Hajee Mohammad Danesh Science and Technology University, Dinajpur
B.Sc. (Engineering) in CSE
Semester Final Examination 2017 (Jul-Dec)
Level 4 Semester II, Course Code: CSE 455, Credit: 3
Course Title: Pattern Recognition

Time: 3 hours

Total Marks: 90

NB: Figures in the right margin indicate full marks. Parts of the same question should be answered together and in the same sequence.

Section-A

Answer any Three

marks

1. (a) What is pattern recognition? Explain the difference between statistical and structural approaches. 1+2

   (b) Describe briefly one the following pattern recognition applications. 6
   - i) Optical character recognition
   - ii) Face recognition

   (c) Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 6*1=6
   - i) What is the mean and median of the data?
   - ii) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
   - iii) What is the midrange of the data?
   - iv) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
   - v) Give the five-number summary of the data.
   - vi) Show a box plot of the data.

2. (a) Briefly outline how to compute the dissimilarity between objects described by the following: 4*1=4
   - i) Nominal attributes
   - ii) Asymmetric binary attributes
   - iii) Numeric attributes
   - iv) Term-frequency vectors.

   (b) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set: 5+3

	A1	A2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

   - i) Consider the data as 2-D data points. Given a new data point, $x = (1.4,1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.
   - ii) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

   (c) State some challenges in pattern and class learning. 3

3. (a) What is meant by dimensionality reduction and why is it important? Describe the strategy of discrete wavelet transform and Principal Components Analysis used for dimensionality reduction. 1+5

   (b) Describe the KNN algorithm. What is the performance of KNN classification when k is very small and k is large? 4+2

   (c) Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the intended use of the data with examples. 3

4. (a) Both Nearest Neighbor Algorithm and Square Error Clustering algorithm are partitional algorithm. Describe the main procedure of each of the algorithms. 5

(b) Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1 (2,10), A2 (2,5), A3 (8,4), B1 (5,8), B2 (7,5), B3 (6,4), C1 (1,2), C2 (4,9). The distance function is the Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only 2+7
(a) The three cluster centers after the first round of execution.
(b) The final three clusters

(c) What are inductive hypotheses and deductive hypotheses? 1

Section B

Answer any Three

marks

1. (a) What is the importance of data visualization techniques? Explain following visualization techniques: 1+2*3=7
   i) Geometric projection visualization techniques
   ii) Pixel-Oriented Visualization Techniques

(b) Use a flowchart to summarize the following procedures for attribute subset selection: 3
(i) stepwise forward selection
(ii) stepwise backward elimination
(iii) a combination of forward selection and backward elimination

(c) Following classifier is generated for the same training set, which has 100 instances. It has the following confusion matrices. Calculate the (i) FN rate (ii)TN rate (iii) Precision (iv) Accuracy, and (v) error rate 1*5=5

		Predicted class	
		+	-
Actual class	+	55	5
	-	5	35

2. (a) Describe a Support Vector Machine. Define the optimization task solved in SVM learning with an example. 2+7

(b) What are the challenges when clustering is applied to a real-world dataset? 3

(c) Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. Partition them into three bins by each of the following methods: (i) equal-frequency (equal-depth) partitioning (ii) equal-width partitioning (iii) clustering 3

3. (a) Why naïve Bayesian classification is called "naïve"? Briefly outline the major ideas of naïve Bayesian classification. 1+3

(b) Explain the following and discuss their significance in pattern recognition with suitable example 4*2=8
i) Mean and Covariance ii) Chi Square Test
iii)Outliers iv) Over fitting

(c) In the table shown below variables x and y represent height and weight of a student 3

Height (x)	160	161	162	163	164	165	166	167	168	189	170
Weight(y)	54	55	55	56	57	56	56	57	56	58	59

Draw a scatter diagram to show the data and find the correlation coefficient and describe the relation between two variables.

4. (a) With suitable example explain the generation of frequent item sets using Apriori algorithm 6

(b) Suppose that we have a database with 5000 transactions and a rule $L \rightarrow R$ with the following support counts: count $(L) = 3400$, count$(R) = 4000$ and count $(L \cup R) = 3000$ What are the values of support and confidence for this rule? 3

(c) Briefly describe and give examples of each of the following approaches to clustering methods: single link technique, MST single link technique, complete link technique, and average link technique. 6