Sample Endterm Exam Data Mining, IBI008

- Write your name and student number on this sheet and your name on all the others.
- The exam is multiple choice.
- All questions have four possible answers, marked by the letters **A**, **B**, **C**, and **D** as well as the answer "Don't know" marked by the letter **E**.
- The correct answer gives 3 points, a wrong answer gives -1 point, and "Don't know" (**E**) gives 0 points.
- Clearly circle your answer.
- NB: in some questions you are asked to identify the *correct* statement and in others the *incorrect* statement.
- All questions are weighted equally.
- Feel free to fill in your answers in the table below (these then count, so be careful and only do so when you have time left).
- Hand in all sheets.
- Good luck!

Name: _____ *Student number:* _____

Question:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Your answer:																	
Score:																	

	Bread	Milk	Beer	Diaper	Juice	Jam
C1	1	1	0	1	1	0
C2	0	1	1	1	0	0
C3	1	1	0	1	1	0
C4	1	0	1	1	0	1
C5	0	1	0	0	1	1

Table 1: Market basket data.

1. In a store five consumers, denoted C1 through C5, have purchased the items given in Table 1, where 1 indicate that a given consumer purchased the item and 0 that the consumer did not purchase the item.

   What are the itemsets with support greater than 50%?

   A. {Milk},{Diaper}.

   B. {Beer}, {Jam}.

   C. {Bread}, {Milk}, {Diaper}, {Juice}.

   D. {Bread}, {Milk}, {Diaper}, {Juice}, {Bread,Diaper}, {Milk,Diaper}, {Milk,Juice}.

   E. Don't know.

2. For the data in Table 1: what is the confidence of the rule {Bread,Milk} → {Diaper} and what is the confidence of the rule {Jam} → {Beer}?

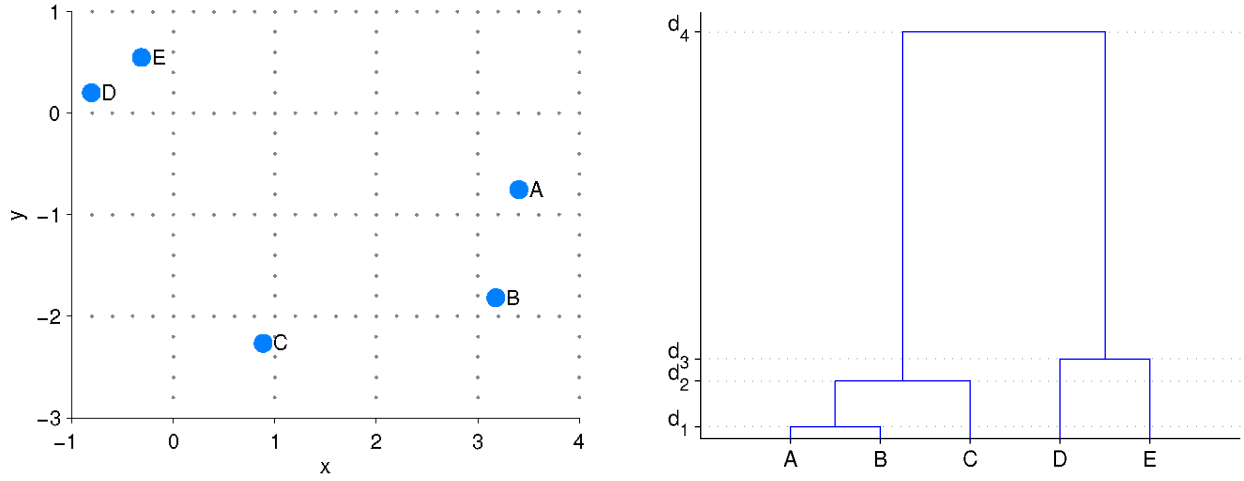
   A. {Bread,Milk} → {Diaper} has a confidence of 40% and {Jam} → {Beer} has a confidence of 20%.

