2.2 Assessing Model Accuracy

Zongyi Liu

2023-05-11

2.2 Assessing Model Accuracy

One of the aims of this book is to obtain the knowledge to assess the accuracy of the model.

2.2.1 Measuring the Quality of Fit

In order to evaluate the performance of the statistical learning results, we need to do have some parameters for assessment, and one of the most common one is MSE: mean squared error.

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

$\hat{f}(x_i)$ stands for the prediction that f-hat gives for the i-th observation. MSE will be small if the predicted responses are close to the true responses, and vice versa.

We do not care about how the model predicted the tested data, but how well it will help to predict future data. We will want to know this:

$$\mathrm{Ave}(y_0 - \hat{f}(x_0))^2,$$

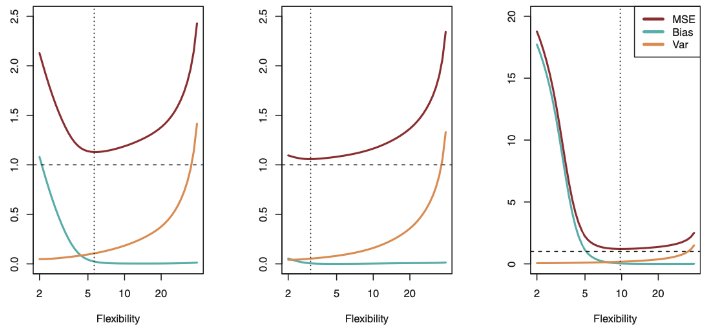
And we would choose a model whose MSE is as small as possible

2.2.2 The Bias-Variance Trade-Off

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

- Variance: the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- Bias: the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model

General rule: as we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases. We can see it in this plot:



2.2.3 The Classification Setting

All above discussions are on the regression settings, whereas many concepts are also applicable in classification setting. Suppose that we seek to estimate f on the basis of training observations {(x_1, y_1),...,(x_n,y_n)}, where now y_1,...,y_n are qualitative. The most common approach for quantifying the accuracy of our estimate f-hat is the training error rate, the proportion of mistakes that are made if we apply our estimate f-hat to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

This equation is referred to as the **training error** rate, the test error associated with a set of test observations of form (x_0, y_0) is

$$\text{Ave}\left(I(y_0 \neq \hat{y}_0)\right)$$

The Bayes Classifier

Its formula is given as

$$\Pr(Y = j | X = x_0)$$

We should assign test observation with predictor vector x0 to the class j where Pr is the largest.

Bayes error rate: the lowest possible test error rate produced by Bayes classifier, which is given by

