Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 12: Decision Trees, Nearest Neighbor Classifier, Regression Trees

QA Session

The latter part of the last lecture slot on February 5th will be dedicated to a QA-session which is intended to give you an opportunity to ask questions about the lecture and exercise contents and also benefit from the discussions of other students' questions. Note that we cannot answer any specific questions regarding exam contents. During the session, we will discuss the questions, which you send to us via e-mail in advance (`berrendorf@dbs.ifi.lmu.de`). Please hand in your questions before February 4th, 12:00, so we have some time to prepare them. Note that in contrast to the lecture, the QA session will not be recorded.

Exercise 12-1 Decision Trees

Predict the risk class of a car driver based on the following attributes:

Attribute	Description	Values
time	time since obtaining a drivers license in years	{1-2, 2-7, >7}
gender	gender	{male, female}
area	residential area	{urban, rural}
risk	the risk class	{low, high}

For your analysis you have the following manually classified training examples:

ID	time	gender	area	risk
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high
7	2-7	f	urban	low
8	2-7	m	urban	low

(a) Construct a decision tree based on this training data. For splitting, use information gain as measure for impurity. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class, you do not need to apply a pruning algorithm.

(b) Apply the decision tree to the following drivers:

ID	time	gender	area
A	1-2	f	rural
B	2-7	m	urban
C	1-2	f	urban

Exercise 12-2 Information gain

In this exercise, we want to look more closely at the information gain measure.

Let T be a set of n training objects with the attributes $A_1, \ldots, A_a$ and the k classes c_1 to c_k.

Let $\{T_i^A \mid i \in \{1, \ldots, m_A\}\}$ be the disjoint, complete partitioning of T produced by a split on attribute A (where m_A is the number of disjoint values of A).

(a) *Uniform distribution*
Compute *entropy*(T), *entropy*(T_i^A) for $i \in \{1 \ldots m_A\}$ as well as *information-gain*(T, A) given the assumption that the class membership of T is uniformly distributed and independent of the values of A. Interpret your result!

(b) *Attributes with many values*
Let A be an attribute with random values, not correlated to the class of the objects. Furthermore, let A have enough values, such than no two instances of the training set share the same value of A. What happens in this situation when building the decision tree? What is problematic with this situation?

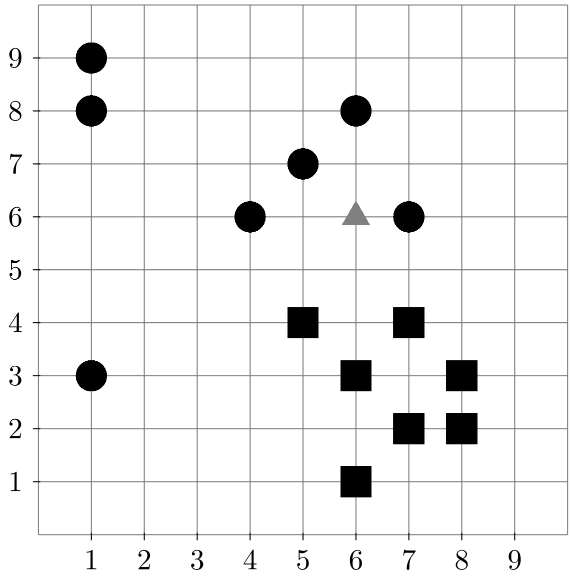
Exercise 12-3 Nearest neighbor classification

The 2D feature vectors in the figure below belong to two different classes (circles and rectangles). Classify the object at $(6, 6)$ — in the image represented using a triangle — using k nearest neighbor classification. Use Manhattan distance (L_1 norm) as distance function, and use the non-weighted class counts in the k-nearest-neighbor set, i.e. the object is assigned to the majority class within the k nearest neighbors. Perform kNN classification for the following values of k and compare the results with your own "intuitive" result.

(a) $k = 4$

(b) $k = 7$

(c) $k = 10$



Exercise 12-4 Regression Trees

Consider the following data samples of the form (x, y), where the input value is $x \in R$ and the output value is $y \in R$:

$$p_1 = (-3, -1), p_2 = (-2, 0), p_3 = (-1, 1), p_4 = (1, 1), p_5 = (2, 0), p_6 = (3, -1)$$

Search for the first best split. If the decision is obvious, you don't have to compute all possible splits. Then decide whether the split is significant or not by using the impurity ratio with $\tau_0 = 0.5$.

(a) Fit constant functions and use the variance of the residuals as impurity measure. Note that an optimal constant regression function always predicts the mean output value over all training samples, such that in this case, the variance of the residuals corresponds to the variance of the outputs.

(b) Fit linear functions and use the variance of the residuals as impurity measure.