   B. {Bread,Milk} → {Diaper} has a confidence of 40% and {Jam} → {Beer} has a confidence of 40%.

   C. {Bread,Milk} → {Diaper} has a confidence of 80% and {Jam} → {Beer} has a confidence of 40%.

   D. {Bread,Milk} → {Diaper} has a confidence of 100% and {Jam} → {Beer} has a confidence of 50%.

   E. Don't know.



(a) (b)

Figure 1: (a) A scatter plot. (b) A dendrogram.

	A	B	C	D	E
A	0.00	0.05	0.44	2.00	1.67
B	0.05	0.00	0.22	1.96	1.86
C	0.44	0.22	0.00	1.58	1.99
D	2.00	1.96	1.58	0.00	0.31
E	1.67	1.86	1.99	0.31	0.00

Table 2: Cosine distances for the data in Figure 1(a).

3. Consider the data illustrated in the scatter plot in Figure 1(a). We want to cluster the data using agglomerative hierarchical clustering based on the cosine distance measure, $d(A, B) = 1 - \cos(A, B)$, using either single linkage (MIN) or complete linkage (MAX). The pairwise distances are shown in Table 2.

   In the agglomerative hierarchical clustering approach, under which of the two cluster similarity measures (MIN and MAX), are A and B the first cluster to be merged?

   A. Neither single linkage (MIN), nor complete linkage (MAX)

   B. Single linkage (MIN), but not complete linkage (MAX).

   C. Complete linkage (MAX), but not single linkage (MIN).

   D. Both single linkage (MIN), and complete linkage (MAX).

   E. Don't know.

4. Continuing question 3: having decided to use *single* linkage (MIN), the resulting clustering is shown as a dendrogram in Figure 1(b). What is the distance between the *second* clusters merged and the *final* clusters merged, that is, what is the value of $d_{24} = |d_4 - d_2|$ in the dendrogram in Figure 1(b)?

   A. $d_{24} = 1.36$.

   B. $d_{24} = 1.58$.

   C. $d_{24} = 1.67$.

   D. $d_{24} = 1.81$.

   E. Don't know.

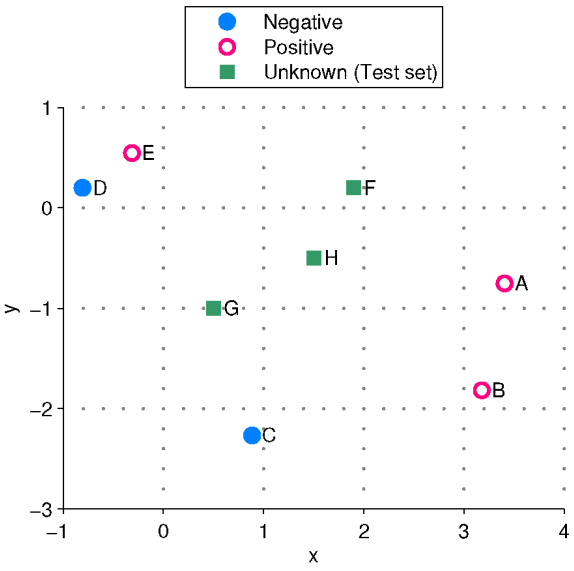


Figure 2: A scatter plot with class labels and test data points.

	A	B	C	D	E
F	0.05	0.19	0.74	1.94	1.40
G	0.37	0.17	0.00	1.65	2.00
H	0.01	0.02	0.36	2.00	1.75

Table 3: Cosine distances for the test data points in Figure 2.

5. In the following we will use the k-nearest neighbors algorithm based on cosine distances to classify the data in Figure 2.

   Using the cosine distances in Table 3, how will F, G, and H be classified using the nearest neighbor classifier (1 nearest neighbor)?

   A. F=Negative, G=Positive, H=Negative.

   B. F=Positive, G=Negative, H=Positive.

   C. F=Positive, G=Positive, H=Positive.

   D. F=Negative, G=Negative, H=Negative.

   E. Don't know.

6. Using again the cosine distances in Table 3, how will F, G, and H be classified using the 3-nearest neighbor classifier?

