MATH 156: Homework 1

Due on Monday, May 6

1. Consider the usual regression data with binary response values $y_1, \ldots, y_n$ and explanatory variable values x_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, p$. The response vector is Y and the matrix of explanatory variable is X. We wish to fit the logistic regression model to the data:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad \text{for } i = 1, \ldots, n,$$

where $y_1, \ldots, y_n$ are independent random variables having Bernoulli distribution with means $p_1, \ldots, p_n$.

(a) Explain why the MLEs of $\beta_0, \ldots, \beta_p$ depend on Y only through the vector $X^T Y$.

(b) Let $\widehat{p}$ be the vector of fitted probabilities with components $\widehat{p}_1, \ldots, \widehat{p}_n$. Express $\widehat{p}_i$ in terms of the MLE $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ and the explanatory variable values.

(c) Express $\sum_{i=1}^n \widehat{p}_i$ in terms of $\#\{i : y_i = 1\}$.

(d) Express the residual deviance in terms of $y_1, \ldots, y_n$ and $\widehat{p}_1, \ldots, \widehat{p}_n$.

(e) We want to obtain $0-1$ valued fitted values $\widehat{y}_1, \ldots \widehat{y}_n$ by putting a threshold $c \in (0, 1)$ across $\widehat{p}_1, \ldots, \widehat{p}_n$. In other words, $\widehat{y}_i = 1$ if $\widehat{p}_i > c$ and 0 otherwise. Express the precision and recall in terms of $y_1, \ldots, y_n$ and $\widehat{y})_1, \ldots \widehat{y}_n$.

2. Consider the binary classification from two scalar exponential priors:

$$p(x|y = j) = \lambda_j e^{-\lambda_j x}, \quad x \geq 0, \ j = 0, 1.$$

Assume $\lambda_0 > \lambda_1$. Let $\widehat{y}$ be the classifier: $\widehat{y} = 1$ if $x \geq t$ and 0 otherwise, where t is a threshold to be determined.

(a) Compute the miss detection probability $P_{MD}(t) := \mathbb{P}(\widehat{y} = 0|y = 1)$ and the false alarm probability $PFA(t) := \mathbb{P}(\widehat{y} = 1|y = 0)$ in terms of t.

(b) For the remainder of the problem, let $\lambda_0 = 1$ and $\lambda_1 = 5$. Plot the ROC (Receiver Operating Characteristics) curve , $P_D := 1 - P_{MD}(t)$ against $P_{FA}(t)$.

(c) Suppose that missed detection is five times as costly as false alarms, and both classes are equiprobable. What is the optimal threshold t ?

3. In this problem you will use various methods discussed in lecture to classify the handwritten digits (MNIST). The data can be found at `http://yann.lecun.com/exdb/mnist/index.html` with detailed descriptions. You can also use API to import the data. You do not need to use all the training data for model training, but you should use all the test data to report the test error.

(a) Use the linear regression of indicator matrix (one-hot encoding). Report the training time, training error and test error.

(b) Use the linear discriminant analysis (LDA). Report the training time, training error and test error.

(c) Use the logistic regression. Report the training time, training error and test error.

(d) Use the random forest. Report the training time, training error and test error.

(e) Use linear SVM, and SVM with Gaussian radial basis kernel. Report the training times, training errors and test errors.

(f) Compare the above methods and conclude.