Machine Learning & Data Mining
Homework 1

Caltech CS/CNS/EE 155
January 12th, 2015

Overview

- This homework is due on Moodle at 5:00 pm on January 20, 2015.
- This homework explores the qualitative effects of different types of regularization and demonstrates how Lasso (ℓ_1) regularization works.
- Students should be able to complete each problem in 2-3 hours.

1 Effects of Regularization [Shenghan Yao, 30 points]

Basic questions:

Question A [4 points]: In order to prevent over-fitting in the least-squares linear regression problem, we add a regularization penalty term. Can adding the penalty term decrease the training (in-sample) error? Also can it always decrease the out-of-sample errors? Please justify your answers.

Question B [2 points]: ℓ_1 regularization is sometimes favored over ℓ_2 regularization due to its ability to generate a sparse w (more zero weights). In fact, ℓ_0 regularization (using ℓ_0 norm instead of ℓ_1 and ℓ_2 norm) can generate a sparser w, which seems favorable in high-dimensional problems. However, it is rarely used. Could you please explain why?

Implementation of L-2 regularization:

We are going to experiment with linear regression for the Red Wine Quality Rating data set.[1] Download the data for training and testing. There are three training data sets wine_training1.txt, wine_training2.txt, and wine_training3.txt (100, 50 and 25 data points respectively) and one testing data set wine_testing.txt (100 data points).

The data in each data set represents how 11 different factors (first 11 columns) affect the wine quality (the 12th column). Each column of data represents a different factor, and is described in brief in the file wine_name.txt. When evaluating training error (E_{in}) and validation error (E_{out}) use the square error:

$$E = \frac{1}{N}\sum_{n=1}^{N}\left(w^T x_n - y_n\right)^2$$

Implement the ℓ_2 regularized least-squares linear regression that minimizes:

$$E = \frac{1}{N}\sum_{n=1}^{N}\left(w^T x_n - y_n\right)^2 + \frac{\lambda}{N} w^T w$$

Train the model with 10 different choices of λ:

$$\lambda = [0.0001, 0.0005, 0.0025, 0.0125, 0.0625, 0.3125, 1.5625, 7.815, 39.0625, 195.3125]$$

(Brief instructions for loading data into Matlab are in loading_data.txt).

Question C [9 points]: Do the following for each of the 3 training data sets and attach your plots in the homework submission (Hint: use semi-log plot):

(1) Plot the training error (E_{in}) versus different λs.

[1] Wine quality Data Set: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

(2) Plot the validation error (E_{out}) versus different λs.

(3) Plot the norm of w versus different λs.

Question D [3 points]: Considering that the data in wine_training3.txt is a subset of the data in wine_training1.txt, compare errors (training and validation) resulting from training with wine_training1.txt (100 data points) versus wine_training3.txt (25 data points). Please briefly explain the difference.

Question E [4 points]: Briefly explain the qualitative behavior (i.e., over-fitting and under-fitting) of the training and validation errors with different λs while training with data in wine_train1.txt.

Question F [4 points]: Briefly explain the qualitative behavior of norm of w with different λs while training with the data in wine_train1.txt.

Question G [4 points]: If the model were trained with wine_train3.txt, which λ would you choose to train your final model? Why?

2 Lasso (ℓ_1) vs. Ridge (ℓ_2) Regularization [Bryan He, 30 points]

Many datasets we now encounter in regression problems are very high dimensional. One way to handle this is to encourage the weights to be sparse, and only allow a small number of features to have non-zero weight. The direct way of encouraging a sparse weight vector is use ℓ_0 regularization and penalize the ℓ_0 norm, which is the number of non-zero elements in the vector. However, the ℓ_0 norm is far from smooth, and as a result, it is hard to optimize.

Two related methods are Lasso (ℓ_1) regression and Ridge (ℓ_2) regression. Although both result in shrinkage estimators, only Lasso regression results in sparse weight vectors. This problem compares the behavior of the two methods.

Question A: Let $\mathcal{D}$ be a set of data points, and let $\mathbf{w}$ be a D-dimensional weight vector. Both Lasso and Ridge regression can be formulated as a maximum a posteriori (MAP) estimate of $p(\mathbf{w}|\mathcal{D})$ with different priors $p(\mathbf{w}|\lambda)$, where λ is a parameter controlling the shape of the prior.

