CS 534: Machine Learning

Homework #2

Due: Feb 24th at 11:59 PM on Canvas

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in Python, R, or Matlab with the hints mostly focusing on those who will be programming in Python. Also be sure to attach the code to ensure you'll get partial / full credit if the answer is not what was expected.

1. **(2 + 2 + 4 = 8 pts)** In many applications, the classifier is allowed to "reject a test example rather than classifying it into one of the classes. Consider, for example, a case in which the cost of misclassification is $10 but the cost of having a human manually take a decision is only $3 and no cost if correct.
   (a) Suppose $P(y = 1 \mid x)$ is predicted to be 0.2. Which decision minimizes the expected loss?
   (b) Now suppose $P(y = 1 \mid x) = 0.4$. Which decision minimizes the expected loss?
   (c) Show that in general, for this loss matrix, for any posterior distribution, there will be two thresholds θ_0 and θ_1, such that the optimal decision is to predict 0 if $p_1 < \theta_0$, reject if $\theta_0 \leq p1 \leq \theta_1$, and predict 1 if $p_1 > \theta_1$, where $p_1 = P(y = 1 \mid x)$. What are these thresholds for the costs above?
2. **(6 + 6 + 6 + 6 + 6 = 30 pts) K-fold Cross validation** Using the email spam dataset from Homework #1, we will explore the effect of different values of K-fold cross validation for finding the optimal regularization parameter for a Lasso logistic regression model. You will want to search for the parameters in log space from -8 to -0.5, with 30 points in between. For parts (a) – (d), plot the CV error curve and report the value of the tuning parameter based on the usual rule and the one standard error rule. Build a new LASSO logistic regression model using the "optimal parameter for the entire training data and report the AUC on both the training and test set.
   (a) K = 2
   (b) K = 5
   (c) K = 10
   (d) LOOCV
   (e) Report the coefficient values for all the features based on the "optimal" parameter values reported in the previous parts (a) – (d). In other words, create a table with the coefficients and the tuning parameter value and each entry contains the coefficient. How do they compare with each other?

3. **(8 + 8 + 8 + 8 = 32 pts) Model Selection using AIC and BIC** Using the same email spam dataset from above, we will explore the use of AIC and BIC to perform forward and backward stepwise regression for building a logistic regression model. Create a table of training AUC and testing AUC, as well as the coefficient values from the four different model selection criterion.
   (a) Use forward stepwise and AIC as the selection criteria to find the right features for a logistic regression model on the training data.
   (b) Use backward stepwise and AIC as the selection criteria to find the right features for a logistic regression model on the training data.
   (c) Use forward stepwise and BIC as the selection criteria to find the right features for a logistic regression model on the training data.
   (d) Use forward stepwise and BIC as the selection criteria to find the right features for a logistic regression model on the training data.
4. **(10 pts) Feature selection Analysis** Compare the results of Problem 2 and Problem 3 both based on generalization error, training error, and the coefficients (e.g., feature selection and estimated coefficients). Also comment on the computational efficiency between the various validation and feature selection models.
5. **(10 + 10 pts) Bootstrapping** Using the same email spam dataset from the last two problems, we will explore the use of bootstrap to find the optimal parameter and calculate the uncertainty of parameters for the coefficient values for ℓ_2–regularized (ridge) logistic regression.
   (a) Use bootstrap to find an "optimal" parameter for the entire training data based on 10 bootstrap samples for the set of parameters specified in Problem 2.
   (b) Using the "optimal" regularization parameter from (a), report the estimated coefficients and the uncertainty based on 10 bootstrap samples.