   A. F=Negative, G=Positive, H=Negative.

   B. F=Positive, G=Negative, H=Positive.

   C. F=Positive, G=Positive, H=Positive.

   D. F=Negative, G=Negative, H=Negative.

   E. Don't know.

7. We will use leave-one-out cross-validation to determine the best value of k for the k-nearest neighbor algorithm.

   Using the labeled objects (A through E) in Figure 2, what is the leave-one-out estimate of the generalization error (classification error rate) for the $k = 1$ nearest neighbor classifier based on the cosine distances given in Table 2?

   A. 0.2.

   B. 0.4.

   C. 0.6.

   D. $\frac{2}{3} \approx 0.667$.

   E. Don't know.

8. Using the labeled objects (A through E) in Figure 2, what is the leave-one-out estimate of the generalization error (classification error rate) for the $k = 3$ nearest neighbor classifier based on the cosine distances given in Table 2?

   A. 0.2.

   B. 0.4.

   C. 0.6.

   D. $\frac{2}{3} \approx 0.667$.

   E. Don't know.

Training set

	s	t	u	v	c
A	1	1	1	0	Positive
B	0	0	1	0	Positive
C	0	1	0	0	Negative
D	1	0	1	1	Negative
E	1	0	0	1	Positive

(a)

Test set

	s	t	u	v	c
F	1	1	1	1	?
G	0	1	0	1	?
H	0	0	1	1	?

(b)

Table 4: Binary attributes, training (a) and test set (b).

9. Consider the training objects (A through E) in Table 4(a), which have four binary attributes, s, t, u, and v. Each object also has an associated class label, c. We will classify the test objects (F through H) using a naïve Bayes classifier:

$$P(c = \text{Positive}|F) = \frac{P(c)P(s=1|c)P(t=1|c)P(u=1|c)P(v=1|c)}{\sum_{c' \in \{\text{Positive,Negative}\}} P(c')P(s=1|c')P(t=1|c')P(u=1|c')P(v=1|c')}.$$

The probabilities needed for the classifier are estimated as the corresponding fraction of training instances in Table 4(a), for example

$$P(c = \text{Positive}) = \frac{3}{5} \quad \text{and} \quad P(s = 1|c = \text{Positive}) = \frac{2}{3}\,.$$

In the naïve Bayes classifier described above, what is the posterior probability that object F from Table 4(b) is Positive?

A. $P(c = \text{Positive}|F) \approx 0.03$.

B. $P(c = \text{Positive}|F) \approx 0.54$.

C. $P(c = \text{Positive}|F) \approx 0.60$.

D. $P(c = \text{Positive}|F) \approx 0.72$.

E. Don't know.

10. Using the naïve Bayes classifier described in Question 9, how will F, G, and H from Table 4(b) be classified?

    A. F=Negative, G=Positive, H=Negative.

    B. F=Positive, G=Negative, H=Positive.

    C. F=Positive, G=Positive, H=Positive.

    D. F=Negative, G=Negative, H=Negative.

    E. Don't know.

11. We would like to combine classifiers in order to improve the classification performance. Which of the following statements is *correct*?

