Assignment #4: Clustering & Ensemble Learning
STA9890
Statistical Learning for Data Mining

Assignment Parameters:

Date Assigned: 2024-04-15
Date Due: 2024-05-07 @ 5:45pm

Submission Mechanism(s):

- Blackboard (strongly preferred)
- Email to instructor: michael.weylandt@baruch.cuny.edu
  Email submissions must be titled *exactly* as `STA9890-S2024-HW4-LASTNAME,FIRSTNAME.pdf`

This extended HW4 is worth 200 points, though the 'technical content' is not quite double that of a typical homework assignment.

Question 1: Review (25 points)

Answer the following questions - you may refer back to earlier homeworks.

1. Compute the bias and variance of OLS. (5 points)
2. Compute the bias and variance of ridge regression. (5 points)
3. Implement a coordinate descent method for *non-negative lasso regression*. Compare your results to those of CVX on the same problem. (To implement non-negative lasso, think about how the updates for NN-Lasso will compare to those of regular Lasso. It's only a minor change) (10 points)
4. Implement *regularized linear discriminant analysis*[1] and compare the accuracy of LDA and RLDA on the `authors` data. (5 points)

Question 2: Clustering (50 points)

For this problem use the combined training AND testing splits of the `authors` dataset. Your goal is to use the word counts to cluster the data (without the author labels) and see if your groups coincide with the true author attribution.

(a) Visualize the data. (You may choose to use one or more methods to visually summarize the data and use for exploratory analysis).

(b) Compare and contrast the following clustering methods:

i. K-means

ii. Hierarchical Clustering (Try at least 4 linkages and at least 3 distances. Which ones did you choose? Why?)

iii. Biclustering. (You may try your choice(s) of biclustering method(s). Hint: Try the NMF and the cluster heatmap.)

Reflect upon your results. Which is the best method to visualize the data? Which distance metric is best? Which clustering method yielded groups that closely coincide with true authorship? Why? Are there any words that are more important for clustering? Which ones?

Question 3: Ensemble Learning (50 points)

For this problem use the provided training and testing splits of the `authors` data. Compare and contrast the following methods for predicting authorship:

[1] Recall, regularized LDA replaces $\hat{\Sigma}$ with $\lambda\hat{\Sigma} + (1-\lambda)\sigma^2\mathbf{I}$.

1. Classification Trees. (Which error measure did you use? Why?)
2. Bagging.
3. Boosting. (Which boosting method did you use? Why?)
4. Random Forests. (Which parameter settings did you use? Why?)

Reflect upon your results. Which method yields the best error rate? Which method yields the most interpretable results? Which words are most important for authorship attribution?

Question 4: Optimized Random Forests (50 points)

As discussed in class, Random Forests are generally constructed as (unweighted) ensembles of decision trees. In this section, we will explore the concept of an *optimized random forest*: that is, we will use *stacking* to learn an optimal combination of tree weights and compare the performance of the optimized RF to the classical RF. For this problem, you may use your favorite decision tree software and `CVXR` - you may not use software that implements this method 'out of the box.'

1. Pick two authors from the `authors` data to reduce this to a binary classification problem.
2. Split the training data further into a 'true training' and an 'ensembling' set.
3. Using an existing software package, generate 100 random forest trees on the 'true training' set.
4. Generate predictions from each tree on the 'ensembling' set to form a data matrix.
5. Using `CVXR`, implement *non-negative ridge-regularized logistic regression* to create a optimized random forest ensemble.
6. Compare the performance of the optimized RF with a standard RF trained on the entire training set.
7. Repeat this process several times over (with new training/ensembling/test data and new author pairs) to compare the predictive accuracy of ORF with classical RF.
8. Suppose we want to use a *sparse* ORF to improve computational performance. Modify your existing pipeline to implement *non-negative lasso-regularized logistic regression* in the stacking step and compare the performance with that of both ORF and classical RF.

Question R: Reflection (25 points)

Reflect on the course and answer the following questions:

- What have you learned in this course that you did not anticipate learning?
- What skills that you learned in this course do you anticipate will be the most helpful in future classes? In your future career?
- What skills or techniques do you want to learn more about?
- If one topic could be added to this course, what would it be?
- How could you demonstrate mastery of the skills learned in this course during an interview?