Unsupervised Learning & Dimensionality Reduction

# The Datasets

* 2 datasets
* Remind the reader of what they are
* Explain why they are interesting
  + Non-trivial amounts of features such that you apply dimensionality reduction.

# Clustering

Run 2 different algorithms on 2 datasets and describe what you see.

* *K-*means clustering
  + How did you choose *k?*
* Expectation Maximization
* Choose your own measure of distance/similarity. JUSTIFY

# Dimensionality Reduction

Apply the following dimensionality reduction to the two datasets and describe what you see

* PCA
  + What is the distribution of eigenvalues?
  + Assuming you only generate *k* projections (i.e. you do dimensionality reduction) how well is the data reconstructed?
* ICA
  + How kurtotic are the distributions?
  + Do the projection axes seem “meaningful”?
* Randomized Projections
  + Assuming you only generate *k* projections (i.e. you do dimensionality reduction) how well is the data reconstructed?
  + How much variation did you get when you re-ran several times?
* Dealer’s choice

# Clustering after Dimensionality Reduction

Reproduce your clustering experiments, but on the data that you’ve run dimensionality reduction. (2 datasets x 4 dimensionality reductions x 2 algorithms = 16). Look at them all but focus on the more interesting findings in your report.

* When you reproduced your clustering experiments on the datasets projected onto the new spaces created by ICA, PCA, and RP, did you get the same clusters as before? Different clusters? Why or why not?

# Neural Network Learner with Dimensionality Reduction applied to Dataset

Apply the dimensionality reduction algorithms to one of your datasets from assignment 1 (already done above (one wine set)) and rerun your neural network learner on the newly projected data.

* Any differences in performance?
* Any differences in speed?
* Any other differences?

# Neural Network Learner with Dimensionality Reduction and Clustering applied to Dataset

Apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms, treating the clusters as if they were new features. In other words, treat the clustering algorithms as if they were dimensionality reduction algorithms.

Clustering -> neural network

Rerun your neural network learner on the newly projected data.

* Any differences in performance?
* Any differences in speed?
* Any other differences?

# Notes

* Explanation of your methods
* How are these algorithms the same as, different from, and interact with your earlier work
* Clusters
  + A description of the clusters that you got
  + Why did you get the clusters you did
  + Do they make “sense”? See isbell paper
  + If you used data that already had labels (For example data from a classification problem from assignment #1) did the clusters line up with the labels?
  + Do they otherwise line up naturally?
  + Why or why not?
  + Compare and contrast the different algorithms?
  + What sort of changes might you make to each of those algorithms to improve performance?
  + How much performance was due to the problems you chose?
  + Justify your analysis with data explicitly
* For PCA, what is the distribution of eigenvalues?
* For ICA, how kurtotic are the distributions?
* Do the projection axes for ICA seem to capture anything “meaningful”?
* It might be difficult to generate the same kinds of graphs for the part of the assignment. Come up with some way to describe the kinds of clusters you get. If you can do it visually, all the better.
* **10 page limit**