Intro to Big Data Science — Spring 2021-2022

Name: 刘乐奇 **ID No.:** 12011327

Quiz 5

To receive credit, this worksheet MUST be handed in at the end of the class.

abd

1. Suppose you run k-means and after the algorithm converges, you have: $c^{(1)} = 3$, $c^{(2)} = 3$, $c^{(3)} = 5$, ..., where $c^{(i)}$ is the index of the class centroid closest to $x^{(i)}$. Which of the following statements are true? (Select all correct answers)

   (a) The third example $x^{(3)}$ has been assigned to cluster 5.

   (b) The first and second training examples $x^{(1)}$ and $x^{(2)}$ have been assigned to the same cluster.

   (c) The second and third training examples have been assigned to the same cluster.

   (d) Out of all the possible values of $k \in \{1, 2, \ldots, K\}$, the values $k = 3$ minimizes $\|x^{(2)} - \mu_k\|^2$ where μ_k is the k-th centroid.

C

2. Suppose you run k-means using $k = 3$ and $k = 5$. You find that the cost function is much higher for $k = 5$ than for $k = 3$. What can you conclude?

   (A) This is mathematically impossible. There must be a bug in the code.

   (B) The correct number of cluster is $k = 5$.

   (C) In the run with $k = 5$, k-means got stuck in a bad local minimum. You should try re-running k-means with multiple random initializations.

   (D) In the run with $k = 3$, k-means got lucky. You should try re-running k-means $k = 3$ and different random initializations until it performs no better than with $k = 5$.

A

3. In hierarchical clustering, in order to produce compact clusters, which of the following methods should you use:

   (A) Single linkage

   (B) Complete linkage

   (C) Average linkage

   (D) Maximum linkage

AB CD

4. Which of the following are good/recommended applications of PCA? (Choose all answers that apply.)

   (A) To compress the data so it takes up less computer memory/disk space.

   (B) To reduce the dimension of the input data so as to speed up a learning algorithm.

   (C) Instead of using regularization, use PCA to reduce the number of features to reduce overfitting.

   (D) To visualize high-dimensional data (by choosing $k = 2$ and $k = 3$).

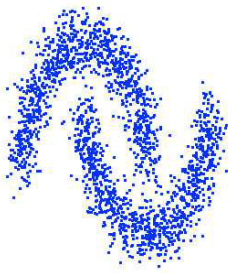
a

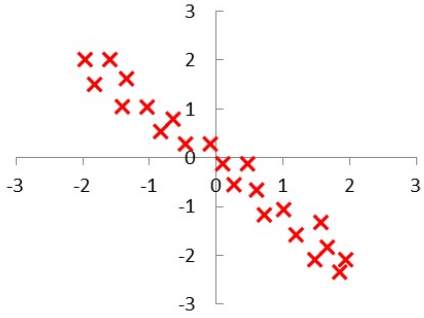
5. Which of the following methods may achieve the poorest clustering for the two moon data set shown below?

   (a) K-Means

   (b) Hierarchical clustering

   (c) DBSCAN





(d) Spectral clustering

6. Suppose you run PCA on the dataset below. Which of the following would be a reasonable vector $u^{(1)}$ (the most principal component) onto which to project the data?

(A) $u^{(1)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

(B) $u^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

(C) $u^{(1)} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

(D) $u^{(1)} = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

7. (Optional) A long time ago there was a village amidst hundreds of lakes. Two types of fish lived in the region, but only one type in each lake. These types of fish both looked exactly the same, smelled exactly the same when cooked, and had the exact same delicious taste - except one was poisonous and would kill any villager who ate it. The only other difference between the fish was their effect on the pH (acidity) of the lake they occupy. The pH for lakes occupied by the non-poisonous type of fish was distributed according to a Gaussian with unknown mean (μ_{safe}) and variance (σ^2_{safe}) and the pH for lakes occupied by the poisonous type was distributed according to a different Gaussian with unknown mean (μ_{deadly}) and variance (σ^2_{deadly}). (Poisonous fish tended to cause slightly more acidic conditions).

   Naturally, the villagers turned to machine learning for help. However, there was much debate about the right way to apply EM to their problem. For each of the following procedures, indicate whether it is an accurate implementation of Expectation-Maximization and will provide a reasonable estimate for parameters μ and σ^2 for each class.

   (a) Guess initial values of μ and σ^2 for each class. (1) For each lake, find the most likely class of fish for the lake. (2) Update the μ and σ^2 values using their maximum likelihood estimates based on these predictions. Iterate (1) and (2) until convergence.

   (b) For each lake, guess an initial probability that it is safe. (1) Using these probabilities, find the maximum likelihood estimates for the μ and σ^2 values for each class. (2) Use these estimates of μ and σ^2 to reestimate lake safety probabilities. Iterate (1) and (2) until convergence.

   (c) Compute the mean and variance of the pH levels across all lakes. Use these values for the μ and σ^2 value of each class of fish. (1) Use the μ and σ^2 values of each class to compute the belief that each lake contains poisonous fish. (2) Find the maximum likelihood values for μ and σ^2. Iterate (1) and (2) until convergence.