ECE 532, Spring 2019
Homework #3
Due: Tue, Mar 5 @ 2:30 pm

Problem 1. The file `bc_data.csv` provides data from the famous "Wisconsin breast cancer" study. The 560 rows correspond to different individuals, with the first column listing the ID number of the individual, the second column recording whether the tumor was benign (0) or malignant (1), and the remaining 30 columns listing measurements related to the tumor.

(a) Use the built-in Matlab function `glmfit`, with option `binomial`, to implement logistic regression for predicting the category of diagnosis (0 vs. 1), when regressed on the five features 8, 21, 22, 28, and 29 (in columns 10, 23, 24, 30, and 31). Report the vector of coefficients $\widehat{\beta}$. Also use $\widehat{\beta}$ to predict each of the labels y_i in the dataset from the feature vectors x_i, and report the proportion of correct predictions.

(b) Now use the iterative formulas for gradient descent derived in class to optimize the MLE objective function. (You may need to play around with the stepsize.) Plot the error of the iterates $\|\beta^t - \widehat{\beta}\|_2$ as a function of t—the error should converge to 0 as t increases. Don't forget to manually insert an intercept in the data matrix X in order to match the Matlab implementation.

(c) Repeat the exercise in part (b), this time using Newton-Raphson iterates.

Problem 2. In this problem, you will work through a derivation of the fact that the MLE for logistic regression does not exist when the data are linearly separable.

(a) Write down the formula $L_\beta(X, y)$ for the likelihood of observing the data $\{(x_i, y_i)\}_{i=1}^n$ when the logistic parameter is β (see Lecture 6).

(b) If the data are linearly separable, there exists some vector $\beta_{sep} \in \mathbb{R}^p$ that perfectly classifies the data. In other words,

$$y_i = \begin{cases} 1 & \text{if } x_i^T \beta_{sep} \geq 0, \\ 0 & \text{if } x_i^T \beta_{sep} < 0. \end{cases}$$

Sketch a cartoon illustration of this when $p = 2$, showing the location of β_{sep}.

(c) Using the formula in (i) and the fact that the logistic function $f(u) = \frac{1}{1+\exp(-u)}$ is strictly increasing, argue that the likelihood expression $L_\beta(X, y)$ can be made arbitrarily large by taking β to be an appropriate multiple of β_{sep}.

Problem 3. We now revisit the breast cancer data from Problem 1.

(a) Perform 5-fold CV (without repermuting the rows of the data matrix—we have already done this for you) to choose between logistic regression, LDA (Matlab function `classify`), and SVM (Matlab function `fitcsvm`). Provide a table with the accuracy of each classifier on each fold of the data (measured in terms of proportion of correct predictions). Which classifier should you select?

(b) We now turn to visualizing the data. Using only features 21 and 28, plot the data in $\mathbb{R}^2$, using markers o/+ for the categories 0/1. Perform logistic regression (with intercept), LDA, and SVM classification by regressing the diagnosis only on these two features, and plot the line corresponding to the decision boundary in each case.

Problem 4. The file `wine.csv` contains data from 178 wines, which are classified into one of three types. The first column lists the type (1, 2, or 3), and the remaining 13 columns list levels of 13 types of attributes. We will build a classifier for categorizing the wines based on attributes. First portion off the last 50 rows (which constitute a random subsample) for the validation set, and use the remaining rows as the training set.

(a) Perform LDA with 1 vs. 1 classification on the training set. Report (i) the proportion of correctly classified points in the validation set, (ii) the proportion of incorrectly classified points, and (iii) the proportion of points with an ambiguous classification.

(b) Now repeat part (a), but using 1 vs. all classification, instead.

(c) Finally, use the Matlab function `mnrfit` to implement multiclass logistic regression directly on the training data. Report the proportions of correctly/incorrectly classified points in the validation set.

(d) Based on the results, would you choose the method in (a), (b), or (c)?