Homework 4

INFX 5

Submitted by: Rajesh Awal

Q.no.1 Answer:

Stochastic gradient decent or mini batch gradient technique linear regression training algorithm can be used for training data with million features.

Q.no.2 Answer:

If the features have very different scale, it will take a longer time to optimize or to converge. to solve this problem, one method could be feature engineering that includes data standardization and normalization. Data standardization is the procedure to convert the data within one standard deviation and normalization is the procedure to convert the data between 0 to 1.

Q.no.3 Answer:

Because the minimization surface in gradient descent is convex, it will not stuck at local minimum

Q.no.4 Answer:

The convergence is not about the number of runs, rather it more about the parameter one chooses. If the parameters chosen are correct, it is very likely that the result converge near toward the global minimum, and for every run with similar correct parameters, similar value is obtained.

Q.no.5 Answer:

We can solve the problem of increasing validation error by minimizing the step size. Minimizing step size increase the probability that convergence is being optimized

Q.no.6 Answer:

It is not a good idea to stop Mini-batch gradient descent immediately when the validation error goes up. It is because it is not considered as optimized. Rather the iteration needs to complete for each batch. Among them, the smallest error parameters are considered.

Q.no.7 Answer:

Stochastic gradient descent will reach the vicinity of optimal solution faster than other gradient descent. Among all gradient descent, batch gradient descent will converge. We can make other converge by changing or minimizing the step size and i\iterating again.

Q.no.8 Answer:

If there is large gap between the training error and the validation error, it means that the model is overfitted. To solve this problem, we can (a)reduce the degree of polynomial in the regressor, (b) add l1 or l2 regularization parameters that reduce the valueless parameters, (c) use n-fold cross validation model

Q.no.9 Answer:

Since the error is very high, it is the problem of high bias as it is deviating away from the actual value. To solve this problem, we should decrease the regularization parameter, α.

Q.no. Answer:

a. We use ridge regression instead of plain linear regression when we have a lot number of features. The regularization parameter on ridge regression prevents overfitting problem.
b. We use lasso regression instead of ridge regression, when there are only few features important to train the model. The L1 parameters reduce the coefficient of unimportant parameters toward zero making its contribution insignificant.
c. Elastic net has L1 and L2 regularization parameters. We use elastic net when we want benefit of both L1 and L2, especially when we have collinear data.

Q.no.11 Answer:

If we want to classify pictures as outdoor/indoor and daytime/nighttime, we should use two logistic regression classifiers. It is because we have two output that are not mutually exclusive and SoftMax are used only in mutually exclusive output.

Q.no.12 Answer:

The linear logit for softmax regressor are:

$$S_k(x) = \beta_{k,o} + \beta_k^t x$$

The probability produce by the logit is given by:

$$P_k = \frac{e^{sk(x)}}{\sum_{j=1}^{k} e^{sj(x)}}$$

For two class softmax regressor, we get the probabilty value for two class as:

$$P_1 = \frac{e^{s1(x)}}{e^{s1(x)} + e^{s2(x)}} = \frac{1}{1 + e^{s1(x)+e^{s2(x)}}}$$

$$P_2 = \frac{e^{s2(x)}}{e^{s1(x)} + e^{s2(x)}} = \frac{e^{s2(x)-e^{s1(x)}}}{1 + e^{1+e^{s1(x)+e^{s2(x)}}}}$$

For the softmax, we have,

$s_1(x) - s_2 = \beta_{1,0} + \beta_1^T x$ -($\beta_{2,0} + \beta_2^T x$) = $\beta_{1,0} - \beta_{2,0} + (\beta_1 - \beta_2)^T x = \gamma_1 - \gamma_2 x$

And,

$$P_1 = \frac{1}{1 + e^{\beta_1 - \beta_2 x}}$$

From this equation, the four unknown $\beta_{1,0}, \beta_1, \beta_{2,0}, \beta_2$

Changes to
γ_1, γ_2

Softmax, only understand one parameters and it helps to solve the early stoping for softmax.