Assignment 2

(COMP3605 - Introduction to Data Analytics, 2022-2023)

Date Available: Sunday, October 23, 2022
Due Date: 11.50 PM, Sunday, November 06, 2022
Total Mark: 100 marks

Answer ALL Questions

INSTRUCTIONS
1. Type or write your answers neatly.
2. Show all working of your answers.
3. Your solutions must be your own. You must not share your working or solutions with your peers.
4. You are not permitted to copy, summarize, or paraphrase the work of others in your solutions.
5. Submit your answers in a single zipped file named A2_ID.zip to the email comp3605@gmail.com, where ID is replaced with your student ID. The file A2_ID.zip contains
- a single PDF file containing all of your typed, handwritten, and screenshots answers.
- a signed and dated UWI Plagiarism Declaration indicating that the work submitted is your own.

Question 1
You are given the transactional data set *D* shown in the table below. The data set *D* has ten transactions. Let the minimum support (*minsup*) be 0.3.

The transactional data set *D*

TID	Items
T01	C, E, M
T02	A, C
T03	A, B, C, G, O
T04	B, E, O
T05	C, G, M
T06	A, C, E, O
T07	B, C, O
T08	C, E, G
T09	B, C, E, G
T10	B, C, G

a. [34 marks] Find all frequent itemsets in *D* using the horizontal Apriori algorithm.
b. [12 marks] Given *minsup* = 0.3, *minconf* = 0.75, show the detailed generation of strong association rules from the frequent 3-itemset $\ell = \{B, C, G\}$.

[**Total mark**: **46**]

PLEASE TURN TO THE NEXT PAGE

Question 2 The basic k-nearest neighbor algorithm is given below.
Algorithm Basic k-NN classification algorithm.
1. Let k be the number of nearest neighbors and D be the set of training examples.
2. **for** each test example $z = (\mathbf{x}', y')$ **do**
3. Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4. Select $D_z \subseteq D$, the set of k closest training examples to z.
5. $y' = \arg\max_v \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$, where $I(a, b) = 1$ if $a = b$ and 0 otherwise.
6. **end for**

Assume that the test instance $z = (\mathbf{x}', y')$ has k nearest neighbors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_k$ (e.g., k = 3, 5, 7, ...) and the class label of $\mathbf{x}_i$ is y_i. Let L be the set of class labels (i.e., $L = \{C_1, C_2\}$, where C_1 is the positive class denoted as + and C_2 is the negative class denoted as –).

• For the majority voting technique: Let $h(v)$ be the number of nearest neighbors with the class label v. That is, we have $h(v) = \sum_{i=1}^{k} I(y_i = v)$, where a class label $v \in L = \{C_1, C_2\}$.

For the class label $v = C_1$ (i.e., class +), we have $h(C_1) = I(y_1 = C_1) + I(y_2 = C_1) + ... + I(y_k = C_1)$ and for the class label $v = C_2$ (i.e., class –), we have $h(C_2) = I(y_1 = C_2) + I(y_2 = C_2) + ... + I(y_k = C_2)$. The class label of the test example z is specified as follows.

$$\text{Majority Voting: } y' = \arg\max_v h(v) = \arg\max_v h(v).$$

That is, if $h(C_1) > h(C_2)$, the class label of the test example z is $y' = C_1$ (i.e., class +). Otherwise, (i.e., $h(C_1) < h(C_2)$), the class label of the test example z is $y' = C_2$ (i.e., class –).

• For the distance-weighted voting technique: Let $f(v) = \sum_{i=1}^{k} w_i \times I(y_i = v)$, where $w_i = 1/d(\mathbf{x}', \mathbf{x}_i)^2$.

For the class label $v = C_1$ (i.e., class +), we have $f(C_1) = w_1 \times I(y_1 = C_1) + w_2 \times I(y_2 = C_1) + ... + w_k \times I(y_k = C_1)$ and for the class label $v = C_2$ (i.e., class –), we have $f(C_2) = w_1 \times I(y_1 = C_2) + w_2 \times I(y_2 = C_2) + ... + w_k \times I(y_k = C_2)$. The class label of the test example z is determined as follows.

$$\text{Distance-Weighted Voting: } y' = \arg\max_v f(v) = \arg\max_v f(v).$$

That is, if $f(C_1) > f(C_2)$, the class label of the test instance z is $y' = C_1$ (i.e., class +). Otherwise (i.e. $f(C_1) < f(C_2)$), the class label of the test instance z is $y' = C_2$ (i.e., class –).

You are given the one-dimensional data set D shown in the table below. The data set D has ten data points.

The one-dimensional data set D

i	1	2	3	4	5	6	7	8	9	10
$\mathbf{x}_i$	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y_i	–	–	+	+	+	–	–	+	–	–

a. [10 marks] Use the majority voting technique to classify the test example $z = 5.0$ using 9-NN (i.e., $k = 9$).

b. [10 marks] Use the distance-weighted voting technique to classify the test example $z = 5.0$ using 9-NN (i.e., $k = 9$).

[Total mark: 20]

PLEASE TURN TO THE NEXT PAGE

Question 3
You are given six two-dimensional points shown in the table below.

Point	*x* coordinate	*y* coordinate
p_1	0.1831	0.1085
p_2	0.9624	0.1916
p_3	0.0732	0.9594
p_4	0.2572	0.6066
p_5	0.4476	0.7871
p_6	0.2292	0.9489

a. [4 marks] Use the Euclidean distance to compute the distance matrix M for the six points.
b. [30 marks] Show the results of the **group-average linkage** version of the basic agglomerative hierarchical clustering algorithm. That is, for each iteration of the algorithm, you need to show the found closest two clusters and the updated distance matrix M.
The average distance between two clusters C_i and C_j is calculated by using the UPGMA (Unweighted Pair Group Method with Arithmetic mean) approach. That is, we have

$$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \| \boldsymbol{p} - \boldsymbol{p}' \|_2$$

where $\|\cdot\|_2$ is Euclidean distance (a.k.a. L_2-norm), $n_i = |C_i|$, $n_j = |C_j|$.

[**Total mark**: **34**]

End of Assignment 2