(1) [3 points] In the case of Lasso regression, the prior is that the weights are independent and identically distributed (i.i.d.) zero-mean Laplacian random variables

$$p(\mathbf{w}|\lambda) = \prod_{j=1}^{D} \text{Lap}(w_j|0, 1/\lambda) = \prod_{j=1}^{D} \frac{\lambda}{2} e^{-\lambda|w_j|}.$$

Show that under this prior,

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmax}}\, p(\mathbf{w}|\mathcal{D})$$

is equivalent to

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmin}} \left(-\log\, p(\mathcal{D}|\mathbf{w}) + \lambda\|\mathbf{w}\|_1\right).$$

(2) [3 points] In the case of Ridge regression, the prior is that the weights are i.i.d. zero-mean Normal random variables

$$p(\mathbf{w}|\lambda) = \prod_{j=1}^{D} \mathcal{N}(w_j|0, 1/2\lambda) = \prod_{j=1}^{D} \sqrt{\frac{\lambda}{\pi}} e^{-\lambda w_j^2}.$$

Show that under this prior,

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w}|\mathcal{D})$$

is equivalent to

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left(-\log\, p(\mathcal{D}|\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \right).$$

(3) [3 points] Suppose that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$ and $\mathcal{D}$ contains $\mathbf{X}$ and $\mathbf{y}$. Show that

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathcal{D}|\mathbf{w})$$

is equivalent to

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

This means that least-squares linear regression is a maximum likelihood estimator when there is white Gaussian noise.

Question B: This section compares the behavior of Lasso and Ridge regression on a synthetic dataset. The dataset consists 1000 independent samples of a 9-dimensional feature vector $\mathbf{x} = [x_1, \ldots, x_9]$ drawn from a uniform distribution on the interval $[-1, 1]$, along with the response

$$y = -4x_1 - 3x_2 - 2x_3 - 1x_4 + 0x_5 + 1x_6 + 2x_7 + 3x_8 + 4x_9 + n = \mathbf{w}^T\mathbf{x} + n,$$

where n is a standard Normal random variable. The file question2data.txt consists of 1000 lines of 10 tab-delimited values. The first 9 columns represent $x_1, \ldots, x_9$, and the last column represents y. Each row consists of one sample. Using MATLAB, you may load the file by running

```
>> data = dlmread('question2data.txt', '\t');
>> X = data(:, 1:9);
>> Y = data(:, 10);
```

You may also generate the dataset using the process specified above. This problem explores the behavior of the estimated weights as the strength of the regularization (λ) varies.

(1) [4 points] Estimate the weights $\mathbf{w}$ using linear regression with Lasso regularization for various choices of λ. For each of the weights, plot the weight as a function of λ (start with $\lambda = 0$ and increase λ until all weights are small). Using a linear scale for λ will allow the plot to be easily interpreted.

(2) [4 points] Estimate the weights $\mathbf{w}$ using linear regression with Ridge regularization for various choices of λ. For each of the weights, plot the weight as a function of λ (start with $\lambda = 0$ and increase λ until all weights are small).

(3) [3 points] As regularization parameter varies, how many of the estimated weights are exactly zero with Lasso regression? How many of the estimated weights are exactly zero with Ridge regression?

Question C: For general choices of $p(D|\mathbf{w})$, an analytic solution for regularized linear regression may not exist. However, when $p(D|\mathbf{w})$ has a standard normal distribution (corresponding to linear regression), an analytic solution exists for 1-dimensional Lasso regression and for Ridge regression in all dimensions.

(1) [4 points] Solve for $\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{w}^T\mathbf{x}\|^2 + \lambda\|\mathbf{w}\|_1$ in the case of a 1-dimensional feature space. This is linear regression with Lasso regression.

(2) [1 point] Suppose that when $\lambda = 0$, $w_1 \neq 0$. Does there exist a value for λ such that $w_1 = 0$? What is the smallest such value?

(3) [4 points] Solve for $\underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{w}^T\mathbf{x}\|^2 + \lambda\|\mathbf{w}\|_2^2$ for an arbitrary number of dimensions. This is linear regression with Ridge regression.

(4) [1 point] Suppose that when $\lambda = 0$, $w_i \neq 0$. Does there exist a value for λ such that $w_i = 0$? What is the smallest such value?