- Linear regression and support vector machine

Goal

The goal of this experiment it:

1. Compare and understand the difference between gradient and batch random stochastic gradient descent.

2. Compare and understand the differences and the relationships between logistic regression and linear classification

3. Further understand the principles of SVM and practice on lager data.

Dataset

For this experiment we used a9a testing of LIBSVM Data, which including 32561/16281(testing) samples and each sample has 123/123 features. That we will divide into training set and validation set.

Environment for experiment:

To realize our experiment we will need to setup the following software and necessary library

python3, at least including following python package: sklearn，numpy, jupyter，matplotlib

It is recommended to install anaconda3 directly, which has built-in python package above.

Experiment step

The experiment will follow two step which are:

Logistic regression and batch stochastic Gradient descent, linear classification and batch stochastic gradient descent.

- **Logistic regression and batch stochastic.**

1. Load the training set and validation set.

2. Initialize logistic regression model parameter (you can consider initializing zeros, random numbers or normal distribution).

3. Select the loss function and calculate its derivation, find more detail in PPT.

4. Determine the size of the batch_size and randomly take some samples, calculate gradient G toward loss function from **partial samples**.

5. Use the **SGD** optimization method to update the parametric model and encourage additional attempts to optimize the **Adam** method.

6. Select the appropriate threshold, mark the sample whose predict scores **greater than the threshold as positive, on the contrary as negative**. Predict under validation set and get the loss *Lvalidation.*

7. Repeat step 4 to 6 for several times, and **drawing graph of** *Lvalidation* **with the number of iterations.**

- Linear classification and batch stochastic gradient

1. Load the training set and validation set.

2. Initialize SVM model parameters (you can consider initializing zeros, random numbers or normal distribution).

3. Select the loss function and calculate its derivation, find more details in PPT.

4. Determine the size of the batch_size and randomly take some samples, calculate gradient G toward loss function from **partial samples**.

5. Use the **SGD** optimization method to update the parametric model and encourage additional attempts to optimize the **Adam** method.

6. Select the appropriate threshold, mark the sample whose predict scores **greater than the threshold as positive, on the contrary as negative**. Predict under validation set and get the loss *Lvalidation.*

