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

Final Exam, Version 1

Thursday, May 12, 2016
180 minutes, 150 points

Name: NetID:

1 [30]	2 [30]	3 [47]	4 [43]	Total

1. [30] Preprocessing Data, Data Cube

   (a) [4'] Present the value range for each of the following measures.

   i. [2'] Jacquard coefficient
   ANSWER: $[0, 1]$

   ii. [2'] Covariance
   ANSWER: $(-\infty, +\infty)$

   (b) [6'] Give three example distance measures for each of the following two kinds.

   i. [3'] The distance between two objects
   ANSWER: **Any three of the following: Euclidean distance, Manhattan distance, Supremum distance** $(i.e., L_\infty norm)$**, Cosine distance, Minkowski distance, ...**

   ii. [3'] The distance between two clusters
   ANSWER: **Any three of the following: Single-link, Complete link, Average link, distance between cluster centroids, distance between cluster medoids**

   (c) [12'] Consider 5 data points in a 2-D space: (-3,3), (-1,1), (0,0), (1,-1), and (3,-3). (For all sub-questions below, correct answers without explanations receive full points; and incorrect answers with explanations may receive partial credit.)

   i. [2'] Calculate the covariance matrix.
   ANSWER: **The covariance matrix can be either one of the following, depending on if you divide the product with** 5 **or** 4**.**

$$\begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix} \begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix}$$

ii. [2'] Calculate the correlation coefficient for the two dimensions.
ANSWER: -1 **because they are completely negatively corrleated. You can also use formula to calculate.**

iii. [3'] Calculate the first and the second principal components (two vectors), and indicate which is the first principal component. (Note: drawing is not enough, and no calculation is needed)
ANSWER: $(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}), (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$**. The negation of them are also correct. Please note that you have to normalize the vectors to make them unit vectors because principal components must be orthonormal.**

iv. [3'] What are the coordinates of the 5 data points, *projected* to the 1-D space corresponding to the first principal component?
ANSWER: **By drawing the graph, we can easily guess they are:** $-3\sqrt{2}, -\sqrt{2}, 0, \sqrt{2}, 3\sqrt{2}$**. You can also do matrix multiplication, and it is not too slow to do that as well.**

v. [2'] Is the projection of data points to the 1-D space in the previous sub-question a lossless or lossy compression? Briefly explain.
ANSWER: **Lossless because we can fully recover the original data points from them. A shorter answer could be: as all the data points are on the same line, we only need one dimension to express the data, which is exactly what PCA does.**

(d) [8'] Suppose the base cuboid of a data cube contains only two cells.
$(a_1, a_2, a_3, ..., a_{10}), (b_1, b_2, b_3, ..., b_{10})$, where $a_i = b_i$ if i is an odd number; otherwise $a_i \neq b_i$.

i. [3'] How many nonempty aggregated (i.e., non-base) cells are there in this data cube?
ANSWER: 2014**. It can be calculated by** $2 \times 2^{10} - 2^5 - 2$**.**

ii. [3'] How many nonempty, closed aggregated cells are there in this data cube?
ANSWER: 1**. There are three closed cells** $(a_1, a_2, a_3, ..., a_{10}) : 1, (b_1, b_2, b_3, ..., b_{10}) : 1, (a_1, *, a_3, *, ..., a_9, *) : 2$**, but only one closed aggregated cell** $(a_1, *, a_3, *, ..., a_9, *) : 2$**.**

iii. [2'] If we set minimum support = 2, with the measure being `count` how many nonempty aggregated cells are there in the corresponding iceberg cube?
ANSWER: 32**.** $(a_1, *, a_3, *, ..., a_9, *) : 2$ **and its further generalizations, so in total** 2^5**.**

2. [30] Frequent Pattern and Association Mining

(a) [8'] The price of each item in a store is nonnegative. For each of the following cases, identify the type of constraint they represent and briefly discuss how to mine such association rules efficiently with frequent pattern mining algorithms.

