Feedback — Week 2 Quiz

Help Center

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Sat 9 May 2015 11:38 AM PDT**. You got a score of **8.75** out of **10.00**.

Question 1

Given Eps and Minpts, if a point p is directly density-reachable from a point q, which of the following statements are correct?

Your Answer		Score	Explanation
☐ Point q is density reachable from p	✔	0.20	
☐ Point p may not be density reachable from q	✔	0.20	
☑ Point p is density-connected to q	✔	0.20	
☑ Point q may not be directly density-reachable from p	✔	0.20	
☑ Point q is density-connected to p	✔	0.20	
Total		1.00 / 1.00	

Question Explanation

The following statements are correct:

- Point q may not be directly density-reachable from p. (Correct since p may not be core point.)
- Point p is density-connected to q. (Correct since both p and q are density-reachable from the point q.)
- Point q is density-connected to p. (Correct since both p and q are density-reachable from the point q.)

The following statements are incorrect:

- Point p may not be density reachable from q. (Incorrect since p must be density-reachable from q, and the chain contains only p and q.)
- Point q is density reachable from p. (Incorrect since p may not be a core point.)

Question 2

Suppose Eps = 1cm and Minpts = 1. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Which of the following statements are correct?

Your Answer		Score	Explanation
☐ Points p and q must be in the same cluster	✔	0.25	
☐ Points p and q may not belong to any clusters	✔	0.25	
☐ Points p and q must be in different clusters	✔	0.25	
☑ Points p and q may be in the different clusters	✔	0.25	
Total		1.00 / 1.00	

Question Explanation

The following statements are correct:

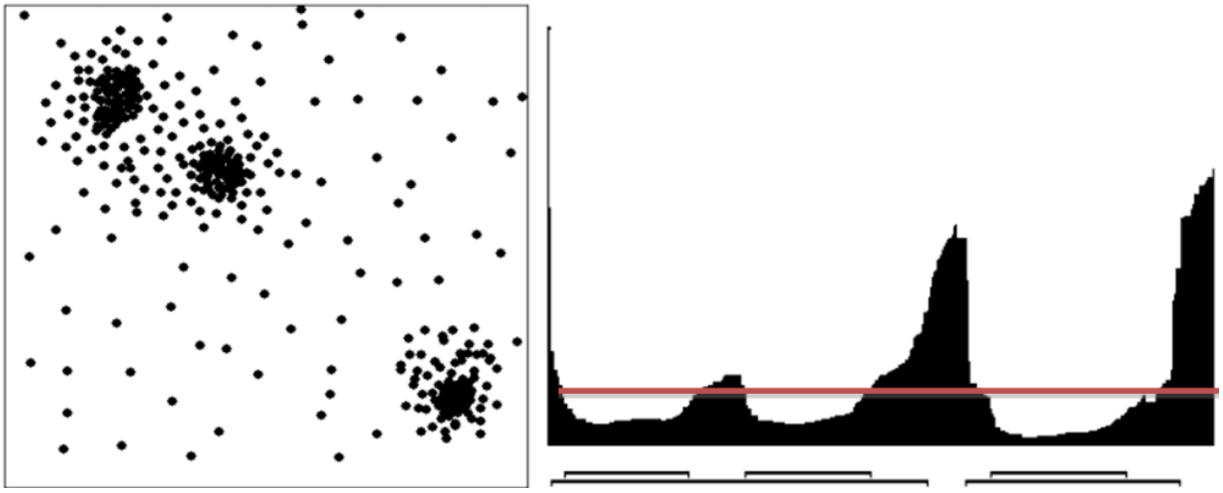
- Points p and q may be in the different clusters. (Correct. For instance, when p and q are the only two points in the dataset, since Minpts = 1, thus p and q form two clusters.)

The following statements are incorrect:

- Points p and q must be in different clusters. (Incorrect when there exists a point o from which both p and q are density-reachable. Note that any point in the dataset for Minpts = 1 is a core point.)
- Points p and q must be in the same cluster. (Incorrect. For instance, there are only two points p and q in the dataset.)
- Points p and q may not belong to any clusters . (Since p and q are both core points, it is impossible for them to be noise/outliers, which implies p and q must belong to some clusters.)

Question 3

Given the following synthetic data set (left) and the reachability-plot (right), how many clusters would there be if we set the threshold to be the redline in the right figure?



Your Answer		Score	Explanation
2			
4			
5			
3	✔	1.00	
1			
Total		1.00 / 1.00	

Question Explanation

Since there are three valleys below the red line, it implies that there would be two clusters.

Question 4

Considering the CLIQUE clustering algorithm, which of the following statements are correct?

