Test Pearl 110 — Intelligent Interaction

Pearls of Computer Science (201300070)
October 19 2018,
Module coordinator: Doina Bucur
Instructors: Christin Seifert and Gwenn Englebienne

- You may use 1 A4 sheet with your own notes for this test, as well as a *simple* calculator
- Scientific or graphical calculators, laptops, mobile phones, books etc. are not allowed. ***Put those in your bag now (with the sound switched off)!***
- Total number of points: 100

20 points **Question 1**

A bag contains 10 fair dice: of which two are 4 sided dice, one is a 6 sided dice, 2 are 8-sided dice, 4 are 12-sided dice and 1 is a 20-sided dice. All the dice are fair, so that for example the probability of throwing a 3 with a 20 sided dice is $1/20$. John randomly draws a dice from the bag and then throws this dice. The outcome is 5. This is the data or observation D.

	H	$P(H)$	$P(D\|H)$	$P(D\|H).P(H)$	$P(H\|D)$
2x	4				
1x	6				
2x	8				
4x	12				
1x	20				

10

(a) Copy the above table to your answer form and fill in the table with the correct entries (all empty entries should be filled in). In the table H stands for the possible dice (hypothesis), $P(H)$ is the prior probability, $P(D|H)$ the likelihood, and $P(H|D)$, the normalized posterior of H given the data D.

(b) What is the most likely value of H when we know D (i.e. that the outcome is 5)?

25 points **Question 2**

Consider the following piece of email, denoted by e:

```
we have display boxes with credit applications that we need to place in the
small owner-operated stores in your area . here is what you do : 1 . introduce
yourself to the store owner or manager . 2 . use our 90 % effective script
which tells them how this little display box will save their customers hundreds
of dollars , be a drawing card for their business
```

and a vocabulary V given by $V =$ (business,display,lifetime,save,text).

(a) How will this piece of text e be coded using a bag of words coding (with multinomial document-wise distributions)?

Now assume that one has the following database with spam and ham emails, only described by the words in V:

Id	words in V	class
1	business, lifetime, save, lifetime, save, text	spam
2	business, display, display, save, save, text	spam
3	business, text, text, save, save	ham
4	display, display, lifetime, text	ham
5	lifetime, save, save, text	ham
6	business, display, text	ham

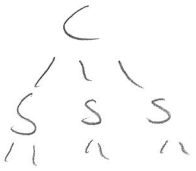
23245

(b) Compute the likelihood that the above email is generated by the spam class, i.e. compute $P(e|spam)$. Assume that no smoothing is applied and one uses a multinomial bag of words coding.
(c) Also compute $P(e|ham)$
(d) How will this email e be classified, *spam* or *ham*? **Clearly explain your answer.**

25 points **Question 3**

Consider the following simple data set about cars with the attributes (features) C (color) and S (speed), with C taking the values b (black), w (white) and g (green) and P (price) has values f (fast) and *slow* (slow). The class label is given in the last column and is y (yes, buy) or n (no, don't buy).

Id	C	S	*Class label*
1	b	f	y
2	w	f	y
3	g	f	y
4	b	s	y
5	w	f	y
6	g	s	y
7	b	s	n
8	w	s	n
9	g	s	n
10	b	s	n
11	w	f	y
12	g	f	y
13	w	f	n
14	w	f	y
15	w	f	y



(a) What is the information gain for attribute C and for attribute S?
A table with values for $-p\log_2(p)$ can be found at the end of this exam.
(b) Which attribute will be at the top of the decision tree and why?
(c) Compute and draw the complete decision tree.
(d) According to this decision tree, should you buy fast, white cars? Why? Should you buy slow, green cars? Why?

15 points **Question 4**

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	10	5	12
	C_2	9	11	1
	C_3	2	0	6

(a) What is the accuracy of this classifier? Round to two digits after the decimal point.
(b) If C_3 is not relevant any more for the task at hand, what is the accuracy of this classifier, that now has only to distinguish classes C_1 and C_2? Round to two digits after the decimal point.
(c) For which class does the classifier have the highest precision? Calculate all precision values and also note down your final decision. Round to two digits after the decimal point.
(d) For which class does the classifier have the highest recall? Calculate all recall values and also note down your final decision. Round to two digits after the decimal point.
(e) Which of the two classes does the classifier confuse most (and in which direction)?

15 points **Question 5**

Assume that we are training a linear classifier and that the current linear classifier is given by the line $-2-2x_1+2x_2=0$. The next feature point in our training set is given by $x=(-1,-1)$.

(a) How will the feature point x be classified, given the current weights $w=(-2,-2,2)$ of the linear classifier, 0 or 1?

(b) Assume that the feature point x is misclassified. How will the weights of the linear classifier be adapted? Assume a learning rate α of 0.2 and only one iteration of adapting the weights.

(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!**

Table for $-p\log_2(p)$

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				