CS189: Introduction to Machine Learning

Homework 4

Due: March 15, 2013

Submission: **bSpace**

Problem: Spam classification using Logistic Regression

The spam dataset given to you as part of the homework in `spamData.mat` (adapted from `http://archive.ics.uci.edu/ml/datasets/Spambase`) consists of 4601 email messages, from which 57 features have been extracted as follows:

- 48 features giving the percentage (0 - 100) of words in a given message which match a given word on the list. The list contains words such as business, free, george, etc. (The data was collected by George Forman, so his name occurs quite a lot!)
- 6 features giving the percentage (0 - 100) of characters in the email that match a given character on the list. The characters are `;( [ ! $ #` .
- Feature 55: The average length of an uninterrupted sequence of capital letters
- Feature 56: The length of the longest uninterrupted sequence of capital letters
- Feature 57: The sum of the lengths of uninterrupted sequence of capital letters

The dataset consists of a training set size 3065 and a test set of size 1536. One can imagine performing several kinds of preprocessing to this data. Try each of the following separately:

i) Standardize the columns so they all have mean 0 and unit variance.

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

iii) Binarize the features using $\mathbb{I}(x_{ij} > 0)$.

For this homework, you need to do the following:

1. Derive the batch gradient descent equations for logistic regression with l_2 regularization. Plot the training loss (the negative log likelihood of the training set) vs the number of iterations. Report errors on the train and test sets for all three feature sets given above.

Note: One iteration here amounts to scanning through the whole training data and computing the full gradient.

2. Derive stochastic gradient descent equations for l_2 regularized logistic regression. Plot the training loss vs number of iterations. Do you see any differences from the corresponding curve from (1)? If so, why? Report final errors on train and test sets for all three feature sets given above.

   Note: One iteration here corresponds to processing just one data point.

3. Instead of a constant learning rate (η), repeat (2) where the learning rate decreases as $\eta \propto 1/t$ for the t^{th} iteration. Plot the training loss vs number of iterations and report the error rates on all the above feature sets. Is this strategy better than having a constant η? Can you think any other strategies that might work well?

4. Do 5-fold cross validation (just for feature set (ii)) for the strength of the l_2 regularizer for stochastic gradient descent. Plot your cross-validation error vs the regularization parameter. Using your cross-validated regularization parameter, report the errors on the train and test sets (for feature set (ii)).

NOTE: You may use any programming language you wish as long as I can run your code with no/minimal setup. You are NOT supposed to use any kind of software package for logistic regression!

Submission Instructions

In your submission, you need to include a write up with answers to all the questions and the plots. You also need to include your code a README with instructions as to how I can run your code. This time I am trying to use bspace for submissions. You need to log in to **bspace** and to the assignments tab under the course and submit it there as **a zip file that would contain your code, README and report**. Your code should be simple MATLAB/python/other code and not include any big libraries etc. **Everyone needs to submit the assignment through their own bSpace account (even if they do it in groups)** so that I dont miss out on anyone while grading. While submitting, include the names of all the people you worked with.