UIUC-CS412 "Introduction to Data Mining" (Summer 2016)

Final Exam

Monday, August 5, 2016
180 minutes, 75 points

Name [1]: NetID [1] :

1 [20]	2 [20]	3 [15]	4 [20]	5 [10]	6 [20]	7 [3]	Total [100]

Instructions

(a) Make sure your exam has **11 printed pages**.

(b) Do **NOT** look through the exam before the start of the exam.

(c) Calculators **with memory** are **NOT** allowed.

(d) Please keep your ID cards on the desk.

1. Short Answer Questions [20 points].

These questions should be answered in **not more than 1-2 lines**. Be as specific as possible.

(a) [1] What is stratified sampling? (**1-2 lines of explanation required**)
Answer: Dividing the given population into groups (or strata) based on some kind of similarity and then sampling from each of the group to have an overall representation.

(b) [1] [True /False] In PCA the first principal component is the direction which has the maximum variance for the given data points.
Answer: True

(c) [1] What is the OLAP operation used for performing aggregations on a data cube?
Answer: Roll up

(d) [1] Which method is better for computing iceberg cubes, Multiway Array Aggregation or BUC? Explain why. (**1-2 lines of explanation required**)
Answer: BUC is better.

(e) [1] [True /False] In graph mining algorithm gSpan, forward edges are prioritized over backward edges since it utilizes DFS.
Answer: False

(f) [1] For a set of values S and value v, Characterize the constraint $max(S) \leq v$ (label if it satisfies antimonotonic, monotonic, and/or succinct constraint category).
Answer: Antimonotonic and succinct.

(g) [1] Describe a type of constraint that is convertible.
Answer: $avg(S.profit) \geq 25$

(h) [1] State one difference between classification and regression?
Answer: Regression generates a continuous value output whereas classification gives discrete output value.

(i) [1] [True / False] If we train the Naive Bayes classifier using infinite training data that satisfies all of its modeling assumptions (e.g. conditional independence), then it will achieve zero *true error* over test examples from the same distribution.
Answer: False

(j) [1] The confidence of the estimate of classification performance increases with increasing the size of (a) Training dataset, (b) Test dataset. Select one of the two options.
Answer: Test dataset

(k) [1] Select the clustering algorithm that aims to optimize an objective function: *AGNES* or *KMeans*
Answer: KMeans

(l) [1] [True / False] The *KMeans* clustering algorithm finds the best value of k as part of its normal operation.
Answer: False

2. Data Processing[10 points].
Consider the ***sample*** data points selected from a pool as shown in table **??**.

X	66	72	77	84	83	71	65	70
Y	8	11	15	20	21	11	8	11

Table 1: Sample Data Points

(a) [6] What is the covariance $c(X, Y)$ for the given data?
Answer: In order to find the covariance we first need to determine the sample mean for each dimension.
$\mu_x = 73.5$, $\mu_y = 13.125$
Since the given data is a **sample** we use the following formula for covariance

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_y)}{N - 1} = 35.928$$

(b) [4] What is the correlation coefficient for the given data?
Answer: The correlation coefficient can be determined using the following formula

$$r_{x,y} = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

The standard deviation for sample is given by

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu_x)^2}{N - 1}} = 7.192$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \mu_y)^2}{N - 1}} = 5.055$$

$$r_{x,y} = 0.9883$$

3. Data warehouse, OLAP, Cube Computation [15 points].

(a) [6] Multiway Array Aggregation and BUC

(i) [3] Which cube computation method drills down from the apex cuboid (i.e., $(*, *, *)$)?
Answer:
BUC

(ii) [3] In what situations would you use Multiway Array Aggregation over BUC?
Answer:
When there are a small number of dimensions. When you want to materialize the entire cube is acceptable as well.