i. [4'] Containing at least one Nintendo game.
ANSWER: **The constraint is succinct and monotonic. This constraint can be mined efficiently using FP-growth as follows.**
• **All frequent Nintendo games are listed at the end of the list of frequent items L.**
• **Only those conditional pattern bases and FP-trees for frequent Nintendo games need to be derived from the global FP-tree and mined recursively.**

ii. [4'] Containing one free item and other items the sum of whose prices is at least \$200.
ANSWER: **The constraint is monotonic. (Or, subconstraints "containing one free item" and "the sum of whose prices is less than \$200" are succinct and monotonic, respectively.) This constraint can be mined efficiently using FP-growth as follows.**
• **Put all frequent free items at the end of the list of frequent items L.**
• **Only conditional pattern bases and FP-trees for frequent free items need to be derived from the global FP-tree and mined recursively. Other free items should be excluded from these conditional pattern bases and FP-trees.**
• **Once a pattern with items the sum of whose prices is at least \$200, no further constraint checking for total price is needed in recursive mining.**
• **A pattern as well as its conditional pattern base can be pruned if the sum of the price of items in the pattern and the frequent ones in the pattern base is less than \$200.**

(b) [10'] 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). Let the minimum support be 3.

customer id	shopping sequence
1	$(bc)(de)f$
2	$bcdef$
3	$(bc)dbegf$

Use Generalized Sequential Patterns(GSP) to mine frequent sequential patterns from this database. You need to list all the steps and the results for mining frequent sequential patterns.
ANSWER: **(1) Scan database once, count support for candidates and prune length-1 sequential patterns by min sup to get $L1$.**
(2) Generate $C2$ by self-joining.

(3) Scan DB again, prune length-2 sequential patterns by min sup to get $L2$.
(4) Generate $C3$ by self-joining.
(5) Scan DB again, prune length-3 sequential patterns by min sup to get $L3$.
(6) $C4$ is null based on Apriori Principle. Terminate.
The frequent sequential patterns are $\{L1, L2, L3\}$. The steps are shown in the following table.

C1	L1	C2	L2	C3	L3	C4
b	b:3	bb	bd:3	bdf	bdf:3	null
c	c:3	bc	be:3	bef	bef:3	
d	d:3	bd	bf:3	cdf	cdf:3	
e	e:3	be	cd:3	cef	cef:3	
f	f:3	bf	ce:3			
g		cb	cf:3			
		cc	df:3			
		cd	ef:3			
		ce				
		cf				
		db				
		dc				
		dd				
		de				
		df				
		eb				
		ec				
		ed				
		ee				
		ef				
		fb				
		fc				
		fd				
		fe				
		ff				
		(bc)				
		(bd)				
		(be)				
		(bf)				
		(cd)				
		(ce)				
		(cf)				
		(de)				
		(df)				
		(ef)				

(c) [12'] For the same database as question (a), use PrefixSpan to mine frequent sequential patterns from this database. You need to list all the steps and the results for mining frequent sequential patterns.

ANSWER: **(1) Scan DB once, count support for candidates and prune length-1 sequential patterns by min sup to get length-1 prefix list $\{a, b, c, d, e, f\}$.**

(2) For each prefix in $P1$, generate its projected database.

(3) For each projected database, we add one more item to the old prefix(length-1) to generate new prefix(length-2) whose support passes min sup. Then we generate its projected database. Similarly, we generate new prefix and its projected database recursively. The algorithm terminates when the projected database is null.

The results of each step is shown in the following table. The frequent patterns are all the prefix.

prefix	projected DB	prefix	projected DB	prefix	projected DB
⟨b⟩	(_c)(de)f	⟨bd⟩	(_e)f	⟨bdf⟩	null
	cdef		ef		
	(_c)dbegf		begf		
		⟨be⟩	f	⟨bef⟩	null
			f		
			gf		
		⟨bf⟩	null		
⟨c⟩	(de)f	⟨cd⟩	(_e)f	⟨cdf⟩	null
	def		ef		
	dbegf		begf		
		⟨ce⟩	f	⟨cef⟩	null
			f		
			gf		
		⟨cf⟩	null		
⟨d⟩	(_e)f	⟨df⟩	null		
	ef				
	begf				
⟨e⟩	f	⟨ef⟩	null		
	f				
	gf				
⟨f⟩	null				

3. [47] Classification

(a) [12'] Suppose we want to predict whether a restaurant is popular based on its price, parking, and cuisine, and we collected training data as in the following table. Popularity is the label, and (Price, Parking, Cuisine) are the features. Answer following questions.

