University of Texas at Arlington
Computer Science and Engineering

CSE5334– QUIZ #2
Data Mining

Instructor: Prof. Won Hwa Kim

Name: _____
Student Number: _____

Distribution of Marks

Question	Points	Score
1	30	
2	10	
3	15	
4	20	
5	25	
Total:	100	

1. (True or False) Identify if the following statments are True or False.
   (a) (3 points) A probability distribution function (pdf) $p(x)$ is always positive.
   (b) (3 points) The sum of a pdf is always equal to 1.
   (c) (3 points) From a joint probability $p(x, y)$, the marginal distribution of x is defined as $\sum_x p(x, y)$.
   (d) (3 points) $p(x, y) = p(x)p(y)$.
   (e) (3 points) k-means is a classification algorithm for supervised learning.
   (f) (3 points) Binomial distribution models the number of successes x in a sequence of n dependent experiments.
   (g) (3 points) Decision tree branches out based on attributes/features of a dataset.
   (h) (3 points) Insufficient data points or too simple model can cause overfitting.
   (i) (3 points) If there are two prediction models that return the same result, then the more complex one is better than the simpler one.
   (j) (3 points) Decision tree cannot learn oblique decision boundaries.

2. Bayes Theorem. Somewhere, 51% of the adults are males. (It doesn't take too much advanced mathematics to deduce that the other 49% are females.) One adult is randomly selected for a survey.
   (a) (2 points) What is the prior probability that the selected person is a male? What is the prior that the person is a female?

   (b) (8 points) It is later learned that the survey was asking whether you smoke a cigar or not. Based on a prior investigation, it is known that 10% of males and 2% of females smoke cigars. Use this additional information to find the probability that the selected adult is a male. Show your work.

3. For data arriving at three different nodes in a decision tree, the class labels are given as the tables below.

class 0	2
class 1	98

class 0	57
class 1	43

class 0	75
class 1	25

(a) (7 points) Calculate GINI Index for each table.

(b) (8 points) Compute Entropy for each table.

4. You are given with a training dataset as below:

index	Refund	Martial Status	Taxable Income	Cheat
1	yes	single	125k	no
2	no	married	100k	no
3	no	single	70k	no
4	yes	married	120k	no
5	no	divorced	95k	yes
6	no	married	60k	no
7	yes	divorced	220k	no
8	no	single	85k	yes
9	no	married	75k	no
10	no	single	90k	yes

(a) (15 points) Construct a decision tree that classifies 'Cheating on Tax' status by considering the attributes in the following order: Martial status (married or not), Refund and Taxble income (greater than 80k or not).

(b) (5 points) A testing object comes in with the following attributes. What would the decision be using the decision tree from (a)?

index	Refund	Martial Status	Taxable Income	Cheat
1	yes	divorced	95k	?
2	no	single	70k	?

5. The probability distribution function of a Bernoulli distribution with a parameter μ as the probability of $x = 1$ is given as

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{1}$$

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

(b) (10 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}) of the i.i.d samples $\mathbf{x} = \{x_1, x_2, \cdots x_N\}$. Show your work.

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