Introduction to Statistical Learning

Ömür UĞUR

Institute of Applied Mathematics
Middle East Technical University, Ankara, Turkey

Outline

1. Introduction to Statistical Learning
   - Modelling the Response
   - Assessing the Model Accuracy
   - Classification Setting
     - The Bayes Classifier
     - The K-Nearest Neighbours Classifier

Outline

1. Introduction to Statistical Learning
   - Modelling the Response
   - Assessing the Model Accuracy
   - Classification Setting
     - The Bayes Classifier
     - The K-Nearest Neighbours Classifier

Outline

1. Introduction to Statistical Learning
   - Modelling the Response
   - Assessing the Model Accuracy
   - Classification Setting
     - The Bayes Classifier
     - The K-Nearest Neighbours Classifier

Predictors and Responses

Statistical Learning is not only understanding the statistics of the data, but infer useful information from the data for possible future outlook.
Suppose that we observe a *quantitative* response Y and p different predictors $X_1, \ldots, X_p$; we assume a relationship of the form

$$Y = f(X) + \epsilon,$$

where f is a *true*, but *fixed* unknown function of $X = (X_1, \ldots, X_p)^\top$, and ϵ is a random *error term*, which is *independent* of X and has mean zero and, possibly, a variance of σ_ϵ^2: $\epsilon \sim \text{wn}(0, \sigma_\epsilon^2)$, and it stands for an *irreducible* error.
We can then predict Y using

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ represents the estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

Reducible & Irreducible Errors

Clearly, the accuracy of $\hat{Y}$ depends on *two* quantities:

$$\begin{aligned}\mathbb{E}\left[(Y-\hat{Y})^2\right] &= \mathbb{E}\left[\left(f(X)+\epsilon-\hat{f}(X)\right)^2\right] \\ &= \underbrace{\left[f(X)-\hat{f}(X)\right]^2}_{\text{reducible}} + \underbrace{\mathbb{V}\text{ar}\left[\epsilon\right]}_{\text{irreducible}},\end{aligned}$$

where $\mathbb{E}\left[\left(Y-\hat{Y}\right)^2\right]$ represents the average, or *expected value*, of the squared distance between the predicted and actual value of Y, and $\mathbb{V}\text{ar}\left[\epsilon\right]$ represents the variance associated with the error term ϵ.

