BG4104 Group Report (20%)
Unsupervised Learning in healthcare problem

1 Objective

Now it is 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. You may program in any language that you wish insofar as you feel the need to program. As always, it is your responsibility to make sure that we can actually recreate your narrative, if necessary.

2 Procedure

You are to implement five algorithms on one dataset that you feel most confident working with. Each of you had a unique dataset for your first assignment, but for this task, using just one dataset is sufficient. Remember to focus on the **interesting aspects or insights your dataset** can offer. Of course, you may run the algorithms separately and choose the best-performing one, but please make sure to manage your time wisely.

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

- Expectation Maximization
- K-Means Clustering

The last three are linear dimensionality reduction algorithms:

- Randomized Projections
- Principle Component Analysis (PCA)
- Independent Component Analysis (ICA)

You are to run several experiments with the goal of disseminating how dimensionality reduction affects your data. Develop hypotheses based on your dataset and the following exploration. These hypotheses should be well-posed and grounded in theory from the lectures and readings.

2.1 Experiments and Analysis

The following should comprise your exploration.

1. Apply the clustering algorithms on the dataset. You will report on each of the clustering algorithms for the dataset, resulting in 2 demonstrations.
2. Apply the dimensionality reduction algorithms on the dataset. You will report on each of the dimensionality reduction algorithms, resulting in 3 demonstrations.
3. Re-apply the clustering algorithms on the set of dimensionality reduction dataset. This will result in 6 combinations of results of dataset, 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 what combination performs the best. Justification will be especially important as space is limited in the report.

4. Re-run your neural network learner from Assignment #1 with each of the dimensionality reduction algorithms applied. You will report on the different linear methods from Step 3 (PCA, ICA, or RP) and what performed the best for the dataset with reason. 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 performance (EM and K-Means) from your NN for comparison. You will need to define how you set up the comparison. Justification will be especially important as space is limited in the report.

Analysis writeup is limited to 8 pages. The page limit does include your citations. Anything past 8 pages will not be read. Please keep your analysis as concise while still covering the requirements of the assignment. As a final check during your submission process, download the submission to double check everything looks correct on Canvas. Try not wait until the last minute to submit as you will only be tempting Murphy's Law.

In addition, your report must be written in LaTeX on Overleaf. You can create an account with your NTU email address. When submitting your report, you are required to include a 'READ ONLY' link to the Overleaf Project. If a link is not provided in the report, 5 points will be deduced from your score. Do not share the project directly with the Instructor or TAs via email. For a starting template, please use the IEEE Conference template.

Analysis writeup is limited to 8 pages. The page limit does include your citations. Anything past 8 pages will not be read. Please keep your analysis as concise while still covering the requirements of the assignment. Try not wait until the last minute to submit as you will only be tempting Murphy's Law.

In addition, your report must be written in LaTeX on Overleaf. You can create an account with your NTU email account. When submitting your report, you are required to include a 'READ ONLY' link to the Overleaf Project. If a link is not provided in the report, 5 points will be deduced from your score. Do not share the project directly with the Instructor or TAs via email. For a starting template, please use the IEEE Conference template.

2.2 Acceptable Libraries

Here are a few examples of acceptable libraries. You can use other libraries as long as they fulfill the conditions mentioned above.

Machine learning libraries:

- scikit-learn (python)
- Weka (java)
- e1071 (R)
- ML toolbox (matlab)

Plotting:

- matplotlib (python)
- seaborn (python)
- yellowbrick (python)
- ggplot2 (R)

3 Submission Details

The due date is indicated on the NTUlearn for this assignment.

Late Due Date [20 point penalty per day]:

Each group only need to submit one report. You must submit:

- A file named README.txt containing instructions for running your code (see note below)
- Your main.py and other file you think is necessary to run your code.
- A file named your group-analysis.pdf

Note: we need to be able to get to your code and your data. Providing entire libraries isn't necessary when a URL would suffice; however, you should at least provide any files you found necessary to change and enough support and explanation so we can reproduce your results on a standard linux machine.