Neural Networks: Practical Issues

Michael

School of Mathematics and Statistics, UCD

School of Economics, University of Nottingham

1 Introduction

When we apply our machine learning (ML) or deep learning (DL) models in different problems, we will encounter so many different practical issues, such as choices of activation functions, value of learning rate, etc. Knowing how to do hyper-parameter tuning, regularization and optimization is essential for improving our ML/DL models. According to Andrew Ng, practicing ML/DL models is a process of iteration, in which you have to learn to find the best choices of hyper-parameters by keeping trying different methods.

In this set of notes, we will discuss the following topics:

- Model assessment
- Regulation
- Optimization
- Hyper-parameters tuning and programming framework

2 Model Assessment

The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is extremely important in practice, since it guides the choice of learning method or model, and gives us a measure of the quality of the ultimately chosen model.

To assess our model, we need split our data into three different datasets:

- Training set
- Validation set
- Test set

When your dataset id not very big, the ratio for those datasets could be: 60/20/20. However, when you have a large dataset (for instance, 1 million), then the ratio for those datasets should be around: 98/1/1.

Before we continue to discuss model assessment, we have to know the difference between model selection and model assessment:

- Model selection: estimating the performance of different models in order to choose the best one.
- Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

The training set is used to fit the models; the validation set is used to estimate prediction error for model selection; the test set is used for assessment of the generalization error of the final chosen model. Ideally, the test set should be kept in a "valut", and be brought out only at the end of the data analysis.

2.1 The Bias-Variance Decomposition

Andrew Ng states that almost every good machine learning practitioner has a very sophisticated view on understanding of bias-variance decomposition. Hence, it is very importance to know what kind of factors will affect the bias-variance trade-off.

Now, if we assume that $Y = f(X) + \epsilon$ where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma_\epsilon^2$, we can derive an expression for the expected prediction error of a regression fit $\hat{f}(X)$ at an input point $X = x_0$, using squared-error loss:

$$\begin{aligned}
\text{Err}(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\
&= \sigma_\epsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}
\end{aligned}$$

The first term is the variance of the target around its true mean $f(x_0)$ and cannot be avoided no matter how well we estimate $f(x_0)$, unless $\sigma_\epsilon^2 = 0$. The second term is the squared bias, the amount by which the average of our estimate differs from the true mean; the last term is the variance; the expected squared deviation of $\hat{f}(x_0)$ around its mean. *Typically the more complex we make the model $\hat{f}$, the lower the (squared) bias but the higher the variance.*

When we had higher variance, we say our model is over-fitting, and when we had higher bias, we say our model is under-fitting.

Problem	Recipe
High bias	big neural network model, or more complex model
High variance	more data, regulation

One should realize that sometimes it is difficult to reduce bias and variance at the same time. In the era of big data, the bias-variance trade-off can be tackled in some sense when you got both complex neural network model and big dataset.

2.2 Regularization

The purpose of doing regularization is mainly to reduce the variance. Generally, we have L1 and L2 regularization. Taking regression models as examples, with L2 regularization, we have the following cost function:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

With L1 regularization, the Lasso regression add s the absolute value of magnitude of coefficient as penalty term to the loss function:

$$\sum_{i=1}^{n}(yi - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In neural network model, the formulas of cost function have to adjusted based on the above regularization. One thing I have to mention is that one should realize that when you change the cost function, the formulas for doing gradient descent will also change.

There is another way for doing regularization, which avoids the optimization process through the cost function. It is called dropout regularization. You can read more on this method of regularization in this website. By the way, dropout regularization is very frequently used by ML practitioners as it is more efficient.

There are other techniques that can help us to get the same effects of regularization, which include:

- Data augmentation
- Early stopping of iteration (no over-fitting)

3 Introduction

This LaTex template is for students who are taking courses from *Chair for the Study of Economic Institutions, Innovation, and East Asian Development* at Goethe University in Frankfurt. The Top and bottom margins are one inch. The left margin is 3cm, and the right is 4cm. It uses Times New Roman font and 12pt size. The text is justified with 1.5 line spacing.

The main advantage of using LaTex is the efficiency of citation. When you write any academic paper, you need to cite sources properly. In WORD, you have to copy the right references in the right format and organize them in alphabetical order. With LaTex, all you need to do is to type:

```
\cite{} % gives inline citation
\citep{} % gives Parenthetical citation
```

The above code would generate the citation and reference entry for you automatically. For instance, according to Christensen et al. (2013) disruptive innovation refers to innovations and technologies that make expensive or sophisticated products and services accessible and more affordable to a broader market (Greenwade, 1993).

Every time you cite, the reference will be added on the reference list.

4 Literature Review

I would like to write a literature review here. This is a citation example.

Nothing is new. It is just a tex compiler with different settings.

Hope you enjoy it.

4.1 Nature of Innovation

4.2 Innovation Strategies

5 Data Description

6 Methodology

7 Results and Discussion

8 Conclusion

References

Christensen, C., Raynor, M. E., and McDonald, R. (2013). *Disruptive innovation.* Harvard Business Review.

Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351.