Reducible Error via Modelling I

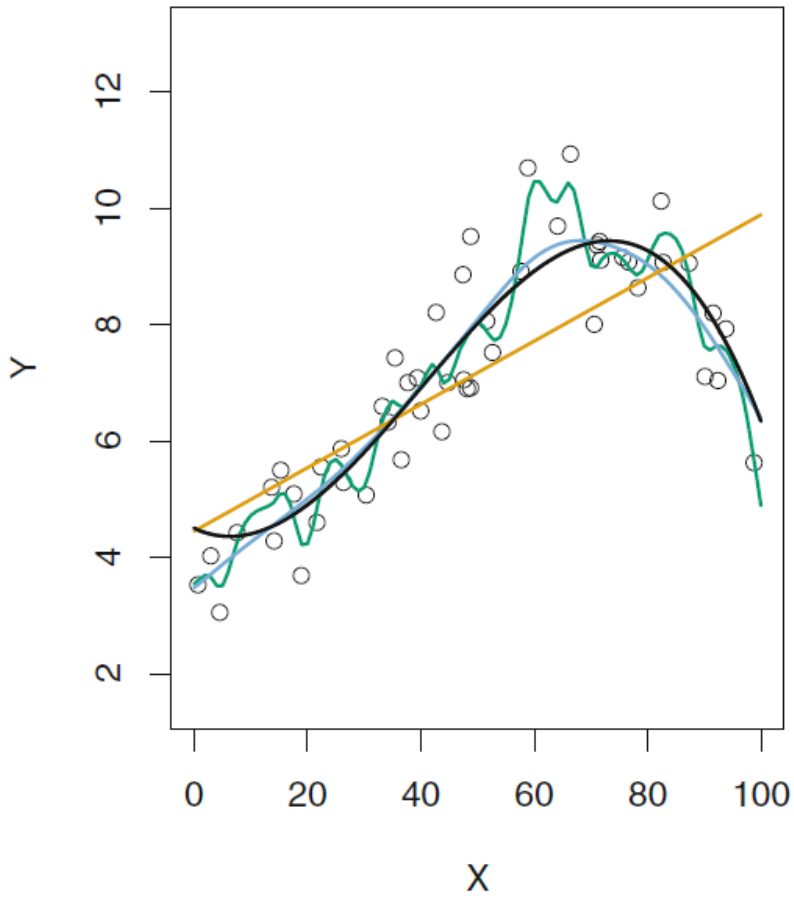

Figure 1: Representation of Reducible Error

Reducible Error via Modelling II

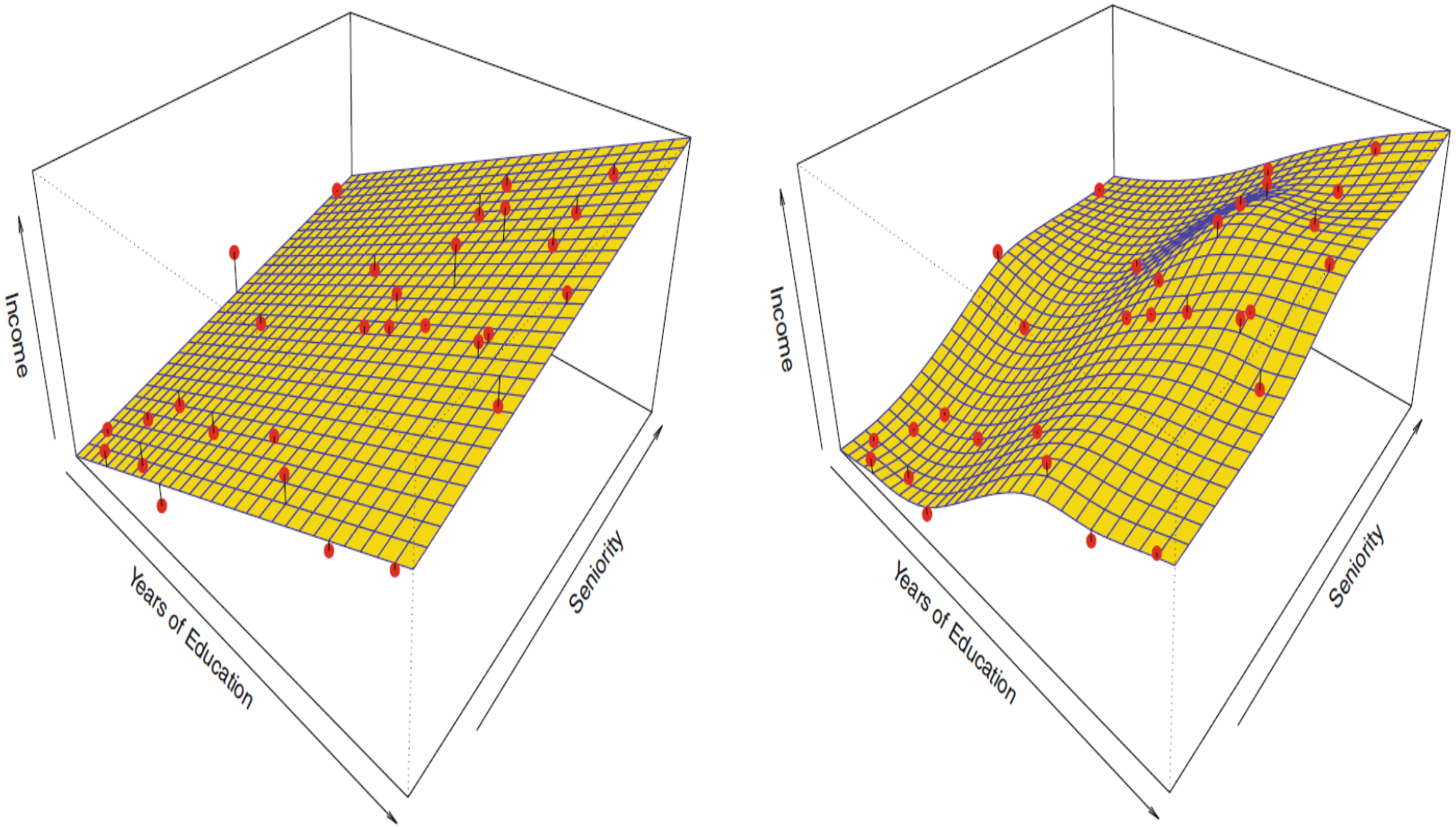

Figure 2: Representation of Reducible Error in 2D

Methods for Estimating f

The goal is to apply statistical (learning) methods to the (training) data to estimate the unknown function f such that

$$Y \approx \hat{f}(X)$$

for any observation (X, Y).

