CS 534: Machine Learning
Homework 2

(Due Oct 12th at 11:59 PM on Gradescope)

Submission Instructions:

The homework must be submitted electronically on Gradescope as a single submission. Each question must be tagged appropriately (you need to set which page the problem is on), otherwise you may lose credit. The answer must also be explicit, if you are exporting an iPython notebook, you will need to add remarks to denote your final answer for the question.

The code can be done in Python, R, or MATLAB. The use of any other programming language must be confirmed with Huan (TA). Source code must be submitted separately on Canvas (the deadline is 15 minutes after the Gradescope submission) and zipped into a single file. There is no need to attach the data, and you can assume the data will be in the same directory as the code. The source code should be executable and there should be no explicit path names for the data files (`joyce/CS534/HW2/train.csv` is an example of hard-coding).

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+4+8+8+5=29 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 $\begin{bmatrix}2\\2.5\end{bmatrix}$, and covariances $\begin{bmatrix}2.5 & 1\\1 & 2.5\end{bmatrix}$ and $\begin{bmatrix}2.5 & 1.5\\1.5 & 1\end{bmatrix}$ 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) Generate 2 datasets from the known distribution. The first one will have 20 training samples (13 points from C1, 7 points from C2), and 10 test samples (6 and 4 points from C1 and C2 respectively). The second dataset will contain 100 training samples (60 from C1, 40 from C2) and 200 test samples (120 from C1, 80 from C2). What is the optimal Bayes error rate on the two test datasets (i.e., how well does the Bayes optimal decision boundary from (a) do)?

   (c) Create your own implementation of LDA, which allow the user to pass in the training samples (features and response). Test your implementation on the two datasets from (b). What are the error rates on the training data and test data?

   (d) Create your own implementation of QDA, similar to the LDA portion. Test your implementation on the two datasets from (b). What are the error rates on the training data and test data?

(e) Suppose the cost of misclassifying an input actually belonging to C1 is twice as expensive as misclassifying an input belonging to C2. Correct classification does not incur any cost. If the objective is to minimize the expected cost rather than expected misclassification rate, how would this change the Bayes optimal decision boundary from (a)?

3. **(4+6+6+5=21 pts) Spam classification using Naive Bayes and Standard Logistic Regression**

Consider the email spam dataset, which contains 4601 e-mail messages that have been split into 3000 training (`spam.train.dat`) and 1601 test emails (`spam.test.dat`). 57 features have been extracted with a binary label in the last column. You can read more about the data at the UCI repository (`http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection`). The features are as follows:

- 48 continuous real [0,100] attributes of type word freq WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.
- 6 continuous real [0,100] attributes of type char freq CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurrences) / total characters in e-mail
- 1 continuous real [1,...] attribute of type capital run length average = average length of uninterrupted sequences of capital letters
- 1 continuous integer [1,...] attribute of type capital run length longest = length of longest uninterrupted sequence of capital letters
- 1 continuous integer [1,...] attribute of type capital run length total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail
- 1 nominal 0,1 class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

(a) You will explore the effects of feature preprocessing and its impact on Naive Bayes and Standard (unregularized) logistic regression. Preprocess your data in the following ways (they are independent of one another and should not build on each step):

i. No preprocessing.

ii. Standardize the columns so they all have mean 0 and unit variance. Note that you want to apply the transformation you learned on the training data to the test data. In other words, the test data may not have mean of 0 and unit variance.

iii. Transform the features using $\log(x_{ij} + 0.1)$.

iv. Binarize the features using $\mathbb{1}_{(xij>0)}$ (Note that $\mathbb{1}$ denotes the indicator function).

You are free to use any preprocessing module (e.g., `sklearn.preprocessing`).

(b) Fit a Naive Bayes model to each of the four preprocessing steps above using only the training data. You are free to use any existing packages. Report the accuracy rate and AUC on the training and test sets.

(c) Fit a standard (no regularization) logistic regression model to each of the four preprocessing steps above using only the training data. You are free to use any existing packages. Report the accuracy rate and AUC on the training and test sets.

(d) Plot the receiver operating characteristic (ROC) curves derived from the test data for each of the eight models (4 Naive Bayes, 4 logistic regression models). Comment on how the models compare with one another with regards to ROC, AUC, and accuracy.

4. **(3+10+6+10+8+4=41 pts) Exploring Model Selection Strategies for Logistic Regression with Regularization**

We will be using the SPAM dataset from the previous part for this problem. You can preprocess the data however you see fit, either based on the results of the previous problem or introducing another preprocessing method. The only requirement is that it is consistent throughout the rest of this problem.

(a) Implement the validation/hold-out technique for logistic regression (regularized model), where the performance is reported for the validation portion. Your implementation should work for any user-specified split ratio (this will help with the later parts of this problem).

(b) Implement the k-fold cross-validation approach for logistic regression (regularized model), where the performance is reported from the k-different validation. Your implementation should work for any user-specified k.

(c) Fit ridge and LASSO using the validation/hold-out technique by searching over a variety of split ratios and regularization parameters. For each unique split ratio, specify the best 'parameter', report the validation error, and the test error generated from training on *all* the training data using the best parameter.

(d) Fit ridge and LASSO using the k-fold cross-validation approach by searching over $k = 2, 5, 10$ and regularization parameters. For value of k, specify the best 'parameter', report the average validation error, and the test error generated from training on *all* the training data using the best parameter.

(e) Fit ridge and LASSO using the Monte Carlo Cross-validation approach with 10 samples (i.e., use the validation/hold-out technique from (a) 10 times). For the different user-specified split ratio, specify the best 'parameter', report the average validation error, and the test error generated from training on *all* the training data using the best parameter.

(f) Comment on how the different model selection techniques compare with one another with regards to AUC and classification error, robustness of the validation estimate, and the computational complexities of the three different hold-out techniques.