Getting Your Bias Variance Right and Regularization

Divisha Bera
Software Engineer

ABSTRACT
It is clear that with ever improving computational power and endless data, there have been more breakthroughs in Machine Learning. Some practices have clearly emerged as promising while building a neural network. A performance metric to judge the model, is to see if it is in the wrong side of bias or variance. While building a classifier, cases with high bias, and high variance crop up. This paper shall attempt to shed some light on the problem of bias-variance, and how to solve them, with some approaches to perform Regularization.

General Terms
Machine Learning, Neural Networks

Keywords
Machine Learning, Bias, Variance, Neural Networks, Regularization.

1. INTRODUCTION
Since the beginning to time humans have analyzed data, recognized patterns, adapted to shifts in change. That is the real intelligence of humans, their ability to learn and then on the basis of that learning make decisions. Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed. While creating the neural networks, often seen is the problem of reducing the bias and variance. A concept of tradeoff between bias and variance was also introduced.

2. BIAS & VARIANCE
The bias is said to be high when there is an under fitting of data. As seen in the figure 4 plotted on for a classifier. The model is just too general. Most of the weights might be zero, therefore a limited number of parameters may be present.

$$y = b + w_1 x_1 + w_2 x_2$$

The model evaluates very few attributes and some important attributes on the subject may be ignored.

The variance is said to be high as when there is an over fitting of data, there are way too many parameters in the model. The model is too specific.

$$y = b + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6$$

As model evaluates a very large number of attributes, so if something is not found in the subject, it classifies wrongly.
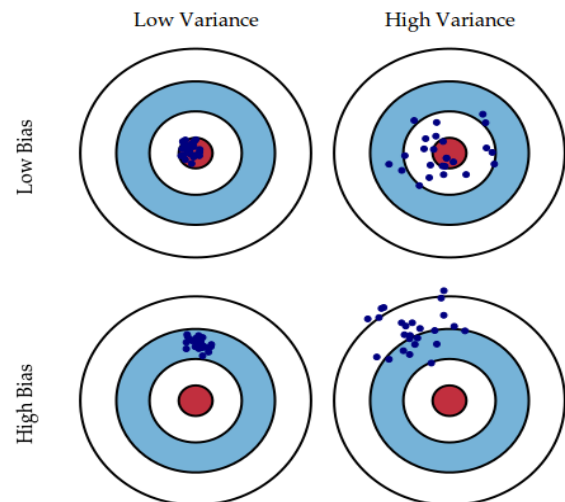

Fig 1: Graphical illustration of bias and variance

3. DETERMINE WHAT IS WRONG IN THE MODEL
In order to determine what is wrong with the model, plot the loss function. The loss function J() is nothing but a function that is used evaluate the difference between the actual and the desired output.

3.1 High Bias
This learning curve shows high error on both the training and test sets, so the algorithm is suffering from high bias:
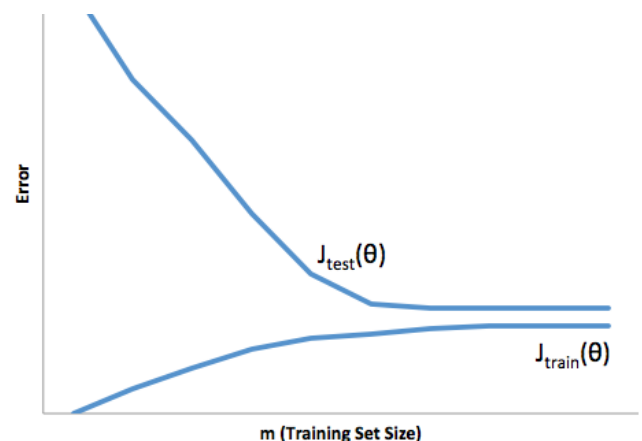

