Homework 4

Statistics S4240: Data Mining
Columbia University
Due Thursday, August 13

For your .R submission, submit a file labeled `hw04.R`. The write up should be saved as a .pdf of size less than 6MB. **DO NOT** submit .rar, .tar, .zip, .docx, or other file types.

Problem 1. (10 Points) James 6.8.1

Problem 2. (10 Points) James 6.8.3

Problem 3. (10 Points) James 6.8.5

Problem 4. (10 Points) James 8.4.5

Problems 5 to 7 use classification trees and logistic regression to classify the Federalist Papers.

Question 5. (20 Points) Use your code from Homework 4 to read in the Federalist Papers and create document term matrices `dtm.hamilton.train`, `dtm.hamilton.test`, `dtm.madison.train`, and `dtm.madison.test`. Create a set of labels for each document term matrix, with Madison documents given values 0 and Hamilton documents given values 1. Combine the document term matrices and labels to create two data frames: one that includes all training data and one that includes all testing data. Make sure that the column labels for the covariates are the dictionary words (hint: use `as.vector(dictionary$words)` to get a vector of words) and the column label for the response is `y`.

a. (10 Points) Use tree classification to predict the author using the training data. Apply the model to the testing data. Specifically, in `R` use `rpart` classification with Gini impurity coefficient splits. Then compute the proportion classified correctly, the proportion of false negatives, and the proportion of false positives. Plot the tree with labeled splits.

b. (10 Points) Now use tree classification again, but this time with information gain splits, to predict the author. Apply the model to the testing data. Then compute the proportion classified correctly, the proportion of false negatives, and the proportion of false positives. Plot the tree with labeled splits. Are there any differences between the two plots? If so, what are they and why do you think they arose?

Question 6. (20 Points) Create centered and scaled versions of your document term matrices. (Do not center and scale the labels.) We will use these for regularized logistic regression with `glmnet`.

a. (2 Points) Could we use an unregularized logistic regression model with this data set? Why or why not?

b. (8 Points) Use `glmnet` to fit a ridge regression model on the training data. Apply the model to the testing data. Then compute the proportion classified correctly, the proportion of false negatives, and the proportion of false positives. Find the 10 most important words according to the model along with their coefficients.

c. (10 Points) Use `glmnet` to fit a lasso regression model on the training data. Apply the model to the testing data. Then compute the proportion classified correctly, the proportion of false negatives, and the proportion of false positives. Find the 10 most important words according to the model along with their coefficients. Compare the "important" words selected by ridge and lasso. Are the words different? What about their relative weights?

Question 7. (20 Points) We can use feature selection to remove features that are irrelevant to classification. Instead of calculating the probability over the entire dictionary, we will simply count the number of times each of the n most relevant features appear and treat the set of features themselves as a dictionary.

a. (10 Points) A common way to select features is by using the mutual information between feature x_k and class label y. Mutual information is expressed in terms of entropies,

$$I(X, Y) = H(X) - H(X \,|\, Y) = H(Y) - H(Y \,|\, X).$$

Show that

$$\begin{aligned} I(Y, x_k) = \sum_{\tilde{y}=0}^{1} & p(x^{test} = k \,|\, y = \tilde{y}) p(y = \tilde{y}) \log \frac{p(x^{test} = k \,|\, y = \tilde{y})}{p(x^{test} = k)} \\ & + \big(1 - p(x^{test} = k \,|\, y = \tilde{y})\big)\, p(y = \tilde{y}) \log \frac{1 - p(x^{test} = k \,|\, y = \tilde{y})}{1 - p(x^{test} = k)}. \end{aligned}$$

b. (10 Points) Compute the mutual information for all features; use this to select the top n features as a dictionary. Use the document term matrices from the resulting dictionary for all four of the methods in questions 5 and 6: tree classification with Gini splits, tree classification with information splits, ridge logistic regression, and lasso logistic regression. (Hint: subset your previously computed matrices/data frames.) For each method use the testing set to compute the proportion classified correctly, the proportion of false negatives, and the proportion of false positives for $n = \{200, 500, 1000, 2500\}$. Display the results in three graphs (each graph will now have four lines). What happens? Why do you think this is?

Question 8. (25 points) James 8.4.10

Note: To begin this problem, you should execute `library(ISLR)` and `data("Hitters")` to load the data set.

Question 9. (20 points) James 10.7.1