- *Parametric Methods*: includes parameters $\beta = (\beta_0, \ldots, \beta_p)^\top$ to be estimated, by $\hat{\beta} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)^\top$; such as

$$\text{(linear)}: \quad f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

or

$$\text{(logistic)}: \quad f(X) = \frac{\exp\left\{\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right\}}{1 + \exp\left\{\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right\}}$$

- *Non-Parametric Methods*: involves no explicit assumptions on the form of f, such as *piece-wise polynomials*, *splines*, etc.

Outline

1. Introduction to Statistical Learning
   - Modelling the Response
   - Assessing the Model Accuracy
   - Classification Setting
     - The Bayes Classifier
     - The K-Nearest Neighbours Classifier

Mean Squared Error (MSE)

In order to evaluate the performance of a statistical learning method on a given data set, the most commonly used measure is the *mean squared error* (MSE):

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

where $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives the ith observation in the *training* data set $\{(x_i, y_i)\}_{i=1}^{n}$.
Hence, expected *test* MSE, for any (x_0, y_0) in the *test* set contains

$$\begin{aligned}\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(f(x_0) + \epsilon - \hat{f}(x_0)\right)^2\right] \\ &= \mathbb{V}\text{ar}\left[\hat{f}(x_0)\right] + \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \mathbb{V}\text{ar}\left[\epsilon\right]\end{aligned}$$

provided that $\mathbb{E}\left[\epsilon\hat{f}(x_0)\right] = \mathbb{E}\left[\epsilon\right]\mathbb{E}\left[\hat{f}(x_0)\right] = 0$, and where

$$\text{Bias}\left(\hat{f}(x_0)\right) = \mathbb{E}\left[\hat{f}(x_0)\right] - f(x_0).$$

Bias-Variance Trade-off I

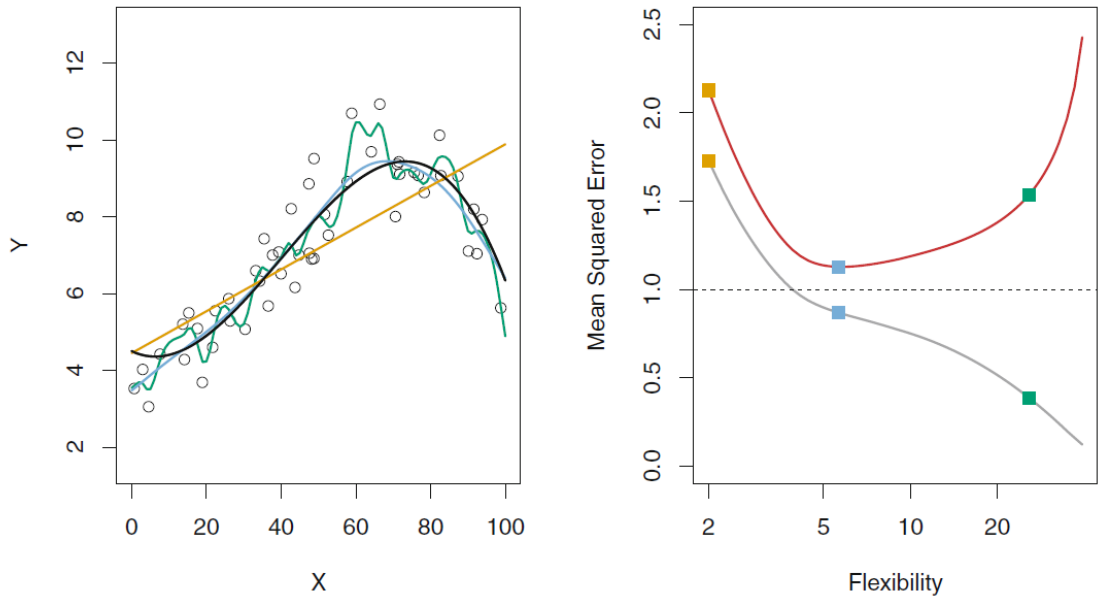

Figure 3: Bias-Variance Trade-off: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