Fig 2: Learning curve on High Bias

3.2 High Variance
This learning curve shows a large gap between training and test set errors, so the algorithm is suffering from high variance:
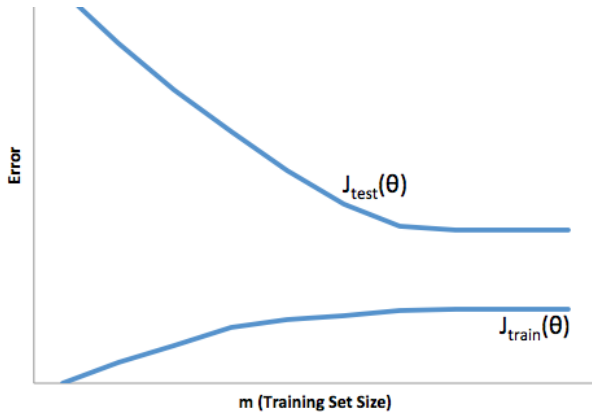
Fig 3: Learning curve on High Variance

Table 1. Example determining the bias/variance

Train Set Error	Dev Set Error	Conclusion
1%	12%	High Variance
14%	15%	High Bias
15%	30%	High Bias & High Variance
0.5%	0.6%	Low Bias & Low Variance

4. HOW TO SOLVE THE PROBLEM

If there is high variance in the model then try:

- Get More Data
- Try Regularization

If there is high bias in the model then try:

- Train a bigger neural network
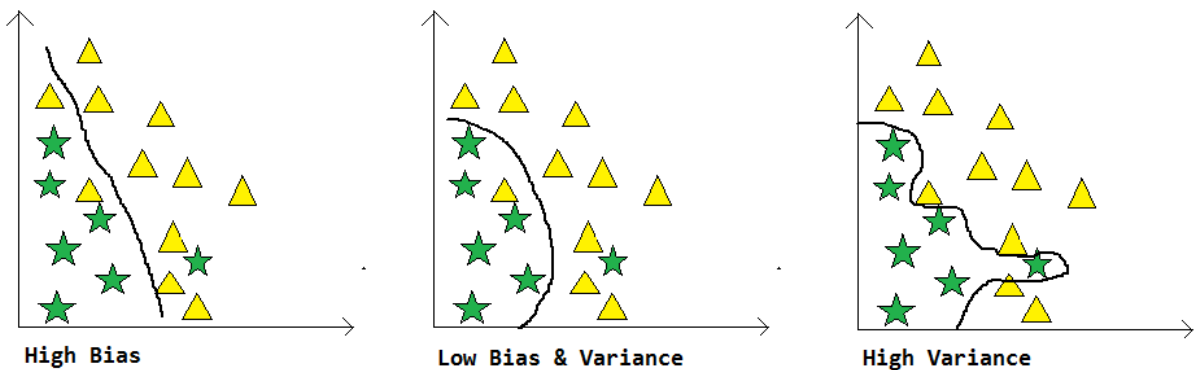- Train it longer
- Try another Neural Network Architecture

Fig 4: Bias Variance Graphs

5. REGULARIZATION

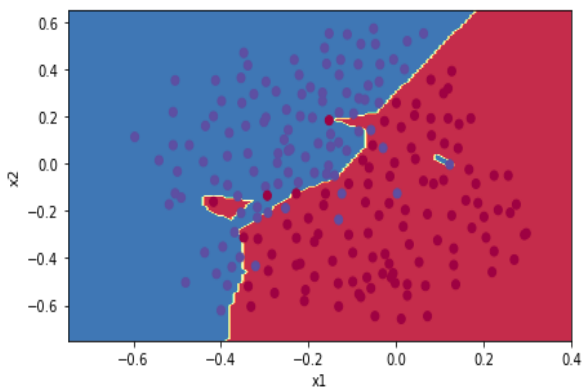Below is how the model performs, without regularization:

Fig 5: Model without Regularization

Clearly the model is overfitting. There are many regularization techniques, paper explores the L2 Regularization and Dropout.

5.1 L2 Regularization

L2 regularization advocates that a model with small weights is simpler than a model with large weights. Penalize the square values of the weights in the cost function, making them smaller. In this the regularization term is the sum of the squares of all the feature weights.

