Numbers

The assignment is worth 10% of your final grade.

Read everything below carefully!

Why?

Now it's time to explore unsupervised learning algorithms. This part of the assignment asks you to use some of the clustering and dimensionality reduction algorithms we've looked at in class and to revisit earlier assignments. The goal is for you to think about how these algorithms are the same as, different from, and interact with your earlier work.

The same ground rules apply for programming languages and libraries.

The Problems Given to You

You are to implement (or find the code for) six algorithms. The first two are clustering algorithms:

- *k*-means clustering
- Expectation Maximization

You can choose your own measures of distance/similarity. Naturally, you'll have to justify your choices, but you're practiced at that sort of thing by now.

The last four algorithms are dimensionality reduction algorithms:

- PCA
- ICA
- Randomized Projections
- Any other feature selection algorithm you desire

You are to run a number of experiments. Come up with at least two datasets. If you'd like (and it makes a lot of sense in this case) you can use the ones you used in the first assignment.

1. Run the clustering algorithms on the datasets and describe what you see.
2. Apply the dimensionality reduction algorithms to the two datasets and describe what you see.
3. Reproduce your clustering experiments, but on the data after you've run dimensionality reduction on it. Yes, that's 16 combinations of datasets, dimensionality reduction, and clustering method. You should look at all of them, but focus on the more interesting findings in your report.
4. Apply the dimensionality reduction algorithms to one of your datasets from assignment #1 (if you've reused the datasets from assignment #1 to do experiments 1-3 above then you've already done this) and rerun your neural network learner on the newly projected data.
5. Apply the clustering algorithms to the same dataset to which you just applied the dimensionality reduction algorithms (you've probably already done this), treating the clusters as if they were new features. In other words, treat the clustering algorithms as if they were dimensionality reduction algorithms. Again, rerun your neural network learner on the newly projected data.

What to Turn In

You must submit:

1. A file named *README.txt* that contains instructions for running your code
2. your code (link only in the README.txt)
3. a file named yourgtaccount-*analysis.pdf* that contains your writeup.

The file yourgtaccount-*analysis*.pdf should contain:

- a discussion of your datasets, and why they're interesting: If you're using the same datasets as before at least briefly remind us of what they are so we don't have to revisit your old assignment write-up... and if you aren't well that's a whole lot of work you're going to have to recreate from assignment 1 isn't it?
- explanations of your methods: for example, how did you choose *k*?
- a description of the kind of clusters that you got.

- analyses of your results. Why did you get the clusters you did? Do they make "sense"? 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? Be creative and think of as many questions you can, and as many answers as you can. Take care to justify your analysis with data explicitly.
- Can you describe how the data look in the new spaces you created with the various algorithms? 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"? Assuming you only generate *k* projections (*i.e.*, you do dimensionality reduction), how well is the data reconstructed by the randomized projections? PCA? How much variation did you get when you re-ran your RP several times (I know I don't have to mention that you might want to run RP many times to see what happens, but I hope you forgive me)?
- 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? Why not?
- When you re-ran your neural network algorithms were there any differences in performance? Speed? Anything at all?

It might be difficult to generate the same kinds of graphs for this part of the assignment as you did before; however, you should come up with some way to describe the kinds of clusters you get. If you can do that visually all the better.

Note: Analysis writeup is limited to 10 pages total.

Grading Criteria

At this point, you are not surprised to read that you are being graded on your analysis more than anything else. I will refer you to this section from assignment #1 for a more detailed explanation. As always, start now.