CS412 "An Introduction to Data Mining" (Fall 2014)
Final Exam

(Friday, Dec. 12, 2014, 180 minutes, 150 marks, two sheets of references, brief answers)

Name: NetID: Score:

1 [30]	2 [10]	3 [20]	4 [30]	5 [40]	6 [20]	Total [150]

1. [30] Short Questions

   (a) [3] Name one visualization technique that may help to decide if using PCA on a particular 2D dataset is beneficial or not. Explain how the technique is helpful for the purpose.

   Answer: Scatterplot, which helps to analyze correlations among the 2 features of the dataset. If the two features are correlated, using PCA would be useful, because the reduction will not lose much information. □

   (b) [3] Give an example when using Manhattan distance is more suitable than using Euclidean distance.

   Answer: When we want to measure walking distance between two destinations in Manhattan. □

   (c) [4] Suppose Dunkin' Donuts wants to study the correlation between the unit price $x_{i,t}$ and net profit $y_{i,t}$ where i denotes the i^{th} store, while t denotes the date. A linear regression model is used that $y = \alpha x + \beta$. (A) Judge which category α and β belong to (distributive, algebraic, or holistic), and (B) explain how each can be computed efficiently in multidimensional space.

   Hint: α and β can be calculated as follows

$$\alpha = \frac{NT\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}y_{i,t} - \sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t}}{NT\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}^2 - (\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t})^2}$$

$$\beta = \frac{1}{NT}(\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t} - \alpha\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t})$$

   Answer: 1. Both of them belong to algebraic: because all of components are algebraic; 2. α can be updated efficiently based on the distributive measures: $\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}y_{i,t}$, $\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}$, $\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t}$, $\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}^2$; β can be updated efficiently based on the distributive measures: $\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t}$ and $\sum_{t=1}^{T}\sum_{i=1}^{N} x_{i,t}$. computation: Pearson correlation co-efficient can be computed efficiently based on the about formula using distributive measures: □

   (d) [4] Suppose a base cuboid has D dimensions and contains m ($m > 0$) nonempty cells. Each dimension i has H_i levels (not including *all*), answer the following questions.

      i. How many *aggregate cuboids* does this cube contain (not including the base cuboid)?

ii. What is the *maximum number* of *nonempty cells* possible in such a materialized cube?

Answer:

i. $\prod_{i=1}^{D}(M_i + 1) - 1$, -1 to exclude the base cuboid.

ii. $(\prod_{i=1}^{D}(M_i + 1) - 1)m + 1$, each cuboid contains m cells except the apex cuboid.

□

(e) [4] Give one example of classification problem where false negative rate is more important than false positive rate, and explain why.

Answer: For example, we want to build a classifier to predict whether a patient has cancer. In this case, false negative rate is the percentage of patients who are diagnosed as healthy among the patients who actually have cancer; the high false negative rate means the classifier is useless. (There are many possible examples; these examples usually have imbalanced distribution, and the positive class is more important.) □

(f) [4] Give one real-world application when semi-supervised learning might be a good solution.

Answer: For example, we want to classify images into predefined categories. Usually we have a large amount of the unlabeled images, and it's impossible to label all of them. So we label a small portion, and train a model using both labeled and unlabeled data. (As long as students understand the concept of semi-supervised learning, and the example is reasonable, they get full points) □

(g) [4] Explain if the following claim is true or not: the result of DBSCAN with $MinPts = 2$ will be the same as of hierarchical agglomerative clustering (AGNES) with the single link metric, with the dendrogram cut at height ϵ.

Answer: Yes, because in that case, DBSCAN will join pairs of nearby points into chains if the distance of points in a pair smaller than ϵ, which is similar to AGNES. □

(h) [4] Among DBSCAN ($MinPts > 2$), K-Means, and AGNES, which one is deterministic, i.e., the output does not depend on random numbers used in the algorithm. Explain why.