L2 regularization term is: $||w||_2^2 = w_1^2 + w_2^2 + \cdots + w_n^2$

The cost function is modified as:

$$J_{regularized} = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\log(a^{[L](i)}) + (1-y^{(i)})\log(1-a^{[L](i)})) + \frac{1}{m}\frac{\lambda}{2}\sum_{l}\sum_{k}\sum_{j}W_{kj}^{[l]2}$$

Where cross entropy part is:

$$-\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\log(a^{[L](i)}) + (1 - y^{(i)})\log(1 - a^{[L](i)}))$$

L2 Regularization cost is:

$$\frac{1}{m}\frac{\lambda}{2}\sum_l\sum_k\sum_j W_{kj}^{[l]2}$$

$\frac{\lambda}{n}W^l$ Is added to the partial derivative of all the weight terms during Backpropagation.
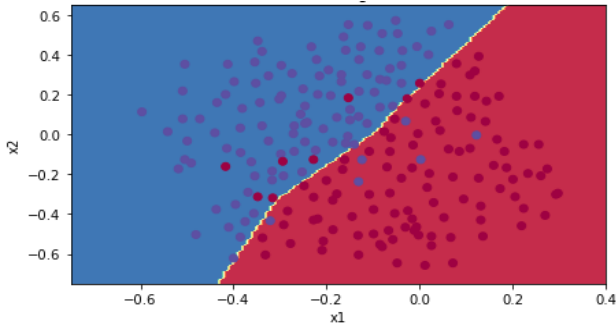
Here is how the model looks after L2 Regularization:

Fig 6: Model with L2 Regularization

5.2 Dropout

In the dropout regularization, randomly shut down some neurons in every backward and forward propagation iteration. That means that the neuron is encouraged to depend lesser on specific activation from other neuron as it may go off anytime. It is important to note that dropout should not be applied to the test set.
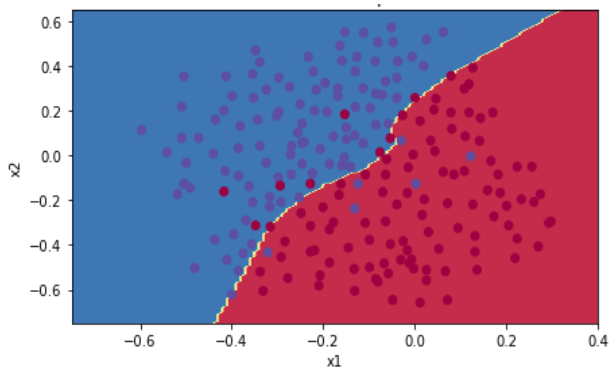
Here is how model looks after Dropout:

Fig 7: Model with Dropout.

6. CONCLUSION

This paper discussed on how to identify which problem the model is suffering from, and how to fix it. While bias variance trade off used to be an approach to tone the model, it is not needed in the current scenario anymore. There are more tools to reduce both variance and bias, without having to affect the other. For high variance reduction, try regularization techniques, if getting more data is too expensive. Thus the problem is solved without any compromise of any sort.

7. REFERENCES

[1] Sebastian Raschka 2016 Model evaluation, model selection, and algorithm selection in machine learning.

[2] Bengio, Yoshua, and Yves Grandvalet. 2004. "No Unbiased Estimator of the Variance of K-Fold Cross-Validation." J. Mach. Learn. Res. 5 (December). JMLR.org: 1089–1105.

[3] Stanford University, Coursera 2018 Machine Learning Computing Systems.

[4] Google 2018 Machine Learning Crash Course.

[5] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. 2005 Classification: Basic Concepts, Decision Trees, and Model Evaluation." In Introduction to Data Mining. Boston: Pearson Addison Wesley.

[6] Hawkins, Douglas M., Subhash C. Basak, and Denise Mills. 2003. "Assessing Model Fit by Cross-Validation." Journal of Chemical Information and Computer Sciences 43 (2). American Chemical Society: 579–86.

[7] Anders Krogh, John A. Hertz. A Simple Weight Decay Can Improve Generalization.

[8] Juergen Schmidhuber 2014 Deep Learning: An Overview

[9] Andrew Ng. Feature selection, L1 vs. L2 regularization

[10] Coursera 2018 deeplearning.ai