Problem Set #3 - Unsupervised Learning

Due Friday October 4th at 5:00 pm

For this problem set you will submit a pdf file and a ipynb/rmd document to the Canvas site with your answers. Make sure comments tie directly to the appropriate figures and tables. Include your code when indicated.

Simulation

1. **Explore Clustering -** Simulate a dataset where $n = 100$ and $p = 50$. Use your ICC function from problem set #2 to create 3 (k) distinct clusters in the dataset. **Hint:** As long as the groups are properly ordered, each variable, p, can be simulated independently with your ICC function. You can repeat this b times with the `replicate()` function.

   (a) Set your ICC to 0.3 and simulate a dataset. Show your simulation code. Apply K-means clustering and evaluate how well your clusters uncover the true groups.

   (b) Simulate additional datasets, varying the ICC. Apply K-means clustering with k = 3. Evaluate how well you are able to uncover the true cluster labels. Visualize your results and comment on the "required" ICC for effective clustering. You can use metrics like Average Silhoutte, Within SS, etc.

   (c) Choose the minimum ICC from above where you think you can uncover the clusters. Keep n and p the same but vary the number of variables that have an ICC. i.e. some $q < p$ will have an ICC where the rest will have an ICC of 0. These variables with $ICC = 0$ are called "noise." Assess the impact of the amount of noise variables on ability to recapture clusters.

   (d) Perform the two analyses above across a full grid of ICCs and noise. Evaluate your results. What patterns do you see? Is there an *interaction* between noise and ICC?

2. **PCA and Clustering -** While PCA solves a different problem than clustering it is also used to find structure in data.

(a) Choose a dataset from 1b where you were able to uncover clusters. Apply PCA. Using a scree plot, how many meaningful PCs do you find? Plot PC1 vs PC2.

(b) Repeat but set the number of groups k to be 4. What relationships do you notice between the number of "true groups" and the number of meaningful components.

Working With Data

Download the Mice Protein Expression data set from the UCI Machine Learning Repository:
`https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression`

This dataset contains 77 protein expression values for 1080 mice.

1. **Exploring/Cleaning Data -**

   (a) Focusing just on the 77 protein measurements provide any relevant descriptive statistics and/or visualizations.

   (b) Some values are missing. Remove any proteins missing $>$ 10% of the time. Perform single mean imputation for any proteins missing $<$ 10% of the time. Show your code.

2. **PCA**

   (a) Perform PCA on your imputed expression measurements. Do not scale the data. How many "meaningful" PCs do you get?

   (b) Scale the data and repeat. Comment on any differences. Point to any descriptive statistics that would explain the source of these differences.

   (c) Columns 79 - 82 provide labels to the data. Using the unscaled data to visualize the first couple of PCs and see whether they are able to separate out any of the groups.

   (d) The PCs can be used in a regression model - this is called *principal components regression.* This can be useful if $p > n$ or if there is high correlation between the feature.

      i. Fit a logistic regression, regressing the "Genotype" variable onto the expression measurements. Comment on your results

      ii. Regress the Genotype variable onto the first few "meaningful" PCs from 2a. Comment on the differences.

3. **Clustering**

   (a) **K-Mediods**

i. Perform K-mediods clustering on the imputed expression data. Based on your results above what would expect to be the optimal k. What do you find as the optimal k.

ii. How do the cluster groups correspond to any of the labeled data?

iii. We can also cluster proteins. Since we are more interested in how proteins move together, a correlation based distance is more meaningful. `dist()` does not provide a correlation based distance as a default so you will have to calculate it yourself. Display results of your clustering.

(b) **Hierarchical Clustering.** Clustering in general can be unstable and sensitive to different choices. Here we are going to cluster proteins so use a correlation distance.

i. **Cluster Method:** Compare the results of single, complete and average linkage. Cut the tree at 4 clusters. How much do the 3 procedures correspond with respect to cluster assignment

ii. **Standardizing Data:** Choose one of the linkage methods from above. Standardize and center your data. How do the results differ.

iii. **Distance Metric:** Instead of using a correlation distance use euclidean distance. Compare cluster membership.

iv. **Sampling:** Sample 20% of the data (observations) and compare cluster membership. Repeat this a few more times.

v. Comment on what you think the most important factors are in cluster stability