LAB DAY 2

**K-means using Python SciKit Learn Package:**

Follow through the following example:

<http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html>

Once you are confident you understand the various steps involved and how to apply the K-means package to a dataset, move to the following set of exercises.

**K-means Lab Problem Dataset:**

Attached are 2 datasets on which you will implement K-means algorithm. Datasets are from Tom Mitchell’s CMU course on Machine learning found here: http://www.cs.cmu.edu/~epxing/Class/10701-10s/HW/

1. Import the two datasets given, as a dataframe in Pandas
2. The first column is x and second column is 7. Visualize the dataset
3. Implement K-means on both datasets (<http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)>
4. Plot the results of the clusters (with two different colors)
5. Comment and discuss on how K-means performed for both datasets. Why do you think this is?

**LINEAR REGRESSION USING GRADIENT DESCENT**

**Relationship between girls age and height**

1. Load the girls train dataset and plot it.
2. Model the height of the girls as

Write a function that will perform batch gradient descent to update the betas. Choose an alpha rate of 0.05 and maximum iterations = 1500

1. What is the MSE (Mean Square Error) of your regression model on the training set?
2. Plot the following 3 graphs:
3. Regression line with the data
4. Contour lines of Betas ( on x-axis and on y-axis
5. Bowl shaped cost function
6. Model Testing:
7. Test your model by predicting the height for a 5 year old girl.
8. Load the girls\_test.csv dataset and calculate the MSE for the test set. How does it compare to the training set’s MSE?

For reference, following are the equations to update Betas and Cost Function:

