STAT 542 Fall 2014

Homework Assignment 3

Due Wednesday, October 29.

In this homework, you are asked to try various classification methods we've learned in class on the `spam` data. Download the file `spamHW3.Rdata` from the assignment page. Run

```
load("spamHW3.Rdata")
```

in `R`, then you'll see the training set (`spam.train`) and the test set (`spam.test`). The last column, named "`Y`" is the class label, and the other 57 columns are the features. You can find the description of this data set from the text book.

Try the following methods, and summarize the following two classification accuracies (on the test set) in a table:

- 0/1 classification accuracy;
- log likelihood (or - deviance), i.e., $\sum_i y_i \log p(\mathbf{x}_i) + (1-y_i)\log(1-p(\mathbf{x}_i))$ where we assume Y is coded as 0 or 1, and $p(\mathbf{x})$ is the estimated probability for $Y = 1$ at a test point $\mathbf{x}$.

Some classification methods are model-based, in the sense that we have made some assumptions on the joint distribution $P(x, y)$ or on the conditional distribution $P(y|x)$. You'll be asked to the list of assumptions for those methods.

Problem 1. Try *logistic regression* with (i) full model, (ii) forward selection via AIC, (iii) forward selection via BIC, and (iv) variable selection via Lasso[1]. What are the assumptions we make for logistic regression. Comment on the variable sets returned by various methods. How many variables are used by each method?

Problem 2. Try (i) *LDA*, (ii) *RDA*, and (iii) *Naive Bayes* (both parametric and nonparametric). List the assumptions for each method.

Try the three methods again, based on the union of variables selected at **Problem 1**, i.e., drop variables, which are not selected by AIC, BIC, or Lasso.

Problem 3. Use SVM. Try (i) linear, (ii) quadratic, and (iii) Gaussian kernels. For each kernel, explain how you select the tuning parameters and report the corresponding value/values. Also for each kernel, report the number of support vectors you use in your final model. Explain how to obtain the predicted probability

[1] The package `glmnet` can deal with a combination of L_1 and L_2 penalties on the logistic coefficient, but you are asked to use just the L_1 penalty here.

$p(\mathbf{x}_*) = P(Y_* = 1|\mathbf{x}_*)$ for a test sample $\mathbf{x}_*$ (see Chap 10.6).

Try the three SVM models again based on the union of variables selected at **Problem 1**, i.e., drop variables that are not selected by AIC, BIC, or Lasso.

Problem 4. Try classification tree. Explain which impurity measure you use when growing the tree, and which you use when pruning the tree. What's the size of the tree (i.e., the number of terminal/leaf nodes) you finally use.

The follow questions are related to the weakest link algorithm. Cut the tree to size 5. Explain which internal node (of this size 5 tree) is the weakest link, i.e., the node whose branches will be removed when the penalty α increases. At what value of α, we'll remove its branches.

Problem 5. Try `randomForest`. Explain how large is your forest. That is, how many trees you grow? When building a tree, instead of picking the best split among all p variables, `randomForest` picks the best one among only $m < p$ variables. What is the value you use (or set up automatically by `R`) for m?

Provide a variable importance plot. How is it different from or similar to the variable selection result you obtained at **Problem 1**.

Problem 6. Use the `gbm` package to implement AdaBoosting. How many boosting iterations you use initially, and how do you choose the optimal numbers of boosting iterations?

Problem 7. Use *Neural Nets.* You decide the number of hidden layers (suggest to be just 1 or 2) and the number of neurons in each hidden layer.