    A. Combining classifiers will in general not improve the classification performance if their errors are independent.

    B. In Boosting the data set is sampled with replacement from a uniform distribution in each Boosting round.

    C. In Bagging the data set is sampled without replacement from a uniform distribution.

    D. The main difference between Bagging and Boosting is the distribution from which the data is sampled.

    E. Don't know.

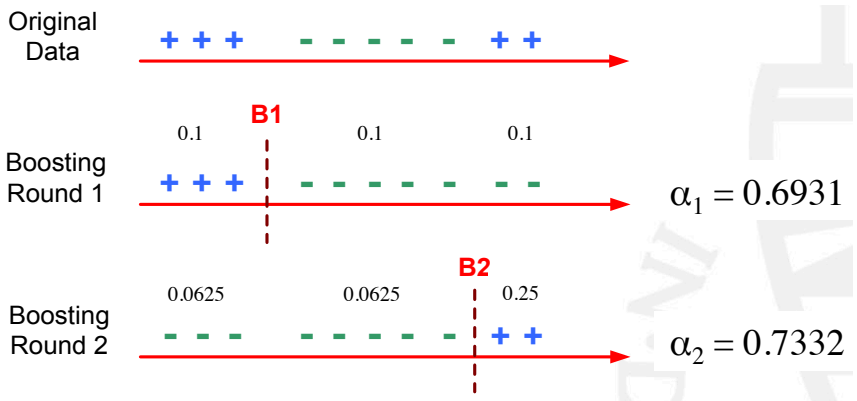


Figure 3: ADAboost after 2 rounds

12. The AdaBoost algorithm is a ensemble classifier that implements the boosting approach. At each round a base classifier is learned, and its relative importance established based on the corresponding error rate. Figure 3 depicts the result of two boosting rounds using binary base classifiers. The Figure shows the weights of the points at each round, as well as the importance α_i of the corresponding base classifier.

    For the AdaBoost example in Figure 3 with importance weighted majority voting scheme, after 2 rounds of boosting, how many points (out of 10) are classified *incorrectly* by the resulting ensemble classifier at that stage?

    A. 0

    B. 1

    C. 2

    D. 3

    E. Don't know.

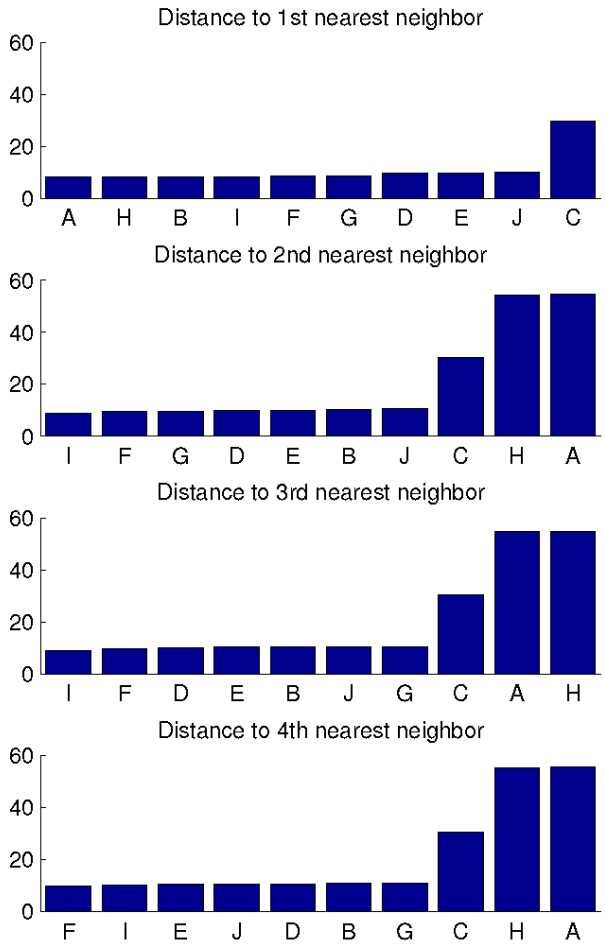


Figure 4: Distance to kth nearest neighbor.

13. We consider a data set with 10 objects, denoted A through J, that theoretically should lie in one cluster. However, we suspect that there might be one or more outliers in the data set. To examine this, we plot in Figure 4 the distance to the kth nearest neighbor for $k = \{1, 2, 3, 4\}$, from top to bottom.

    Based on these plots, which statement is *incorrect*?

