华南理工大学

South China University of Technology

The Experiment Report of Machine Learning

SCHOOL: SCHOOL OF SOFTWARE ENGINEERING

SUBJECT: SOFTWARE ENGINEERING

Author:
Zequan Zeng

Supervisor:
Qingyao Wu

Student ID：201720144917

Grade:
Graduate

December 14, 2017

Linear Regression, Linear Classification and Gradient Descent

Abstract—Regression and Classification are two basic techniques of machine learning.In this report,We will introduce Linear Regression and Linear Classification .And we will update two models parameters using Gradient Descent.Among our experiments,two models can finally get low loss using Gradient Descent.

I. INTRODUCTION

In this report, we will introduce Linear Regression and Linear Classification.We use the mean square error (MSE) as the loss function of the Linear Regression.We will use support vector machine(SVM) to solve Linear Classification.

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.Then we will update two models parameters using Gradient Descent.We will also explore how the different learning-rate of the Gradient Descent influence the models.

II. METHODS AND THEORY

1. Linear Regression

Linear Regression takes the form

$$\mathrm{y_i} = w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im} + b_i = X_i^T W + b_i$$

where $i = 1,2,...,n$,and $^{\mathrm{T}}$ denotes the transpose,so that $\mathrm{X}_i^T W$ is the inner product between vectors $\mathrm{X_i}$ and W .

We can also stack these n equations together and get a vector form as

$$\mathrm{Y} = X^T W$$

where

$$\mathrm{Y} = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \mathrm{X} = \begin{pmatrix} x_{11}, x_{21}, ..., x_{n1} \\ x_{12}, x_{22}, ..., x_{n2} \\ ..... \\ x_{1m}, x_{2m}, ..., x_{nm} \end{pmatrix}, \mathrm{W} = \begin{pmatrix} w_1 \\ w_2 \\ ... \\ w_m \end{pmatrix}$$

We use the mean square error (MSE) as the loss function of the Linear Regression.The mean square error takes the form

$$\mathrm{MSE} = \frac{1}{2n} \sum_{i=1}^{n} (y_{\mathrm{i}} - \hat{y}_i)^2$$

where $\hat{\mathrm{y}}_{\mathrm{i}}$ is the predictive value of sample i, and y_i is the real value of sample i.

2. Linear Classification

Linear Classification takes the same form as Linear Regression. But assume that $\mathrm{y}_i = \{-1,1\}$, we will set a threshold that if the predictive value is higher than the threshold , then we put the classification result to 1,and if the predictive value is lower than the threshold , then we put the classification result to -1.

We use the Support Vector Machine (SVM) model to solve Linear Classification.Then the loss function of the model takes the form

$$\frac{\| \mathrm{W} \|^2}{2} + \frac{\mathrm{C}}{n} \sum_{i=1}^{n} \max(0, 1 - y_i (W^T x_i + b))$$

where W is the weight vector of the decision variables, $\mathrm{y_i}$ is the real value of the sample i, $\mathrm{x}_i \in \mathrm{R}^{\mathrm{m}}$ is the decision variables, b is the bias, C is a hyper parameter.

3. Gradient Descent

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.

We will update both two models parameters using Gradient Descent.We use the follow formula to calculate the gradient of the loss function and update the parameters:

$$W = W - \frac{\alpha}{n} \sum_{i=1}^{n} \frac{\partial \mathrm{J}(W_{\mathrm{x}_i})}{\partial W_{x_i}}$$

where W is the weight vector of the decision variables, α is the learning-rate and different value of learning-rate will affect the convergence , $\frac{\partial \mathrm{J}(W_{\mathrm{x}_i})}{\partial W_{x_i}}$ is the gradient of sample i.Here we use the whole samples to calculate gradient and update W parameters.

III. EXPERIMENT

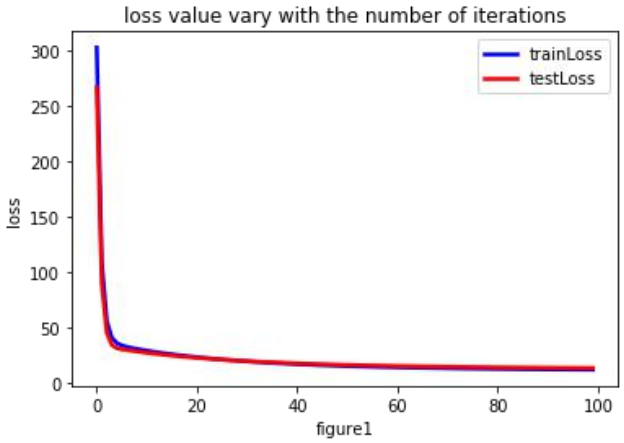
1. Dataset

Linear Regression uses the scaled edition of Housing in LIBSVM Data,including 506 samples and each sample has 13 features.And we divide the data to training data including 404 samples and validation data including 102 samples.

Linear classification uses the scaled edition of Australian in LIBSVM Data,including 690 samples and each sample has 14 features.And we divide the data to training data including 552 samples and validation data including 138 samples.

