CHEAT SHEET

Model Debugging

If a model is not performing well, there are several ways to improve its performance. To determine which of the many techniques to use, the first step is to identify the root of the problem.

	High Variance	High Bias
Description	Models with high variance are capable of memorizing many more properties of the training data and do not do well on unseen data, resulting in low training error but high test error.	Models with high bias are too simplistic or make ill-suited assumptions and therefore cannot even achieve low error on the training data set.
Symptoms	Training error is much lower than test error.	Training error is higher than a desired error threshold.
Remedies	• Add more training data • Reduce model complexity (complex models are prone to high variance) • Bagging	• Use a more complex model (or use nonlinear models) • Add features • Boosting

Visualize Variance and Bias

The graph below exemplifies data with high/low variance and high/low bias as "darts" thrown at a target. The bullseye at the center of the target is the location of the perfect classifier on the testing data. The blue dots illustrate the darts, which represent classifiers trained on different training data sets.

High variance/low bias models perform well when performance is averaged over large data sets. In expectation they are close to the bullseye but can perform very differently on any two particular data sets. We say that these models are overfit; that they have learned to predict meaningless noise patterns in the data.

High bias/low variance settings lead to models that are very similar across different training data sets (the blue dots are close together); however, they are systematically off-target (i.e., they make wrong assumptions).

The worst case is **high bias and high variance**. The goal is to achieve a model with low bias and low variance.