ID	Price	Parking	Cuisine	Popularity
1	Medium	Available	Mexican	P
2	High	Available	Italian	NP
3	Low	Available	American	P
4	Medium	No	Mexican	NP
5	Low	Available	American	P
6	Medium	No	Italian	P
7	High	Available	Italian	NP
8	High	Available	Mexican	P
9	High	No	American	NP
10	Low	No	Italian	P
11	Low	Available	Mexican	NP
12	Low	Available	Italian	P

i. [3'] What's the information gain for the 'Price' attribute? Please show your calculation. (Note: If you don't have a calculator, listing the valid equation is OK.)
ANSWER: **The entropy of the whole dataset D is:**
$Info(D) = I(7,5) = -\frac{7}{12} * log(\frac{7}{12}) - \frac{5}{12} * log(\frac{5}{12}) = 0.9798$
If we choose the attribute 'price' to split in the root node, the information will be:
$Info_{price}(D) = \frac{5}{12} * I(4,1) + \frac{3}{12} * I(2,1) + \frac{4}{12} * I(3,1) = 0.8008$
Then the information gain for the attribute can be computed as:
$Gain(price) = Info(D) - Info_{price}(D) = 0.179$

ii. [3'] Now suppose we want to use Gini Index as attribute selection measure. What's the Gini index for the attribute 'Parking'? What's the reduction in impurity in terms of Gini Index? Please show your calculation. (Note: You can write the answer and intermediate values as fractions in simplest form.)
ANSWER: **Let A = 'Available', and D be the whole dataset.**
$Gini_A(D) = \frac{4}{12} * (1 - (\frac{2}{4})^2 - (\frac{2}{4})^2) + \frac{8}{12} * (1 - (\frac{5}{8})^2 - (\frac{3}{8})^2) = 0.479$
$Gini(D) = 1 - (\frac{7}{12})^2 - (\frac{5}{12})^2 = 0.4861$
$\Delta Gini = Gini(D) - Gini_A(D) = 0.0071$

iii. [4'] Based on the training data, we want to construct a Naive Bayes classifier. Please estimate the following terms. (No smoothing is required, and please show your calculation) (Note: You can write the answers and intermediate values as fractions in simplest form.)

- Pr(Popularity = 'P')
  ANSWER: $\frac{7}{12}$

- Pr(Popularity = 'NP')
  ANSWER: $\frac{5}{12}$

- Pr(Price = 'Low', Parking = 'Available', Cuisine = 'Mexican' | Popularity = 'P')
  ANSWER: $Pr(Price =' Low' \mid Popularity = `P')(Parking = `Available' \mid Popularity = `P')(Cuisine = `Mexican' \mid Popularity = `P') = \frac{4}{7} * \frac{5}{7} * \frac{2}{7} = \frac{40}{343}$

- Pr(Price = 'Low', Parking = 'Available', Cuisine = 'Mexican' | Popularity = 'NP')
  ANSWER: $Pr(Price =' Low' \mid Popularity = `NP')(Parking = `Available' \mid Popularity = `NP')(Cuisine = `Mexican' \mid Popularity = `NP') = \frac{1}{5} * \frac{3}{5} * \frac{2}{5} = \frac{6}{125}$

iv. [2'] Suppose a restaurant has the values: Price = 'Low', Parking = 'Available' , Cuisine = 'Mexican'. Based on the calculation in (iii), is this restaurant classified as popular? Please show your reasons.
ANSWER: **Since $\frac{7}{12} * \frac{40}{343} > \frac{5}{12} * \frac{6}{125}$, the restaurant is classified as popular.**

(b) [6'] Suppose we build a classifier to predict whether a patient has cancer or not. We want to evaluate the performance of the classifier, and collect its confusion matrix as in the following table. In the given confusion matrix, 'A' represents 'Actual class', and 'P' represents 'Predicted class'. Please answer following questions.