Answer: DBSCAN and K-Means are not deterministic, which are shown in the lecture and homework respectively. AGNES is generally deterministic because it generally does not use random numbers. AGNES is sometimes random: for example, if one point has the same distances to the other two, then ties are broken at random... □

2. [10] Data preprocessing. Consider 10 data points in 2-D space, as specified in Table 1.

X1	-2	-1	0	1	2	-2	-1	0	1	2
X2	-1	-1	-1	-1	-1	1	1	1	1	1
Class	no	no	no	no	no	yes	yes	yes	yes	yes

Table 1: Data points in 2-D space.

(a) [3] One of the principal component is $\begin{bmatrix}0\\1\end{bmatrix}$, which is visualized and labeled as $v1$ on Figure 1. Please write down the vector representation of the remaining principal component, as well as visualize and label it as $v2$ on Figure 1.

Answer: As 2 principal components should be orthonormal, the second one should be $\begin{bmatrix}1\\0\end{bmatrix}$ □. Its visualization is a line orthogonal with the given principal component.

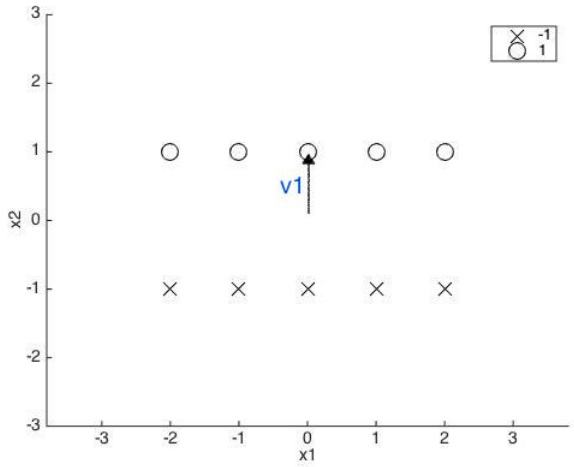

Figure 1: Visualization of data points and principal components

(b) [3] Which one should be the first principal component, i.e., the most important one? Briefly explain.

Answer: The one found in the question above ($v2$), because the projected data points on it has higher variance than the projected data points on the original one ($v1$). □

(c) [2] Use formula to find the projected point for [-2,1] on the given principal component. Please visualize it on Figure 1.

Answer: $\begin{bmatrix}0 & 1\end{bmatrix} * \begin{bmatrix}-2\\1\end{bmatrix} = \begin{bmatrix}1\end{bmatrix}$ □

(d) [2] If we want to use the dataset to train a model to predict the class label, does it make more sense to project data points on the first or the second component? Briefly explain.

Answer: The second component ($v1$), because the projected data points, which corresponds to feature X2, can classify the training set perfectly. □

3. [20] Data Warehousing, OLAP and Data Cube Computation

(a) [8] Suppose an online social website consists of five dimensions: user, location, posts information, friends, and education, and one measure: overall post count. Education refers to the schools the user has attended by 2014. Overall post count measures how active the user is.

i. Draw a **snowflake** schema diagram (sketch it, do not have to mark every possible level, and make your implicit assumptions on the levels of a dimension when you draw it).

ii. If one would like to start at the Apex cuboid and find top 3 active users in each university within Illinois state at year 2014, what are the specific OLAP operations (e.g., roll-up on which dimension from which level to which level) that one should perform based on your design?

Answer:

i. There could be many different answers in the design. One possible answer could be as follows. Dimensions:

- user(name, id, gender, age, location, ...);
- location(country, state);
- posts(content, time, source, ...);
- friends(id, time);
- education(university, time, major, ...);

The operations are as follows

- Drill-down on Location Dimension to state-level;
- Drill-down on Time Dimension to year-level;
- Dice on (i.e., select) state = "illinois" and Year = "2014";
- Drill-down on Education Dimension to University-level;
- Drill down on Student dimension to user name (or ID);
- Select top 10 overall post counts, and print the corresponding user names.

