Tradeoff analysis

1 Problem setting

We ae given a dataset $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, awn i.i.d. from some distribution P(X,Y).

Let H_D denote the set of all hypotheses we have considered over D, and $h_D \in H_D$ is the best model we have obtained in the hypothesis set. Furthermore, let f_D denote the best estimator for the dataset D. Then the MSE between the best estimator h_D in our hypothesis and the best estimator f_D is

$$\varepsilon^2 = E\left[h_D(x) - f_D(x)\right]^2 = E\left[h_D(x) - E[h_D(x)] + E[h_D(x)] - f_D(x)\right]^2$$

$$= \underbrace{E\left[h_D(x) - E[h_D(x)]\right]^2}_{\text{Variance}} + \underbrace{\left(E[h_D(x)] - f_D(x)\right)^2}_{\text{bias}}$$

More specifically, the MSE between the best estimator h_D and the true value y is given as

$$\underbrace{E_{\mathbf{x},y,D}\left[(h_D(\mathbf{x}) - y)^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{\mathbf{x},D}\left[(h_D(\mathbf{x}) - \bar{h}(\mathbf{x}))^2\right]}_{\text{Variance}} + \underbrace{E_{\mathbf{x},y}\left[(\bar{y}(\mathbf{x}) - y)^2\right]}_{\text{Noise}} + \underbrace{E_{\mathbf{x}}\left[(\bar{h}(\mathbf{x}) - \bar{y}(\mathbf{x}))^2\right]}_{\text{Bias}^2}$$

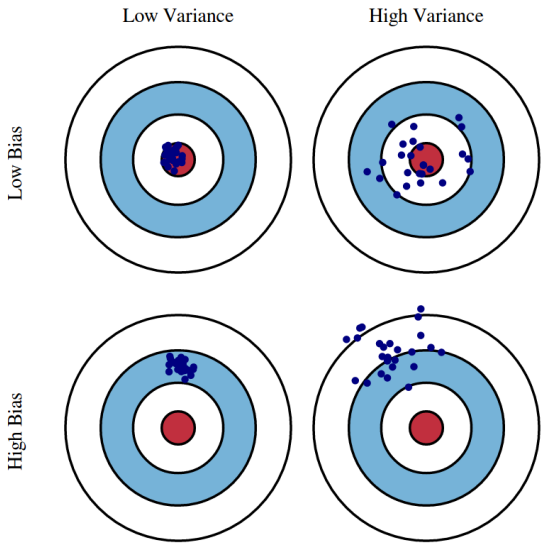
Where $\bar{h}(x) = E[h_D(x)]$ is the expected value of the optimal estimator in our hypothesis set, $\bar{y}(x) = f_D$ is the value obtained by the best estimator, and y is the true value.

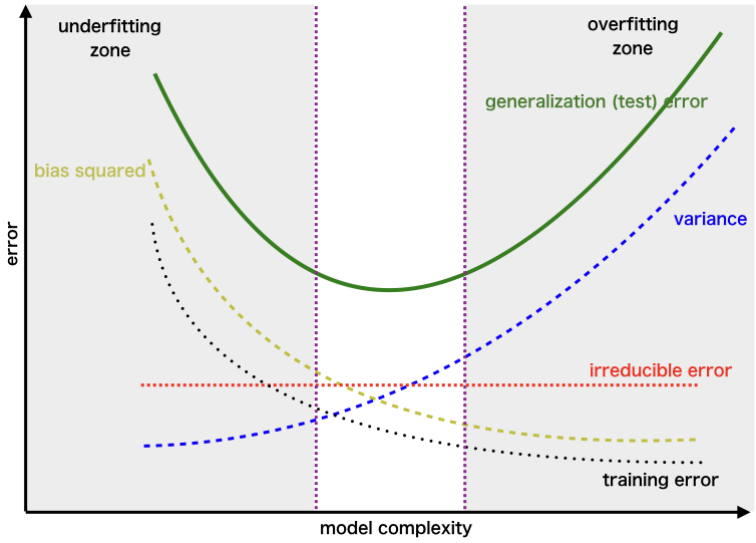
2 Interpretation

- **Variance**: how a hypothesis learned from a specific dataset D diverges from the average prediction $E[h_D(x)]$? The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn't seen before. As a result, such models perform very well on training data but have high error rates on test data.
  When a model is high on variance, it is then said to as Overfitting of Data. Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.
  While training a data model variance should be kept low.
- **Bias:** how far the expected prediction $E[h_D(x)]$ diverges from the optimal predictor f_D? Being high in biasing gives a large error in training as well as testing data. It recommended that an algorithm should always be low-biased to avoid the problem of **underfitting**.
  By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as Underfitting of Data. This happens when the hypothesis is too simple or linear in nature.

- **(Irreducible) Noise**: How big is the data-intrinsic noise? This error measures ambiguity due to your data distribution and feature representation. You can never beat this, it is an aspect of the data.

The relationship between variance and bias is shown by following figures.





2 Conclusion

- High bias and low variance (Underfitting)
  High bias in general represents very simple model, such as linear model.
- Low bias and high variance (Overfitting)
  High variance in general represents very complex model, such as high-order polynomial models.
- Low bias and low variance (good)
- The hypothesis with high variance and high bias has the worst performance.

Therefore, our goal is to find a hypothesis with comparatively low bias and variance.

Reference

https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

https://www.quantstart.com/articles/The-Bias-Variance-Tradeoff-in-Statistical-Machine-Learning-The-Regression-Setting/

https://www.geeksforgeeks.org/ml-bias-variance-trade-off/