Your Answer		Score	Explanation
Any point that does not belongs to a cluster in a 1-D subspace A may belong to some cluster (dense region) in any 2-D subspaces that includes A	✔	0.25	

☑ Any point that does not belongs to a cluster in a 1-D subspace A must not belong to any clusters (dense regions) in any 2-D subspaces that include A.	✔	0.25
☑ Any point that belongs to a cluster in a 1-D subspace A may not belong to any clusters (dense regions) in any 2-D subspaces that include A.	✔	0.25
☐ Any point that belongs to a cluster in a 1-D subspace A must belong to some cluster (dense region) in any 2-D subspaces that includes A	✔	0.25
Total		1.00 / 1.00

Question Explanation

The following statements are correct:

- Any point that belongs to a cluster in a 1-D subspace A may not belong to any clusters (dense regions) in any 2-D subspaces that include A. (Correct since the other dimension in a 2-D subspace may belong to sparse regions, i.e., does not belong to any clusters.)
- Any point that does not belongs to a cluster in a 1-D subspace A must not belong to any clusters (dense regions) in any 2-D subspaces that include A. (Correct for any data point, as long there is one dimension that belongs to the sparse region, the data point will not belong to any dense region in higher dimensions that includes A.)

The following statements are incorrect:

- Any point that belongs to a cluster in a 1-D subspace A must belong to some cluster (dense region) in any 2-D subspaces that includes A; (Incorrect since the other dimension in a 2-D subspace may belong to sparse regions, i.e., does not belong to any clusters.)
- Any point that does not belongs to a cluster in a 1-D subspace A may belong to some cluster (dense region) in any 2-D subspaces that includes A; . (Incorrect for any data point, as long there is one dimension that belong to the sparse region, the data point will not belong to any dense region in higher dimension that includes A.)

Question 5

In Gaussian Mixture Model, the random variables are univariate. For i-th cluster, where i = 1… k, the data points are drawn from Gaussian distribution,

$x \sim N(\mu_i, \sigma_i)$, i.e., $P(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$. Suppose we have learned two clusters from some dataset with the parameters $\mu_1=0, \sigma_1=1, \mu_2=3, \sigma_2=2, P(C_1)=0.6, P(C_2)=0.4$. For the new data point x=1, what is the probability that it belongs to the first cluster, i.e. $P(C_1|x)$?

Your Answer		Score	Explanation
0.5			
0.75			
0.25			
0.6	✖	0.00	
Total		0.00 / 1.00	

Question Explanation

Please note the texts of this question have been updated.

$P(C_1|x) \propto P(x|C_1 = N(\mu_1, \sigma_1))P(C_1) = 0.6 * \frac{1}{\sqrt{2\pi}*1} * e^{-\frac{(1-0)^2}{2*1^2}}$. Similarly, $P(C_2|x) \propto 0.4 * \frac{1}{\sqrt{2\pi}*2} * e^{-\frac{(1-3)^2}{2*2^2}}$. Because $P(C_1|x) + P(C_2|x) = 1$, we have $P(C_1|x) = \frac{0.6}{0.6+0.2} = \frac{3}{4} = 0.75$.

Question 6

Which of the following statements about Gaussian Mixture Models (GMM) are correct?

Your Answer		Score	Explanation
GMM assumes the data points are generated by some Poisson distributions.	✔	0.25	
To speed up GMM, one can take the clustering results from other efficient clustering methods, such as k-means, to set up a suitable initialization.	✔	0.25	
GMM cannot automatically estimate the number of clusters.	✖	0.00	

☑ GMM usually converges much slower than K-means because it updates the parameters more smoothly.	✔	0.25
Total		0.75 / 1.00

Question Explanation

GMM assumes the data points are generated by some Gaussian distributions instead of Poisson.