□

(b) [6] Suppose the base cuboid of a data cube contains only four cells

$$(a_1, a_2, a_3, a_4, a_5, \ldots, a_{2k})$$
$$(b_1, a_2, b_3, a_4, b_5, \ldots, a_{2k})$$
$$(c_1, a_2, c_3, a_4, c_5, \ldots, a_{2k})$$
$$(d_1, d_2, d_3, d_4, d_5, \ldots, d_{2k})$$

where $k \in \mathbb{N}_+$ (k is positive), $a_i \neq b_i \neq c_i \neq d_i$, $\forall i = 1, \ldots 2k$.

i. [2] If we set minimum support = 2, how many nonempty aggregate cells are there in the corresponding *iceberg cube*?

ii. [2] How many closed cells are there in the *iceberg cube*?

Note: Please show essential calculation steps.

Answer:

i. 2^k;

ii. 2, which are $(*, a_2, *, a_4, \ldots, a_{2k}), and (*, *, *, *, \ldots, *)$.

□

(c) [6]

i. [3] For the following tasks, which cube implementation method is better? multiway array cubing, BUC (bottom-up computation), or **neither**? Briefly explain.

A. Considering a data cube about sales in a small store, with five dimensions (customer, time, product, unit, price), fully materialize this data cube.

B. Computing a large iceberg cube of around 830 dimensions.

Answer:

A. multiway array cubing.

B. Neither. It is a problem with high dimension.

□

ii. [3] In the study of TV ratings, as it is impossible to gather the opinion of everyone in the population, the ratings analysis relies on a sample of the data for analysis. The sampling cube is a data cube structure that stores the sample data and their multidimensional aggregates. In the context of sampling cube, what does "confidence interval" mean? What can we do to boost the reliability of query answers?
Hint: The confidence interval determines the reliability of query answers. Consider the factors that influence the confidence interval.
Answer:

A. confidence interval is an estimated range of values with a given high probability of covering the true population value.

B. 1. Reduce variance, Intra-cuboid query expansion; 2. Increase sample size, Inter-cuboid query expansion.

□

4. [30] Mining Frequent Patterns

(a) [14] Given a database of 4 transactions, you are to use MaxMiner to mine the frequent max patterns. Here $min\_sup = 2$.

customer_id	shopping items
1	$abcde$
2	bde
3	aef
4	$bcde$

Table 2: Transaction Database to Mine Max Patterns

i. [4] Show the nodes you generate from the root node ($abcdef$) of your set-enumeration tree.
Answer: One possible answer: (the order of items may change.)
a(bcde), b(cde), c(de), d(e), e() □

ii. [4] When do you apply global pruning principle? Show an example from the given database. You need to show what are pruned.
Answer: When we check the frequency of the union of the head $h(N)$ and $t(N)$, i.e. $h(N) \cup t(N)$ and find its frequency pass min_sup, we do global pruning by pruning all nodes on $N's$ right at the same level. Example: When we check the frequency of $bcde$ in node $b(cde)$, we prune nodes $c(de)$, $d(e)$ and $e()$ on its right. □

iii. [4] When do you apply local pruning principle? Show an example from the given database. You need to show what are pruned.
Answer: When we check the frequency of the union of the head $h(N)$ and each item in $t(N)$, we do local pruning by pruning the itemsets whose frequencies do not pass min_sup. Example: When we check the frequency of ab in node $a(bcde)$, we prune ab because its frequency is 1. □

iv. [2] Show the max patterns in the given database.
Answer: ae, bcde □

(b) [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.

customer_id	shopping sequence
1	$a(bc)(de)f$
2	$bc(ad)ef$
3	$a(bc)d(ab)ef$

Table 3: Transaction Database to Mine Sequential Patterns

i. [4] Show $\langle b \rangle$-projected database.
Answer: (_c)(de)f, c(ad)ef, (_c)d(ab)ef □

ii. [4] What frequent patterns will you get from $\langle b \rangle$-projected database?
Answer: be, bd, bf, bdf, bef □

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

ii. [4] Containing items the sum of whose prices is less than $50.
Answer: The constraint is antimonotonic. This constraint can be mined efficiently using Apriori as follows. Only candidates the sum of whose prices is less than $150 need to be checked. □

(c) [8] Suppose a sequential 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. You are going to use PrefixSpan to mine the frequent sequential patterns.

