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

Knowledge Discovery in Databases

SS 2012

Übungsblatt 4: Classification

Aufgabe 4-1 *Naive Bayes*

The skiing season is open. To reliably decide when to go skiing and when not, you could use a classifier such as Naive Bayes. The classifier will be trained with your observations from the last year. Your notes include the following attributes:
The weather: The attribute `weather` can have the following three values: `sunny`, `rainy` and `snow`.
The snow level: The attribute `snow level` can have the following two values: ≥ 50 (There are at least 50 cm of snow) and < 50 (There are less than 50 cm of snow).

Assume you wanted to go skiing 8 times during the previous year. Here is the table with your decisions:

weather	**snow level**	**ski ?**
sunny	< 50	no
rainy	< 50	no
rainy	≥ 50	no
snow	≥ 50	yes
snow	< 50	no
sunny	≥ 50	yes
snow	≥ 50	yes
rainy	< 50	yes

(a) Compute the *a priori* probabilities for both classes `ski = yes` and `ski = no` (on the training set)!

(b) Compute the conditional distributions for the two classes for each attribute.

(c) Decide for the following weather and snow conditions, whether to go skiing or not! Use the Naive Bayes classificator for finding the decision.

	weather	**snow level**
day A	sunny	≥ 50
day B	rainy	< 50
day C	snow	< 50

Aufgabe 4-2 *Nearest Neighbor classification*

Give a set of points, consisting of at least four points in 2 dimensions, such that the Nearest Neighbor classification ($k = 1$) only gives incorrect classification results. Use Euclidean distance as distance function.

Aufgabe 4-3 *Evaluation of classifiers*

Given a data set D with objects from classes A and B ($D = A \cup B$) where the class assignments are *random* (not related to the attribute values). Therefore, the best classifier is to always give the majority class. Furthermore, let the two classes have the same size $|A| = |B|$.

- Which true *error rate* is to be expected from such an *optimal* (for this data set) classifier?
- Which error rates are to be expected when evaluating the optimal classifier using a leave-one-out test and the 0.632 Bootstrap method? Interpret these results.