COMP 7745/8745: Midterm Exam

March 3, 2016

Allocated Time: 1 hour and 25 minutes

96/100

NAME: SAYMA AKTHER

STUDENT-ID:

- **Open Books and Open notes**
- **Please write your answers directly below the questions in the space provided. If you run out of space, you can use the back side of the exam.**
- **State any assumptions if you need to make them.**
- **Best of luck!**

Maximum Points: 100

Number of pages: 11

1) State whether true or false (16 points)

a. Given a dataset with no noisy examples (i.e., it is never the case that for 2 examples, the attribute values match but the class value does not), the training error for decision trees is always equal to 0.

True

b. Given a linearly separable dataset, the leave-one-out cross-validation error for logistic regression is always equal to 0.

False

c. Given a dataset with no noisy examples, Naive Bayes is guaranteed to have a training error equal to 0.

False

d. Perceptrons and the logistic regression classifier both have the same decision boundary

True

e. As we increase the number of examples, we can approximate MAP learning as Max-Likelihood learning since they tend to be similar to each other.

True

f. As we increase the number of hidden layers in the neural network, the leave-one-out cross validation error is guaranteed to decrease.

False

g. The ID3 algorithm has an inductive bias that it favors small-depth trees over larger-depth trees

True

h. $(\neg X \vee Y) \wedge (\neg Y \vee X)$ can be represented by a linear unit perceptron

False

2) **Naive Bayes Classification:** Following is a dataset on burglaries

Alarm	Dog	Large-House	Burglary
0	1	0	0
1	0	0	0
0	0	0	1
1	1	1	0
0	1	1	1

a) **Predict the value of Burglary given an instance (Alarm=0; Dog=0; Large-House = 1) (11 points)**

$P(B=0) = 3/5 \qquad P(B=1) = 2/5$

$A=0,\ D=0,\ LH=1$

$P(B=0|X) \propto P(X|B=0)\cdot P(B=0)$

$\propto P(A=0, D=0, LH=1 | B=0)\cdot P(B=0)$

$\cong P(A=0|B=0)P(D=0|B=0)\ P(LH=1|B=0$

$P(B=0$

$\cong 1/3 \times 1/3 \times 1/3 \times 3/5 \cong \frac{1}{45}$

$P(B=1|X) \propto P(X|B=1)\cdot P(B=1)$

$\propto P(A=0, D=0, LH=1|B=1)\cdot P(B=1)$

$\cong P(A=0|B=1)\cdot P(D=0|B=1)\cdot P(LH=1|B$

$\cong 2/2 \times 1/2 \times 1/2 \times 2/5 \cong \frac{1}{10}$

$P(B=0|X) < P(B=1|X)$

$\therefore$ predict is $\boxed{1}$

2) Naive Bayes Classification: Following is a dataset on burglaries

Alarm	Dog	Large-House	Burglary
0	1	0	0
1	0	0	0
0	0	0	1
1	1	1	0
0	1	1	1

a) Predict the value of Burglary given an instance (Alarm=0; Dog=0; Large-House = 1) (11 points)

$P(B=0) = 3/5 \qquad P(B=1) = 2/5$

$A=0,\ D=0,\ LH=1$

$P(B=0|X) \propto P(X|B=0)\cdot P(B=0)$

$\propto P(A=0, D=0, LH=1|B=0)\cdot P(B=0)$

$\cong P(A=0|B=0)\,P(D=0|B=0)\,P(LH=1|B=0)\cdot P(B=0)$

$\cong 1/3 \times 1/3 \times 1/3 \times 3/5 \cong \frac{1}{45}$

$P(B=1|X) \propto P(X|B=1)\cdot P(B=1)$

$\propto P(A=0, D=0, LH=1|B=1)\cdot P(B=1)$

$\cong P(A=0|B=1)\cdot P(D=0|B=1)\cdot P(LH=1|B=1)\cdot P(B=1)$

$\cong 2/2 \times 1/2 \times 1/2 \times 2/5 \cong \frac{1}{10}$

$P(B=0|X) < P(B=1|X)$

$\therefore$ predict is $\boxed{1}$

b) The county records office told you that there are 100 more records, but since part of the records were not available, they only told you that of the 100 records, 80 of them were records with burglary = 0 and 20 of them were records with burglary = 1. Using this information, what is your new prediction for (Alarm=0; Dog=0; Large-House = 1)? (6 points)

$3 = 0 \quad 2 = 1$
$80 = 0 \quad 20 = 1$

$P(B=0) = \frac{83}{105} \quad P(B=1) = \frac{22}{105}$

$P(B=0|X) \propto P(A=0|B=0)\,P(D=0|B=0)\,P(LH=1|B=0) \cdot P(B=0)$

$\simeq \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{83}{105}$

$\simeq \frac{83}{2835} \simeq 0.02927$

$P(B=1|X) \propto P(A=0|B=1)\,P(D=0|B=1)\,P(LH=1|B=1) \cdot P(B=1)$

$\simeq \frac{2}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{22}{105}$

$\simeq \frac{22}{420} \simeq 0.05238$

$P(B=0|X) < P(B=1|X)$

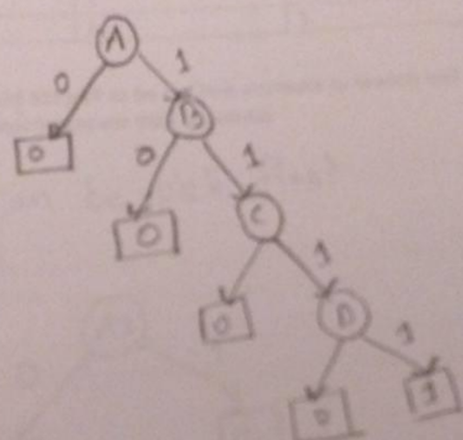
$\therefore$ prediction is $\boxed{1}$

3) Decision Trees

a) Represent the following Boolean function using a decision tree (Hint: You should be able to do this without using a brute force strategy of enumerating each of the 16 possibilities) (11 points)

A ^ B ^ C ^ D

6

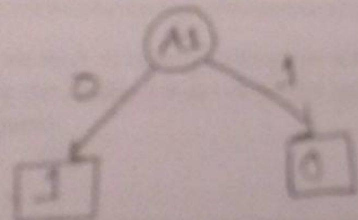


b) Consider the following dataset.