7. Repeat step 4 to 6 for several times, and **draw graph of** *Lvalidation* **with the number of iterations.**

- Conclusion

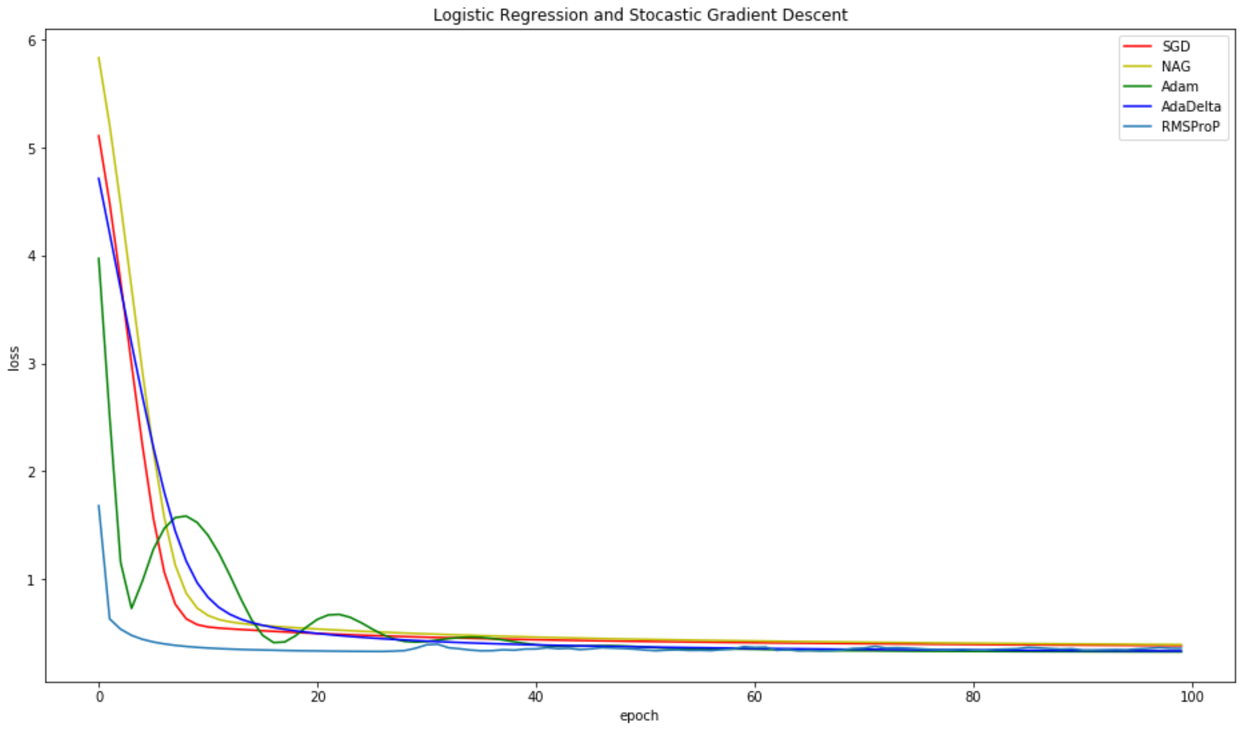

Fig 1: logistic regression and stochastic gradient descent.

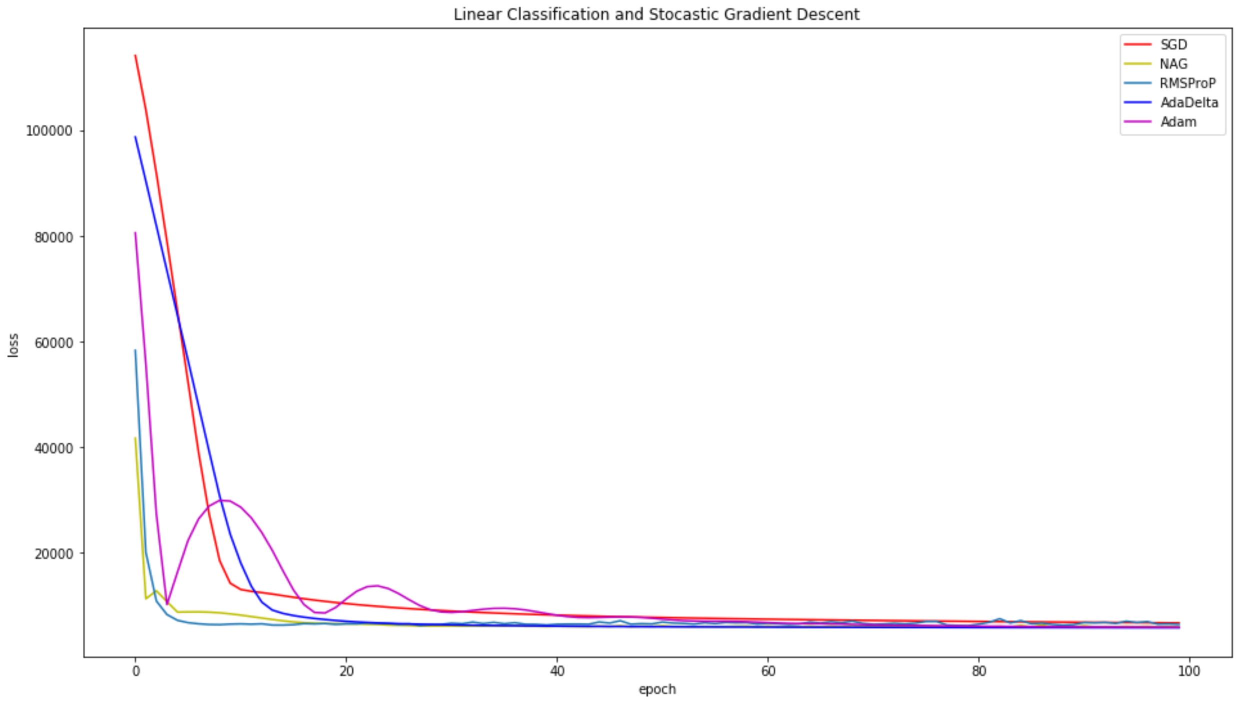

Fig 2: linear regression and stochastic gradient descent.

We therefore assume that: in the logistic regression, data used can be either categorical or quantitative, but the result is always categorical.

In the linear regression, a linear relation between the explanatory variable and the response variable is assumed and parameters satisfying the model are found by analysis, to give the exact relationship. Linear regression is carried out for the quantitative variables, and the resulting function is quantitative.