Test Pearl ~~100~~ — Intelligent Interaction

October 2 2015

The test consists of 5 questions. The grade is the number of achieved points divided 100.

☞ **1** *(20 points)* A bag H_1 contains **10** marbles: 2 red, 3 white and 5 blue. Bag H_2 contains also **10** marbles: 4 red, 2 white and 4 blue. Someone throws a fair dice and if the outcome is divisible by 3 then he chooses bag H_1, else he chooses bag H_2. After choosing a bag he draws 5 marbles with replacement. The outcome D is 2 red, 2 white and 1 blue marble, so $D =< 2,2,1 >$.

(a) Compute $P(D|H_1)$.
(b) What is the most likely bag from which the marbles are drawn; H_1 or H_2? Motivate your answer by a computation using Bayes law.

Antwoord op 1.

1. $P(D|H_1) = \frac{5!}{2!1!2!}(\frac{2}{10})^2(\frac{3}{10})^2(\frac{5}{10}) = 0.054$
2. $P(D|H_2) = \frac{5!}{2!1!2!}(\frac{4}{10})^2(\frac{2}{10})^2(\frac{4}{10}) = 0.0768.$ And $P(H_1|D)/P(H_2|D)$ equals $[(\frac{3}{10})^2(\frac{5}{10})(\frac{1}{3})]/[(\frac{4}{10})^2(\frac{4}{10})(\frac{2}{3})] = 45/128 = 0.35$. Hence H_2 is the most likely bag.

☞ **2** *(20 points)* Given the following piece of text from an email:

attention if you are in debt. if you are then we can help. qualifying is now at your fingertips and there are no long distance calls

(a) Assume that we use as vocabulary $V =$\{attention, adult, debt, publications, qualifying, xxx \}. How would this piece of text be coded using a binary coding and this vocabulary V?
(b) For convenience consider a smaller vocabulary $V =$\{attention, adult, debt\} and assume that we have a dataset consisting of 100 emails of which 30 are spam and with the following vocabulary frequency list:

Word	Ham	Spam
attention	30	10
adult	0	22
debt	4	20

This means for instance that the word "attention" occurs in 30 ham emails and in 10 spam emails. Assume that a new email arrives with binary coding $< 1,0,1 >$. Compute the likelihood that this email is from the spam class. in other words compute $P(< 1,0,1 > |Spam)$.
(c) How is this new email with coding $< 1,0,1 >$ classified; *Ham* or *Spam*, if one uses a Naive Bayes approach with no smoothing?

Antwoord op 2.

1. $< 1,0,1,0,1,0 >$
2. $P(< 1,0,1 > |Spam) = \frac{10}{30}\frac{8}{30}\frac{20}{30} = \frac{16}{270} = 0.0593$
3. $P(< 1,0,1 > |Ham) = \frac{12}{490} = 0.0245.$
4. It is easily computed that $P(< 1,0,1 > |Ham)\frac{7}{10}/P(< 1,0,1 > |Spam)\frac{3}{10} = 0.964 < 1$. Hence email is classified as Spam.

☞ **3** *(20 points)* Consider a dataset with attributes (features) A and B, attribute A can have values a_1 or a_2 and attribute B can have values b_1 and b_2. The class label is given by P (from positive) or N (from negative).

Ex.	A	B	*Class*
1	a_1	b_2	P
2	a_1	b_1	P
3	a_1	b_1	P
4	a_1	b_2	P
5	a_1	b_1	P
6	a_2	b_2	N
7	a_2	b_2	N
8	a_2	b_2	N
9	a_1	b_1	N
10	a_1	b_2	N

A data analyst wants to construct a decision tree from this data set using information gain.

(a) What is the information gain of attribute A?
A table with values for $-p\log_2(p)$ can be found at the end of this test.
(b) What is the information gain of attribute B?
A table with values for $-p\log_2(p)$ can be found at the end of this test.
(c) Which attribute will be at the root (top) of the decision tree? Explain your answer.
(d) Construct the complete decision tree.

Antwoord op 3.

1. Initial entropy is 1. For $A = a_1$ we have 5 P and 2 N, for $A = a_2$ we have 3 N. Hence entropy for $A = a_2$ is 0 and the average entropy after splitting on A is $\frac{7}{10}[-\frac{5}{7}\log_2(\frac{5}{7}) - \frac{2}{7}\log_2(\frac{2}{7})] = \frac{7}{10} * 0.86 = 0.602$. Hence information gain is 0.398
2. Average entropy after splitting on attribute B is $\frac{4}{10}[-\frac{3}{4}\log_2(-\frac{3}{4}) - -\frac{1}{4}\log_2(-\frac{1}{4})] + \frac{6}{10}[-\frac{2}{6}\log_2(\frac{2}{6}) - \frac{4}{6}\log_2(\frac{4}{6}]$ which equals $0.4[0.31+0.50]+0.6[0.53+0.39] = 0.876$ So the gain is 0.124.
3. Attribute A will be the top node of the DT.

☞ **4** *(20 points)* A certain classifier was tested on a test, resulting in 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

(a) What is the accuracy of this classifier?
(b) What is the recall of this classifier for class C_2?
(c) What is the precision of this classifier for class C_3?

Antwoord op 4.

1. Accuracy: $[120+150+130]/[120+150+130+15+20+10+16+22+3] = 0.82$
2. Recall C_2: $150/[16+150+10] = 0.85$
3. Precision C_3: $130/[130+10+20] = 0.81$

☞ **5** *(20 points)* Consider the training a linear classifier. Assume that the current linear classifier is given by the line $3-2x_1+2x_2 = 0$. The next feature point in our training set is given by $x = (-2,2)$.

(a) How will the feature point x be classified, given the current weights $w = (3,-2,2)$ of the linear classifier, 0 or 1?
(b) Assume that the feature point x is misclassified How will the weights $w = (3,-2,2)$ of the linear classifier be adapted. Assume a learning rate α of 0.3.

(c) How will x be classified after the above adaptation of the weight vectors w? Is this adaptation a step in the right direction? **Motivate your answer!**

Antwoord op 5.

1. $3+2*2+2*2=11$ hence point is classified as 1.
2. Adaptation: $(3,-2,2)-0.3*(1,-2,2)=(2.7,-1.4,1.4)$
3. $2.7+1.4*2+1.4*2=8.3$ Point is till not classified correct but $8.3<11$ hence it is a step in the right direction.

Table for $-p\log_2(p)$

p	$-p\log_2(p)$	p	$-p\log_2(p)$
0	0	1/6	0.43
1	0	2/6	0.53
1/2	0.50	3/6	0.50
1/3	0.53	4/6	0.39
2/3	0.39	5/6	0.22
1/4	0.50	1/7	0.40
2/4	0.50	2/7	0.51
3/4	0.31	3/7	0.52
1/5	0.46	4/7	0.46
2/5	0.53	5/7	0.35
3/5	0.44	6/7	0.19
4/5	0.26		