Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Matthias Schubert

Knowledge Discovery in Databases

WS 2008/09

Übungsblatt 2

Aufgabe 2-1 Hausaufgabe
Metrische Distanzfunktionen

Seien $x = (x_1, \ldots, x_d)$ und $y = (y_1, \ldots, y_d)$, mit $d \in \mathbb{N}$ und $x, y \in \mathbb{R}^d$. Zeigen oder widerlegen Sie, dass die folgende Distanzfunktion eine Metrik ist:

(a) $dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$

Aufgabe 2-2 Überkreuzvalidierung

(a) Geben Sie einen Algorithmus in Pseudocode an, der eine gegebene Datenmenge in k stratifizierte Faltungen aufteilt. Dabei bedeutet stratifiziert, dass der relative Anteil der Trainingsbeispiele jeder Klasse in jeder Faltung möglichst der Verteilung in der gesamten Datenmenge entsprechen soll. Sie können für Ihren Algorithmus davon ausgehen, daß die Trainingsbeispiele einer Klasse jeweils in einer separaten Datei gespeichert sind.

(b) Wie würde man die erzeugten Faltungen jetzt für eine k-fache Überkreuzvalidierung benutzen. Geben Sie auch hierfür einen Algorithmus in Pseudocode an.

Aufgabe 2-3 Bewertung eines Klassifikationsergebnisses

Gegeben sei folgende Konfusionsmatrix, die durch 10-fache Überkreuzvalidierung erzeugt wurde.

klassifiziert als →	**Klasse A**	**Klasse B**
Klasse A	84	1
Klasse B	5	10

(a) Berechnen Sie den Klassifikationsfehler, die Klassifikationsgenauigkeit, sowie Precision und Recall für beide Klassen.

(b) 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 ist wie folgt definiert:

$$F_1(A) = \frac{2 \cdot Recall_A \cdot Precision_A}{Recall_A + Precision_A}$$

Berechnen Sie das F_1-Measure für beide Klassen.

In welchen Fällen ist die Verwendung des F_1-Measures der Berechung der Klassifikationgüte vorzuziehen ?

(c) Da das F_1-Measure nur klassenweise definiert ist, ist es nicht direkt dazu geeignet einen Überblick über die Gesamtgüte des Klassifikators zu geben. Daher mittelt man es häufig über alle Klassen auf. Dabei wird zwischen den folgenden 2 Methoden unterschieden:

- Micro Average F_1-Measure: Hierbei wird zuerst für jede Klasse einzeln die folgenden 3 Werte ermittelt: Anzahl der richtig klassifizierten Objekte (true positives"), Anzahl der falsch zu einer Klasse zugordneten Objekte (false positives") und Anzahl der Objekte, die zur Klasse gehören und zu einer anderen zugeordnet wurden(false negatives"). Diese 3 Werte werden dann für jede Klasse aufsummiert und anschließend werden Precision, Recall und F_1-Measure gebildet.
- Macro Average F_1-Measure: Precision and Recall werden für jede Klasse gebildet und anschließend wird der Durchschnitt über alle Klassen gebildet. Danach kann dann das F_1-Measure gebildet werden.

Bilden Sie im obigen Beispiel das Micro-Average und Macro-Average F_1-Measure.

Wiso unterscheiden sich beide Methoden?