    A. C is an outlier.

    B. A is an outlier that is closer to C than to H.

    C. H is an outlier that is closer to A than to C.

    D. I is not an outlier.

    E. Don't know.

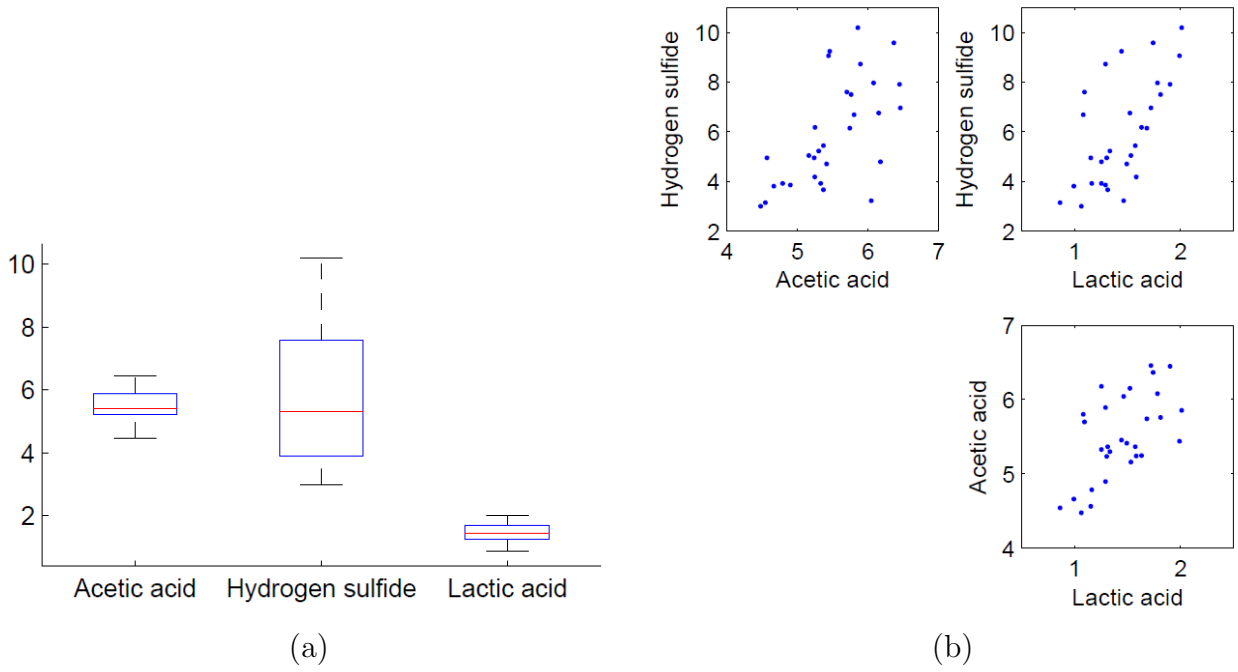


Figure 5: Box plots (a) and scatter plots (b) of a data set with measurements of three attributes (acetic acid, hydrogen sulfide, and lactic acid) for 30 cheddar cheeses.

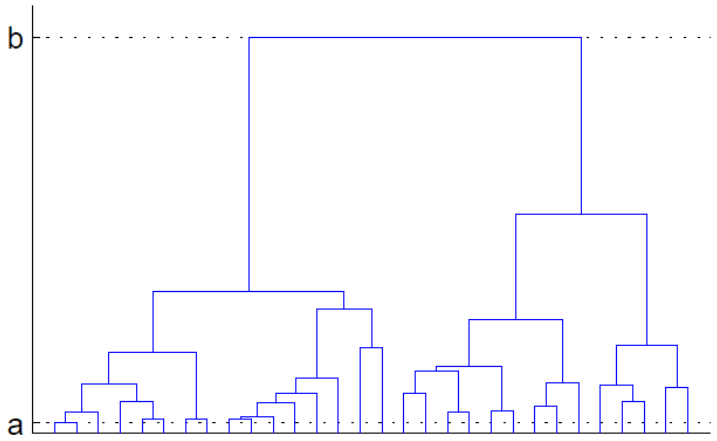


Figure 6: Dendrogram obtained by running hierarchical clustering on the data from Figure 5.

14. We consider a data set containing measurements of three attributes (acetic acid, hydrogen sulfide, and lactic acid) for 30 cheddar cheeses. Box plots of the attributes are shown in Figure 5(a), and scatter plots of the attributes are shown in Figure 5(b).