$A \setminus P$	Yes (has cancer)	No	Total
Yes	90	210	300
No	140	9560	9700
Total	230	9770	10000

i. [3'] Please compute the sensitivity, specificity and precision.
ANSWER: $TP = 90, FN = 210, FP = 140, TN = 9560$
$Sensitivity = \frac{TP}{TP+FN} = \frac{90}{90+210} = 0.3$
$Specificity = \frac{TN}{FP+TN} = \frac{9560}{140+9560} = 0.9856$
$Precision = \frac{TP}{TP+FP} = \frac{90}{90+140} = 0.3913$

ii. [3'] Please interpret the sensitivity and specificity in this particular setting. Based on the interpretation, is this a useful classifier?
ANSWER: **In this particular setting:**
• Sensitivity is the percentage of patients who are predicted as "having cancer" among the patients who indeed have cancer.
• Specificity is the percentage of patients who are predicted as "not having cancer" among the patients who are healthy (without cancer).
Though this classifier has high specificity(0.9856), it has quite low sensitivity(0.3), which is quite undesirable, since it means in many cases, this classifier fails to identify people who need treatment.

(c) [4'] Assume that you are working on designing a two class classifier, with equal priors (i.e. the probability of each class is the same). This is an image classification problem, and you are give image histograms as a feature. Assume that you are told that the Bayes Error is 0.4. (a) In this scenario, what is the classification accuracy of the optimum classifier?
ANSWER: **The classification accuracy for the optimum classifier is 60%.**

(b) Given your answer to part (a), how can you further reduce the classification error?
ANSWER: **Since the Bayes error is large, the best way to improve the accuracy of the classifier is to change the features used in the classification, since different features will have different Bayes error—this is why feature selection is such an important aspect of the classifier design workflow.**

(d) [4'] What are the major differences among the two methods for increasing the accuracy of a classifier: (1) **bagging** and (2) **boosting**?
ANSWER: **(1) bagging: Averaging the prediction over a collection of classifiers.**
(2) boosting: Use weighted vote with a collection of classifiers, and assign more weights to those examples who were misclassified by the previous classifier in the next round of induction.

(e) [6'] What are the major differences among the three methods for the evaluation of the accuracy of a classifier : (1) **hold-out method**, (2) **cross-validation**, and (3) **bootstrap**?
ANSWER: **(1) hold-out method: use part of the data (e.g., $2/3$) for training and the remaining for testing.**
(2) cross-validation: partition data into (relatively even) k portions, $D_1, ..., D_k$, use D_i for testing and the other $k-1$ portions for training for any i, and then merge the results.
(3) bootstrap: works for small data set, it samples the given training data uniformly with replacement, e.g., 0.632 bootstrap.

(f) [4'] The Gini Index is a measure of impurity of a class; when there are n classes in the Data D, the gini index is defined as:$gini(D) = 1 - \sum_{i=1}^{n} p_i^2$. (a) what is the smallest value of the index? When will it occur?
ANSWER: **The smallest value of the index is 0, when the class is pure; that is $p_i = 1$ for some i.**

(b) What is the largest value of this index? When does it occur?
ANSWER: **The largest value of the index is $1 - 1/n$, where n is the number of classes; this is the case when all classes are equally likely.**

(g) [4'] Consider the following set of ROC curves drawn in Figure 1 for classifiers M_1 and M_2. (a) Which of the two classifiers is better? Why?
ANSWER: **M_1 is the better classifier since it has a larger area under the curve.**

(b) Draw a ROC curve onto Figure 1 of a classifier that has better performance than M_1 and M_2.
ANSWER: **Any curve that is monotonically increasing and is drawn above M_1.**