5. [40] Classification

ID	Message	Label
1	save money coupon	Yes
2	coupon money visit	Yes
3	pay you money	No
4	save you	No
5	pay you visit	No

Table 4: Training data for Question 5a and 5b

(a) [7] Suppose we want to build a **Naive Bayes** classifier to filter spam messages. We collect five messages with labels as in **Table 4**. As the first step, you need to do pre-processing on the raw data. Please turn the data into a format upon which you can build a Naive Bayes classifier. Show the pre-processed results. (Hint: Generate a feature vector for each message).
Answer: We can use each word in the data as an attribute, and pre-process data into the following format: X_1 indicates whether the word 'save' appears in the message; similarly, $X_2 \sim X_6$ for 'money', 'coupon', 'visit', 'pay' and 'you'. □

(b) [7] Based on the pre-processed data in **Question 5a**, we want to construct a Naive Bayes classifier. If no smoothing is applied, we'll have a problem in classifying the short message "save you coupon". State the problem via calculation, and outline a solution to the problem.

X_1	X_2	X_3	X_4	X_5	X_6	Label
1	1	1	0	0	0	1
0	1	1	1	0	0	1
0	1	0	0	1	1	-1
1	0	0	0	0	1	-1
0	0	0	1	1	1	-1

Table 5: Solution to question 5a

Answer: Since the word "you" only appears in non-spam messages, and "coupon" only appears in spam messages. As a results, Pr("save you coupon" | "spam") = Pr("save you coupon" | "non-spam") = 0.

To avoid the zero-probability problem, we can apply smoothing on each attribute. Students need to outline the procedures of smoothing □

A\P	yes	no	Total
yes	180	20	200
no	320	480	800
Total	500	500	1,000

Table 6: A confusion matrix for the classes *spam* = yes and *spam* = no. 'A' represents "Actual class", and 'P' represents "Predicted class".

(c) [6] Suppose we have built a classifier to filter spam messages, and evaluated its performance. We summarize the evaluation results in the form of confusion matrix as in **Table 6**. Please answer the following questions:

i. Compute the precision and recall of the classifier.

ii. Based on the calculation, is it a useful classifier? Explain why or why not for this particular task.

Answer: TP = 180, FN = 20, FP = 320, TN = 480

- Precision = $\frac{TP}{TP+FP} = \frac{180}{180+320} = 36\%$
- Recall = $\frac{TP}{FP+FN} = \frac{180}{180+20} = 90\%$

No. It's not a useful classifier, since the precision is quite low. It means that many normal text messages are classified as "spam", which is not desirable. □

x_1	x_2	y
1	0	+1
−1	0	+1
0	1	−1
0	−1	−1

Table 7: Training data for Question 5d and 5e

(d) [14] Suppose we have four training points, which are listed in **Table 7**. Each point has two attributes, i.e., x_1 and x_2, and a label y. You'll apply Adaboost Algorithm in this question. Suppose, initially, the weight for each point is uniform, i.e., $w_1(j) = \frac{1}{4}, j = 1, ..., 4$. In each round, we sample with replacement according to the weights, and get a training datset $D_i, i = 1, 2, 3$. Suppose, based on D_i, we learn a classifier $M_i, i = 1, 2, 3$, which has the following rule:

$$M_1: \hat{y} = \begin{cases} +1 & \text{if } x_1 \leq -0.5 \\ -1 & \text{if } x_1 > -0.5 \end{cases} \quad M_2: \hat{y} = \begin{cases} +1 & \text{if } x_2 \leq 0.5 \\ -1 & \text{if } x_2 > 0.5 \end{cases} \quad M_3: \hat{y} = \begin{cases} +1 & \text{if } x_1 \geq 0.5 \\ -1 & \text{if } x_1 < 0.5 \end{cases}$$

Please answer:

i. Compute the (weighted) error ϵ_i of model M_i $(i = 1, 2, 3)$, and the weights for each point after the first and second round, i.e., w_2 and w_3.

ii. Combine the three classifiers M_1 to M_3 based on your calculation in the above step. What's the error of this combined classifier on the training dataset?

