Q1 Information and Honor Code

0 Points

In this assignment, you complete the Colab3 worksheet and obtain results from it. Your answer would be an integer, or a float number. The float value should be a decimal number rounded to the **nearest 0.001**. For example, 0.2435 would become `0.244`.

You can submit as many times as you want, and the last submission will be graded. Only the fully corrected answer will receive 1 point. No late day is allowed for any Colab assignment.

Please verify that you have read the above instructions and the Stanford Honor Code and that you have not given or received unpermitted aid while completing this assignment.

If you have any questions about how the Honor Code applies to Colab assignments or other parts of the course, please contact the teaching staff for clarification.

⊙ I have read and understood the above information

Q2 K-means

3 Points

You would like to explore the breast cancer dataset with Spark clustering.

Q2.1 Distance

1 Point

What is the default distance metric used by k-means in https://spark.apache.org/docs/latest/mllib-clustering.html#k-means? Select all that apply.

- manhattan
- euclidean

Q2.2 Silhouette score

1 Point

You fit the dataset with k-means clustering, where k = 2,. You make prediction on the same dataset, and evaluate the clustering by computing the Silhouette score (squared euclidean distance). Make sure you use seed = 1 in the clustering algorithm.

What is the Silhouette score for your prediction on the dataset? (Float)

Q2.3 Prediction accuracy
1 Point

How many data points in the dataset have been clustered correctly? (Integer)

Q3 PCA
3 Points

Next, you want to reduce your feature dimensions with principal component analysis (PCA).

Q3.1 Output vector
1 Point

First, you map the row features in the breast cancer dataset into dense vectors, and create the dataframe for it.

Then, you preform PCA to compute the top 2 principle components, and visualize the output of the first 20 rows.

Which one is the PCA output for the first row?

-2368.993755782054, 121.58742425815576
-2095.6652015478608, 145.11398565870167
-2030.2124927427058, 295.29798399279264
-2260.0138862925405, -187.96030122263656

Q3.2 Clustering after PCA
1 Point

After PCA, you perform clustering using k-means with the same parameter as in Q2.

What is the Silhouette score for your prediction on the dataset using features with PCA? (Float)

Q3.3 Prediction accuracy after PCA
1 Point

How many data points in the dataset have been clustered correctly using features with PCA? (Integer)