Homework 4 Basic Machine Learning
For the deadline see Canvas
Version: Fri 21st Sept, 2018 at 13:26.

Introduction

1. This is a group assignment, so sign up in groups of two students.
2. Each group has to submit a **pdf** with their answers and explanation. **Please put your names and group number at the top of the hand in**.
3. For questions about this homework assignment use the Discussion Board on Canvas.
4. Of course you may use a calculator or a programming environment such as Matlab or Python. **But your report should not contain any code. Explain your computations and results in English!**
5. It is allowed to incorporate handwritten notes or derivations or drawings in your submission as long as these are readable!
6. Explain your answers!

Exercise 1: Decision trees (10 points)

Read the handout on decision trees (Handout_DT.pdf)

Part a (to be adapted)

Do Exercise 3 of Section 4.8 of the handout on Decision Trees (Handout_DT.pdf).

Part b

A data analyst has collected data (see table below) about customer loans. The goals is to predict, based on the customer profile, if a loan for a customer has a high risk or not.

payment history	debt	guarantee	income	risk
bad	high	no	0-15 KEuro	high
average	high	no	15-35 KEuro	high
average	low	no	15-35 KEuro	low
average	low	no	0-15 KEuro	high
average	low	no	$>$ 35 KEuro	low
average	low	sufficient	$>$ 35 KEuro	low
bad	low	no	0-15 KEuro	high
bad	low	sufficient	$>$ 35 KEuro	low
good	high	sufficient	$>$ 35 KEuro	low
good	high	no	0-15 KEuro	high
good	high	no	15-35 KEuro	low
good	high	no	$>$ 35 KEuro	low
bad	high	no	15-35 KEuro	high

What is the Gini index of the attribute *payment history*?

Part c

Assume that we test a Decision Tree classifier on a test set consisting of 220 samples and the measured accuracy is 0.87. What is the 95% confidence interval for classifier?

Exercise 2: (10 points)

For a certain three class classification problem a classifier has following confusion matrix.

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	110	8	7
	C_2	16	130	10
	C_3	26	5	120

In this confusion matrix the columns correspond to the predicted class and the rows the actual class.

Part a

What is the accuracy of the classifier?

Part b

What is the precision for class C_2?

Part c

What is the recall for class C_1?

Exercise 3: Pruning (10 points)

Consider the following situation in process of pruning a decision tree, cf. Figure 1:

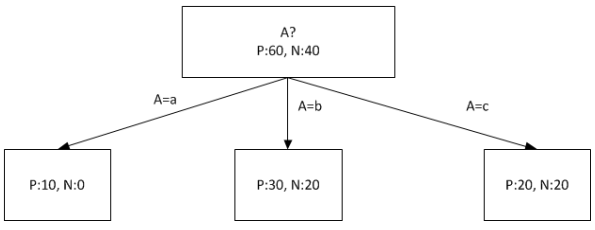

Figure 1: Part of a DT examined for pruning.

Part a

Calculate the value for the Δ which is used in χ^2 pruning. (See and read the slides for more info.)

Part b

For which confidence levels will the top node in Figure 1 not be pruned? Explain how you derived these confidence levels.

Exercise 4: Multiple choice questions (3 bonus points)

Design two multiple choice (MC) questions concerning the material of week 4. Clearly indicate what knowledge or skill you want to test with the MC questions and what the correct answers are.