Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert, Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases II

WS 2012/2013

Übungsblatt 1: Feature Selektion

Aufgabe 1-1 *Warum Feature-Selektion?*

Wie wir in der Vorlesung gelernt haben, besteht Feature-Selektion darin, aus einer Menge von gegebenen Features eine informative Untermenge auszuwählen. In dieser Aufgabe wollen wir uns anhand der folgenden Fragen überlegen, warum man Feature-Selektion benötigt:

(a) Warum ist Feature-Selektion aus *experimenteller* Sicht sinnvoll?

(b) Warum ist Feature-Selektion aus *statistischer* Sicht sinnvoll?

(c) Warum ist Feature-Selektion auf *naturwissenschaftlicher* Sicht sinnvoll?

Aufgabe 1-2 *XOR-Problem: Kombination von Features*

Als nächstes wollen wir klären, ob zwei Features, die für sich genommen irrelevant sind, gemeinsam informativ seien können. Dazu betrachten wir das folgende 2-dimensionale Klassifikationsbeispiel. Gegeben seien:

- Klasse 1: Punkt A $(1, 1)$ und Punkt C $(-1, -1)$
- Klasse 2: Punkt B $(1, -1)$ und Punkt D $(-1, 1)$

Jeder Punkt besitzt eine x- und eine y-Koordinate als Feature.

(a) Kann man Klasse 1 und Klasse 2 anhand der x-Koordinate, anhand der y-Koordinate, oder mit beiden gemeinsam separieren?

(b) Würde ein Forward Selection oder ein Backward Elimination Algorithmus x und y als informative Features auswählen?

Aufgabe 1-3 *Unterraumselektion mittels Inkonsistenz*

Wir haben in der Vorlesung konsistenzbasierte Kriterien für Feature-Selektion kennengelernt.

Bestimmen Sie für das folgende Klassifikationsproblem den informativsten Unterraum mittels Branch-and-Bound mit Inkonsistenzkriterium.

ID	Attribut X	Attribut Y	Attribut Z	Klasse
A	2	rot	ja	1
B	3	rot	ja	1
C	3	grün	ja	1
D	4	grün	ja	2
E	1	rot	ja	2
F	1	grün	ja	2

Aufgabe 1-4 *Potential für Inkonsistenz bei unterschiedlichen Wertedomänen*

Seien Attribute $A_i \in \mathbb{N}$, Attribute $B_i \in \{\text{rot}, \text{grün}, \text{blau}\}$ und Attribute $C_i \in 0, 1$.

Können in einem Datensatz mit n Elementen alle Elemente paarweise unterschiedlich sein, wenn wir einen Featureraum mit folgenden Attributen haben:

- A_1
- B_1
- C_1
- $C_1 \times C_2 \times C_3$
- $B_1 \times C_2$
- $B_i^k \times C_j^l$
- $B_1 \times C_2 \times A_3$

Aufgabe 1-5 *Statistische Maße zur Featurebewertung*

Wir haben in der Vorlesung mehrere statistische Maße zur Bewertung von Features kennengelernt, darunter Informationsgewinn und Mutual Information.

Berechnen Sie für folgendes binäre Klassifikationsproblem mit 3-dimensionalen Vektoren ein Ranking der einzelnen Features

(a) basierend auf dem Informationsgewinn,

(b) basierend auf χ^2 (splitten Sie X bei $X \leq 2$)

(c) basierend auf der Mutual Information.

Jede Klasse umfasst 4 Punkte:

(a) Klasse 1: $A(2, 1, 0)$, $B(2, 0, 0)$, $C(1, 1, 0)$ $D(1, 0, 0)$

(b) Klasse 2: $E(4, 0, 1)$, $F(4, 1, 0)$, $G(3, 0, 1)$ $H(3, 1, 0)$