CSC 4760/6760 Big Data Programming

Assignment 6

Due Date: 11:59 pm, April 14, 2021

Problem 1. (100 points) Implement the k-means algorithm in Spark. Please use the following dataset in your experiments.

Input Dataset : kmeans_input.txt (Uploaded into iCollege)

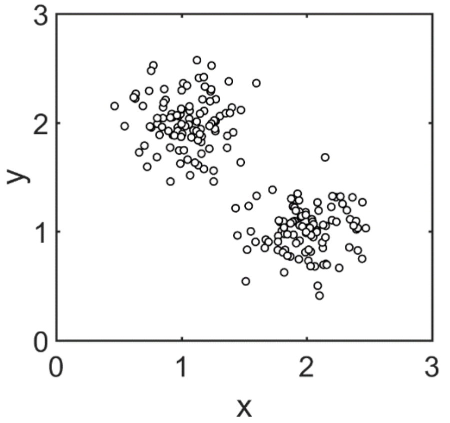
In Windows, please use “WordPad” to open it.

The dataset is in “libsvm” format, please use the following sentence to read it in Spark.

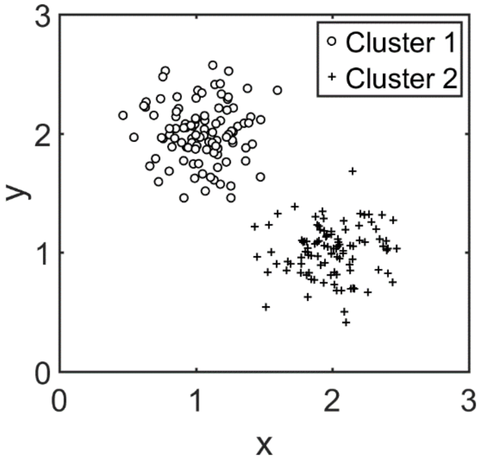
```
dataset = spark.read.format("libsvm").load("/home/rob/data/kmeans_input.txt")
```

In the dataset, each row represents a data point. In total, there are 200 rows, which means there are 200 data points. Each data point contains two features, which represents the x and y coordinates.

The following figure visualizes the dataset.



We can see that there are two clusters. The following figure shows the two clusters.



Problem: You are required to use K-Means algorithm to compute the two clusters. The input is the raw data in “kmeans_input.txt”, the output is the data point with cluster labels.

Implementation:

Design and implement a PySpark program to solve the problem. We did not provide any template python file this time. You may want to create one python file from scratch.

You are required to use the k-means function Spark Machine Learning library to implement this function. Please refer to the following webpage for more details.

Spark MLlib Clustering - K-means

https://spark.apache.org/docs/latest/ml-clustering.html

Refer to the Python API docs for more details.

https://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#pyspark.ml.clustering.KMeans

Report:

Please write a report illustrating your experiments. You need to explain your basic idea about how to call the k-means function in your program. You may add comments to the source code such that the source code can be read and understood by the graders.

In the report, you should include the answers to the following questions.

1) Explanation of the source code

2) Experimental Results: The screenshot of the terminal output of your program. In the output, you need to output the centers of the two clusters. I uploaded the screenshot of my program in iCollege. You may refer to that image.

Submission Materials:

a) Your report

b) Source code (.py file)

c) The screenshot of the terminal showing the output results (cluster centers) of your program