CSE 847 (Spring 2020): Machine Learning— Homework 4

Instructor: Jiayu Zhou

Due on Friday, March 27th.

1 Logistic Regression: Experiment

In this part you will implement logistic regression. You are to write a MATLAB function called logistic_train.m that takes an input data set, a set of binary training labels, and an optional argument that specifies the convergence criterion, and returns a set of logistic weights. Specifically the function should have the following form:

```
function [weights] = logistic_train(data, labels, epsilon, maxiter)
% code to train a logistic regression classifier
%
% INPUTS:
%   data    = n * (d+1) matrix withn samples and d features, where
%             column d+1 is all ones (corresponding to the intercept term)
%   labels  = n * 1 vector of class labels (taking values 0 or 1)
%   epsilon = optional argument specifying the convergence
%             criterion - if the change in the absolute difference in
%             predictions, from one iteration to the next, averaged across
%             input features, is less than epsilon, then halt
%             (if unspecified, use a default value of 1e-5)
%   maxiter = optional argument that specifies the maximum number of
%             iterations to execute (useful when debugging in case your
%             code is not converging correctly!). If not specified set to 1000.
%
% OUTPUT:
%    weights = (d+1) * 1 vector of weights where the weights correspond to
%              the columns of "data"
```

The classifier should be trained using the gradient descent procedure as described in class, using the log-likelihood objective function (not the Newton-Raphson (IRLS) iterative procedure). You can initialize all the weights at 0. You will test the algorithm on the Spam Email data set[1]. There are 57 features and 2 class labels. Please read the README before you use this data. All features have been converted from counts into binary features (by splitting above and below the mean count).

Create a separate test data set consisting of all rows in the file from row 2001 to 4601 inclusive (and corresponding labels). You now have 2 data sets, a training data set with 2000 rows (the first 2000 rows of the original file) and a test data set with 2601 rows. Train your logistic regression classifier on the first n rows of the training data, n = 200; 500; 800, 1000; 1500, 2000 and report the accuracy on the test data as a function of n.

2 Sparse Logistic Regression: Experiment

In this part, you will perform experiments using sparse logistic regression (ℓ_1-regularized logistic regression). Use the Alzheimer's disease dataset as described in `https://github.com/jiayuzhou/CSE847/tree/master/data/alzheimers`. Sparse logistic regression is then applied to train a linear

[1] `https://github.com/jiayuzhou/CSE847/tree/master/data/spam_email`

model on the given training set and prediction is then performed on the given test set. You should use the implementation in SLEP[2], where the sparse logistic regression is the function `LogisticR`[3]. An example of using the sparse logistic regression is as follows:

```
function [w, c] = logistic_l1_train(data, labels, par)
% OUTPUT w is equivalent to the first d dimension of weights in logistic_train
%        c is the bias term, equivalent to the last dimension in weights in logistic_train.

% Specify the options (use without modification).
opts.rFlag = 1;  % range of par within [0, 1].
opts.tol = 1e-6; % optimization precision
opts.tFlag = 4;  % termination options.
opts.maxIter = 5000; % maximum iterations.

[w, c] = LogisticR(data, labels, par, opts);
```

The input `par` is the ℓ_1 regularization parameter, which scales from 0 to 1. Try different values of regularization parameter and report both the AUC (use Matlab code `perfcurve`) and the number of features selected (number of non-zero entries in `w`). A suggested list of parameters is [0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1], but other choices of parameters are also encouraged. Note that when parameter is 0, the formulation is equivalent to the classical logistic regression.

For both experiments in this homework, submit a brief report. In addition to the report, submit the hard copy of the MATLAB code (do add some comments in your code for others to understand your code).

[2] `https://github.com/jiayuzhou/SLEP/`

[3] Located at `https://github.com/jiayuzhou/SLEP/tree/master/SLEP/functions/L1/L1R`