Artificial Intelligence
Tutorial 5 on Machine Learning

Introduction

The following multiple choice questions are examples of typical questions one can expect on the AI exam. The questions on the AI exam are also multiple choice, but for this tutorial one has to explain the answers given. Moreover at the end one can find some open questions.
After the tutorial the answers to the MC will be available on Canvas.

Questions on ML

1. For marketing purposes a retailer wants to distinguish between costumers younger than 35 (class Y) and customers older than 35 (class O). The following table summarizes the data set in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by A with values $a1$, $a2$ and $a3$, B with values $b1$ and $b2$, C with values $c1$ and $c2$ and D with values $d1$ and $d2$

A	B	C	D	Number of Instances	
				Y	O
a1	b1	c1	d1	12	4
a2	b1	c1	d2	4	6
a3	b1	c1	d1	6	0
a1	b2	c1	d2	0	12
a2	b2	c1	d1	4	2
a3	b2	c1	d2	0	4
a1	b1	c2	d1	0	8
a2	b1	c2	d2	8	0
a3	b1	c2	d1	4	0
a1	b2	c2	d2	0	4
a2	b2	c2	d1	8	0
a3	b2	c2	d2	4	0

Given a new customer x, what is the information content in the answer to the question: "What is the class (Y or O) of x"? Or in other words what is entropy of the data set above with respect to the class labels Y and O. Choose the alternative which

is closest to your answer.
You can find a table of $-p \log_2(p)$ values at the last page of this tutorial.

(a) 1.00

(b) 0.00

(c) 0.99

(d) 0.01

2. The analyst wants to learn the above classification problem using decision trees. If he uses "information gain" as selection criteria what will be the information gain of attribute A? Choose the alternative which is closest to your answer.
You can find a table of $-p \log_2(p)$ values at the last page of this exam.

(a) 0.83

(b) 0.16

(c) 0.82

(d) 0.17

3. 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
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	high
good	high	no	> 35 KEuro	low
bad	high	no	15-35 KEuro	high

What is the information gain of the attribute *payment history* ?

4. Consider the following dataset with attributes (features) A en B, both with values T en F. The class label is given by + en $-$.

Ex.	A	B	*Class label*
1	T	F	+
2	T	T	+
3	T	T	+
4	T	F	+
5	T	T	+
6	F	F	-
7	F	F	-
8	F	F	-
9	T	T	-
10	T	F	-

What is the entropy of this dataset with respect to the class label?

(a) 1

(b) 0

(c) 1/2

(d) None of the above.

5. Once again consider the dataset of the previous question. What is the information gain of attribute A?

(a) 0.86

(b) 0.30

(c) 0.60

(d) 0.40

6. Consider the following confusion matrix

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	120	15	20
	C_2	16	150	10
	C_3	22	3	130

What is the accuracy of this classifier?

(a) 150/486

(b) 120/158+150/168+130/160

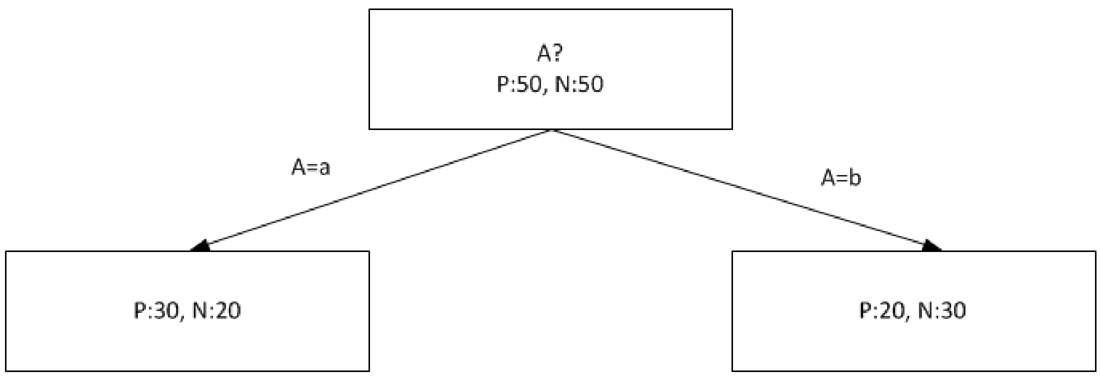
(c) 120/135+150/176+130/155

(d) 400/486

7. Once again consider the confusion matrix of the previous question. What is the precision for class C_2?

    (a) 150/486
    (b) 150/168
    (c) 150/176
    (d) 150/400

8. Consider the following part of the decision tree, with two leaf nodes and one parent node which splits on attribute A. The notation P:x N:y means that the node has x positive examples and y negative examples.



In order to apply χ^2 pruning one has to calculate the value of Δ:

$$\Delta = \sum_{i=1}^{i=d} \left[\frac{(p_i - \widehat{p_i})^2}{\widehat{p_i}} + \frac{(n_i - \widehat{n_i})^2}{\widehat{n_i}} \right]$$

What is the value of Δ in this case?

    (a) 4
    (b) 8
    (c) 12
    (d) None of the above

Table for $-p\ \log(p)$

p	$-p\ \log_2(p)$	p	$-p\ \log_2(p)$	p	$-p\ \log_2(p)$
0	0	1/8	0.38	1/10	0.33
1	0	2/8	0.50	2/10	0.46
1/2	0.50	3/8	0.53	3/10	0.52
1/3	0.53	4/8	0.50	4/10	0.53
2/3	0.39	5/8	0.42	5/10	0.50
1/4	0.50	6/8	0.31	6/10	0.44
2/4	0.50	7/8	0.17	7/10	0.36
3/4	0.31	1/9	0.35	8/10	0.26
1/5	0.46	2/9	0.48	9/10	0.14
2/5	0.53	3/9	0.53	1/11	0.31
3/5	0.44	4/9	0.52	2/11	0.45
4/5	0.26	5/9	0.47	3/11	0.51
1/6	0.43	6/9	0.39	4/11	0.53
2/6	0.53	7/9	0.28	5/11	0.52
3/6	0.50	8/9	0.15	6/11	0.48
4/6	0.39			7/11	0.42
5/6	0.22			8/11	0.33
1/7	0.40			9/11	0.24
2/7	0.51			10/11	0.13
3/7	0.52				
4/7	0.46				
5/7	0.35				
6/7	0.19				