Problem Set 2
Due Sunday, October 29, 2017 at 11:55pm

How to Submit

Create one .zip file (**not** .rar or something else) of your code and written answers and submit it via `ilearn.ucr.edu`. Your zip file should contain

- Problem 1: ans1.[pdf|txt]
- Problem 2: learnlogreg.m, testlogreg.m, ans2.[pdf|txt]
- Problem 3: plotlogregerr.m & q3plot.pdf

Submit answers in a pdf or ascii text file, not an MS Word document.

Each file should include at the top (in comments if necessary)

- Your name & UCR student ID number
- The date
- The course (CS 171) & assignment number (PS 2)

Problem 1. [5 pts]

In class, we derived the gradient descent rule for logistic regression. In this problem, modify the loss function for logistic regression to include a regularization term with strength λ that penalizes the sum of the squares of the weights.

Write the new loss function. Derive the gradient descent rule for this new loss function.

Problem 2. [10 pts]

Write a function `learnlogreg` that uses the above gradient descent optimization.

Your function should have the signature `w = learnlogreg(X,Y,lambda)`. Use a constant step size, η, of 0.1. As a stopping condition, stop when the size of the gradient (as a vector) falls below 10^{-3}. Assume that the first feature of X (the first column) is a constant (all 1s) and no regularization should be applied to it. In class we assumed that the two classes were +1 and -1. You'll need to write this function to accept `Y` of +1 and 0.

Write a second function `err = testlogreg(X,Y,w)` which accepts a testing set and learned weights and returns the fraction of examples incorrectly classified.

Use the supplied `trainspam.txt`, `testspam.txt`, and `loadspdata.m` (all the same as last assignment) and test your logistic regression on this testing data. **What is the test error rate when training with $\lambda = 0.01$?** Note, you'll have to add the column of 1s; the data do not already contain it.

Problem 3. [10 pts]

Plot a comparison of the cross-validation estimate of the error rate with the testing error rate, as a function of λ.

Write a function with signature `plotlogregerr(trainX,trainY,testX,testY,lambdaset)` that will perform 5-fold cross validation for each of the λ values in `lambdaset` on the training set. It will also use the entire training set and calculate the error rate for the testing set on the same set of λs. Your code should plot each of these error rates (as separate curves on the same plot) as functions of λ. Use `loglog` (instead of `plot`) to plot on a log-log scale (which will make the differences more apparent).

Run `plotlogregerr` on the spam data with `lambdaset = logspace(-8,-2,10)`. Note, this could take a few minutes to complete. **Save your plot as `q3plot.pdf`**