Answer: The computation is shown in Table 8 to 10.

x_1	x_2	y	w_1		$M_1: \hat{y}$	w_2	normalized w_2
1	0	+1	$\frac{1}{4}$		-1	$\frac{1}{4}$	$\frac{1}{2}$
-1	0	+1	$\frac{1}{4}$		+1	$\frac{1}{12}$	$\frac{1}{6}$
0	1	-1	$\frac{1}{4}$		-1	$\frac{1}{12}$	$\frac{1}{6}$
0	-1	-1	$\frac{1}{4}$		-1	$\frac{1}{12}$	$\frac{1}{6}$

Table 8: The error of round 1 $\epsilon_1 = \frac{1}{4}$ and $\frac{\epsilon_1}{1-\epsilon_1} = \frac{1}{3}$

x_1	x_2	y	w_2		$M_2: \hat{y}$	w_3	normalized w_3
1	0	+1	$\frac{1}{2}$		+1	$\frac{1}{10}$	$\frac{3}{10}$
-1	0	+1	$\frac{1}{6}$		+1	$\frac{1}{30}$	$\frac{1}{10}$
0	1	-1	$\frac{1}{6}$		-1	$\frac{1}{30}$	$\frac{1}{10}$
0	-1	-1	$\frac{1}{6}$		+1	$\frac{1}{6}$	$\frac{1}{2}$

Table 9: The error of round 2 $\epsilon_2 = \frac{1}{6}$ and $\frac{\epsilon_2}{1-\epsilon_2} = \frac{1}{5}$

x_1	x_2	y	w_3		$M_3: \hat{y}$
1	0	+1	$\frac{3}{10}$		+1
-1	0	+1	$\frac{1}{10}$		-1
0	1	-1	$\frac{1}{10}$		-1
0	-1	-1	$\frac{1}{2}$		-1

Table 10: The error of round 3 $\epsilon_3 = \frac{1}{10}$ and $\frac{\epsilon_3}{1-\epsilon_3} = \frac{1}{9}$

The weight for each classifier is $\alpha_1 = log(3), \alpha_2 = log(5), \alpha_3 = log(9)$, so the combined classifier is $M(x) = \text{sign}(\log(3)M_1(x) + \log(5)M_2(x) + \log(9)M_3(x))$. If we applied the combined classifier

on the training dataset, we have the following Table 11. We can conclude from Table 11 that the combined classifier M can correctly classify the training dataset.

y		M1	M2	M3	M
$+1$		-1	$+1$	$+1$	$+1$
$+1$		$+1$	$+1$	-1	$+1$
-1		-1	-1	-1	-1
-1		-1	$+1$	-1	-1

Table 11: Apply the combined classifier to the training dataset

□

(e) [6] Perceptron algorithm iteratively updates the weights of a linear classifier until certain condition is satisfied. Please answer:

i. State this stopping condition.

ii. For the training data in **Table 7**, will the Perceptron algorithm terminate? Why or why not? (Hint: plot the points)

Answer: The Perceptron algorithm will stop if it finds a linear classifier that can correctly classify all the training data. Since the training data in **Table 5d** is not linearly separable, the algorithm won't terminate. You can find the plot in Figure 2.

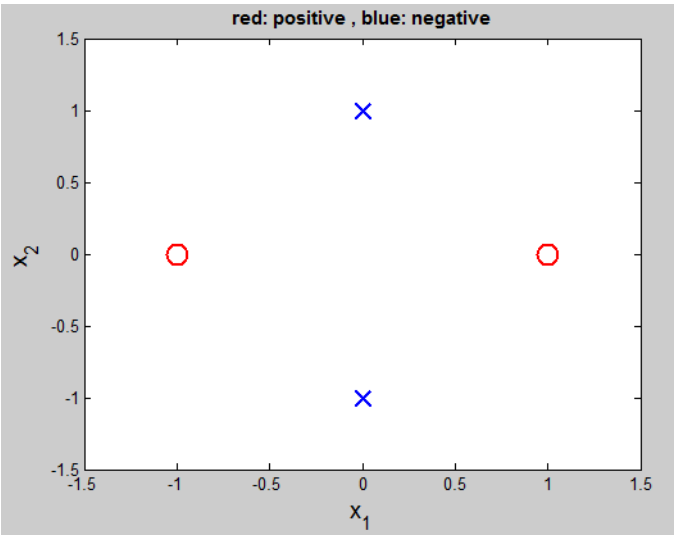

Figure 2: Solution to Question 5e

.