(b) [9] Cube Computation
Consider a base cuboid of 4 dimensions that contains 4 base cells:

$$(b_1, a_2, a_3, a_4)$$
$$(a_1, b_2, a_3, a_4)$$
$$(a_1, a_2, b_3, a_4)$$
$$(a_1, a_2, a_3, b_4)$$

where a_i, b_i, and c_i are distinct for $i = 1, 2, 3, 4$. There is *no* dimension with concept hierarchy. The measure of the cube is *count*. The count of each base cell is 1.

(i) [4] How many nonempty aggregated cells are there in the full cube?
Answer:
This cube has a total of 3^4 combinations possible, and there are the following empty cells.
All four dimensions are a, 1;
There are two dimensions with value b, $\binom{4}{2} \times 2^2 = 24$;
There are three dimensions with value b, $\binom{4}{3} \times 2 = 8$;
All four dimensions are b, 1;
After subtracting the 4 base cells, you are left with $3^4 - 1 - 24 - 8 - 1 - 4 = 43$ cells.

(ii) [4] **List all** nonempty aggregated cells that have *count* = 2.
Answer:
$\{a_1, a_2, *, *\}, \{*, a_2, a_3, *\}, \{*, *, a_3, a_4\}, \{a_1, *, *, a_4\}.\{a_1, *, a_3, *\}, \{*, a_2, *, a_4\}$

(iii) [1] Is $(*, *, *, *)$ an aggregate closed cell?
Answer:
Yes.

4. Mining Frequent Patterns [25 points].

(a) [8] Frequent Pattern Mining
Consider the following transactions:

Transaction Number	**Items bought**
1	Beer, Diaper, Tylenol
2	Beer, Diaper, Milk, Tylenol
3	Diaper, Milk
4	Beer, Diaper, Tylenol
5	Beer

Table 2: Transactions records

(i) [2] List all the frequent items. $min\_support = 3$.
Answer: $\{T : 3\}, \{B : 4\}, \{D : 4\}$

(ii) [2] List all the **closed** frequent itemsets. $min\_support = 3$.
Answer: $\{BDT : 3\}, \{B : 4\}, \{D : 4\}$

(iii) [2] Why is mining **closed** frequent itemsets important?
Answer: Mining only closed frequent itemsets can substantially reduce the number of patterns generated in frequent itemset mining while preserving the complete information regarding the set of frequent itemsets.

(iv) [2] List all the maximal frequent itemsets (max itemsets). $min\_support = 3$.
Answer: $\{BDT : 3\}$

(b) [6] Sequential Pattern Mining
Suppose a sequence database D contains three sequences as follows. Note (bc) means that items b and c are purchased at the same time (i.e., in the same transaction). Minimum support is 1. You are going to use PrefixSpan to mine the frequent sequential patterns.

Customer ID	**Shopping sequence**
1	$(fd)a(bc)a(bc)$
2	$dbc(fad)$
3	$(da)(bc)(ef)a$

Table 3: Transaction database to mine sequential patterns

(i) [2] Use PrefixSpan. Show <d>'s projected database.
Answer: $a(bc)a(bc), bc(fad), (\_a)(bc)(ef)a$

(ii) [2] What is the support count of <ab>in <d>'s projected database?
Answer: 2

(iii) [2] Use PrefixSpan. Show <df>'s projected database.
Answer: $(\_ad), a$

(c) [10] Graph Pattern Mining
Use Figure 1 to answer the following questions. The graph has 3 types of vertices (A, B, C) and threes type of edges (X, Y, Z).