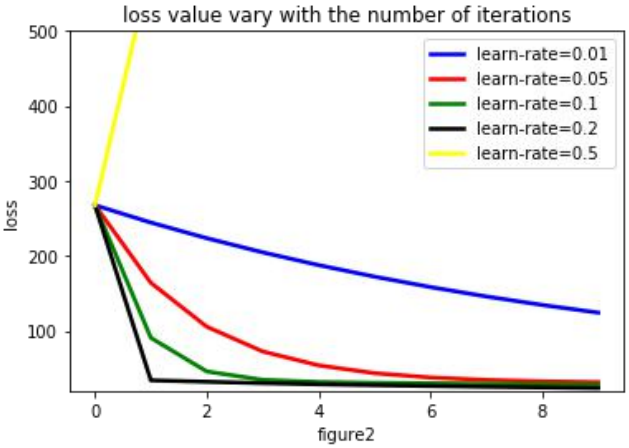
2. Experiment of Linear Regression

1) We set the learning-rate to 0.1, the iterations to 100,and initialize linear model parameters to zeros.The figure1 shows that train-loss and test-loss vary with the number of iterations.



From the figure1,we can find that both train-loss and test-loss decrease with the number of iterations, which is shown that the gradient descent can effectively minimize the loss function.It is also shown that the model has a good robustness.

2) We set the learning-rate to 0.01、0.05、0.1、0.2、0.5 respectively, the iterations to 10,and initialize linear model parameters to zeros.The figure2 shows that the loss vary with the number of iterations under different learning-rate.



From the figure2,we can find that the lower learning-rate, the slower the convergence of the model.But if the learning-rate is too high, the gradient descent can't minimize the loss function, even it will get the worse result.

3) We set the learning-rate to 0.05, the iterations to 10、50、100、1000 respectively, and initialize linear model parameters to zeros.

Table1

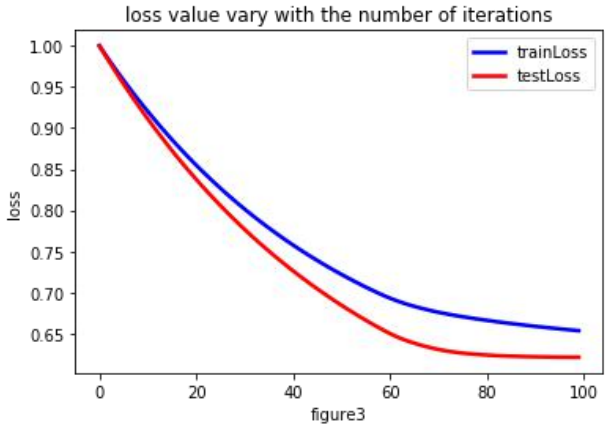
Test-loss vary with the number of iterations

iterations	10	50	100	1000
test-loss	31.438	21.519	16.657	12.287

From Table1, we can find that the more iterations, the lower test-loss.It is shown that the effect of gradient descent is related to the number of iterations.

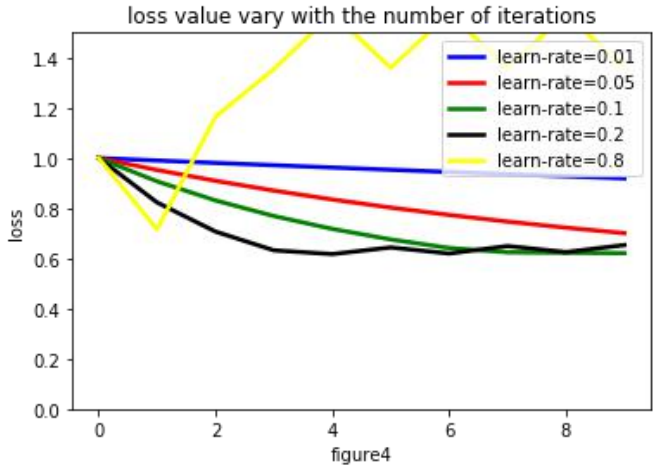
3. Experiment of Linear Classification

1) We set the learning-rate to 0.01, the C to 1,the iterations to 100,and initialize linear model parameters to zeros.The figure3 shows that train-loss and test-loss vary with the number of iterations.



From the figure3, we can find the similar results between Linear Regression and Linear Classification.

2)We set the learning-rate to 0.01、0.05、0.1、0.2、0.8 respectively,the C to 1, the iterations to 10,and initialize linear model parameters to zeros.The figure4 shows that the loss vary with the number of iterations under different learning-rate.



From the figure4, we can also find the similar results that learning-rate is related to the effect of gradient descent.

IV. CONCLUSION

Among our experiments, we can find that Gradient descent is the good algorithm to minimize the loss function, when we use the MSE as the loss function of Linear Regression and use SVM to solve Linear Classification.

It is obviously shown that the lower learning-rate, the slower convergence of the model.But if the learning-rate is too high, the model might get the worse result.

It is also shown that the effect of gradient descent is related to the number of iterations.The more iterations, the lower loss.