□

6. [20] Clustering

Suppose we have 11 data points, which are listed and plotted in Figure 3. The ground truth (the correct cluster output) is also provided.

(a) [3] If we perform AGNES, a hierarchical clustering algorithm on the points above, using single link method and adopting Euclidean distance as the dissimilarity measure, how many levels will

Point	x	y	Cluster
P1	1	1	C3
P2	1	3	C1
P3	1	4	C1
P4	1	5	C1
P5	0	4	C1
P6	2	4	C1
P7	3	4	C2
P8	4	4	C2
P9	5	4	C2
P10	4	3	C2
P11	4	5	C2

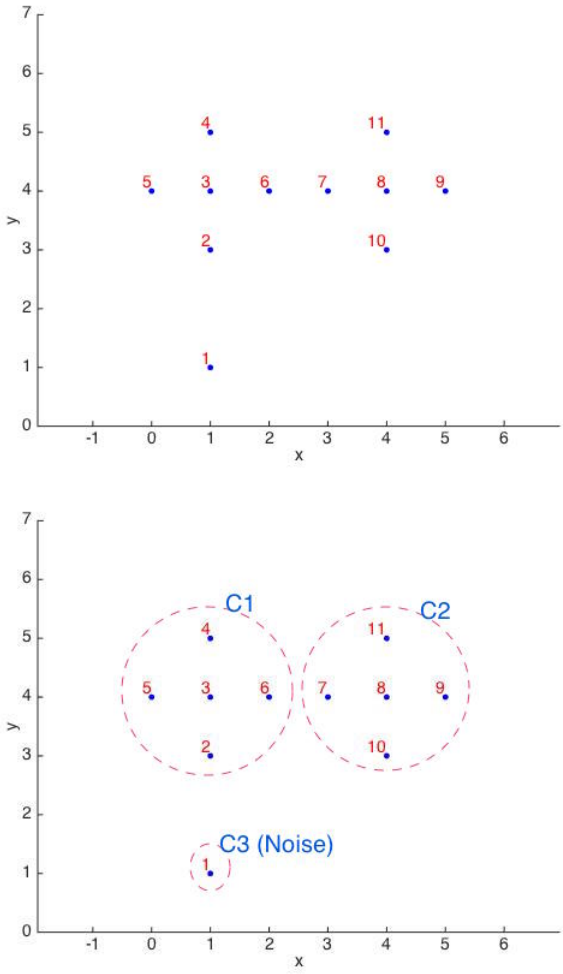

Figure 3: Clustering data and ground truth

the output diagram have? What are the corresponding values? Hint: each level of a dendrogram corresponds to a dissimilarity value, and you do not have to show how you perform AGNES.

Answer: Two levels: 1.0 and 2.0. □

(b) [2] If we don't know the ground truth, and want to cluster the data set into 2 groups, based on the result above, what are the members of the 2 groups?

Answer: Clusters:

- C1: P2, P3, P4, P5, P6, P7, P8, P9, P10, P11
- C2: P1

□

(c) [4] Based on the given ground truth, what are the B-Cubed precision and recall of the output? Show your computation.

