Data Science

11. Exercise / Practical

Theoretical Exercises

Excercise 11.1: (Theoretical) Decision Tree[a]

[a] https://www.dbs.ifi.lmu.de/Lehre/KDD/WS1819/tutorials/exercise_12.pdf

We want to predict the risk class of car driver based on the attributes time (years the driver has a license), gender and area (type of area the driver lives in):

	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) Created a decision tree based classified with the method given in the lecture

b) Apply the decision tree classifier on the following data

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

Practical Exercises

Excercise 11.2: (Practical) Ames Housing Classification

We consider the Ames Housing dataset and want to predict the SalePrice as classes (cheap, expensive) based on the given variables.

a) Prepare the data for a machine learning model by

- compute the new variables ”years since built” and ”years since remod/add”
- categorize the values for SalePrice in cheap (0 - 160000) and expensive (160000 - ∞)
- analyze all variables and add classes ”other” for values with few occurrences
- remove unneeded variables and variables with extreme imbalanced data (e.g. 2 classes, one with 99% of the data and the other with 1% of the data)

b) Create a train / test splitting

c) Create a decision tree classifier

d) Analyze the quality of the classifier by computing the different measures given in lecture

e) Train different versions of the classifier (by varying the parameters) and compare these

Hint: A decision tree classifier can be created with sklearn: `https://scikit-learn.org/1.5/modules/generated/sklearn.tree.DecisionTreeClassifier.html`