A1	A2	A3	Class
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	0
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1

Which one would ID3 pick to be the first attribute to branch on? (Note you should not have to use a calculator to answer this) (6 points)

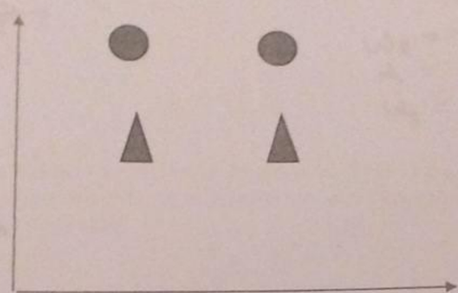
$\boxed{A1}$ can be pick first



4) Perceptrons and Neural Networks

a)

- Consider the following dataset where we want to classify circles from triangles.



Assume in all the following cases that we have an arbitrarily small learning rate and we ~~are allowed to~~ make infinitely many iterations through the examples (12 points)

i) Suppose we wish to learn a perceptron with linear activation function, 2 weights and no bias, using the gradient descent weight update rule ~~for~~, will we converge such that we get 0 training error?

NO

ii) Suppose we wish to learn a perceptron with linear activation function, 2 weights and a bias, using the gradient descent weight update rule, will we converge such that we get 0 training error?

Yes

iii) Suppose we wish to learn a perceptron with threshold function, 2 weights and no bias, using the perceptron weight update rule, will we converge such that we get 0 training error?

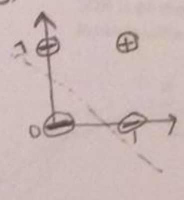
NO

iv) Suppose we wish to learn a perceptron with threshold function, 2 weights and ~~no~~ bias, using the perceptron weight update rule ~~for~~ will we converge such that we get 0 training error?

Yes

c) Can we learn the function $X1 \wedge X2$ using a perceptron? What should we set the weights of the perceptron such that all instances of this function are classified correctly (8 points)

Yes,



$$\begin{cases} \text{if } w_0 + w_1 A + w_2 B > 0 & 1 \\ \text{else} & -1 \end{cases}$$

A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

$w_0 = -0.6$
$w_1 = 0.5$
$w_2 = 0.5$

5) Model Selection

a)

Somebody submits a research paper for your review with the following experiment.

Learn a logistic regression based classification model on the full dataset. Let us denote the model as LR. Learn a Naïve Bayes classification model on the full dataset. Let us denote the model as NB. They then define a new classifier as follows: classify each data instance as a +ve instance if both NB and LR label the data instance as +ve, else classify the instance as –ve. They report leave-one-out cross-validation error as their generalization error.

Do you think the authors have a valid experiment methodology and should their results be accepted? Justify your answer in 1 or 2 sentences. (6 points)

Yes, Naive Bayes has training and test data. Its Error generation and Logistic regression in leave-one-out cross validation generation may be same.

4

b) Let us consider a naïve learning algorithm which simply functions like a lookup table. That is, given an example to classify, it labels the example as +ve and if that exact example is present in the training set, else, it labels it as -ve. (12 points)

a. Given a dataset with no duplicate entries and 50% of the dataset is of the +ve class and 50% is of the –ve class, what is the leave-one-out cross validation error. (error = #misclassified/total)

$\frac{1}{2}$

b. Given a dataset with no duplicate entries and 1% of the dataset is of the +ve class and 99% is of the –ve class, what is leave-one-out cross validation error. (error = #misclassified/total)

$\frac{1}{100}$

c. Given a dataset with no duplicate entries and 50% of the dataset is of the +ve class and 50% is of the –ve class, what is the precision and recall w.r.t the +ve class.

precision = 0 Because all entries will be classified as (-) ve

recall = 0

d. Given a dataset with no duplicate entries and 1% of the dataset is of the +ve class and 99% is of the –ve class, what is the precision and recall w.r.t the +ve class.

precision = 0 Because all entrie will be classi as (-)ve.

recall = 0

6) General Question

a) Your boss gives you a dataset that consists of 2 attributes (X1, X2) which are both real-valued. Your task is to learn a concept: $a1 \leq X1 < b1;\ a2 \leq X2 < b2$, where a1,a2,b1 and b2 are all real-valued parameters that you learn from the dataset. What learning algorithm among the following would you select as the best one for this task? Give a 1 line justification for your choice. (6 points)

   a. Decision Trees
   b. Naïve Bayes
   c. Neural Networks

I think Neural Networks would be the best one since Naive Bayes and Decision trees are good with discrete but Neural Network is good with both discrete and continuous.

b) Happy with your performance in the previous question, your boss gives you a new dataset with several duplicate entries. That is, there are several examples that are identical to each other. Your boss asks you to apply Naïve Bayes and verifies whether you would need to pre-process the data by removing duplicates. Will the Naïve Bayes classifier that we learn before and after we remove all duplicates in the dataset be identical to each other? Give a 1 line justification. (6 points)

Since Naive Bayes uses count. So Naive Bayes can count the changes. Here probability changes, classifier changes.
So, may not be identical to each other.