvolle Begründung gibt bereits Teilpunkte!

b) (0.5 Punkte pro Teilaufgabe) Zeigen oder widerlegen Sie folgenden Aussagen.
Hinweis: Sie dürfen die Anti-monotonie Eigenschaft aus Teilaufgabe a) verwenden.

1) $s(A \rightarrow B) \geq s(A \rightarrow \emptyset) \cdot s(B \rightarrow \emptyset)$
2) $c(A \rightarrow B) = c(B \rightarrow A) \Rightarrow supp(A) = supp(B)$
3) $c(A \rightarrow B) \cdot c(B \rightarrow C) = c(A \rightarrow C)$
4) $c(A \rightarrow B) \geq c(A \rightarrow C)$ mit $B \subseteq C$
5) $c(\emptyset \rightarrow A) \cdot c(\emptyset \rightarrow B) \geq c(A \rightarrow B)$
6) $lift(A \rightarrow B) \geq lift(A \rightarrow C)$ mit $B \subseteq C$

c) (0.5 Punkte pro Teilaufgabe) Betrachten Sie folgende Transaktionsdatenbank:

$$\begin{aligned} T_1 &= \{Bier, Grillkohle\} \\ T_2 &= \{Bier, Grillkohle, Zahnpasta\} \\ T_3 &= \{Bier, Zahnpasta\} \\ T_4 &= \{Zahnpasta\} \end{aligned}$$

Beantworten Sie folgende Fragen:

1) Wie stehen $\{Bier\}$ und $\{Grillkohle\}$ zueinander, d.h. wie ist die Korrelation von $\{Bier\} \rightarrow \{Grillkohle\}$?
2) Wie stehen $\{Bier, Grillkohle\}$ und $\{Zahnpasta\}$ zueinander, d.h. wie ist die Korrelation von $\{Bier, Grillkohle\} \rightarrow \{Zahnpasta\}$?