Chapter 24

Model Selection

24.2 Simple Strategies

Given some data, our primary goal is to fit the best possible model that "makes sense". However, there are a plethora of models that all seem as good a fit as any other. This begs the obvious question: How are we supposed to decide which model is to be selected? In this section, we discuss a few strategies which are to be followed while comparing between models.

The first thing to consider when analysing data is **domain knowledge**. We must ask ourselves whether there is any theoretical justification supporting the model that we have chosen. For example, suppose we find that a cubic polynomial fit is better than a quadratic polynomial fit for some data obtained from a physics experiment. However, if the laws of physics dictate that the correct fit is quadratic, then we must select the quadratic polynomial fit over the cubic fit.

Another thing that must be kept in mind is the actual loss incurred when choosing one model over another. Unfortunately, since this is not a hypothesis test (more on this in later sections), fitting each model separately remains the only possible way to determine the loss in information.

An alternative to the above is the **cross-validated** version of the loss. Here, we split the data set into two parts, one larger and one smaller, and model using the larger subset. The data in the smaller subset is reserved as test data for our model, i.e, it serves as a proxy for future data and we decide which model is better based on its performance on this test data.

24.3 Strategies for Practical Applications

We ask the reader to be **cautious** with domain knowledge as usually, no obvious hints are apparent from the domain. So we are only left with the data and the models we wish to compare. Hence in practice, we must come up with a different set of strategies. The best way of proceeding is to try and think of a way to answer the following questions.

1. How good is the fit?
2. What is the cost of the fit?

We want to maximise the goodness of fit while minimising the loss in information incurred in achieving that fit. Usually, an increase(decrease) in one is characterised by an increase(decrease) in the other.

There are multiple ways to balance this trade-off. The most popular set-up where this is encountered is when both the candidate models are fitted using Maximum Likelihood Estimation. There can be other cases, such as when the models are fitted on an ad hoc basis or by some mathematically motivated methods such as Least Squares Estimation without any underlying probabilistic assumptions. However, we are going to restrict our discussion to the Maximum Likelihood case.

Let us try to answer the above questions for this set-up.

Here, how good the fit is can be conveniently measured by the maximum value of the log likelihood that we have attained.

Similarly, the price can again be easily measured by the number of free parameters that we have to estimate in that model.