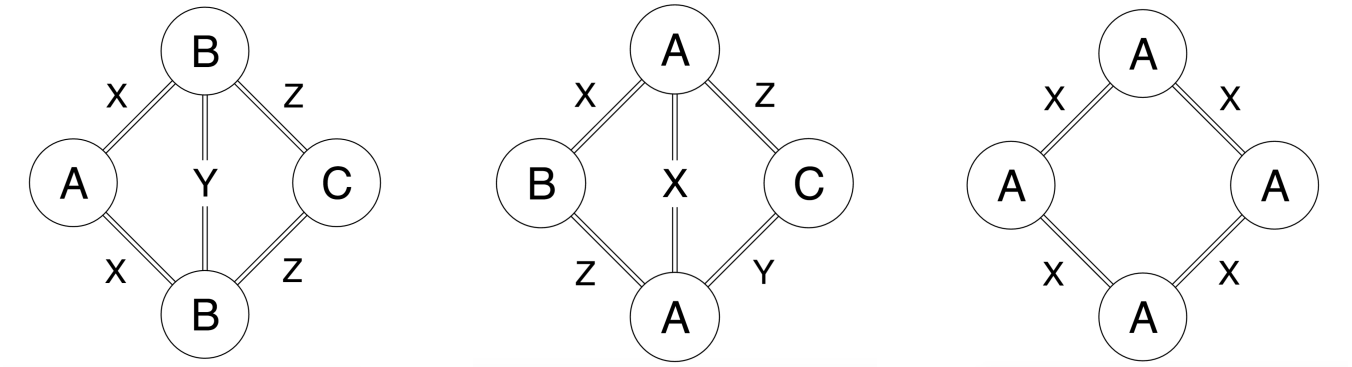


Figure 1: Three separate graphs.

(i) [2] Identify the frequent edges. Represent the edges using the following notation (V_1, E, V_2). If node C is connected to another node C with edge X, it will be (C, X, C). Minimum support is 2.
Answer: $(A, X, A), (A, X, B)$

(ii) [4] Draw the resulting graphs (all three of them) after removing the frequent subgraph resulting from growing from one of the frequent edges.
Answer:

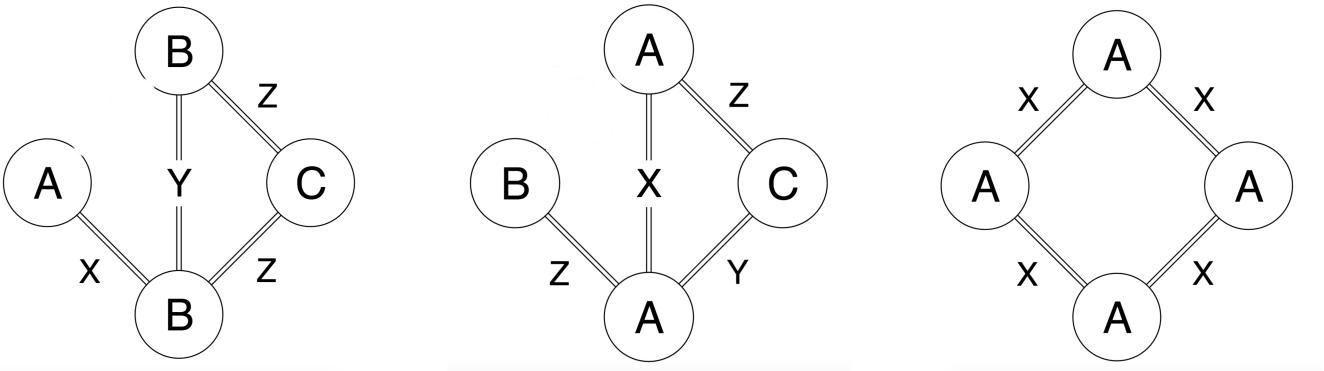


Figure 2: Three separate graphs after removing the frequent subgraph resulting from growing from (A, X, B).

This is one possible answer. Other answers were also accepted if the frequent subgraph was correctly identified and removed from the original graphs.

(iii) [2] What is the significance of minimum DFS code? Why do we use it?
Answer: It makes it possible to find the canonical form of a graph G. To identify graphs that have been already mined.

(iv) [2] Name how gSpan can save memory usage over Apriori-based algorithms.
Answer: By discovering frequent subgraphs without **candidate generation**.