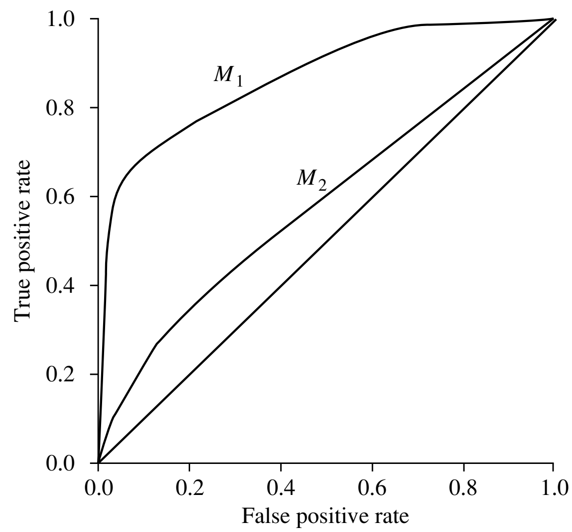


Figure 1: Annotate the output the ROC question on this figure

(h) [2'] Assume that you are supervising John on an binary image classification problem; the classes have equal priors. John comes back with three different classifiers and associated classification error results: classifier 1 with 25% error, classifier 2 with 50% error and classifier 3 with 80% error. What classifier will you choose to work with? Why?
ANSWER: **The one with 80% error; since this is a two class problem, we can simply flip the labels and create a classifier with 20% error.**

(i) [3'] To speed up calculation in his neural network which is being used for a binary classification problem with equal class priors, John changes the activation function f at each from a sigmoid function at node j ($f = \frac{1}{1+\exp(-\sum_k w_k + \theta_j)}$) to that of a linear function ($f = -\sum_k w_k + \theta_j$). He finds that his error rate increases. What could be a possible explanation?
ANSWER: **When $f = -\sum_k w_k + \theta_j$, the net result is an activation function that is a linear discriminant. We know that linear discriminants cannot separate linearly inseparable classes such as the classes in the XOR problem.**

(j) [2'] Can one make causal inferences with a Bayesian Belief Network?
ANSWER: **No. the belief network efficiently models a joint probability distribution and cannot model causality.**

4. [43] Clustering

(a) [3'] Assume that we want to partition N points into k clusters. What is the computational complexity of the optimal algorithm? Provide a brief explanation.
ANSWER: $O(k^n)$, **since the problem is NP-hard, every assignment of points to clusters needed to be tried to identify the optimum clustering result.**

(b) [3'] Following up on the previous question, is a $k-$means algorithm guaranteed to determine the global optimum partition of k clusters? Explain.
ANSWER: **No; the $k-$means algorithm finds a local optimum.**

(c) [3'] What is one disadvantage of partitioning based algorithms such as $k-$means that prompts the development of density based algorithms such as DBSCAN?
ANSWER: **The fact that partitioning based techniques that use Euclidean distances result in spherical clusters, implies that arbitrary shaped clusters may not be discovered at all.**

(d) [3'] Divisive and Agglomerative algorithms are top-down and bottom up approaches to clustering data. Why are agglomerative algorithms much more popular than are divisive algorithms?
ANSWER: **This is because in divisive algorithm, every combination of points to produce the division needs to be tried out making divisive algorithms impractical for large datasets.**

(e) [3'] Assume that we want to cluster N points; these points have 3 ground truth clusters. What will the output number of clusters be, if we run the $k-$means algorithm on this dataset which has 3 ground truth clusters, with parameter $k = 5$?
ANSWER: **There will be 5 clusters output since $k = 5$.**

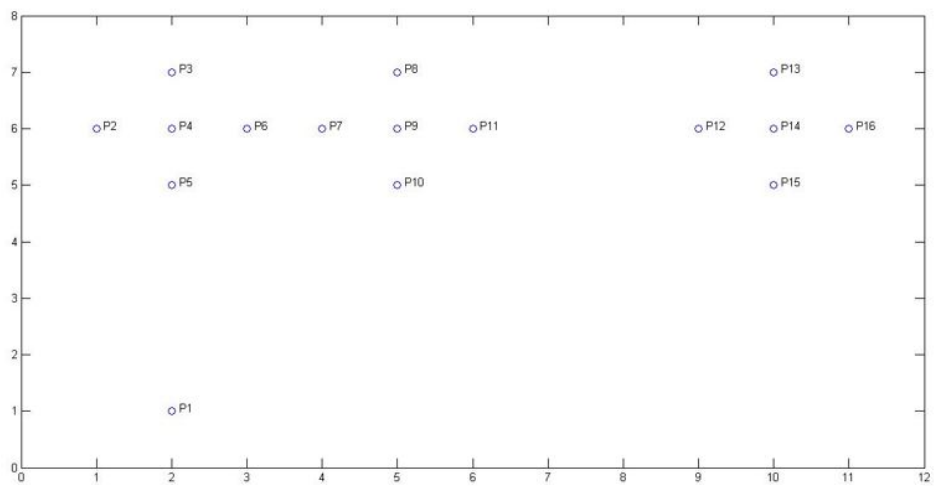
(f) [6'] Assume that you are give a file containing information about N individuals, with each line of the file containing two numbers: age, and income. You are asked to find "good" clusters. You have no information on the data besides the file.

