DALHOUSIE UNIVERSITY
Inspiring Minds

CSC6515: Machine Learning for Big Data
Final Examination Fall 2015
December 16, 2015

Instructor: Dr. Stan Matwin

Closed Text Exam; Time: 2.5 hrs. Total points: 100

Your answers will be on the question sheet AS WELL AS in the exam booklet provided. Pleae hand in BOTH.

Use one booklet for rough work, and the other for proper answers.
Good luck, and have a nice Christmas!

Name (PRINT):

B #

Question	Mark	Max. mark	Question	Mark	Max. mark
1a		2	6		6
1b		2	7		5
1c		2	8		7
2a		7	9		12
2b		4	10		5
3		4	11		5
4		8	12		9
5		5	13		9
			14		8
TOTAL=		34			66

1. [Big Data]
   a. How many bytes is an Exabyte?

   b. How many petabytes are there in a exabyte?

   c. how many petabytes are there in a yottabyte?

2. [Decision trees] You are given a dataset S concerning opinions of moviegoers. Each instance describes a movie seen by a person, and the class attribute is the Likes/Not. Attributes are:

   Age: Young, Old
   Sex: Male, Female
   Violence in the movie : Yes/No

Age	Sex	Violence	Drama	Likes/Not
Y	M	N	Y	N
Y	F	Y	N	L
Y	M	N	Y	N
Y	F	Y	N	L
Y	M	Y	N	L
O	F	Y	N	L
O	F	N	Y	N
O	M	Y	N	L
O	M	Y	N	L
O	F	N	N	N

a) which attribute will be chosen as the root by the Information Gain criterion?
b) Which is the second best attribute?

3. [Ensembles, Random Forest]
   In the Random Forest algorithm,

a. Attributes are sampled
b. Instances are sampled
c. Both attributes and instances are sampled
d. Attributes and instances are subject to bagging

4. [Naïve Bayes] Suppose the following new instance of the data from Question 1 is to be classified by an Naïve Bayes classifier:

Age	Sex	Violence	Drama	Likes/Not
O	M	N	N	?

5.

The NB formula is $v_{NB} = \text{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

Show the calculations and the predicted class. If one of the probabilities is computed from the data as =0, use a simple form of Laplace smoothing: add 1 to the numerator and to the denominator of the fraction representing this conditional probability.

5. [Learning tasks] Researchers wanted to investigate whether there is a relationship between a subject's number of years of education and his or her driving record, recorded as total tickets in a lifetime. In other words, they wanted to predict the number of tickets a person will get in their life. Which method would you use? (Circle all that apply)

a) Linear regression
b) Logistic regression
c) Perceptron
d) Locally weighted regression
e) SVM with linear kernel

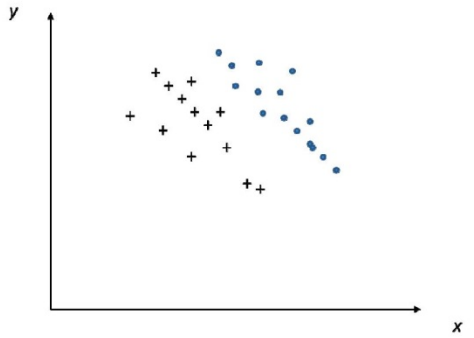
6. [Linear models] Suppose we have a dataset with only one input variable. So, the training examples are of the form (x, y) where x is scalar. We want to apply regression on the data. Consider the following two scenarios

$$A: \quad y = \theta_1 x + \theta_0$$
$$B: \quad y = \theta_1 x^2 + \theta_0$$

Which of the following is correct? (choose the answer that best describes the outcome)
a) There are datasets for which A would perform better than B.
b) There are datasets for which B would perform better than A.
c) Both (a) and (b) are correct.
d) Since x^2 can be directly obtained from x, they would perform equally well on all datasets.

7. [Feature selection/modification] Assume the following distribution of instances of two classes in 2-attribute space *x*,*y*:

Draw how the new attributes *x'*, *y'* that PCA would produce for this distribution.



8. [theory] VC dimensions is related to the capacity of learning machines. Is it good to have a very large or unbounded VC dimension (i.e. infinite)? Why? [20 words max]

9. [theory] Assume two data sets sampled from the same distribution where the number of observations for each data is 5,000 and 100,000 respectively. The train and test set is randomly constructed by dividing the data 80%-20%.

   With y-axis denoting the error and x-axis denoting the model complexity, draw two curves for training error and test error for each data set. You should have total of 4 curves: one training error and one test error curve for each of train and set dataset.

- Draw all 4 of them in the same diagram.
- Clearly mark all your curves. Use 5K train, 5k test, 100k train, 100k test as legends for your curves.

10. [Cloud/Hadoop] Please circle among the Spark functions below those that are <u>actions</u>;

map	first	collect
filter	groupBy	union
count	reduceByKey	take

11. [Cloud/Hadoop] For the given code below please write down the value of x and y.

```
d = sc.parallelize(["It was a bright cold day in April, and the clocks were striking thirteen."])
mappedRDD      = d.map(lambda x: x.split(" "))
flatMappedRDD  = d.flatMap(lambda x: x.split(" "))
x = len(mappedRDD.collect())
y = len(flatMappedRDD.collect())
```

12. [Clustering]. : What is the main difference between k-Medoids and k-Means? When is the use of the latter preferred over the use of the former? What is the main computational difference? [30 words max]

13. [Clustering] what are the main DB-SCAN parameters? What is the main advantage of DBSCAN compared to k-means? How does DBSCAN classify points as noise. [30 words max]

14.[LDA] In the following plate diagram of the LDA, explain the meaning of **all** the symbols, i.e.

η =
β_k =
....