modeling assumptions used in statistics, learning theory, information theory etc. The second approach (called *resampling*) is purely data-driven, and it makes no statistical assumptions about the data or the type of unknown model being estimated. For both approaches, analytic and resampling, several models are estimated from data using different values of the complexity parameter, and then the optimal selected model corresponds to the smallest estimated prediction risk.

Several analytical model selection criteria (i.e., estimates of prediction risk) for linear regression estimators are described below. These estimates of prediction risk (2.4) can be written as a function of the MSE fitting error, or empirical risk, penalized by some analytic factor depending on model complexity:

$$R \cong r\left(\frac{DoF}{n}\right) R_{emp} \tag{2.10}$$

where $r(p)$ is a monotonically increasing function of the ratio of degrees of freedom (DoF) and the training sample size n, $p = DoF / n$. The number of degrees of freedom (DoF) is simply the number of parameters (in a linear model). The empirical risk R_{emp} is the mean squared error (MSE) for training data. The function r is often called a penalization factor, because it penalizes the empirical risk for increasingly complex models. The following forms of r have been proposed in the statistical literature:

Akaike's final prediction error (*fpe*): $r(p) = (1+p)(1-p)^{-1}$ (2.11)

Schwartz' criterion (*sc*): $r(p,n) = 1 + p(1-p)^{-1} \ln n$ (2.12)

Generalized cross-validation (*gcv*): $r(p) = (1-p)^{-2}$ (2.13)

These criteria are derived under asymptotic assumptions (as sample size $n \to \infty$) for linear estimators and work well only for large training sets.

Next we discuss resampling methods for estimating prediction error in order to select optimal model complexity. The basic idea is first to estimate a model using a portion of available data and then to use the remaining samples to estimate the prediction risk for this model. The first portion of the data (nl samples used for model estimation or

learning) is called a *learning (or training)* set, and the second portion of the data with $nv = n - nl$ samples is a *validation* set. The various implementations of resampling differ according to strategies used to divide the data.

The simplest approach is to split the data (say, randomly) into two portions, i.e. 80% for training and 20% for validation. Then the model $f_\lambda(\mathbf{x}, \omega^*)$ is estimated by minimizing empirical risk (2.3) on the training data set. Here the subscript λ denotes the 'complexity' parameter of a learning method, i.e. the polynomial degree for polynomial regression, or the value of k for k-nearest neighbors. The prediction risk (2.4) for this model is then estimated as an average loss on the validation set:

$$R(\lambda) \cong R_v(\lambda) = \frac{1}{nv} \sum_{i=1}^{nv} L(y_i, f_\lambda(\mathbf{x}_i, \omega^*)) \qquad (2.14)$$

The predictive model $f_\lambda(\mathbf{x}, \omega^*)$ is estimated for many different values of complexity parameter λ, and the optimal model complexity λ_{opt} is selected such that it yields the smallest validation error (2.14).

This strategy of splitting the data into two subsets (training and validation) is simple, but may be sensitive to random partitioning of the data. So in practice, more robust partitioning strategies are used. For example, the method known as *leave-one-out cross-validation* uses $nv = 1$ (one validation sample) and tries all n possible partitions of data into n-1 training samples and 1 validation sample. Then n separate models are estimated, one for each partitioning, and average validation error is used as an estimate of prediction risk. A more practical approach

FIGURE 2.13 Five-fold cross-validation.

called K-*fold cross-validation* is to divide the data into K (randomly selected) disjoint subsets of equal size $n_v = n/K$. Partitioning of the data into learning and validation subsets is shown in Fig. 2.13 for 5-fold cross-validation. An algorithmic description of K-fold cross-validation procedure, assuming the squared error loss, is presented next. For given training data $Z = [\mathbf{X}, \mathbf{y}]$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$:

1. Divide the training data Z into K disjoint subsets of equal size, $Z_1, Z_2, \ldots, Z_K$.
2. For each validation sample Z_i of size n/K
   (a) Use the remaining data, $Z_l = \bigcup_{j \neq i} Z_j$ to estimate model $\hat{f}_i(\mathbf{x})$.
   (b) For the regression estimate $\hat{f}_i(\mathbf{x})$, calculate the empirical risk for the "left out" validation data Z_i:

$$r_i = \frac{K}{n} \sum_{Z_i} (\hat{f}_i(\mathbf{x}) - y)^2$$

3. Compute the estimate for the prediction risk by averaging the empirical risk sums for $Z_1, Z_2, \ldots, Z_K$:

$$R(\omega) \cong R_{cv}(\omega) = \frac{1}{K} \sum_{i=1}^{K} r_i$$

Typical choices for K are 5 or 10. Note that leave-one-out cross-validation is a special case of K-fold cross-validation, for $K=n$. The leave-one-out approach is computationally more expensive than K-fold cross-validation (by a factor n/K).

