Hajee Mohammad Danesh Science and Technology University, Dinajpur
Department of Computer Science and Engineering
B. Sc. (Engineering) in Computer Science and Engineering
Special Repeat Examination 2018
Level 4 Semester II, Credit: 3.0
Course Code: CSE 455, Course Title: Pattern Recognition

Time: 03 Hours

Total Marks: 90

[N.B. The figure in the right margin indicates the marks allocated for respective question. Split answer of any question will not be accepted.]

SECTION-A

*(Answer any **three** from the following questions.)*

1. a) What are the primary difference between pattern recognition, machine learning, and data mining? Which tasks are important in each of these areas? 3+2

   b) Describe the basic steps that must be followed in order to develop a pattern recognition task. 5

   c) What are the necessities of dimensionality reduction? Briefly describe the strategy of attribute subset selection used for dimensionality reduction. 2+3

2. a) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): 3+2
      (i) Compute the Manhattan distance between the two objects.
      (ii) Compute the Minkowski distance between the two objects, using q=4.

   b) What is spatial data? For the vector **c**= (0,-1, 0, 1), and **y** = (1, 0, -1, 0), calculate the cosine and Jaccard similarity. 1+4

   c) A typical pattern recognition task involves the recognition of 2D shapes. For simplicity, consider a word of 4 classes: squares, triangles (each angle = 60^0), hexagon (each angle is 120^0) and circles. Propose a feature vector that can differentiate between any of the previous 4 shapes. Justify the appropriateness of it for the classification task. 3+2

3. a) How can sampling be used as data reduction technique? What are the key differences among simple random sample, cluster sample and stratified sample? 2+3

   b) Suppose a hospital tested the age and body fat data for the following randomly selected adults with results: 4×2

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

      (i) Calculate the mean, median, and standard deviation of % fat.
      (ii) Draw the boxplot for age.
      (iii) Normalize the two variables based on z-score normalization.
      (iv) Calculate the correlation coefficient. Are these two variables positively or negatively correlated?

   c) What does motivate the attempts to "Clean" the training set? 2

4. a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 5

   b) Discuss the performance of k-NN classifier when (i) k is very small, and (ii) k is large. 4+4

   c) Find the centroid, radius, and diameter for the following set of patterns: (1, 1), (1, 3), (1, 4), (2, 2), (2, 3). 2

SECTION-B

*(Answer any **three** from the following questions.)*

1. a) What are outliers? How might you determine outliers in the data? 2+3

   b) Both Nearest Neigbour Algorithm and Square Error Clustering Algorithm are partitional algorithm. Describe the main procedure of each of the algorithms. 5

   c) Suppose you have the following training examples described by three attributes, x1; x2; x3, and labeled by classes c1and c2. Using the 3-NN algorithm, decide whether the new pattern p = [4, 3, 3] should belong to c1 or c2. 5

x1	x2	x3	Class
2	0	3	c1
3	1	2	c1
2	1	3	c1
0	5	0	c2
1	4	0	c2

2. a) What is curse of dimensionality? Following classifier is generated for the same training set, which has 100 instances. It has the following confusion matrix. Calculate the (i) Precision, (ii) Accuracy, and (ii) error rate. 2+3

		Predicted class	
		+	-
Actual class	+	50	10
	-	10	30

   b) What is linear discriminate function? Explain how it can be used to find a linear classifier to discriminate between two classes? 2+3

   c) When do you need non-linear classifiers? How can SVM be used to handle this classifier? 2+3

3. a) What are inductive hypotheses and deductive hypotheses? How is non-separability of data handled in SVM learning? 2+3

   b) A major problem with the single link algorithm is that clusters consisting of long chains may be created. Describe and illustrate this concept. 5

   c) What is the gradient descent method? Suppose $f(x, y) = 3x^2 + 2y^3 - 2xy$. Find the minimum stationary points of this function. 2+3

4. a) Why native Bayesian classification is called "naive"? Briefly outline the major ideas of naive Bayesian classification. 2+3

   b) Write short note on any two of the following concepts: 5

      (i) ROC Curves
      (ii) Cross validation
      (iv) PCA
      (iv) Over-fitting the data

   c) Explain the generation of frequent item sets using APRIORI algorithm with suitable example. 5