Programming Assignment 4

Due: May 23

K-Means

1. Implement K-Means and run on the Abalone dataset for K = 1, 2, 4, 8, 16
   (a) Be sure to Z-scale your input variables
   (b) Choose the initial clusters by randomly selecting K observations
   (c) Terminate when no more observations switch clusters
2. Output the centroids, and Within Cluster Sum of Squares (WCSS) for each run
3. Calculate and output the Mean and Standard Deviation for each feature within each cluster
4. Plot the WCSS vs K
5. Note: Lectures 19 and 20 discuss the algorithms for K Means in quite a bit of detail

K-Means and QR

1. Train a QR Model for observations for each of clusters above
2. Run your test set against this compound model
   (a) For each test observation choose the cluster whose centroid is nearest this point
   (b) Calculate the RMSE for all points
3. Output
   (a) Plot the RMSE against K

Notes and hints

1. Lectures 19 and 20 detailed the algorithms and pseudo code
2. Make sure you use the Means and SD's calculated from your training set when Z-Scaling in your test set
3. You may use library provided routines for sampling and QR, although if you successfully implemented the other assignments and have design / infrastructure in place it will be faster to use those
4. As in other assignments, randomly sample (non-biased) 90% for training and the other 10% for testing