DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment 3

Classification

ENGG 5103: Techniques for data mining Deadline: 5:00 pm, Dec 11th, 2016

Submission: On CUHK e-learning

Name: ___________________________

SID: ___________________________

1. **(Bayesian method, 15%)** Suppose we have a new inspection method for a certain kind of disease. For a patient infected with the disease, the new method will report positive result with probability 0.99. While, for a normal person, the probability of positive result is 0.01. Assume the incidence of the disease is 0.001, please answer the following questions:

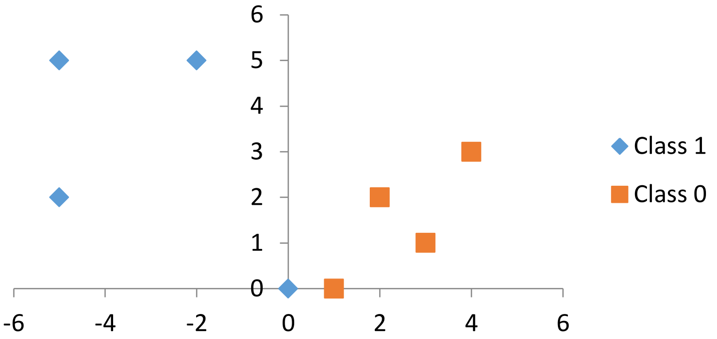
    a. Calculate the probability of reporting positive result.
    b. Use Bayes' Theorem to calculate the probability of being infected with the disease given that the inspection result is positive.

2. **(Neural networks, 35%)** Given the following 2D data points belonging to 2 different classes (shown in Table 1 and Figure 1), please complete the following tasks:

Table 1

X Axis	Y Axis	Class
-5	5	1
-5	2	1
-2	5	1
0	0	1
1	0	0
4	3	0
2	2	0
3	1	0

Figure 1



a. Use **Minimum Distance Classifier** (MDC) to classify these data points.

i. Calculate centroid points m1 and m0 for both class 1 and class 0.

ii. Write down the equation of decision boundary and plot it on the picture with all the data points.

iii. Given data points [0, 1], which class dose it belong to?

b. Perform **Single Layer Perceptron** (SLP) to classify these data points.

i. Let $net = xw_1 + yw_2 + b$ and $class\_label = f(net)$

$$f(net) = \begin{cases} 1, & if\ net\ \geq 0 \\ 0, & otherwise \end{cases}$$

Denote target class as t, predicted class as p, learning rate as α, please write down the formulas for updating the intercept b and the weights w_1 and w_2

ii. Let the initial values of the parameters be $w_1 = -2$, $w_2 = 1$ and $b = -1$. Set the learning rate α = 0.5, please complete the following update table for the parameters:

Parameter Update Table for SLP

X Axis	Y Axis	Target Class	Predicted Class	w_1	w_2	b	net
-5	5	1	1	-2	1	-1	14
-5	2	1					
-2	5	1					
0	0	1					
1	0	0					
4	3	0					
2	2	0					
3	1	0					

3. **(Nearest neighbor classifier, 20%)** Suppose we have the following 2D data points belonging to class C1 and C2 (Table 2)

Table 2

X Axis	Y Axis	Class
1	1	C1
1	2	C1
1	0	C1
2	0	C1
0.5	0.5	C1
0	0	C2
4	0	C2
1	3	C2
2	2	C2
3	1	C2

a. Sketch the decision boundary of the **1-nearest neighbor classifier** between class C1 and C2 on a 2D graph with both axes ranging from -5 to 5.

b. Predict class label of data points (-3, 0) and (0, 4) using **3-nearest neighbor classifier**.

4. **(Bayesian classifier, 30%)** The following dataset stores the information for various kinds of mushrooms. Your task is to learn a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odor.

Table 3

Shape	Color	Odor	Edible
C	B	1	Yes
C	W	1	Yes
C	W	1	Yes
D	W	2	Yes
D	W	2	Yes
D	B	3	No
C	B	3	No
C	W	3	No
D	B	2	No
D	B	2	No

a. What's the conditional entropy H(Color | Shape = D)?
b. Which attribute would the ID3 algorithm choose to serve as the root of the decision tree (no pruning)?
c. Draw the full decision tree learned for this data by ID3 (no pruning).
d. Suppose we have the following validation set in Table 4. What's the training set error and validation set error of this unpruned tree? Express the error as the number of misclassified samples. (For example, if there is one misclassified sample in the training set, the error for the training set would be 1.)

Table 4

Shape	Color	Odor	Edible
D	B	3	No
C	W	2	No
C	B	3	Yes