Lab #4
Data Science I

Unsupervised techniques are often used in the analysis of genomic data. In particular, PCA and hierarchical clustering are popular tools. We will use these techniques on the NCI60 cancer cell line microarray data, which consists of 6,830 gene expression measurements on 64 cancer cell lines. Each cell line is labeled with a cancer type. We do not make use of the cancer types in performing PCA and clustering, as these are unsupervised techniques. But after performing PCA and clustering, we will check to see the extent to which these cancer types agree with the results of these unsupervised techniques. This lab is adapted from a lab found in An Introduction to Statistical Learning (http://faculty.marshall.usc.edu/gareth-james/ISL/)

1. Read in both the microarray data (NCI60_data.csv) and the labels of the cancer type (NCI60_labs.csv).
2. We will first do a PCA analysis. Scale, perform PCA, and plot the variance explained and the cumulative variance explained by the PCs. How many PCs are produced in the analysis? Approximately how many PCs are required to explain 90% of the variance in the data?
3. Plot the scores on the first versus second PC and the scores on the first versus third PC. Color the scores by cancer type. Interpret your results.
4. Next we will perform clustering on the data. Visualize the dendogram for hierarchical clustering with complete, average, and single linkage. Label the dendogram with the cancer types. Does the choice of linkage impact the results?
5. Use complete linkage for the remainder of the analysis. Cluster the observations into 4 groups. Look at the labels for each of the groups and comment on the success of the clustering.
6. Rather than performing hierarchical clustering on the entire data matrix, we can simply perform hierarchical clustering on the first few principal component score vectors. Sometimes performing clustering on the first few principal component score vectors can give better results than performing clustering on the full data. Perform hierarchical clustering on the first 5 principal component scores and comment on the results.