Homework 6

Statistics W4240: Data Mining
Columbia University
Due Monday, December 8 (All Sections)

For your .R submission, submit files labeled `hw06_q1.R`, `hw06_q3.R`, `hw06_q4.R`, and so on. The write up should be saved as a .pdf of size less than 4MB. **DO NOT** submit .rar, .tar, .zip, .docx, or other file types.

Question 1. (25 points) James 8.4.10

Note: To begin this problem, you should execute `library(ISLR)` and `data("Hitters")` to load the data set. You may also want to see the following link to familiarize yourself with the dataset itself: `http://rgm.ogalab.net/RGM/R_rdfile?f=vcd/man/Hitters.Rd&d=R_CC`

Question 2. (15 points) James 9.7.3

Question 3. (20 points) As in the previous two homeworks, we will use the `Federalist` dataset. 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 the previous homework: 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 4. (20 points) Begin by setting up the document term matrices as in the previous homeworks, and standardize the training and testing data as in the most recent homework (as done for the LASSO problem, for example).

a. (5 points). Using the first 100 words in the document term matrices, fit an SVM with a linear kernel to the `Federalist` training data. Use the function `sum` in the `e1071` library that we have demonstrated in class. It is also demonstrated in Chapter 9 of James. Note that your classification variable (for example, y, the binary authorship variable) should be input as a factor to ensure classification. Use this linear SVM to classify your testing data with `predict()`. What is the percentage of correct classification? Is this particularly good, given your previous experience? If not, why might this be underperforming?

b. (5 points). Repeat part (a) with a linear SVM using the first 5, 10, 15, 20, ... , 100 words. Plot the classification performance as a function of the number of words. Interpret the result.

c. (5 points). Repeat part (b) with a SVM with an RBF kernel. Do you reach any different conclusions? Why or why not?

d. (5 points). Use an SVM with an RBF (Gaussian) kernel to classify authorship based on the top two features that you found in problem 3 (`"upon"` and `"depart"`). Plot and interpret the resulting figure. Include the figure in your submission.

Question 5. (20 points) James 10.7.1