University of Southern Denmark
IMADA
Arthur Zimek

DM868/DM870/DS804: Data Mining and Machine Learning

Spring term 2023

Exercise 12: Decision Trees, Practical Exploration of Classifiers

Exercise 12-1 Decision trees (1 point)

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

- Time since getting the driving license ($1-2$ years, $2-7$ years, > 7 years)
- Gender (male, female)
- Residential area (urban, rural)

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

Person	Time since license	Gender	Area	Risk class
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 dataset. Use information gain for selecting the split attributes. 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:
Person A: 1-2, f, rural
Person B: 2-7, m , urban
Person C: 1-2, f, urban

Exercise 12-2 Information gain (1 point)

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 $\{T_i^A \,|\, i \in \{1, \ldots, m_A\}\}$ 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) *Additional uniform distribution*
We want to analyze how the number of different values influences the information gain. For this, we compare two attributes, attribute A with m_A values and attribute A' with $m_{A'} = m_A + 1$ values, where the relative frequencies in A' in values 1 to m_A are identical to that of A and in the additional value $m_{A'}$ there is a uniform distribution of the classes.
How does *information-gain*(T, A) differ from *information-gain*(T, A')? Interpret your result!

(c) *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 not any 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 Decision trees, naïve Bayes, and k-nn classification – Practical

(a) Work with some toolbox for classification (e.g., R, Python, WEKA) to study the impact of different settings on the behavior of decision trees, the naïve Bayes classifier, and the k nearest neighbor classifier on some dataset (e.g., Iris).

(b) How does the behavior of the k nearest neighbor classifier change with the choice of k?

(c) What is the impact of parameter choices on the quality of decision trees?

(d) How does the behavior of the three classifiers change with the amount of training data (e.g., choice of training-test-splits)?