COP 4521 - Spring 20203

Total Points: 100
Due: Thursday, 03/26/2020

1 Objective

The objective for this assignment is to make sure

- You can use some popular Python machine learning libraries.
- You are familiar with the k-means clustering algorithm and how it works.
- You can use the matplotlib library to plot results.
- You are familair with map-reduce and distributed computing techniques.

2 K-Means Clustering

K-Means clustering is a method used in unsupervised learning techniques to cluster or group the most similar data points together. Clustering is done based on a distance function. Unlike other methods, like single linkage, k-means would give you slightly different results on different runs due to inherent randomness, which is why the algorithm is run several times before the result is produced.

The scikit library implements several clustering methods and contains several test datasets on which you can experiment. For this homework, we are going to run k-means clustering on the Handwritten Digit Dataset.

You can find more information on k-means at `https://en.wikipedia.org/wiki/K-means_clustering`

3 Libraries

For this homework, you will need to use the following libraries

- numpy
- scipy
- matplotlib
- pandas
- scikit

While you will not explicitly use numpy and scipy, the scikit library needs them to be installed. You should already have working installations of numpy and scipy since you probably used them for your third homework.

3.1 scikit

The scikit library contains a lot of the machine learning algorithms, including k-means, which we will be using. It also comes with a bunch of preloaded datasets. The easiest way to install scikit is to use pip, as follows:

```
pip install -U scikit-learn
```

This will take a while but at the end, you should be good to go.

3.2 pandas

pandas is a library that provides a lot of useful data structures and data analysis tools. You can get pandas from PyPi as follows:

```
pip install pandas
```

This may require installation of a few dependencies, like numpy, but, at the end, you should be good to go.

4 Specifications

For this homework, you will analyze the intro level clustering program we demonstrated in class (`iris.py`) and adapt it to perform clustering on the Handwritten Digit dataset.

1. **Simple Implementation**
    - Run the sample program on the iris dataset to familiarize yourself with scikit and clustering.
    - Load the digits dataset instead of the iris dataset.
    - Run a Principal Component Analysis (PCA) on the dataset to reduce the number of features (components) from 64 to 2.
    - Run k-means on this datatset to cluster the data into 10 classes.
    - Plot the results using matplotlib. Choose a set of 10 fairly well-separated colors for the scatterplot.
    - This is only one run if the k-means algorithm. In the real world, we run it several times and assign the point to a majority label.
2. **Distributed Implementation**
    - For the distributed implementation, you will need a mapper program and a reducer program.
    - For the mapper, load the same dataset, perform PCA (choosing the same features every time) and run k-means.
    - Turn the result into a dictionary, where the keys are the data points and the values are the cluster labels.
    - In the reducer function, collate the labels for each point, counting how many times they are assigned to a particular label. Some points may only have one label for all runs, while some might have several different labels.
    - The final label for a point would be the one that is assigned the most number of times (majority poll of labels). Plot the same graph, but with the new labels. To do this, just save the point and the label as a CSV file (directly or by redirecting from stdout) then, write another python program to read that into a list and plot it.
    - Set up a Hadoop cluster using the guide (to be posted soon). Test it with 1 mapper and 1 reducer first. Then, spawn 20 mappers and reduce it to one plot.

5 Dataset

The Handwritten Digits dataset is a set of handwritten digits that have been reduced to pixels. You can find more information about the dataset here: `http://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits`

You do not have to get the datatset from the website. It is available automatically from the sklearn datasets, just like we loaded the iris dataset in class.

6 Sample Output

Your result might come out looking slightly different due to the inherent randomness of k-means and the dimensions you are using to plot the result.

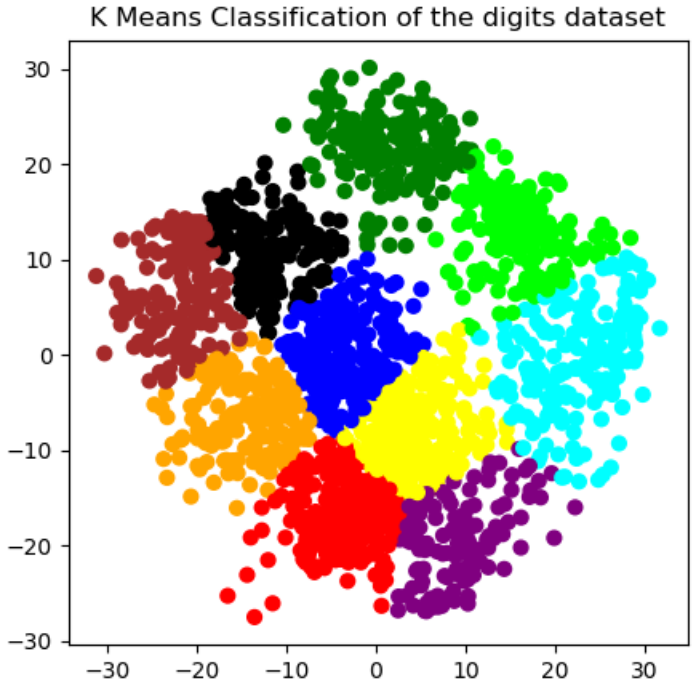

Figure 1: Digits clustered using k-means

7 Submission

You are required to submit a tarball that contains the following:

- The python files containing your code (digits.py, mapper.py, reducer.py, plotter.py)
- A word or PDF document that contains a small description of the k-means algorithm (a couple of paragraphs) and a screenshot of your output. Please write the description by yourself. Copy-pasting from Wikipedia or other sites of the internet is not acceptable.

Please make sure to include your name and FSUID on all the files you are turning in. Please turn in the tarball, named HW3.tar, on Canvas.