Assignment 4

March 25, 2016

1 LOGISTIC REGRESSION AND SVM

Problem 1: Suppose we collect data for a group of students in a class with variables X_1=hours studied, X_2=undergrad GPA, and Y=receive an A (So 1 for getting an A and 0 for not). Fit a logistic regression and produce estimated coefficient: $\beta_0 = -6, \beta_1 = 0.05, \beta_2 = 1$.

1. Estimate the probability that a student who studies for 40 hours and has GPA of 3.5 gets an A in the class. (5pts)

2. How many hours would the student in the previous part need to study to have a 50% chance of getting an A in the class. (5pts)

Problem 2: Recall that the formulation for non-separable SVM is:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} a_i \text{ such that:}$$
$$a_i \geq 0, a_i \geq 1 - y_i(b + w^T x_i), i = 1, 2, \cdots, n.$$

The dual formulation is:

$$\max_{\lambda,\mu} \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\lambda^T Y^T X X^T Y \lambda \text{ such that:}$$
$$\sum_{i=1}^{n} y_i \lambda_i = 0$$
$$0 \leq \lambda \leq C.$$

Let γ be the margin (the distant between hyperplane 1 and -1 in the lecture $\gamma = 2/\|w\|$). Suppose the dual problem is solved and you get solution $\lambda^* = (\lambda_1^*, \lambda_2^*, \cdots, \lambda_n^*)$. Show that:

$$4\gamma^{-2} = \sum_{i=1}^{n} \lambda_i^*.(10pts)$$

Problem 3: Given a data set X and response labels y. First choose value of the tuning parameters of non-separable SVM C=1. You see that on training set, you correctly classify A% of the data. After that you choose C=5, you correctly classify B% of the data. For which value of C, you get higher classification accuracy? Support your answer. (10pts)

Problem 4: The hinge loss is defined to be: $f(x) = max(0, 1-x)$. You can assume $x \in \mathscr{R}$. Using the definition of convex function, show that this loss function is convex.(10pts)

Problem 5: Consider logistic regression with intercept term. So the probablility of success is defined as:

$$P(y = \pm 1|\beta, \beta_0, x) = \frac{1}{e^{-y_i(\beta^T x_i + \beta_0)}}$$

β_0 plays the role of the intercept term similar to multiple linear regression. Then the negative loglikelihood function will be:

$$\min_{\beta,\beta_0}(f(\beta, \beta_0) = \sum_{i=1}^{n} log(1 + e^{-y_i(\beta^T x_i + \beta_0)}.$$

Suppose you are going to use gradient descent method to find the solution.

1. Calculate the derivate of the function in terms of β_0 (5pts).
2. Calculate the derivate of the function in terms of β (5pts).

2 MATLAB PROGRAMMING

Notice: Make sure you name data files correctly (training, validating, and testing...) If I can not run my code on your data to verify the results, you will automatically lose points.
Problem 1: In this problem, you will experiment with part of handwritten digit data set and logistic regression. The file digit.mat contains images of handwritten numbers 3 and 8.

- For logistic regression function in Matlab, the output y has to be 1 and 2. First you will need to change the label vector y. For observation labeled 3, change it to 1, otherwise change it to 2. Now you get a data set with labels 1 and 0. (so 1 for 3 and 2 for 8) (5pts).
- Randomly split the data into 3 parts using crossvalind function in Matlab. You get 3 data sets for training, validating, and testing. Name them: X_train,y_train,X_val,y_val,X_test,y_test. Save these data sets in to a Matlab data file using command : (5pts)

  save(yourname_Problem1_data.mat,'X_train','y_train','X_val','y_val','X_test','y_test')

- Fit a logistic regression on the training data using function mnrfit. Plot the coefficient using the function plot(coefficient). What do you think about the coefficient(5pts).

- Using function glmval to make prediction of success probabilities on the validating and testing data set. Once you obtain these success probabilities, go on and make prediction on the responses. So observations with success probabilities > 0.5, can be assigned to have response 1 (digit 3), otherwise it is assigned to response 2(digit 8). The accuracy is defined as the number of correctly predicted observations / total number of observations in that data set. Report the accuracy on validating set and testing set. (5pts)

Problem 2: In this problem, you will experiment with logistic regression with lasso penalty. The file digit.mat contains images of handwritten numbers 3 and 8.

- For lasso logistic regression function in Matlab, the output y has to be 1 and 0. First you will need to change the label vector y. For observation labeled 3, change it to 1, otherwise change it to 0. Now you get a data set with labels 1 and 0. (so 1 for 3 and 0 for 8) (5pts).

- Randomly split the data into 3 parts using crossvalind function in Matlab. You get 3 data sets for training, validating, and testing. Name them: X_train,y_train,X_val,y_val,X_test,y_test. Save these data sets in to a Matlab data file using command : (5pts)

  save(yourname_Problem2_data.mat,'X_train','y_train','X_val','y_val','X_test','y_test')

- Perform lasso logistic regression on the training data set using 30 values of λ. You can use function lassoglm with options for 30 different values ofr λ. (5pts)

- FitInfo of the model will give you information about values of λ chosen by the algorithm and corressponding deviance values (basicly the likelihood fit to the trainin data).Report the value of λ that gives you the minimum deviance. Fit the lasso logistic regression model using that value of λ. (5pts). (You can access deviance by syntax: FitInfo.Deviance, same goes for Lambda).

- Make prediction on validating and testing sets. Report accuracy (5pts). How is it compared with the previous problem?

Problem 3: In this problem, you will experiment with SVM with the same data set.

- For svm function in Matlab, the output y has to be 1 and -1. First you will need to change the label vector y. For observation labeled 3, change it to 1, otherwise change it to -1. Now you get a data set with labels 1 and -1. (so 1 for 3 and -1 for 8) (5pts).

- Randomly split the data into 3 parts using crossvalind function in Matlab. You get 3 data sets for training, validating, and testing. Name them: X_train,y_train,X_val,y_val,X_test,y_test. Save these data sets in to a Matlab data file using command : (5pts)

  save(yourname_Problem3_data.mat,'X_train','y_train','X_val','y_val','X_test','y_test')

- Choose a sequence of 30 values of parameter C (similar to when we do lasso). Something like C=-6:12/30:6; C=exp(C) would do. For each value of parameter C, fit a SVM model with linear kernel on the training data. Use that model to make prediction on the validating set and record the accuracy. (10pts) You can use function fitcsvm in matlab. To make prediction, use function: predict.

- For which value of C, you obtain highest accuracy of prediction on the validating set? Use that value of C to fit the SVM model on the training data. Make prediction using that model on the testing data. Report the accuracy on testing set. (5pts)

- Choose a sequence of 30 values of parameter C (similar to when we do lasso). Something like C=-6:12/30:6; C=exp(C) would do. For each value of parameter C, fit a SVM model with Gaussian kernel on the training data. Use that model to make prediction on the validating set and record the accuracy. (10pts)

- For which value of C, you obtain highest accuracy of prediction on the validating set? Use that value of C to fit the SVM model on the training data. Make prediction using that model on the testing data. Report the accuracy on testing set. (5pts) How is the accuracy compared to when you used linear kernel?