W3WI DS304.1 Applied Machine Learning Fundamentals

Exercise Sheet # 7 - Decision Trees and Ensemble Methods

Question 1 EX 2020 (ID3 algorithm)

You are given the dataset listed in table 1 below. The set consists of $m = 3$ attributes A_1, A_2, and A_3. The data points belong to either of the two classes $\oplus$ (the positive class) or $\ominus$ (the negative class).

Derive a decision tree classifier from the given dataset using the **information gain** heuristic. Write down all computations necessary and draw the final decision tree!

A_1	A_2	A_3	$\mathcal{C}$
a	p	x	$\oplus$
a	m	x	$\oplus$
b	m	x	$\ominus$
b	p	x	$\ominus$
a	p	y	$\oplus$
a	p	z	$\ominus$
a	m	z	$\ominus$
b	m	z	$\ominus$
b	m	y	$\ominus$
a	m	y	$\oplus$

Table 1: Training data for question 1.

Question 2 (ID3 algorithm)

The following labeled dataset is presented to you (see table 2). Construct a decision tree classifier from the training data using the **information gain** splitting heuristic. To avoid overfitting the training data we specify a maximal depth of 2. Draw the final decision tree you have computed. *(Please show your calculations.)*

Outlook	Temperature	Humidity	Wind	Sport
sunny	cold	high	weak	soccer
cloudy	cold	low	strong	soccer
sunny	warm	low	weak	soccer
rainy	cold	high	weak	squash
sunny	cold	high	weak	squash
rainy	warm	high	strong	squash
cloudy	cold	high	weak	squash
rainy	warm	high	weak	squash
cloudy	warm	high	weak	tennis
cloudy	cold	low	strong	tennis
sunny	cold	low	strong	tennis
cloudy	cold	high	weak	tennis

Table 2: Training data for question 2.

Question 3 EX 2020

Generally speaking, which class distribution maximizes the entropy function? Consider two classes.

Question 4 EX 2020, modified (Ensemble methods)

Which of the following algorithms is **not** an example of ensemble learning?

- ☐ Random forest
- ☐ AdaBoost
- ☐ Logistic regression
- ☐ ExtraTrees
- ☐ k-nearest neighbors
- ☐ All algorithms are ensemble methods.

Question 5 EX 2021, modified (Entropy and Gini-index)

Let the dataset $\mathcal{D} := \{A, A, B, C, B, A, C, C, A, C\}$ consisting of three possible classes be given. Please work through the following tasks:

1. Calculate the entropy of the dataset.
2. Calculate the Gini-index of the dataset.
3. Do entropy and Gini index always lead to the same decision tree?

Question 6 EX 2023 (Random forests)

Briefly outline what a random forest is. Which steps have to be done in the training phase? What is the advantage of a random forest compared to a single decision tree?