$$1 - E\left(\max_j \Pr(Y = j|X)\right),$$

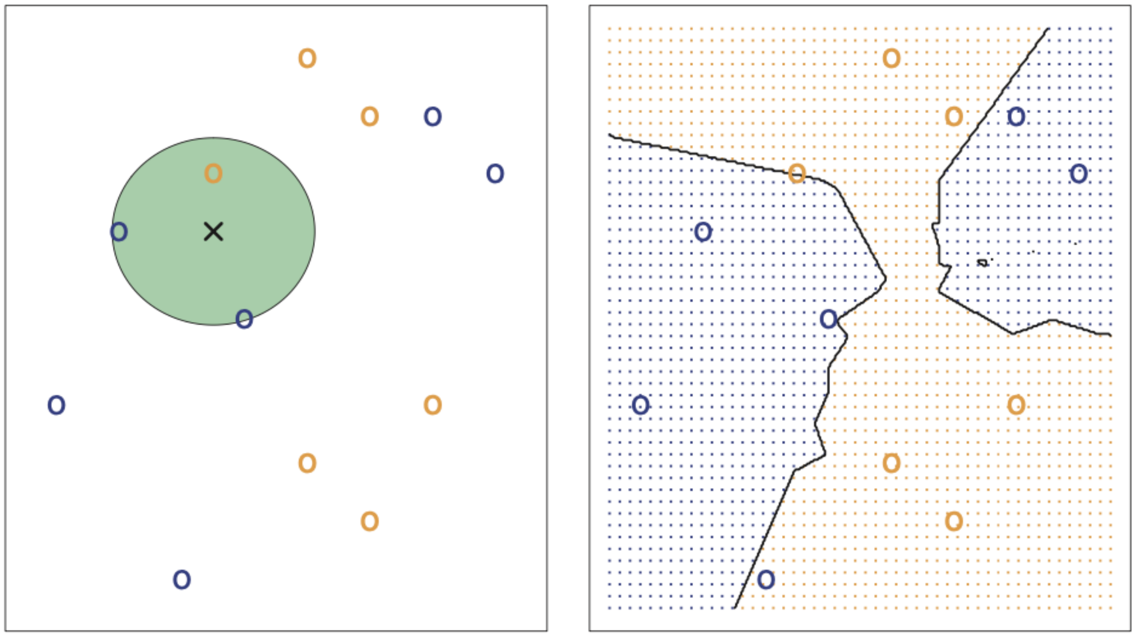
K-Nearest Neighbors

Mechanism: estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability.

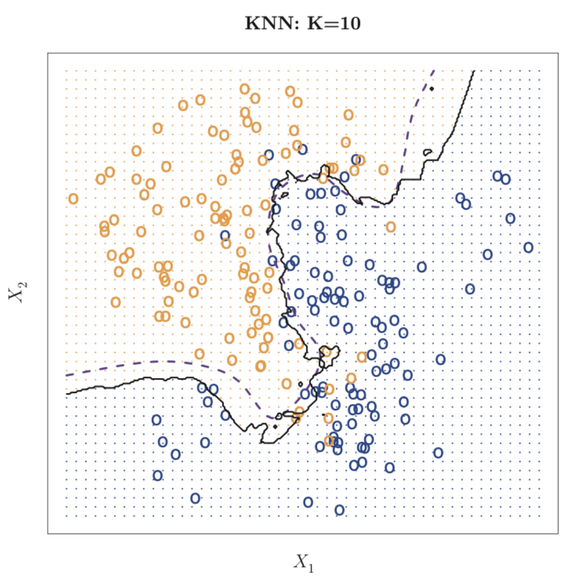
It is given by

$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j).$$

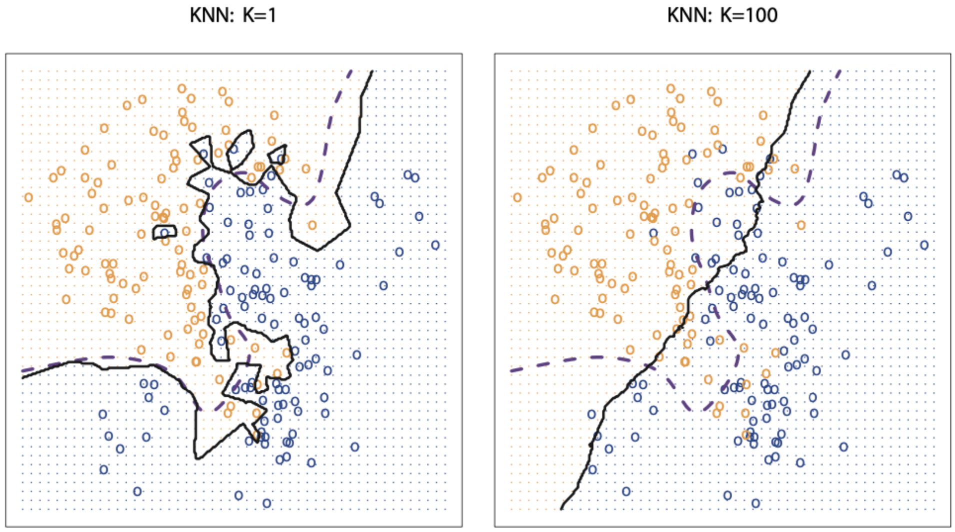
We can see the KNN approach in the following graph:



The KNN approach, using K = 3, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.



The black curve indicates the KNN decision boundary on the data from Figure 2.13, using K = 10. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.



A comparison of the KNN decision boundaries (solid black curves) obtained using K = 1 and K = 100 on the data from Figure 2.13. With K = 1, the decision boundary is overly flexible, while with K = 100 it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.