(1) [3'] What will you do to determine the number of clusters?
ANSWER: **The elbow method is one option; once could also use the minimum description length (MDL) based approaches to identify the number of clusters.**

(2) [3'] Should you use a measure like the Silhouette coefficient (an internal validity measure) or Normalized Mutual Information (NMI, an external validation measure) to validate the clusters?
ANSWER: **Internal validity measures such as the Silhouette coefficient should be used since no ground truth labels are available. Thus NMI cannot be used.**

(g) [22'] Suppose we have 16 data points, which are listed in the following table and figure. The ground truth (correct cluster output) is also provided.

Point	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
x	2	1	2	2	2	3	4	5	5	5	6	9	10	10	10	11
y	1	6	7	6	5	6	6	7	6	5	6	6	7	6	5	6
Cluster	C3	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1	C2	C2	C2	C2	C2



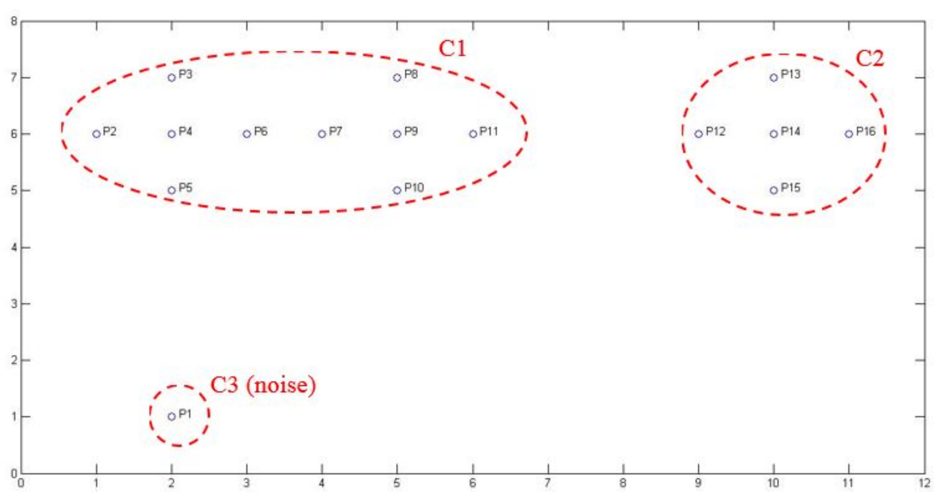


Figure 2: Data for clustering and ground truth

(1) [5'] Using the given data, perform DBSCAN, a density-based clustering algorithm, with parameters $MinPts = 4$, $\epsilon = 1.1$, and random points: P1, then P4, then P7, then P9, then P14. Show how clusters grow, and explicitly give the final clustering result.

ANSWER: **As the ϵ-neighborhoods of P1 have smaller number of objects than $MinPts$, it is marked as noise.**

$C1 = \{P4\}$ $N = \{P2, P3, P5, P6\}$
$C1 = \{P4, P2\}$ $N = \{P3, P5, P6\}$
$C1 = \{P4, P2, P3\}$ $N = \{P5, P6\}$
$C1 = \{P4, P2, P3, P5\}$ $N = \{P6\}$
$C1 = \{P4, P2, P3, P5, P6\}$ $N = \{\}$

P7 is marked as noise for the same reason.

