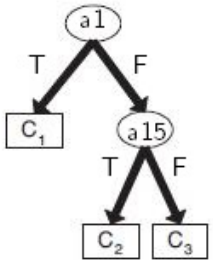
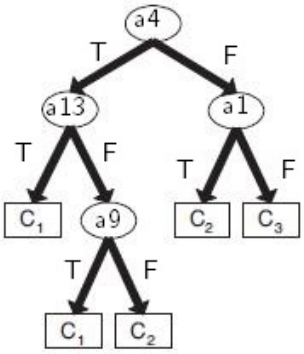
Data Mining	Fall 2021
Tu,Th 2:00 pm – 3:15 pm	HW 4 UG

Chapter 3

Loosely based on Exercise 10 [3pts] Consider the decision trees shown in Figure 4.3. Assume they are generated from a data set that contains 16 binary attributes and 3 classes, $C1$, $C2$, and $C3$. Let the data contain total of n data points classified by the trees. Compute the total description length of each decision tree according to the minimum description length principle.



(a) Decision tree with 7 errors



(b) Decision tree with 4 errors

- sample size is 200 examples.
- Each tree is described by tree size (number of nodes) and description of nodes in lexicographic order (=BFS order)
- all domains are Boolean
- Each node of the tree is described by number of children (0-no or 1-two), ID of the splitting attribute, and junior child value (T or F if internal node) or class (Y no N if leaf). So the encoding of each vertex is based on number of bits necessary to encode attribute number+1+1.
- Cost(tree) is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- all items are encoded using gamma encoding (i.e. if k bits need to be encoded then $2(k-1)+1$ bits of gamma encoding must be generated).

1 pt Show left tree cost as function of sample size n

1 pt Show right tree cost as function of sample size n

3 pts Which one of these trees is MDL-better for for $\delta = 0.99$ at given sample size?

Hint: Use MDL formulas slide 2 lecture 7-2. If necessary, solve equations (use R, Mathematica or Wolfram alpha).

Chapter 4

Exercise 6a [1pt] Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

Exercise 6c [1pt] Given the information in part (a), is a randomly chosen college student who is a smoker more likely to be a graduate or undergraduate student?

Exercise 6d [1pt] Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

Exercise 7a [1pt] Consider the data set shown in Table 5.1. Estimate the conditional probabilities for $P(A|+),\ P(B|+),\ P(C|+), P(A|-),\ P(B|-),\ and P(C|-)$.

Table 5.1. Data set for Exercise 7.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

Exercise 7b [1pt] Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$ using the naive Bayes approach.

exercise 7c [1pt] Estimate the conditional probabilities using the m-estimate approach, with $p = 1/2$ and $m = 4$.

exercise 7d [1pt] Repeat part (b) using the conditional probabilities given in part (c).

Exercise 7e [1pt] . Compare the two methods for estimating probabilities. Which method is better and why?

Bonus

An axis-aligned n-dimesional rectangle classifier $h_{(\bar{l},\overline{u})}$ is given by two vectors $\bar{l}, \overline{u} \in \mathbb{R}^n$ such that $\bar{l} < \overline{u}$ (i.e. $l_i < u_i$ for all $1 \leq i \leq n$). A vector $\overline{x} \in \mathbb{R}^n$ is labeled 1 by this classifier $h_{(\bar{l},\overline{u})}$ if $\bar{l} < \overline{x} < \overline{u}$, i.e. for every i holds $l_i < x_i < u_i$. Otherwise $\overline{x}$ is labeled 0.

Problem B [4 pts] You are tasked with prediction of a heart attack inpatients. A training set consists of data points (patient) that have the following features:

- blood pressure (BP),
- body-mass index (BMI),
- age (A),
- level of physical activity (P),
- income (I).

For legacy reasons you have to pick between 2 algorithms: the first algorithm ERM picks an axis aligned rectangle in the two dimensional space spanned by the features BP and BMI and the other algorithm ERM picks an axis aligned rectangle in the five-dimensional space spanned by all features.

2pts Explain the pros and cons of each choice

2pts Explain how the number of available labeled training samples will affect your choice

Hint: Think of SRM paradigm when you are trying to give your reasons