STA 141C: Homework 2

- Homework due in Canvas: 05/27/2019 at 11:59PM. Please follow the instructions provided in Canvas about homeworks, carefully.

1. **Supervised Learning Basics: 5 points**. Answer true or false for each of the question below and give justification.

   (a) Gradient descent is a supervised learning algorithm.

   (b) Gradient descent is an algorithm to minimize or maximize a function.

   (c) Logistic regression cannot be performed after linear PCA.

   (d) Support vector machine is non-linear classification algorithm.

   (e) Any regression algorithm can be *modified* to be used for classification as well.

2. **Amazon Review Classification: 10 points.** In this question, we will take use the preprocessed Amazon review data from HW2 (in both Document-Term matrix and TF-IDF matrix representation format) to do classification.

   (a) How many reviews of each rating value are there in the entire dataset? Our goal is to build a classifier that reads the review and classifies whether the review was "good" (rating = 5) or "bad" (rating = 1). What is the best performance of a "constant classifier", a classifier that ignores the review and blindly assigns a constant classification?

   (b) Now we'll run $L1$-regularized logistic regression on our dataset. Split the data into two parts: the training set (which is 70%) and testing set (which is the rest). Fit a $L1$-regularized logistic regression model by letting python chose the regularization parameter itself.

   (c) How many covariates have non-zero coefficients in the model selected by python ? List the twenty words with the most positive coefficients and twenty words with most negative coefficients.

   (d) Now run the fitted logisitic model on your testing data and report the misclassification rate. How does this compare with the "constant classifier" we originally discussed before (i.e., is it better or worse)?

3. (**Boston Housing Prediction: 10 points.**) In this example, you will work with the Boston Housing Data set. The goal is to predict the median value of housing based on the values of 13 covariates. In the HOUSING_TRAIN.TXT file, the first 13 columns correspond to covariate ($X \in \mathbb{R}^d$) and the last column corresponds to median value of housing ($Y \in \mathbb{R}$). Assume that they you are using the model

$$Y = \beta^\top X = \sum_{i=1}^{13} X_i \beta_i$$

   to do the prediction.

   (a) Use linear regression on HOUSING_TRAIN.TXT to estimate β (do not use HOUSING_TEST.TXT). Calculate the mean square prediction error.

(b) Now use HOUSING_TEST.TXT (note that this data was not used when estimating β) and the estimate for β from above to predict the housing prices in the test dataset. Calculate the mean-square prediction error.

(c) Change the model in Question 3, to

$$Y = \gamma^\top \tilde{X} = \sum_{i=1}^{13} X_i \beta_i + \sum_{i=1}^{13} \sum_{j=1}^{13} \beta_{i,j} X_i X_j$$

and repeat the same process in question 3. Note that this could still be considered as a linear model with data $\tilde{X} = [X_1, X_2, \ldots, X_{13}, X_1^2, X_1 X_2, \ldots, X_{13}^2]$. Does it give any improvement (in terms of mean squared prediction error) compared to the previous model from question 3 ?