$C2 = \{P9\}$ $N = \{P7, P8, P10, P11\}$
$C2 = \{P9, P7\}$ $N = \{P8, P10, P11\}$
$C2 = \{P9, P7, P8\}$ $N = \{P10, P11\}$
$C2 = \{P9, P7, P8, P10\}$ $N = \{P11\}$
$C2 = \{P9, P7, P8, P10, P11\}$ $N = \{\}$
$C3 = \{P14\}$ $N = \{P12, P13, P15, P16\}$
$C3 = \{P14, P12\}$ $N = \{P13, P15, P16\}$
$C3 = \{P14, P12, P13\}$ $N = \{P15, P16\}$
$C3 = \{P14, P12, P13, P15\}$ $N = \{P16\}$
$C3 = \{P14, P12, P13, P15, P16\}$ $N = \{\}$

No unvisited points left → stop → final clusters:

C1: {P2, P3, P4, P5, P6} C2: {P7, P8, P9, P10, P11}
C3: {P12, P13, P14, P15, P16} C4(Noise): P1

(2) [3'] Do the same thing as in question (1), but with different parameters $MinPts = 4$, $\epsilon = 1.7$, and different random points: P1, then P2, then P4, then P14. Directly show the final clustering result by annotating the following figure. (Note: You don't need to show how clusters grow.)

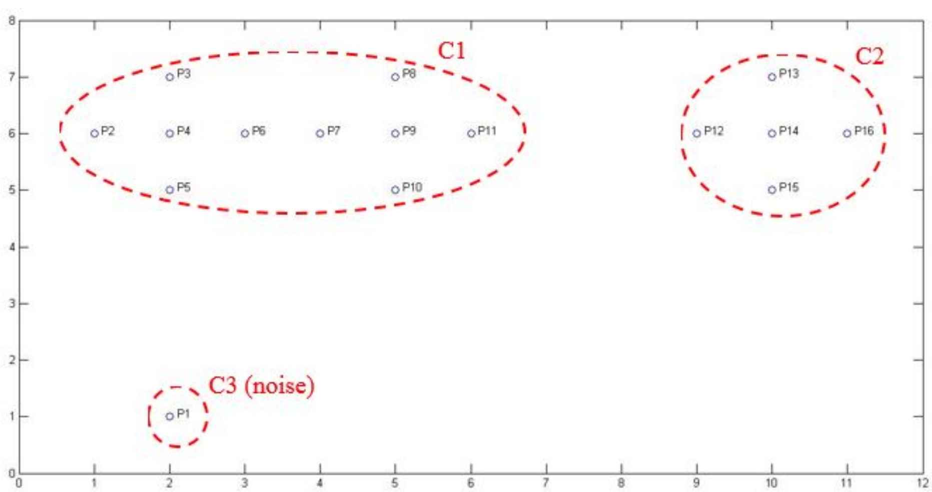
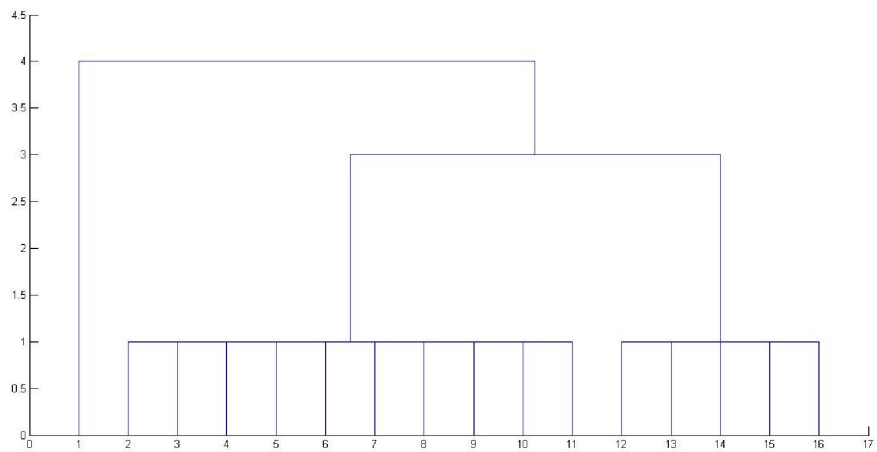


Figure 3: Annotate the output of DBSCAN in question (2) here

(3) [3'] Based on the results of questions (a) and (b), you can see that the output of DBSCAN depends on the parameters. Suggest a general method (independent of specific datasets) to choose the best parameters for DBSCAN. Explain why your method is good.
ANSWER: **The most intuitive method is to run DBSCAN several times with different parameters, and choose the best according to some quality measure. You get full points as long as your explanation makes sense.**

(4) [4'] Using the given data, perform AGNES, a hierarchical clustering algorithm, and assume that the single-link method and Euclidean distance as the dissimilarity measure are used. (Note: You only need to draw the dendrogram.)