Example 2.7: Model selection via 5-fold cross-validation.
A training data set of 25 samples was generated according to the target function

$$y = \sin^2(2\pi x) + \xi \tag{2.15}$$

where the noise ξ is zero mean Gaussian with variance $\sigma^2 = 0.1$. The input x was sampled from a uniform random distribution in [0,1] interval. The squared loss function $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ is used to measure discrepancy between the model estimate and observed output y.

Let us specify a set of possible models (functions) to be algebraic polynomials:

$$f_m(x, \mathbf{w}) = w_0 + \sum_{i=1}^{m} w_i x^i \tag{2.16}$$

Here the set of parameters take the form of vectors $\mathbf{w} = [w_0, ..., w_m]$ for polynomial of degree m. For practical purposes we limit the polynomial degree to $m \leq 9$. For any value of m the polynomial coefficients $\mathbf{w} = [w_0, ..., w_m]$ are estimated by least squares fitting. The task of model selection is to choose the value of m that provides the lowest estimated prediction risk. That is, polynomial degree is the complexity parameter of this learning method. This complexity parameter is selected via 5-fold cross-validation. According to 5-fold cross-validation procedure, the training data is divided (randomly) into 5 subsets (of 5 samples each) as shown in Fig. 2.13. Then each polynomial model (of fixed degree m) is estimated using the learning set and its resampling error is evaluated on the validation set. The average resampling error is recorded and shown in Fig. 2.14. The polynomial

$m+1$	Estimated R via cross validation
1	0.2748
2	0.3013
3	0.3277
4	0.1889
5	0.3022
6	**0.1223**
7	0.2401
8	4.4959
9	1.2492
10	3.3433

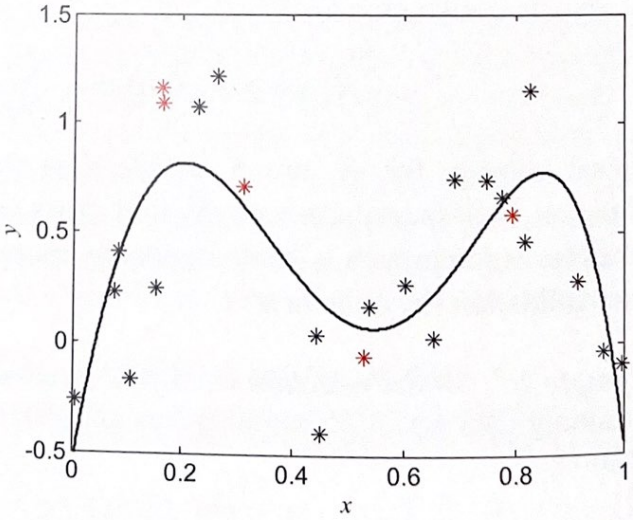

FIGURE 2.14 Polynomial model (of 5-th degree) estimated from all 25 samples. Also shown is one partitioning of the data into 20 training and 5 validation samples (in red).

k	Estimated R via cross-validation
1	0.1964
2	0.2173
3	**0.1333**
4	0.1557
5	0.1919
6	0.2175
7	0.2445
8	0.2676
9	0.2838
10	0.2878

