**Chapter 2 – Statistical learning**

* In essence, statistical learning refers to a set of approaches for estimating *f* (function)
* The accuracy of a function depends on two quantiles:
  + *Reduicable error*: An error where we can potentially improve the accuracy of f by using the most appropriate statistical methods;
  + *Irreducable error*: There are some errors that are *irreductable*, because we cannot by definition measure absolutely everything that can influence Y. This error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.
* Broadly speaking, most statistical learning methods can be characterised as either *parametric* or *non-parametric*
  + Paramaetric models make assumptions about the practical form, or shape, of *f*. This generally simplifies the problem of estimating *f* because it is generally much more easier to estimate a set of paramaters. However, one potential disadvantage of a paramaetric approach is that the model we choose may not actually match the true unknown form of *f*. We can try to avoid this by choosing flexible models that can fit as many different possible functional forms for f. But this involves estimating a greater number of paraamters, and can lead to a phenomonan known as *overfitting* the data, which essentially means they follow the errors, or *noise*, too closely.
  + *Non-parametric* models do not make explicit assumptions about the functional for of *f*. Instead, they seek an estimate of *f* that gets as close to the data points as possible without being too rough or wiggly. One disadvantage of non-parametric models is that a very large number of observsations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for *f*.
* If we are mainly interested in inference, than restrictive models are much more interpreative.
* In terms of measuring the quality of fit, in the regression setting, the most commonly-used measure is the *mean squared error* (MSE). Note that, when using training and test data, we want choose the method that gives the lowest *test* MSE, rather than the lowest *training* MSE.
* The degrees of freedom is a quantity that summarises the flexibility of a curve
* As model flexibility increases, training MSE will decrese, but the test MSE may not. When a given method yields a small training MSE but a large test MSE, we are said to be overfitting the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function *f*.
* Though the mathetmatical proof is beyond the scope of this book, it is possible to show that the expected test MSE, for a given value xo, can always be decomposed into the sum of three fundamental quantities: the *variance* of f, the squred *bias* of f, and the variance of the error terms
* *Variance* refers to bte amount by which *f* would change if we estimated it using a different training data set.
* In general, more flexible statistical approaches have a higher variance.
* On the other hand, *bias* refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
* Generally, more flexible methods result in less bias
* The most common approach for quantifiying the accuracy of our estimate f is the training error rate, the proportion of mistakes that are made if we apply our estimate *f* to the training observations
* It is possible to show, although the proof is outside of this book, that the test error rate is minimised, on average, by a simpler classifer that assigns each observation to the most likely class, given its predictor values. This very simple classifer is called the *Bayes classifer*
* The *Bayes decision boundary* represents the points where the probability is exactly 50%.
* The K nearest neighbours (KNN) classifier applies the *Bayes* rules to k number of observations (pre set) and classifies each to the class with the largest probability. When k =1, the decision boundary is overly flexible, corresponding to a classifer that has low bias but very high variance. As K grows, the method becomes less flexible and produces a decision boundary that is close to linear.

**Exercises – Chapter 2**

1. The samples size *n* is extrnemly large, and the number of predictors *p* is small

*A flexible method is best, as generally flexible methods perform better with require more data and fewer predictors*

1. The number of predictors *p* is extremely large, and the number of observations *n* is small

*Inflexible (parametric) methods work better with smaller numbers of n, and can accommodate many predictors also*

1. The relationship between the predictors and the response is highly non-linear

*This is more appropriate for a flexible statistical method, as inflexible methods rely on the assumption of linearity*

1. The variance of the error terms, ie 02 = Var(e), is extremely high

In general, more flexible methods have a higher variance (and a lower bias), whilst inflexible methods perform better with have a higher bias, and lower variance.

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach?

* *A flexible approach results in high variance and low bias, and an inflexible approach leads to low variance but high bias*
* *A more flexible approach is appropriate for those cases where there is high variance in the data, and the n is high*
* *A inflexible approach, conversely, is more appropriate for those cases where*

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression and classification (as opposed to a non-parametric approach?) What are its disadvantages?