CS434 — Implementation Assignment 4 — Due 11:59PM -, 2020

General instructions.

1. Please add onto the Python starter code (version 3.6+). Note: you may need to install several packages (pip install –user), but try not to require anything else. I'll run it with the packages already there, and you should have everything you need. (Things like 'os' are fine though if they're packaged with Python 3.6).
2. You can work in a team of up to 3 people. Each team will only need to submit one copy of the source code and report. You need to explicitly state each member's contribution in percentages (a rough estimate).
3. Your source code and report will be submitted through Canvas
4. You need to submit a readme file that contains the commands to run your code for requested experiments (e.g. python main.py args).
5. Please make sure that you can run code remotely on the class server (vm-cs434-1).
6. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. In particular, **the clarity and quality of the report will be worth 10 pts.** So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.
7. In your report, **the results should always be accompanied by discussions of the results.** Do the results follow your expectation? Any surprises? What kind of explanation can you provide?

Unsupervised Learning for Human Activity Recognition (total points: 80 pts + 10 report pts + 10 result pts)

In this assignment we will work with the Samsung Human Activity Recognition dataset to practice unsupervised and also dimensional reduction methods. More specifically we are interested in applying k-means clustering and Principal Component Analysis (PCA) methods.

The data for this assignment comes from accelerometers and gyros of Samsung Galaxy S3 mobile phones. Each instance in the dataset is a set of feature values extracted from sensor signals and pre-processed to classify the type of activities a person has with his/her phone in his pockets. There are six activities WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING, LAYING.

The label of each instance (type of activities) are given, however we assume that they are unknown and thus we try to apply an unsupervised k-means clustering to cluster the people based on the available features. We only use the labels to evaluate the clusters with purity measure.

Data. The specifications of the datasets is as follows:

1. **(x_train.txt):** Contains 7352 rows (samples) each with 561 features.
2. **(y_train.txt):** Contains 7352 rows of labels corresponding to the rows in the **(x_train.txt)**.

Important Guidelines. Please note that a skeleton code (starter) is provided and the students should only fill the code where is specified.

Part 1 (50 pts) : k-means clustering. For this part please implement the following steps in the starter skeleton:

1. Please fill the places marked in the "clustering" module to complete the k-means class. In method "init-center" the center of clusters will be initialized. You don't need to implement a sophisticated method to initialize the clusters. A random selection from the input instances will be sufficient. If you implement method "predict" efficiently (using vector based operations) the running time will reduce significantly.
2. The k-means class will be used in module main. In this assignment we only consider value 10 as the maximum number of classes. The maximum number of iterations is set to 20 by default, however we might need more iterations until the k-means converges. Please update function "apply-kmeans" of main as follows:
   - To reduce the sensitivity of the k-means to the initialization, update this function to produce the average of SSE and purity vs k and iterations for 5 different runs of k-means. Note that each run should generate different initial centers in method "init-center".
   - Plot the average (over 5 runs) of SSE versus iterations for $k = 6$. You could use the plot functions provided in the main or change them if needed to show the observation more properly.
   - Plot the average (over 5 runs) of the SSE versus k for $k \in 1 \dots 10$. Apply elbow on the curve of SSE versus k for $k \in 1 \dots 10$, to select the best k. Please report the best k you found.
   - Plot the average of purity versus k for $k \in 1 \dots 10$ for the train set and make observation on this.

Part 2 (40 pts) : dimension reduction (PCA). In this part we are interested in reducing the dimension of the data (which is currently 561) to a smaller number. Please implement the following steps:

1. Implement the marked area in module "decompose" to implement the PCA class. In your implementation please use the mean, cov and eig functions provided in the class.
2. In the main module, complete the "visualize" function to visualize the data points in the first two principle component directions, and color each class with a distinct color.
3. The retain ratio r is the percentage of variance we are interested to maintain and is defined as follows:

$$\sum_{i=1}^{d} \lambda_i \geq r \times \sum_{i=1}^{m} \lambda_i \tag{1}$$

   where d and m are the reduced and original dimensions respectively and $d \leq m$. By default this value is set to 0.9. Please report the d you will find for this ratio.
4. Apply the k-means for $k \in 1 \dots 10$ described in part 1 (average over 5 runs) on the data with reduce dimension for retain ratio $r = 0.9$. Plot the purity of the train for this experiment. Do you observe harmful effect due to dimension reduction. If it hurts the purity, please increase r to a higher values (with a 2 or 3 trials) and report the best r which still reduces the dimension but does not hurt the performance.

Submission. Your submission should include the following:

1. The modified source code with a short instruction on how to run the code for your experiments in a readme.txt.
2. Your report (**only in PDF format**), which begins with a general introduction section, followed by one section for each part of the assignment.
3. Please note that all the files should be in one folder and compressed only by .zip.