**Homework 5 Clustering with Python (100 pts total)**

This assignment is to use k-means, single-link, and complete-link clustering algorithms to cluster the wine dataset. The dataset and its description is given on blackboard from our previous assignment. You need to first load the dataset and separate the class label and features. You need then standardize the features and run k-means, single-link and complete-link hierarchical clustering algorithm on the standardized features. Use PCA to find the first two principal components, and plot clustering results on the plane defined by these two components. The detail requirements are described as below.

**1) Data Preparation. (20 pts - 10 pts for standardizing data and code, 10 points for PCA – Principle Component Analysis and code)**

Load the data, and standardize the data using the following standardization equation:

Where µ is the expectation of feature x, and σ is the standard deviation of feature x.

Run PCA on all features and find the first two components.

**2) Implement k-means algorithm. (40 points – code must be documented – 20 points for algorithm implementation, 10 points for the plots, 10 points for comparison and coloring, and 10 points for the 10 runs and result)**

For this assignment, k = 3, which is given. This part of the assignment requires you to implement the k-means algorithm.

To write your k-means, you need to randomly pick 3 data points to be the initial centroids, and run the clustering steps until it converges or the number of iterations reaches a maximum value (you can define this as a reasonable value, e.g. 500). For different runs, your initial centroids have to be different (random).

Randomly pick 3 data points to be the initial centroids and run your k-means once. Plot the clustering results on the 2D plane defined by the two principle components and use different colors for different clusters. Compare your result to the true labels (which means that when you clustered the data you removed the true labels from the results), and highlight the data points which have been clustered wrong.

Run your k-means 10 times and pick the results which has the lowest within-cluster distance. The within-cluster distance can be calculated by the sum of distances of data points to their closest cluster center. Plot this result and compare it to the first result.

**3) Run the k-means using sklearn. (30 pts – 20 pts for code using sklearn and 10 pts for comparison discussion – 1 paragraph)**

For this part, k = 3, which is also given. You need to use the [sklearn.cluster.KMeans](http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) function in the sklearn package in python to run k-means on the wine dataset and compare the results to your implementation of k-means.

Read this page <http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> to learn how to use this function.

Set the init = ‘random’, n\_init = 10, max\_iter = 500 (or the value you used in your k-means) and run this function. Plot the results, highlight errors, and compare this clustering results to your k-means’ results.

**Homework 6 (100 pts)**

**4) Implement single-link and complete-link hierarchical clustering. (60 pts - 30 pts for code and documentation for single-link, 30 pts for code and documentation complete link, 20 pts for plots and comparison to k-means, 20 pts for the intermediate states and plots)**

Implement single-link and complete-link hierarchical clustering algorithm and run your algorithm on the wine dataset. Stop merging when you have three clusters. Plot the clustering results of this two algorithm and compare the result of single-link and complete-link. Compare the result of them two k-means.

Find the intermediate state when you have 4 clusters, and plot this state. Identify the two clusters that were merged in the last step where 4 clusters became 3 clusters. Use the same hue but different lightness to color these two clusters (e.g. dark blue and light blue), and use different hues for the other two clusters.