Project 3: Unsupervised Learning and Dimensionality Reduction

The assignment is worth 10% of your final grade.

Read everything below carefully, this assignment has changed from previously posted material.

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 six algorithms.

The first two are clustering algorithms. You can choose your own measures of distance/similarity. Justify your choices.

- Expectation Maximization
- Clustering Algorithm of Your Choice

The last four are dimensionality reduction algorithms:

Linear:

- PCA
- ICA
- Randomized Projections

Non-linear:

- Manifold Learning Algorithm of Your Choice

You are to run several experiments with the goal of disseminating how dimensionality reduction affects your data. Come up with at least two datasets. You can use the datasets from the first assignment. Develop hypotheses based on your datasets and the following exploration. These should be well-posed and grounded in theory from the lectures and readings.

The following should comprise your exploration.

1. Apply the clustering algorithms on the datasets. You will report on each of the clustering algorithms for each dataset, resulting in 4 demonstrations.
2. Apply the dimensionality reduction algorithms on the datasets. You will report on each of the dimensionality reduction algorithms, resulting in 8 demonstrations.
3. Re-apply the clustering algorithms on the set of dimensionality reduction datasets. This will result in 16 combinations of results of datasets, dimensionality reduction, and clustering methods. You should look at the full scope of the results and note how they might pertain to your hypotheses. In particular, focus on more interesting findings. You will be reporting one of your clustering algorithms on your datasets with one linear method (PCA, ICA, or RP) and the manifold learning algorithm of your choice, resulting in 4 total demonstrations between the two datasets. Justification will be especially important as space is limited in the report.
4. Choose one of your datasets. Re-run your neural network learner from Assignment #1 with each of the dimensionality reduction algorithms applied. You will report on a different linear method from Step 3 (PCA, ICA, or RP) and on the manifold learning algorithm of your choice, resulting in 2 total demonstrations. Justification will be especially important as space is limited in the report.
5. Using the same dataset as Step 4, use both previously generated clusters from Step 1 as new features in your dataset. Again, rerun your neural network learner on the newly projected data and note the findings. You will report on each of the clustering algorithms, resulting in 2 demonstrations. Justification will be especially important as space is limited in the report.

What to Turn In

You must submit:

1. A file named *README.txt* that contains instructions for running your code, including a link of some sort to your code.
2. a file named yourgtaccount-*analysis.pdf* that contains your writeup.

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

- Brief description of your datasets, and hypotheses you want to highlight in your report.
- Explanations of methods. This is your opportunity to demonstrate nuances needed to support your hypotheses.
- Grounded descriptions of resulting clusters. Support descriptions with data-driven evidence.
- 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 looks in the new spaces you created with the various dimensionality reduction 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? How much variation did you get when you re-ran your random projections several times? How does noise affect each algorithm? What is the rank of your data? Can you qualitatively and quantitatively describe how colinear your data might be? How might specific properties of your data influence outputs of various algorithms?
- 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? Remember to justify why one output might be more interesting when choosing your demonstrations.
- When you re-ran your neural network algorithms were there any differences in performance? Speed? Anything at all? Consider how you might judge differences in performances and include these notes in your discussion.

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. However, a note of caution. Figures should remain legible as we are asking for several demonstrations in many sections. Do not try to squish figures together in specific sections where axis labels become 8pt font or less. We are looking for clear and concise demonstration of knowledge and synthesis of results in your demonstrations. Any paper that solely has figures without formal writing will not be graded. Be methodical with your space.

Note: The report is limited to 8 pages total. We will not grade anything past page 8. This includes any title pages (I do not recommend anyone use a title page in this class), figures, and references. This is different from the first two assignments. This is standard IEEE conference length and good practice in conciseness and brevity.

Grading Criteria

You are being graded on your analysis more than anything else. I will refer you to the grading sections from Assignment #1 for a more detailed explanation. As always, start now and have fun!