Stats 141C	High Performance Statistical Computing	Spring 2017

Homework 3

Lecturer: Cho-Jui Hsieh — Date Due: June 1, 13:30pm (in class), 2017

Keywords: *Classification, Regression.*

For this homework, we will try classification and regression using datasets listed in `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. The format of data is called "SVMLight" format:

```
<label> <index>:<value1> <index>:<value2>  . . .
```

Each line contains an instance. For classification, <label> is an integer indicating the class label. For regression, <label> is the target value which can be any real number. Indices are in ascending order.

Problem 1. Regression [50 pt]

1. Download the "cpusmall" data from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html#cpusmall`. Randomly split the file into training set and testing set. Training set contains 80% instances and testing set contains 20% instances. Solve the ridge regression problem on the training set:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \{ \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i^T \boldsymbol{w}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \} := f(\boldsymbol{w}), \tag{1}$$

   where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th training sample, and $y_i \in \mathbb{R}$ is the i-th target value. Set $\lambda = 1$, solve (1) to get the model $\boldsymbol{w}^*$, and compute the Mean Square Error (MSE) on the testing set. The MSE is defined by

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\boldsymbol{x}_i^T \boldsymbol{w}^* - y_i)^2,$$

   where n_{test} is the number of testing instances. Report the MSE.

2. Run the ridge regression for $\lambda = 0.01, 0.1, 1, 10, 100$, and report the test MSE for each λ value.

3. Write the gradient descent algorithm with fixed step size for solving (1). The gradient descent algorithm is in Algorithm 1. Set $\lambda = 1, \epsilon = 0.001$ and test the algorithm for $\eta = 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$. Report your findings.

4. Run gradient descent (with fixed step size) on "E2006-tfidf" data (see `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html#E2006-tfidf`). The training data can be downloaded from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression/E2006.train.bz2`. The testing data can be downloaded from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression/E2006.test.bz2`. Download both training and testing data from the website. Run your gradient descent implementation. Set $\epsilon = 0.001$ and $\lambda = 1$, try to select a good step size, and report your step size and the prediction accuracy you get.

Algorithm 1 Gradient Descent with Fixed Step Size

- Input: η: step size, ϵ: Stopping condition, $\boldsymbol{w}_0$: initial solution
- $\boldsymbol{w} \leftarrow \boldsymbol{w}_0$
- $r_0 \leftarrow \|\nabla f(\boldsymbol{w}_0)\|$
- For iter $= 1, 2, \ldots, 50$ (Maximum 50 iterations)
  - $\boldsymbol{g} = \nabla f(\boldsymbol{w})$
  - If $\|\boldsymbol{g}\| \leq \epsilon r_0$: Break (End program)
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \boldsymbol{g}$

Algorithm 2 Gradient Descent with Line Search

- Input: ϵ: Stopping condition, $\boldsymbol{w}_0$: initial solution
- $\boldsymbol{w} \leftarrow \boldsymbol{w}_0$
- $r_0 \leftarrow \|\nabla f(\boldsymbol{w}_0)\|$
- For iter $= 1, 2, \ldots, 50$ (Maximum 50 iterations)
  - $\boldsymbol{g} = \nabla f(\boldsymbol{w})$
  - If $\|\boldsymbol{g}\| \leq \epsilon r_0$: Break (End program)
  - $\eta \leftarrow 1$
  - While $(f(\boldsymbol{w} - \eta \boldsymbol{g}) \geq f(\boldsymbol{w}))$

    $\eta \leftarrow \eta/2$
  - $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \boldsymbol{g}$

Problem 2. Classification (Logistic Regression) [50pt]

In this problem, you will write your own code for logistic regression. Given training data $\{\boldsymbol{x}_i, y_i\}$ for $i = 1, 2, \cdots, n$. Each $\boldsymbol{x}_i$ is a feature vector and each y_i is the +1/-1 label. logistic regression model can be learned by solving

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \{\sum_{i=1}^{n} \log(1 + e^{-y_i \boldsymbol{w}^T \boldsymbol{x}_i}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\} := f(\boldsymbol{w}). \tag{2}$$

The model $\boldsymbol{w}^*$ can then be used for prediction.

1. Derive the gradient of (2).
2. Implement gradient descent with fixed step size to solve (2). Test it on news20 binary classification dataset (`https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/news20.binary.bz2`). Note that there will be only one file (news20.binary). Split it into 80% training and 20% testing. Solve the logistic regression problem using $\lambda = 1$ on the training set, and report the prediction accuracy on test set.
3. Implement gradient descent with line search (see Algorithm 2). Report the testing accuracy for $\lambda = 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$.

Problem 3. Bonus [30pt]

Try some of the scikit-learn classification/regression packages for these two problems, and compare the result with your implementation. Report your findings.