



Figure 1.21 (a) Misclassification rate vs K in a K-nearest neighbor classifier. On the left, where K is small, the model is complex and hence we overfit. On the right, where K is large, the model is simple and we underfit. Dotted blue line: training set (size 200). Solid red line: test set (size 500). (b) Schematic of 5-fold cross validation. Figure generated by `knnClassifyDemo`.

won't have enough data to train on, and we won't have enough data to make a reliable estimate of the future performance.

A simple but popular solution to this is to use **cross validation (CV)**. The idea is simple: we split the training data into K **folds**; then, for each fold $k \in \{1, \dots, K\}$, we train on all the folds but the k 'th, and test on the k 'th, in a round-robin fashion, as sketched in Figure 1.21(b). We then compute the error averaged over all the folds, and use this as a proxy for the test error. (Note that each point gets predicted only once, although it will be used for training $K - 1$ times.) It is common to use $K = 5$; this is called 5-fold CV. If we set $K = N$, then we get a method called **leave-one out cross validation**, or **LOOCV**, since in fold i , we train on all the data cases except for i , and then test on i . Exercise 1.3 asks you to compute the 5-fold CV estimate of the test error vs K , and to compare it to the empirical test error in Figure 1.21(a).

Choosing K for a KNN classifier is a special case of a more general problem known as **model selection**, where we have to choose between models with different degrees of flexibility. Cross-validation is widely used for solving such problems, although we will discuss other approaches later in the book.

1.4.9 No free lunch theorem

All models are wrong, but some models are useful. — George Box (Box and Draper 1987, p424).¹²

Much of machine learning is concerned with devising different models, and different algorithms to fit them. We can use methods such as cross validation to empirically choose the best method for our particular problem. However, there is no universally best model — this is sometimes called the **no free lunch theorem** (Wolpert 1996). The reason for this is that a set of assumptions that works well in one domain may work poorly in another.

¹² George Box is a retired statistics professor at the University of Wisconsin.