

Knowledge Discovery in Databases
SS 2013

Übungsblatt 7: Klassifikation I

Aufgabe 7-1 Klassifikation vs. Clusteranalyse

Bei welchen der folgenden Aufgabenstellungen handelt es sich um Klassifikationsprobleme, bei welchen um Clusteranalyse?

- (a) Emails im Posteingang sollen nach Spam und nicht Spam sortiert werden.
- (b) Eine Datenbank von Nutzern soll nach ihrem Kaufverhalten gruppiert werden.
- (c) In einem Supermark sollen Produkte die oft zusammen gekauft werden in einem Regal nebeneinander platziert werden, um so die Verkäufe zu steigern.
- (d) Das Spam-Vorkommen soll analysiert werden, um zu erkennen, ob es darin unterschiedliche Gruppen / Typen von Werbung gibt.
- (e) Basierend auf der DNA einer Person soll vorhergesagt werden, ob sie in den nächsten 10 Jahren an Diabetes leiden wird.
- (f) Daten von Patienten mit Herzkrankheiten sollen analysiert werden, ob es darin Gruppen gibt für die spezielle Therapien besser funktionieren als für andere.
- (g) Einteilung von Webseiten in Kategorien wie "Sport", "Wirtschaft", "Unterhaltung".

Aufgabe 7-2 Bewertung von Klassifikatoren

Gegeben sei ein Datensatz mit bekannter Klassenzugehörigkeit der Objekte. Um die Qualität eines Klassifikators K zu ermitteln wurden die Objekte mittels K klassifiziert. Die Klassifikationsergebnisse sind in der folgenden Tabelle dargestellt.

ID	Objektklasse	$K(o)$
O_1	A	A
O_2	B	A
O_3	A	C
O_4	C	C
O_5	C	B
O_6	B	B
O_7	A	A
O_8	A	A
O_9	A	A
O_{10}	B	C
O_{11}	B	A
O_{12}	C	A
O_{13}	C	C
O_{14}	C	C
O_{15}	B	B

- Berechnen Sie anhand der tabellierten Ergebnisse Precision und Recall jeder Klasse.
- Um ein vollständiges Maß für die Güte der Klassifikation bezüglich einer Klasse zu haben, wird häufig auch das sogenannte F_1 -Measure (harmonisches Mittel zwischen Precision und Recall) verwendet. Das F_1 -Measure für Klasse i ist wie folgt definiert:

$$F_1(K, i) = \frac{2 \cdot \text{Recall}(K, i) \cdot \text{Precision}(K, i)}{\text{Recall}(K, i) + \text{Precision}(K, i)}$$

Berechnen Sie das F_1 -Measure für alle Klassen.

- Berechnen Sie die durchschnittliche Precision, den durchschnittlichen Recall und daraus das F_1 -Measure.

Aufgabe 7-3 Bewertung von Klassifikatoren

Gegeben ein Datensatz D mit Objekten aus zwei Klassen A und B ($D = A \cup B$), die *völlig zufällig* erzeugt wurden. Zudem gibt es in diesem Datensatz für beide Klassen jeweils die gleiche Anzahl an Objekten, d.h. $|A| = |B|$. Der beste Klassifikator kann daher immer nur die Klasse mit den meisten Objekten vorhersagen.

- Welche echte *Fehlerrate* ist von so einem *optimalen* Klassifikator zu erwarten?
- Welche Fehlerraten sind bei der Evaluation dieses Klassifikators beim Leave-one-out Test und dem 0.632 Bootstrap Verfahren zu erwarten? Interpretieren Sie die Resultate.