ISLR Chapter 2

Anirudh Prasanna

June 2025

Chapter 2:(2.2) Assessing the quality of a model

Measuring the quality of a model
In order to measure how well the performance of the model matches the data, we use the mean squared error:

MSE=$\sum_{i=1}^{n}(y_i - \hat{f}(x))^2$

The MSE when computed using training data is the training MSE and the MSE when computed using test data is called test MSE. Usually, the test MSE is greater than the training MSE, as when we minimize training MSE, we fit our parameters/estimators to the training MSE, which may not work well for the test MSE.

Bias-Variance Tradeoff:
Variance: sensitivity of the model to training data
Bias: error due to using a simple model to fit a more complex model.

The expected test MSE can be decomposed as Variance +Bias^2+σ^2. By minimizing the variance and bias, we can minimize the expected test MSE. σ^2 is irreducible error due to noise.

More flexible methods have higher variance since they tend to overfit the data and also fit the noise. So, changing one observation may cause the estimate to change by a lot. They tend to have a low bias since they have more degrees of freedom so they can fit a more complex relation.

Less flexible methods have lower variance since they tend to underfit the data and do not fit the noise. So, changing one observation does not change the data by much. They have a higher bias since they use a more simple estimate with lesser degrees of freedom to model a more complex relation.

Generally, as flexibility increases, variance increases and bias decreases. This tradeoff is called as the Bias-Variance tradeoff.

Classification:
In the case of classification regression, the metric we use to measure the quality of the fit is the error rate.

Error rate = $\sum_{i=1}^{n} \frac{I(y_i \neq \hat{y_i})}{n}$
$I(y_i \neq \hat{y_i})$ is 1 if $y_i \neq \hat{y_i}$, 0 otherwise. It is the fraction of incorrect predictions.

Error rate = 1 - fraction of correct predictions. In a hand-wavy way, we can think of the fraction of correct predictions as being related to $E[P(y_i = \hat{y_i}|X = X_0)]$

The error is minimized if $P(y_i = \hat{y_i}|X = X_0)$ is maximized. This is called as the Bayes Classifier. $P(Y = s_k|X = X_0) > P(Y = s_i|X = X_0)$, i≠k where s_i are the classifications, then the classifier assigns the observation to class s_k. It assigns each observation to the most likely class. Bayes Decision Boundary: separates the various classes.

KNN (K Nearest Neighbours):
We usually don't know $P(Y = s|X = X_0)$. An alternate approach in such cases is KNN. KNN checks the classification of the k nearest neighbours of a given point. The most common class amongst the neighbours is given to the point.

The accuracy of the model depends on the choice of K. For small values of K, the model is overfit, overly flexible and for larger values of K, the model is underfit not very flexible. For small values of K there is low training error as the points are closer, and are mostly of the same class, for large values of K there is a high training error, as the neighbours are much further away and may not have the same class. For both small and high values of K, there is a high test error