CS 534: Machine Learning
Homework 2

(Due Oct 6th at 11:59 PM on Canvas)

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in any language that `Euler` supports.

1. **(2×4=8 pts) Bias-Variance Trade-off of LASSO**

   While it is hard to write the explicit formula for the bias and variance of using LASSO, we can quantify the expected general trend. Make sure you justify the answers to the following questions for full points:

   (a) What is the general trend of the bias as λ increases?

   (b) What about the general trend of the variance as λ increases?

   (c) What is the bias at $\lambda = 0$?

   (d) What about the variance at $\lambda = \infty$?

2. **(5+2+6+6+6+7=32 pts) Discriminant Analysis**

   Suppose points in $\mathbb{R}^2$ are being obtained from two classes, C1 and C2, both of which are well described by bivariate Gaussians with means at $\begin{bmatrix}0\\0\end{bmatrix}$ and $\begin{bmatrix}2\\4\end{bmatrix}$, and covariances $\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}$ and $\begin{bmatrix}2 & 1.5\\1.5 & 2\end{bmatrix}$ respectively.

   (a) If the priors of C1 and C2 are 0.6 and 0.4 respectively, what is the ideal (i.e. Bayes Optimal) decision boundary?

   (b) Suppose the cost of misclassifying an input actually belonging to C2 is four times as expensive as misclassifying an input belonging to C1. Correct classification does not incur any cost. If the objective is to minimize the expected cost rather than expected misclassification rate, what would be the best decision boundary?

   (c) Generate 20 training samples (13 points from C1 and 7 points from C2) based on the distributions specified in (a) as your training data [Hint for Python programmers: consider using the `multivariate_normal` function in numpy]. Generate another 10 test samples (6 and 4 points from C1 and C2 respectively). What is the optimal Bayes error rate on the test data (i.e., how well does the Bayes optimal decision boundary do)?

   (d) Build your own LDA estimator using the training samples – what are the discriminant functions for the two classes? What is the error rate on the test data?

   (e) Build your own QDA estimator using the training samples – what are the discriminant functions for the two classes? What is the error rate on the test data?

   (f) Implement the regularized discriminant analysis. Plot the misclassification rate of both the train data and the test data (on the same graph) as a function of α. What conclusions can you draw from this graph?

3. **(23+7+5=35 pts) Quantifying Uncertainty of Blog Feedback Parameters via Bag of Little Bootstraps**

(a) Implement the bag of little bootstraps (BLB) for standard linear regression and fixed values of b, s, r or the subset size, number of sampled subsets, and number of inner iterations respectively.

(b) Plot the relative error of the estimated coefficients ($\hat{\boldsymbol{\beta}}$) against the "true estimates" ($\tilde{\boldsymbol{\beta}}$) that are obtained when running linear regression on the entire training dataset ($\sum_i |\hat{\beta}_i - \tilde{\beta}_i|$) as a function of s for $b = n^{0.7}$ and $r = 100$.

(c) For the s where the estimated coefficients have converged, how does the confidence intervals of the top 5 coefficients compare to the "true estimates"?

4. **(5+8+8+4=25 pts) Spam classification using Logistic Regression**

Consider the Ling-Spam email spam dataset from the first homework – we will be exploring the effect of pre–processing on the data. You will want to report the AUC and classification error in a table for all your models + different pre-processing techniques.

(a) Fit three logistic regression models (standard, ridge, and LASSO) on the training data by standardizing the columns to have 0 mean and unit variance. For the regularization parameters, justify your final selection.

(b) Convert your feature matrix such that each element reflects the term frequency in the document. Note that if a row, denoted as a vector $\mathbf{x}$ represents an email, then the ith feature is $x_i = \frac{x_i}{\sum_k x_k}$. Train three logistic regression models (standard, ridge, and LASSO) on this new data. For the regularization parameters, justify your final selection. Train three logistic regression models (standard, ridge, and LASSO) on this new data. For the regularization parameters, justify your final selection.

(c) Transform the features again using $\log(x_{ij} + 0.1)$ on your term frequency matrix from part (b).

(d) Comment on how the models compare with one another with regards to AUC and classification error