5. Classification [30 points].

I. *Naive Bayes* [10 points]
You are provided with two classes for a set of phrases as shown in table **??**.

Phrase	**Class**
A Perfect World	*Movie*
A Perfect Day	*Song*
Electric Storm	*Song*
My Perfect Woman	*Movie*
Pretty Woman	*Movie*
Another Rainy Day	*Song*

Table 4: Training samples for phrase classification

(i) [2] Find the prior probabilities for the classes *Movie* and *Song*?
Answer: $P(Movie) = \frac{3}{6} = 0.5$, $P(Song) = \frac{3}{6} = 0.5$

(ii) [1] Determine the conditional probability of the word *Perfect* for both the classes.
Answer: There are 11 unique words in the provided training dataset: *(A, PERFECT, WORLD, DAY, ELECTRIC, STORM, MY, WOMAN, PRETTY, ANOTHER, RAINY)*
The distribution of the words for each class is as given below

Word	**Movie**	**Song**
A	1	1
PERFECT	2	1
WORLD	1	0
DAY	0	2
ELECTRIC	0	1
STORM	0	1
MY	1	0
WOMAN	2	0
PRETTY	1	0
ANOTHER	0	1
RAINY	0	1
Total	8	8

Using the above distribution we can determine the conditional probability of *Perfect* for each class

$$P(Perfect|Movie) = \frac{2}{8}, P(Perfect|Song) = \frac{1}{8}$$

(iii) [1] Determine the conditional probability of the word *Storm* for both the classes.
Using the above distribution we can determine the conditional probability of *Storm* for each class

$$P(Storm|Movie) = \frac{0}{8}, P(Storm|Song) = \frac{1}{8}$$

(iv) [4] What are values of all the conditional probabilities found in (ii) and (iii) after laplacian smoothing?
Answer: Since there are 11 unique words we add count=1 for each of them within each class, so the conditional probability values change to following

$$P(Perfect|Movie) = \frac{3}{19}, P(Perfect|Song) = \frac{2}{19}$$

$$P(Storm|Movie) = \frac{1}{19}, P(Storm|Song) = \frac{2}{19}$$

(v) [2] Determine the class of the phrase *Perfect Storm* based on the above calculations with laplacian smoothing.
Answer:

$$P(Movie|\text{Perfect Storm}) = \frac{1}{2} * \frac{3}{19} * \frac{1}{19}$$

$$P(Song|\text{Perfect Storm}) = \frac{1}{2} * \frac{2}{19} * \frac{2}{19}$$

Thus the given query belongs to class **Song**

II. *Decision Tree* [12 points]
You are required to construct a decision tree using the *Information Gain* metric for the data provided in table **??**. There are 8 samples each with three attributes (M, N, and Q) along with the $Class$ label. M takes two possible values (m_1, m_2), N takes three possible values (n_1, n_2, n_3) and Q takes three possible values (q_1, q_2, q_3). There are only two possible classes for this problem.

id	M	N	Q	Class
1	m_1	n_3	q_2	2
2	m_2	n_3	q_2	1
3	m_2	n_2	q_1	1
4	m_1	n_2	q_1	2
5	m_2	n_1	q_3	2
6	m_1	n_1	q_3	2
7	m_2	n_3	q_1	1
8	m_1	n_2	q_2	2

Table 5: Training samples for decision tree

Note: For this problem you are always required to use log with base 2

(i) [2] What is the entropy for the given training dataset?
Answer: The entropy for the given dataset is obtained using the following

$$Info(D) = -\sum_i p_i log p_i$$

$$= -(\frac{5}{8}log\frac{5}{8} + \frac{3}{8}log\frac{3}{8}) = 0.954$$

(ii) [2] Find the *Information Gain* for the attribute M.
Answer:

$$Info(M) = \frac{4}{8}Info(0,4) + \frac{4}{8}Info(3,1) = 0.4056$$

$$InfoGain(M) = 0.54875$$

(iii) [2] Find the *Information Gain* for the attribute N.
Answer:

$$Info(N) = \frac{2}{8}Info(0,2) + \frac{3}{8}Info(2,1) + \frac{3}{8}Info(1,2) = 0.68822$$

$$InfoGain(N) = 0.26615$$

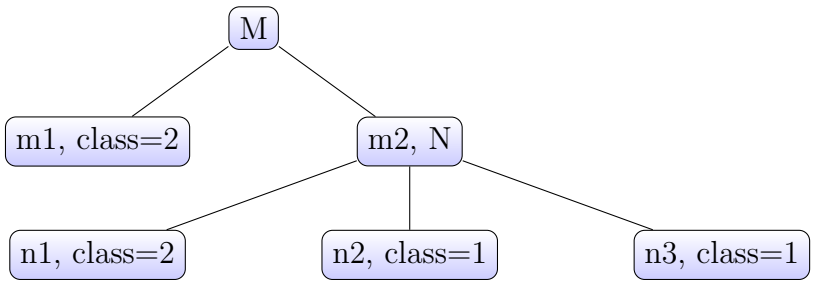
(iv) [2] Find the *Information Gain* for the attribute Q.
Answer:

$$Info(Q) = \frac{3}{8}Info(2,1) + \frac{3}{8}Info(1,2) + \frac{2}{8}Info(0,2) = 0.68822$$

$$InfoGain(Q) = 0.26615$$

(v) [1] Which attribute will you select for the first split?
Answer: Since M has the highest info gain, we split using M

(vi) [3] Draw the final Decision Tree by showing the required calculation for further splits. In case of tie with the Information Gain scores use the attribute which comes first in the alphabetical order.



III. *Perceptron* [8 points]
For this problem you are required to solve the perceptron classifier with the help of the data points provided in table **??**

x_1	8	0	4	0.5	4	2
x_2	4	0	8	0.5	3	5
$Class$	+	-	+	-	+	+

Table 6: Data points for perceptron

The classifier produces the output class symbol according to the following definition

$$y = \begin{cases} +, & \text{if } sign(w_0 + w_1 x_1 + w_2 x_2) \geq 0 \\ -, & \text{otherwise} \end{cases} \tag{0.1}$$

You can start with the initial weights as $w_0 = 0$, $w_1 = 1$, and $w_2 = 1$. Also assume that the learning factor $\eta = 1$.
Answer: We iterate over the given points one at a time and update the weights whenever there is a misclassification. Also use the updated weights for new iterations.
Iteration 1: (8,4) No change
Iteration 2: (0,0) Update

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$$

Iteration 3: (4,8) No change
Iteration 4: (0.5,0.5) Update

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} -2 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Iteration 5: (4,2) No change
Iteration 6: (2,5) No change

Stop here because all the points are correctly classified using the new weights.

6. Clustering [23 points].

I. *AGNES* [9 points]
Suppose we have 9 points which are listed and shown in figure **??**. The ground truth for these points is also provided in the table.

	x	y	*Ground Truth*
P1	2	8	C1
P2	2	7	C1
P3	3	6	C2
P4	3	5	C2
P5	4	5	C2
P6	7	5	C3
P7	7	4	C3
P8	6	4	C3
P9	6	2	C4

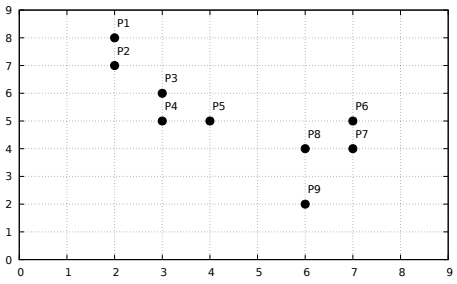
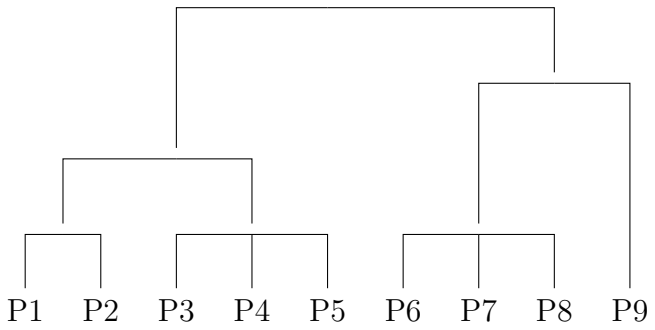