(5) [3'] Using the given data, show the clustering result of K-Means for K = 2 by annotating the following figure, with P10 and P15 as the initial centroids and using Euclidean distance as the distance function. (Note: There is no need to show your computation or to illustrate the means.)

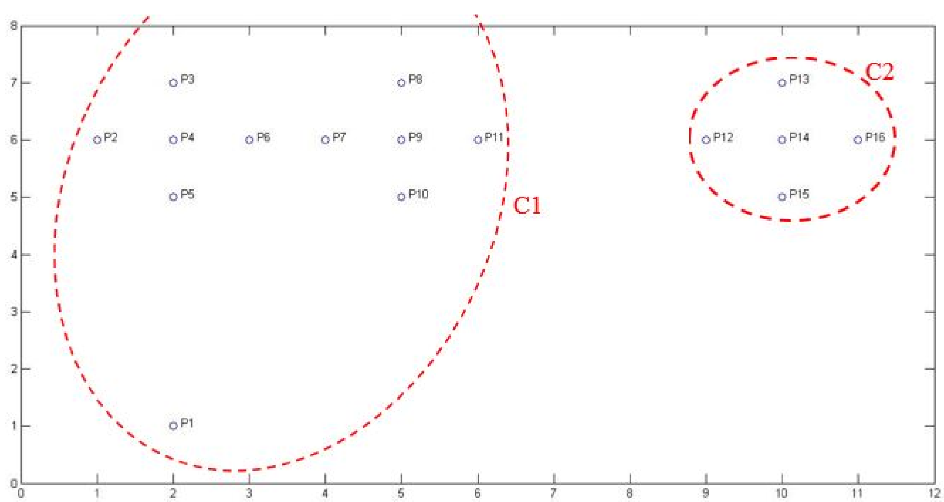


Figure 4: Annotate the output of K-Means in question (5) here

(6) [4'] Suppose a particular algorithm outputs a clustering as shown below. Based on the given ground truth, what are the B-Cubed precision and recall of the clustering? Show your calculations.

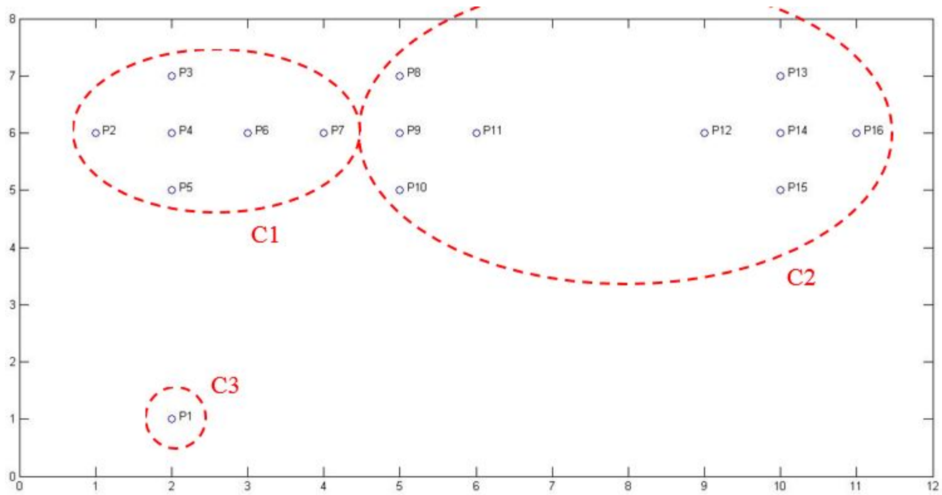


Figure 5: Output of a particular clustering algorithm

ANSWER:

Point	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P_i	$\frac{1}{1}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{6}{6}$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{5}{9}$
R_i	$\frac{1}{1}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{6}{10}$	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{4}{10}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$

$$Precision = \frac{1}{16} * \sum_{i=1}^{16} P_i = \frac{13}{18} \qquad Recall = \frac{1}{16} * \sum_{i=1}^{16} R_i = \frac{7}{10}$$