CHAPTER 4

Exercise Question & Answers

1. What Linear Regression training algorithm can you use if you have a training set with millions of features?

Answer: Normal Equations cannot be used because to the large number of features and it will be highly computationally expensive. Gradient Descent is an alternative.

2. Suppose the features in your training set have very different scales. What algorithms might suffer from this, and how? What can you do about it?

Answer: Because the model will take longer to achieve the global maximum, the Gradient Descent suffers from scale-related issues. This difficulty can always be solved by scaling the features.

3. Can Gradient Descent get stuck in a local minimum when training a Logistic Regression model?

Answer: There is no local minimum since the cost function of the Logistic Regression Model is convex.

4. Do all Gradient Descent algorithms lead to the same model provided you let them run long enough?

Answer: No. The model can diverge if the learning rate is too high. Based on where the initialization is, it can only reach the local minimum.

5. Suppose you use Batch Gradient Descent and you plot the validation error at every epoch. If you notice that the validation error consistently goes up, what is likely going on? How can you fix this?

Answer: If the validation error continuously rises, the model may be diverging as a result of the high learning rate. Divergence is shown when the training error also increases. This may be remedied by slowing down the learning rate and then retraining. If the training error does not increase, our model is overfitting, and we will need to retrain with a new model.

6. Is it a good idea to stop Mini-batch Gradient Descent immediately when the validation error goes up?

Answer: No, since reaching the minimum will be unpredictable (just like Stochastic Gradient Descent, but to less degree). If the mistake does not improve after a time, we may always go back to the best case.

7. Which Gradient Descent algorithm (among those we discussed) will reach the vicinity of the optimal solution the fastest? Which will actually converge? How can you make the others converge as well?

Answer: Because we are utilizing one random training data for each iteration, the Stochastic Gradient Descent will reach the quickest. The Batch Gradient Descent, on the other hand, is the only one that really converges. We won't be able to force the others to converge; they'll just get close to the global minimum.

8. Suppose you are using Polynomial Regression. You plot the learning curves and you notice that there is a large gap between the training error and the validation error. What is happening? What are three ways to solve this?

Answer: The model is overfitting the data if there is a difference between the training and validation error. One of three options for avoiding overfitting the data is to: More data should be used to train the model, it should be regularized, and the model's complexity should be reduced (degree of freedom)

9. Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization hyperparameter α or reduce it?

Answer: Because the errors are both significant, the model has a strong bias, suggesting incorrect assumptions and hence underfitting. We must minimize alpha in order to reduce high bias.

10. Why would you want to use:
• Ridge Regression instead of plain Linear Regression (i.e., without any regularization)?
• Lasso instead of Ridge Regression?
• Elastic Net instead of Lasso?

Answer: Ridge Regression instead of plain Linear Regression- To avoid overfitting, the Ridge Regression regularizes the Linear Regression.
Lasso instead of Ridge Regression- Lasso, which employs L1 norm regularization, conducts feature selection by automatically removing the weights of the least essential features.
Elastic Net instead of Lasso- If there are many characteristics or features that are closely linked, Elastic Net is recommended over Lasso.

11. Suppose you want to classify pictures as outdoor/indoor and daytime/nighttime. Should you implement two Logistic Regression classifiers or one Softmax Regression classifier?

Answer: Because there are two separate binary classifications (outside vs indoor, daylight vs nighttime), we should construct two Logistic Regression classifiers. Each classifier can be represented by a single image. Out of all of them, Softmax Regression falls under only one category.

12. Implement Batch Gradient Descent with early stopping for Softmax Regression (without using Scikit-Learn).

Answer: https://github.com/arafatnoor/Assignment-4