CS 534: Machine Learning

Homework #1

Due: Feb 10th 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. **(5 + 5 + 5 + 5 + 5 + 5 = 30 pts) Data exploration and MLR**

   The "Boston Housing" data set located in the UCI repository `https://archive.ics.uci.edu/ml/datasets/Housing`, records properties of 506 housing zones in the Greater Boston area. Of particular interest is the median home value (MEDV) for the various suburbs of Boston based on the other attributes found in the dataset. For detailed description of the data, please read the data set description found on the website.

   (a) Generate box–plots for the LSTAT (% of lower status in the population) and MEDV (median home value) attributes separately. Identify the cutoff values for outliers using the 25^{th} and 75^{th} percentiles (Q1, Q3) and the interquartile range (IQR):

   $$\begin{aligned} \text{IQR} &= \text{Q3} - \text{Q1} \\ \text{LB} &= \text{Q1} - 1.5 \times \text{IQR} \\ \text{UP} &= \text{Q3} + 1.5 \times \text{IQR} \end{aligned}$$

   (b) Generate a scatterplot of MEDV against LSTAT; comment on how inclusion of the outliers would affect a predictive model of median home value as a function of % of lower status in the population. (Hint: Such effects may be easier to visualize if the outliers are a different color or symbol than the other data).

   (c) Plot the histogram of MEDV. Should there be a transform to make it look more Gaussian-like? If so, what should it be?

   (d) Fit a multivariate linear regression to this dataset with MEDV as the dependent variable. If your answer to the previous part was that a transformation should be used, predict the transformed variable instead. Keep the first 300 records as a training set to fit your model and the remaining 206 samples should be used as a test set. Use only the following variables: LSTAT, RM, CRIM, ZN, and CHAS.

      i. Report the coefficients obtained by your model. Would you drop any of the variables used in your model?

      ii. Report the mean squared error (MSE) obtained on your training set. How much does this increase when you score your model on the test set?

(e) Do you think your MLR model is reasonable for this problem? You should look at the distribution of residuals to provide an informed answer.

2. **(5 + 10 + 10 + 5 + 5 + 5 = 40 pts) Discriminant Analysis**

   (a) 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 (0,0) and (1,3) and covariances $\mathbf{I}$ and $2\mathbf{I}$ respectively. $\mathbf{I}$ is the (2x2) identity matrix. If the priors of C1 and C2 are 0.4 and 0.6 respectively, what is the ideal (i.e. Bayes Optimal) decision boundary (derive the equation for this boundary)?

   (b) Derive the classification boundary for LDA and QDA and express it in terms of the estimated means and covariance matrices. You may want to derive QDA first as LDA is a special case.

   (c) Generate 10 points from C1 and 15 points from C2 based on the distributions specified in (a). For Python programmers, consider using the multivariate_normal function in numpy. Plot these 25 points (scatter plot) along with the Bayes optimal boundary that you obtained from (a), and the two boundaries from QDA and LDA. The Circle and Line2D modules in the Matplotlib Artist class might be useful for drawing the boundaries.

   (d) Estimate the true error rate of the LDA and QDA classifiers that you obtained (using an adequate sample of fresh data from the two distributions).

   (e) Repeat (c), and (d) except now the size of the training samples should be 100 and 150 points from C1 and C2 respectively.

   (f) 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, what would be the best decision boundary? Obtain the equation describing this boundary.

3. **(10 pts) Multivariate Gaussian and Logistic Regression** Suppose that you have data in $\mathbb{R}^d$ from two classes, 1 and 2. The data from each class is distributed as a multivariate gaussian. Let us now create new feature vectors $\mathbf{z}$ that contain all the linear and quadratic terms of the original variables $\mathbf{x}$, (like $x_1, x_2, \cdots, x_1^2, x_2^2, ..., x_1x_2, x_1x_3$, etc). Show that the actual probability of a point belonging to Class 1, $\Pr(y|\mathbf{z})$, can be represented as a logistic regression model in this new feature space.

4. **(15 + 5 = 20 pts) Spam classification using 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) Fit a logistic regression model (if using scikit-learn, use the default regularization parameter) to each of the following types of data preprocessing using only the training data. Report the accuracy rate on the training and test sets. If you are using Python, sckit-learn has a preprocessing module with useful classes such as scale, FunctionTransformer, and Binarizer.

i. 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.

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

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

(b) Plot the receiver operating characteristic (ROC) curves derived from the test data for each of the three logistic regression models on the same graph. Report the area under the ROC curve (AUC) for all three models. Comment on how the models compare with one another with regards to ROC, AUC, and accuracy.