    We wish to examine if there exists some group structure in this data. To do this, we run hierarchical clustering with complete linkage using the Euclidean distance measure. The result of the clustering is shown in Figure 6 as a dendrogram.

    What values of a and b in the dendrogam in Figure 6 agree with the data visualized in Figure 5(b)? Hint: by looking at the data in Figure 5(b), you should be able to rule out the three incorrect answers.

    A. a = 0.3 and b = 19.2.

    B. a = 0.2 and b = 7.4.

    C. a = 2.3 and b = 5.4.

    D. a = 2.5 and b = 8.2.

    E. Don't know.

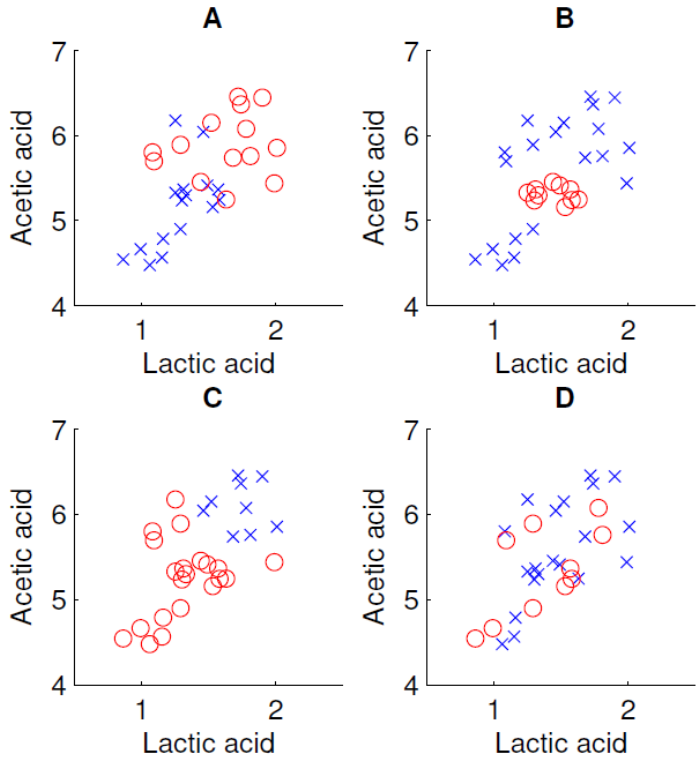


Figure 7: Four different clusterings on the data from Figure 5.

15. Figure 7 shows four different clusterings as scatter plots of lactic acid versus acetic acid for the cheese data set described in Question 14. One of the clusterings shown in Figure 7 corresponds to the clustering that was found by the top level of the hierarchical clustering shown in Figure 6 (where the data is clustered into two groups). Which clustering is this?

    A. Clustering A.

    B. Clustering B.

    C. Clustering C.

    D. Clustering D.

    E. Don't know.

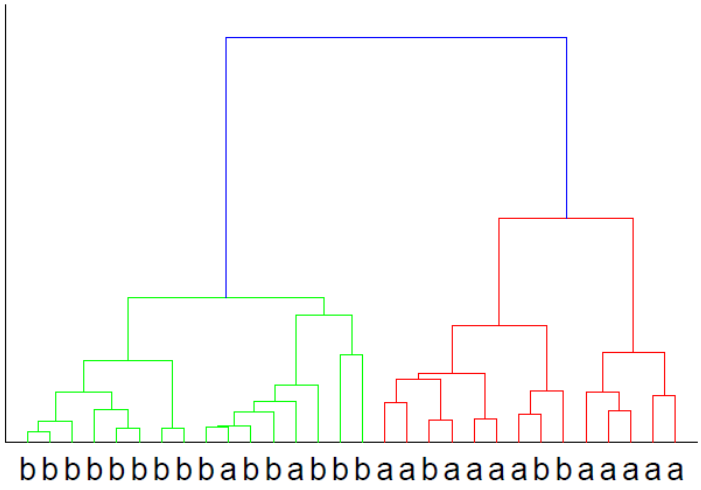


Figure 8: Dendrogram on the data from Figure 5, with labels **a** (above average) and **b** (below average) corresponding to the average taste score given to each cheese by a group of experts.