Figure 3: 2D plot of points

(i) [3] If we use the AGNES hierarchical clustering algorithm on the points above using single linkage for Euclidean distance measure, show the dendrogram with the corresponding levels. You do not have to show any calculation to perform AGNES but indicate the distance in the levels of the dendrogram.



Level 1: Dist(1), Level 2: Dist($\sqrt{2}$), Level 3: Dist(2), Level 4: Dist($\sqrt{5}$)

(ii) [2] If we don't know the ground truth and want to cluster the points into 2 groups then what are the members of the 2 groups based on the dendrogram?
Answer: (P1,P2,P3,P4,P5) and (P6,P7,P8,P9)

(iii) [4] Based on the given ground truth what are the B-Cubed precision and recall for the output you got in the previous part?

Average Precision = 0.567
Average Recall = 1

P_i	P1	P2	P3	P4	P5	P6	P7	P8	P9
Precision	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{1}{4}$
Recall	$\frac{2}{2}$	$\frac{2}{2}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{1}{1}$

II. *K-Means* [10 points]
Consider the same data points as shown in figure **??**. We will now apply K-Means algorithm for $K = 2$ to group the points into two clusters.

(i) [5] Show the steps for the algorithm using $P3$ and $P8$ as the initial centroids. You don't need to show the distance calculation within each iteration. The centroids can be obtained just by observing the provided figure. List all the intermediate clusters and the final cluster.
Iteration 1
C1 (P1,P2,P3,P4,P5): (2.8, 6.2)
C2 (P6,P7,P8,P9): (6.5, 3.75)

Iteration 2
No Change

(ii) [5] Show the steps for the algorithm using $P4$ and $P5$ as the initial centroids. You don't need to show the distance calculation within each iteration. The centroids can be obtained just by observing the provided figure. List all the intermediate clusters and the final cluster.
Iteration 1
C1 (P1,P2,P3,P4): (2.5, 6.5)
C2 (P5,P6,P7,P8,P9): (6, 4)

Iteration 2
C1 (P1,P2,P3,P4,P5): (2.8, 6.2)
C2 (P6,P7,P8,P9): (6.5, 3.75)

Iteration 3
No change

III. *Hierarchical Clustering* [4 points]
Consider the data points as shown below in figure **??**. You need to perform a hierarchical clustering starting with data points as leaf nodes to form groups until you have one big cluster. You need to use *Complete linkage* for grouping together the clusters at each step. The correct answer can be obtained visually without any calculations. Just indicate the clusters and the corresponding members you would expect as you process the data points.

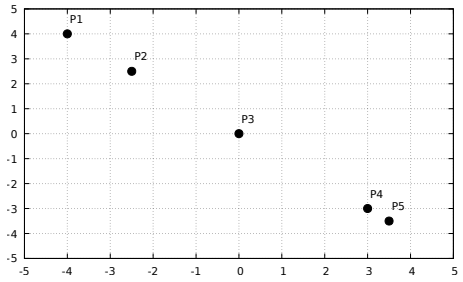


Figure 4: Data to cluster

Answer: You can show by drawing the clusters in the given figure or list the steps
1: {P1,P2},P3,P4,P5
2: {P1,P2},P3,{P4,P5}
3: {P1,P2},{P3,P4,P5}
4: {P1,P2,P3,P4,P5}