University of Texas at Arlington
Computer Science and Engineering

CSE5334– Final Exam
Data Mining

Instructor: Prof. Won Hwa Kim

2020/5/6

Name: ____

Student Number: ____________________________________

Distribution of Marks

Question	Points	Score
1	30	
2	12	
3	6	
4	20	
5	10	
6	10	
7	12	
Total:	100	

1. (True or False) Identify if the following statements are True or False.

   (a) (2 points) Minkowski distance is a generalization of Euclidean distance.

   (b) (2 points) Insufficient data samples or noise in the data cause overfitting.

   (c) (2 points) The sum of a cumulative distribution function (cdf) has to be equal to 1.

   (d) (2 points) From a joint probability $p(x, y)$, the marginal distribution of x is defined as $\sum_x p(x, y)$.

   (e) (2 points) If x and y are independent, $p(x, y) = p(x)p(y)$.

   (f) (2 points) k-means is a clustering algorithm for supervised learning.

   (g) (2 points) Binomial distribution models the number of successes in a sequence of n independent experiments.

   (h) (2 points) Naive Bayes and Logistic regression have the same hypothesis space bias.

   (i) (2 points) Naive Bayes is expected to outperform Logistic regression when given a large number of training samples.

   (j) (2 points) Finding the optimal solution on the training dataset will always yield the optimal solution for testing dataset.

   (k) (2 points) Decision tree cannot learn decision boundaries that are not orthogonal to an axis in its feature space.

   (l) (2 points) Principle component analysis (PCA) performs dimension reduction by selecting the most important attributes.

   (m) (2 points) Principle components correspond to the eigenvectors of a covariance matrix with the largest eigenvalues.

   (n) (2 points) Logistic regression is a designed for regression tasks.

   (o) (2 points) A Neural Network can only learn a linear classifier.

2. Explain the following concepts.

   (a) (3 points) Problem with Boolean search: Feast vs. Famine.

   (b) (3 points) Overfitting.

   (c) (3 points) Autoencoder.

   (d) (3 points) Sigmoid function vs. Rectified Linear Unit (ReLU)

3. Compute Jaccard coefficient between the following sets or sentences. Show your work.

   (a) (3 points) Batman and Robin are taking these courses.

   - Batman: {CSE5334, CSE6363, CSE1325, CSE4334}
   - Robin: {CSE4334, CSE5335, CSE5334, CSE6363, CSE1325}

   (b) (3 points) Compute Jaccard coefficient after processing stop words ‘from’, ‘of’, ‘at’, ’-’.

   - S1: Dr Kim graduated from University of Wisconsin - Madison
   - S2: Prof Kim now works at University of Texas at Arlington

4. The probability distribution function over x (Gaussian distribution) is given as

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

   (a) (5 points) Given i.i.d samples $x = \{x_1, x_2, \cdots x_N\}$ from a Gaussian distribution, construct a likelihood function. Show your work.

   (b) (5 points) What is the log-likelihood function of the likelihood function from (a)? Show your work.

(c) (10 points) Compute the maximum likelihood estimator (i.e., μ_{mle} and σ^2_{mle}) of the i.i.d samples $x = \{x_1, x_2, \cdots x_N\}$ from the Gaussian distribution. Show your work.

5. Consider Naive Bayes classifier for a binary classification task.

   (a) (5 points) It is inference is made using Bayes rule as

$$p(y=1|x_1,\cdots,x_n) = \frac{p(y=1)\Pi_{i=1}^{n} p(x_i|y=1)}{p(x_1,\cdots,x_n)}.$$

   Reformulate the equation above to the formulation of prediction in logistic regression given as

$$f(x) = \frac{1}{1+e^{-w_0+\sum_{i=1}^{n} w_i x_i}}.$$

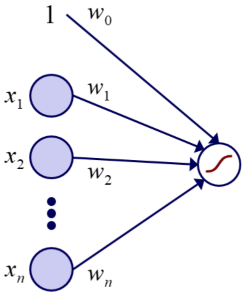
   (b) (5 points) Explain how these two different methods are similar and different.

6. Assume your probabilistic learner produces the following test-set result:

Index	Probability	True Label
1	0.99	neg
2	0.80	pos
3	0.4	pos
4	0.12	neg
5	0.1	neg

(a) (10 points) Draw a Precision-Recall (PR) curve for this result.Show confusion matrix for each threshold and be sure to label your axes. ("Connecting the dots" with straight lines is not correct, but it is ok to do so in this exam.)

7. Consider a single layer neural network with one output unit and no hidden units for a binary classification task. You are trying to minimize cross-entropy $E(w) = \sum_i -y_i ln(o_i) - (1-y_i) ln(1-o_i)$ where i is the data index, o is the output from the perceptron and y is the true label. Sigmoid function $\sigma(z) = \frac{1}{1+e^{(-z)}}$ is used as the activation function.



(a) (7 points) Suppose you are performing online training using gradient descent, what would be the gradient to update your model for each iteration? Show your work. (Hint: $\frac{\partial\sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$)

(b) (5 points) Given an initial $w = [0.5, 1, 0.5]$ (including a bias term), perform 1 iteration of backpropagation with a training data $x = (-2, 2)$ with a label $y = 1$ and a learning rate of $\eta = 0.01$.

This page is intentionally left blank to accommodate work that wouldn't fit elsewhere and/or scratch work.