BUAN 6341
APPLIED MACHINE LEARNING
ASSIGNMENT 4
Due date: May 1, 11:59 pm

In this assignment, you will implement the following clustering algorithms:

1. K-means
2. Expectation Maximization

In addition, you will implement the following feature dimensionality reduction algorithms

1. Any one feature selection algorithm (decision tree, forward selection, backward elimination, etc.)
2. PCA
3. ICA
4. Randomized Projections

You must use R or Python for this assignment. You **can** use **any** publicly available R or Python library/package. You can even use H2O.

For the following tasks, use the same two datasets that you used in assignment 3 (online news and the one you picked).

Tasks:

1. Run the clustering algorithms on your datasets and describe your observations (with plots).
2. Apply the dimensionality reduction algorithms on your datasets and describe your observations (with plots).
3. Run the clustering algorithms again, this time after applying dimensionality reduction. Describe the difference compared to previous experimentation (with plots).
4. Run your neural network learner from assignment 3 on the data after dimensionality reduction (from task 2). Explain and plot your observations (error rates, etc.)
5. Use the clustering results from task 1 as the new features and apply neural network learner on this new data consisting of only clustering results as features and class label as the output. Again, plot and explain your results.

Deliverables:
You are required to submit the following:

- Your code file(s)
- A readme file explaining how to run your code.
- Report (must not exceed 8 pages total)
- Any supporting files (data sets, etc. If the data sets are too large, submit the url)

Your report should be both thorough and concise and contain at the very least the following:

- Explanations of your methods. For example, how did you choose k?
- What type of clusters did you get? Did they line up with the class labels? If not, did they line up naturally? Were they compact or not? Why do you think you got these types of clusters? How

did you pick the features generated by ICA and RP? Compare and contrast the different 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? Compare and contrast the different algorithms.
- When you re-ran your neural network algorithms were there any differences in performance? Speed? Anything at all?
- Can you describe how the data looks in the transformed domains you created with the various algorithms?

It might be difficult to generate the same kind of graphs for this part of the assignment as you did before; however, you should come up with some way to describe the clusters you get. Be creative and think of various experiments that you can come up with for this assignment. You need to give clear description of all your experiments and analysis. Why did you get the results that you did? How did you pick various parameters? Think of as many questions as you can! This assignment will take time. So get started on it today!

Grading:
Total weightage: 15% of final grade

Breakdown:

Code: 0 points (Code should execute and produce the results presented in the report with minimum effort – We will run the code and if it doesn't run or has errors, points will be deducted from the report).

Report: 100 points
Points will be awarded based not only on how good your results are, but also on how well you describe them as well as underlying experimentation. Any plots without explanation = 0 points. Similarly, explanation without plots = 0 points. Keep in mind that you are graded on your analysis and description, as well as creativity. Have fun!