16. Experts have tasted the 30 cheddar cheeses from Question 14 and given them a taste score. For each cheese an average score has been computed, and finally the cheeses have been categorized into two groups: those that scored above average (denoted **a**) and those that scored below average (denoted **b**). The expert categories are indicated in the dendrogram in Figure 8.

    We now want to examine if our clustering reflects these two taste categories. We consider the level of the hierarchical clustering in Figure 8 where the data is clustered into two groups. We recall that the purity (a supervised measure of cluster validity) is given by

$$\text{purity} = \sum_{i=1}^{K} \frac{m_i}{m} \max_j p_{ij} \ ,$$

    where K is the number of clusters, m_i is the number of objects in the ith cluster, m is the total number of objects, and p_{ij} is the fraction of objects in cluster i that belong to class j.

    What is the purity of the clustering?

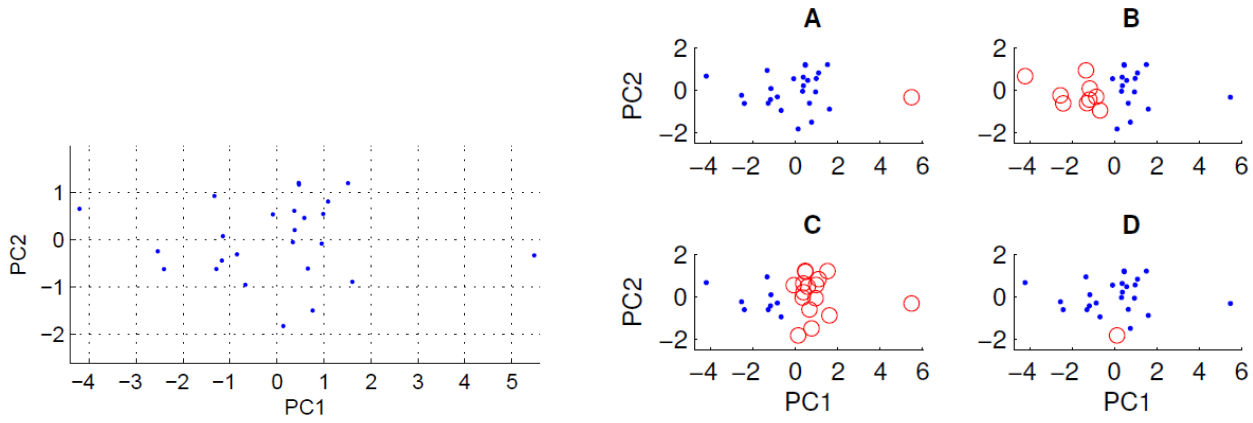
    A. $\frac{27}{32} \approx 0.844$.

    B. $\frac{187}{224} \approx 0.835$.

    C. $\frac{93}{112} \approx 0.830$.

    D. $\frac{5}{6} \approx 0.833$.

    E. Don't know.



(a) (b)

Figure 9: (a) Scatter plot of two-dimensional data with attributes PC1 and PC2. (b) Four different clusterings of the data in (a).

17. Figure 9(a) visualizes data projected onto the first two principal components, PC1 and PC2. We wish to examine if there exists some group structure in the data set. To do this, we have run the k-means algorithm with k = 2 using the Euclidean distance measure on the two-dimensional data set with PC1 and PC2 as attributes. The result of the k-means algorithm can vary between runs depending on the random initial conditions. We therefore ran the algorithm several times and got some different clustering results.

    Only three of the four clusterings shown in Figure 9(b) can be obtained by running the k-means algorithm on the cigarette data. Which one can*not* be obtained by k-means?

    A. A

    B. B

    C. C

    D. D

    E. Don't know.

DONE!