Answer:

Point i	1	2	3	4	5	6	7	8	9	10	11
P_i	1/1	5/10	5/10	5/10	5/10	5/10	5/10	5/10	5/10	5/10	5/10
R_i	1/1	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	5/5	

$$Precision = (1/11) * \sum_{i=1}^{11} P_i = 1/11 * (1/1 + 5 * 5/10 + 5 * 5/10) = 6/11$$

$$Recall = (1/11) * \sum_{i=1}^{11} R_i = 1/11 * (1/1 + 5 * 5/5 + 5 * 5/5) = 1$$

□

(d) [3] Using the same data as in Figure 3, show the clustering result of K-Means by annotating Figure 4, with $K = 2$, initial means: P3 and P8, and Euclidean distance as the distance function. Note: You do not have to show your computation or to illustrate the means.

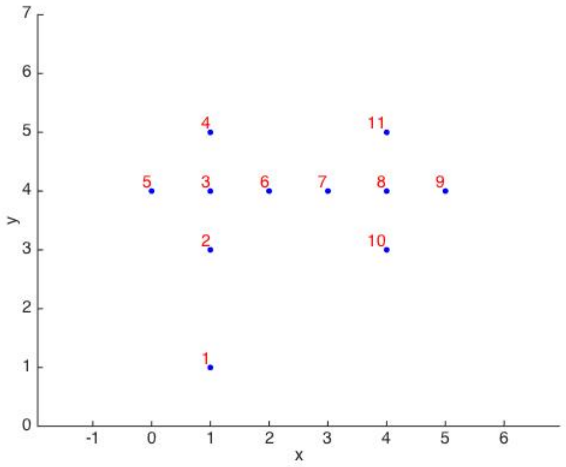

Figure 4: Output of K-Means

Answer: As shown in Figure 5 □

(e) [5] Using the same data as in Figure 3, perform DBSCAN, a density-based algorithm, with $MinPts = 4$, $\epsilon = 1.1$, and random points: P1, then P5, then P3, then P8. You need to list all the steps, and the final clusters.

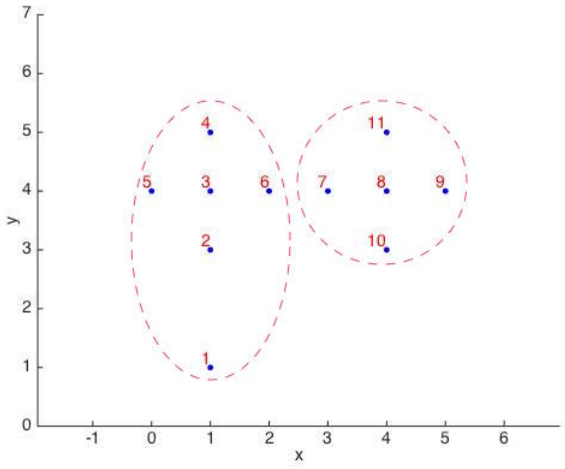

Figure 5: Solution for output of K-Means

Answer: As the ϵ-neighborhoods of P1 and P5 have smaller number of objects than $MinPts$, they are marked as noise.

Cluster 1:

C1 = {P3}	N = {P2, P4, P5, P6}
C1 = {P3, P2}	N = {P4, P5, P6}
C1 = {P3, P2, P4}	N = {P5, P6}
C1 = {P3, P2, P4, P5}	N = {P6}
C1 = {P3, P2, P4, P5, P6}	N = {}
C2 = {P8}	N = {P7, P9, P10, P11}
C2 = {P8, P7}	N = {P9, P10, P11}
C2 = {P8, P7, P9}	N = {P10, P11}
C2 = {P8, P7, P9, P10}	N = {P11}
C2 = {P8, P7, P9, P10, P11}	N = {}

No unvisited points left → stop → final clusters:

- C1: P2, P3, P4, P5, P6
- C2: P7, P8, P9, P10, P11
- C3 (Noise): P1

Note: As shown in the practice questions, students might save some time by writing only the new member of the cluster in each step. For example, in line 3, instead of writing C1 = {P3, P2, P4}, students might write C1 = {..., P4}. □

(f) [3] Which clustering result is better, the one from K-Means or the one from DBSCAN? Which characteristics of the dataset causes that difference?

Answer: The noise of the dataset. DBSCAN can detect it, thanks to $MinPts$ parameter. □