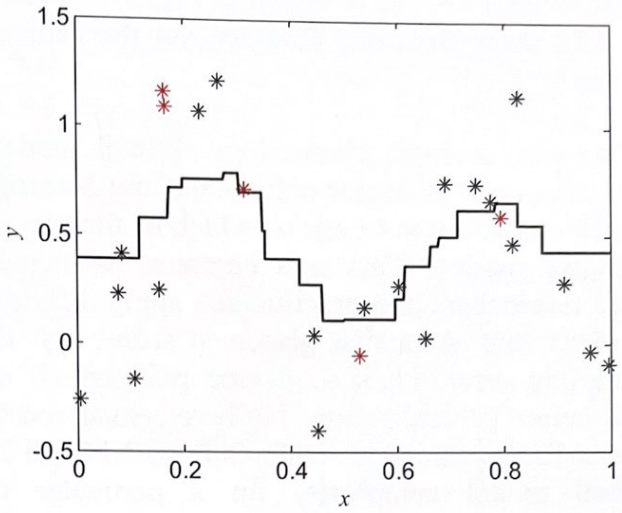

FIGURE 2.15 k-nearest neighbor regression model (k=3) estimated from all 25 samples. Also shown is one partitioning of the data into 20 training and 5 validation samples (in red).

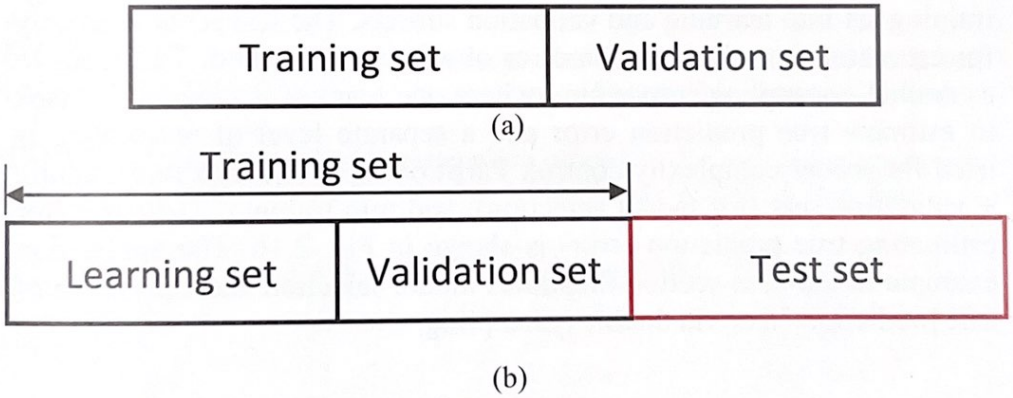

FIGURE 2.16 Use of resampling for:
(a) model complexity control;
(b) estimating prediction error of a learning method.

model providing smallest validation error (in this case, for m=5) is selected as the best predictive model. The final model, shown in Fig 2.14 is estimated by fitting the 5-th degree polynomial to all 25 samples. Next, consider application of k-nearest neighbor regression to the same data set. In this case, model complexity is controlled by the value of k. Application of 5-fold cross validation to k-nearest neighbors regression

yields optimal k=3, as is shown in Fig. 2.15. Note that both Figures 2.14 and 2.15 show the same data set, yet the estimated models look quite different.

Previous example shows two optimal models estimated from the same data, i.e. 5-th degree polynomial and 3-nearest neighbor regression. An obvious question to ask is: which of the two methods yields a better predictive model? This is a common problem in many applications where researchers and practitioners apply different learning methods to the same data. At a first glance, it seems that the method with lower resampling error, i.e. 5-th degree polynomial, should be expected to yield better generalization. However, such reasoning may be flawed, because model selection results in Figs. 2.14 and 2.15 aim at selecting an optimal model complexity for a particular method. Because the validation set was actually used to select the model complexity, the resampling errors shown in these figures may under-estimate the true test error. The true prediction accuracy of a learning method can be estimated by dividing available data into two sets, called *training* and *test* sets, so that the model is estimated using *only* training data. An optimal model complexity can be determined by further dividing the training set into learning and validation subsets. The test set is used only for estimating 'true' prediction error of a learning method. This leads to a 'double resampling' procedure, where one level of resampling is used to estimate true prediction error and a separate level of resampling is used for model complexity control. Partitioning of the data into training + validation sets (for model selection), and into training + test sets (for estimating true prediction error) is shown in Fig. 2.16. The application example in the next section illustrates model selection and estimation of true prediction error via double resampling.

2.5 APPLICATION EXAMPLE

Haberman's Survival Data Set, taken from the UCI Machine Learning Repository at http://archive.ics.uci.edu/ml/, contains 306 cases on the survival of female patients who had undergone surgery for breast cancer, from a study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital. Each of the 306 patient records has a class label (alive/dead) indicating whether the patient survived 5 years or longer after surgery, or died within 5 years. Each patient record has three