University of Michigan, Dept of Statistics

Stat 503, Instructor: Long Nguyen

Homework Assignment 4
(due Tuesday, March 20, 2018)
in class or Canvas dropbox by 1pm

1. **Logistic regression revisited and surrogates to 0-1 loss**

   (a) Explain how the logistic regression estimation method, which involves maximizing the (log) conditional likelihood of the logistic regression model, can be viewed as minimization with respect to the logistic loss
   $$L(y, f(x)) = \log(1 + e^{-yf(x)}),$$
   where $y \in \{\pm 1\}$. (Hint: the answer should be almost immediate, if you realized that at when logistic regression was originally introduced, we assumed the class label $y \in \{0, 1\}$ instead of $y \in \{\pm 1\}$.)

   (b) Recall the decision-theoretic formulation of classification problem: let (X, Y) be a pair of random variables where $X \in \mathbb{R}^d$, $Y \in \{-1, 1\}$ represents the class label. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a discriminant function, which is obtained by minimizing the expected risk:
   $$\min_f \mathbb{E}\{L(Y, f(X))\}$$
   We have established that if L is 0-1 loss, i.e., $L(y, f(x)) = 1(yf(x) < 0)$, an optimal function f takes the form $f(x) = P(Y = 1|X = x) - P(Y = -1|X = x)$. Tracing the same proof technique to answer the following: What is an optimal function f if (i) we take L to be the logistic loss function given above, or (ii) the hinge loss associated with the support vector machines?

2. **Spam classification.** Consider the email spam data set given in Canvas. This consists of 4601 email messages, from which 57 features have been extracted. These are as follows:

   - 48 features giving the percentage of certain words (e.g., ”business”, ”free”, ”george”) in a given message
   - 6 features giving the percentage of certain characters (; ([! $ #)
   - feature 55: the average length of an uninterrupted sequence of capital letters
   - feature 56: the length of the longest uninterrupted sequence of capital letters
   - feature 57: the sum of the lengths of uninterrupted sequences of captial letters

   The data set contains a training set of size 3065, and a test set of size 1536. One can imagine performing several kinds of preprocessing to this data. For instance: (i) Standardize the columns so they all have mean 0 and unit variance; (ii) Transform the features using $\log(x_{ij} + 1)$; (iii) Discretize each features using $\mathbb{I}(x_{ij} > 0)$.

   In this assignment, you will apply SVM, decision trees, and neural networks to predict the type of email from email features. (You only need to report your finding with a particular kind of email features among the three preprocessing methods, but you are encouraged to play with different methods and pick the one that works best).

   (a) For SVM, investigate how sensitive your results are to the choice of kernels (linear, polynomial and Gaussian kernels), the tuning parameter that defines some of these kernels, and the tuning parameter controlling the total amount of slack allowed. Use cross-validation for the selection of tuning parameters for SVM as well as subsequent methods.

(b) For neural networks, investigate how sensitive your results are to the choice of number of hidden layers (try at least one and two hidden layers), and the number of hidden nodes per layer.

(c) For decision trees, investigate how sensitive your results are to the tree sizes.

3. **Class-imbalance data.** In the spam data set, there are about 40% spam instances in the training set, so the two classes (spam/not spam) are fairly balanced. In practice, such balanced data may not be available. To assess the robustness of classification techniques to imbalanced data, please subsample from the spam instances in the training set, so as to obtain more skewed class ratios (specifically, 3:7, 2:8, and 1:9).

   (a) Repeat the classification analysis on these newly created imbalanced data sets and report your results.

   (b) A simple way to ameliorate the class imbalance stiuation is to randomly generate (e.g., via bootstrap) for the underrepresented class with more instances. Please report whether such a technique helps to improve the classification performance of each of the three methods that you have considered.

Instructions The solution to problem 1 can be either typed up or written by hand. The solution to data analysis questions (Problem 2, 3) may be written as a data analysis report. The report needs to be clear, concise, and to the point. There should be no graphs or tables that are not commented on in the text. Please include your R code (at least the main parts) in a *separate* appendix at the end of the report.