Question 7

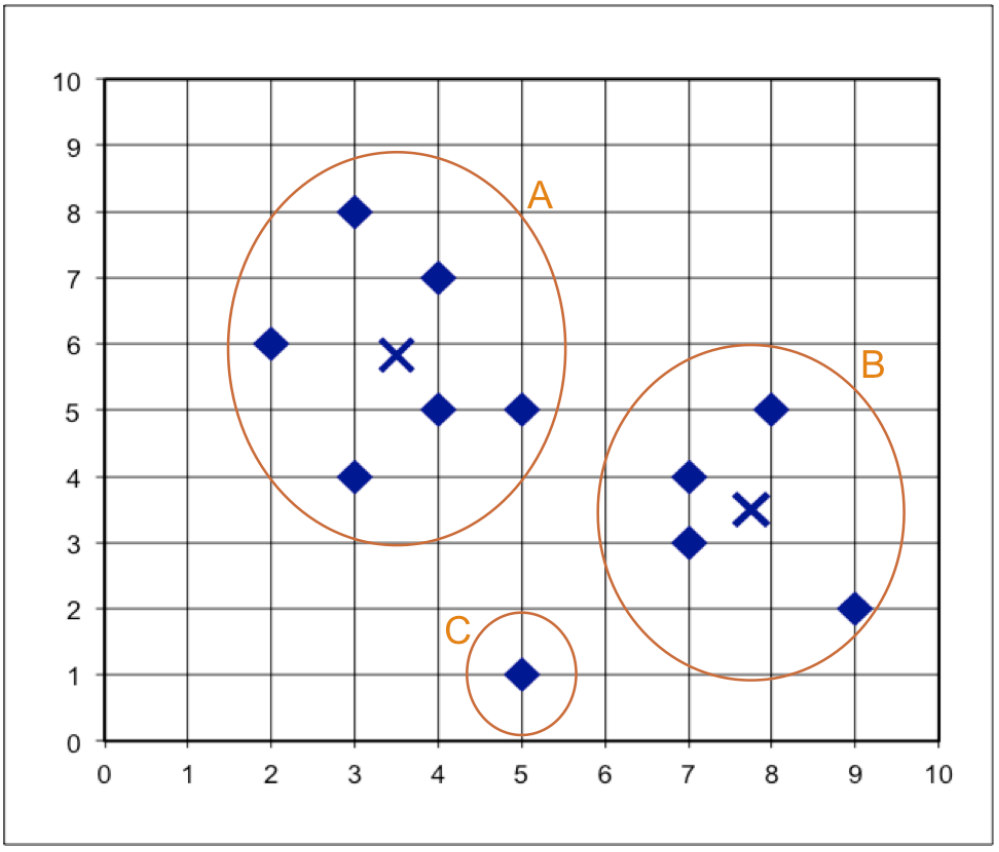
In which ONE of the following scenarios may Gaussian Mixture Models (GMM) not work properly?

Your Answer		Score	Explanation
◉ Build 2 clusters from 2000 10-dimensional data points randomly generated from 10 Gaussian distributions.	✔	1.00	
○ Build 10 clusters from 2000 10-dimensional data points randomly generated from 10 Gaussian distributions.			
Total		1.00 / 1.00	

Question Explanation

Setting the number of clusters in GMM is a little tricky.

Question 8



Consider the three clusters A, B, and C shown in Figure 1. Using Euclidean distance as the similarity measure, which two clusters would be merged first in agglomerative clustering using complete link (diameter)?

Your Answer		Score	Explanation
○ A and B			
○ All three options above are tied			
◉ B and C	✔	1.00	
○ A and C			
Total		1.00 / 1.00	

Question Explanation

B and C. The farthest point in A to B is (3,8). The farthest point in A to C is (3,8). The farthest point in B to A is (9, 2). The farthest point in B to C is (8, 5). Complete link (diameter) between clusters is defined as the similarity between the most dissimilar members. We want to merge two clusters that produce the smallest diameter. Since (5,1) and (8,5) have the smallest distance between them of all possible pairs, B and C should be merged first.

Question 9

Recall from Lecture 4-8 that the objective of learning generative models is to find the parameters that maximize the likelihood of the observed data. Suppose we have a set of points D drawn from Gaussian distribution. For D = {-5, 5, 15}, which of the following set of parameters (μ, σ) produces the maximum L(N(μ, σ2): D)?

Your Answer		Score	Explanation
μ = 5, σ = 5			
μ = 0, σ = 5			
μ = 10, σ = 5			
μ = 5, σ = 10	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

μ = 5, σ = 10. The maximum likelihood estimator for μ in a Gaussian distribution is the sample mean, and the maximum likelihood estimator for σ2 is the sample variance. Something similar in spirit is done in the M-step of the k-means algorithm, where the average of all points in a cluster becomes the new centroid in order to minimize the sum of distance.

Question 10

Consider the three hierarchical clustering algorithms introduced in Lecture 4, BIRCH, CURE, and CHAMELEON. Which of the following statements about these algorithms is TRUE? (Select all that apply)

Your Answer		Score	Explanation
All three algorithms can only work with Euclidean distance as the similarity metric.	✔	0.25	
CURE and CHAMELEON will always produce identical clustering on the same input dataset.	✔	0.25	

☑ BIRCH and CHAMELEON both use a two-phase algorithm where small clusters are first formed via a divisive mechanism before some other clustering algorithm is used to merge them into the final clusters.	✔	0.25
☐ All three algorithms are good at detecting irregular shaped clusters	✔	0.25
Total		1.00 / 1.00

Question Explanation

BIRCH and CHAMELEON both use a two-phase algorithm where small clusters are first formed via a divisive mechanism before some other clustering algorithm is used to merge them into the final clusters. (True. As per the description of the algorithms.)

All three algorithms are good at detecting irregular shaped clusters. (False. While CHAMELEON and CURE are able to identify irregular shared clusters, BIRCH tends to produce spherical clusters.)

All three algorithms can only work with Euclidean distance as the similarity metric. (False. They can work with any similarity metric.)

CURE and CHAMELEON will always produce identical clustering on the same input dataset. (False. As the two algorithms use fairly different mechanisms for identifying clusters, it is unlikely the results will be identical.)