Bias-Variance Trade-off II

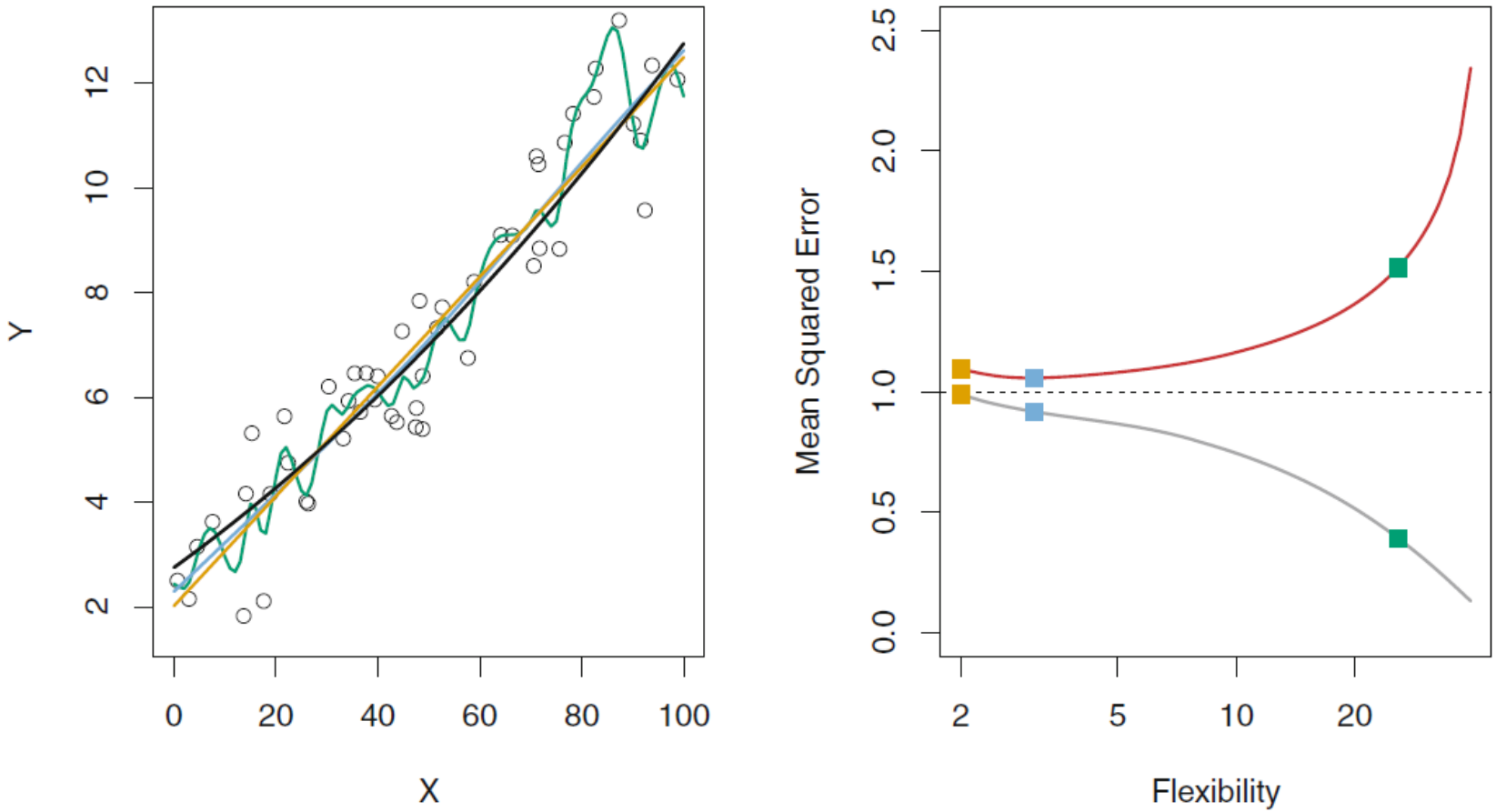

Figure 4: Bias-Variance Trade-off

Bias-Variance Trade-off III

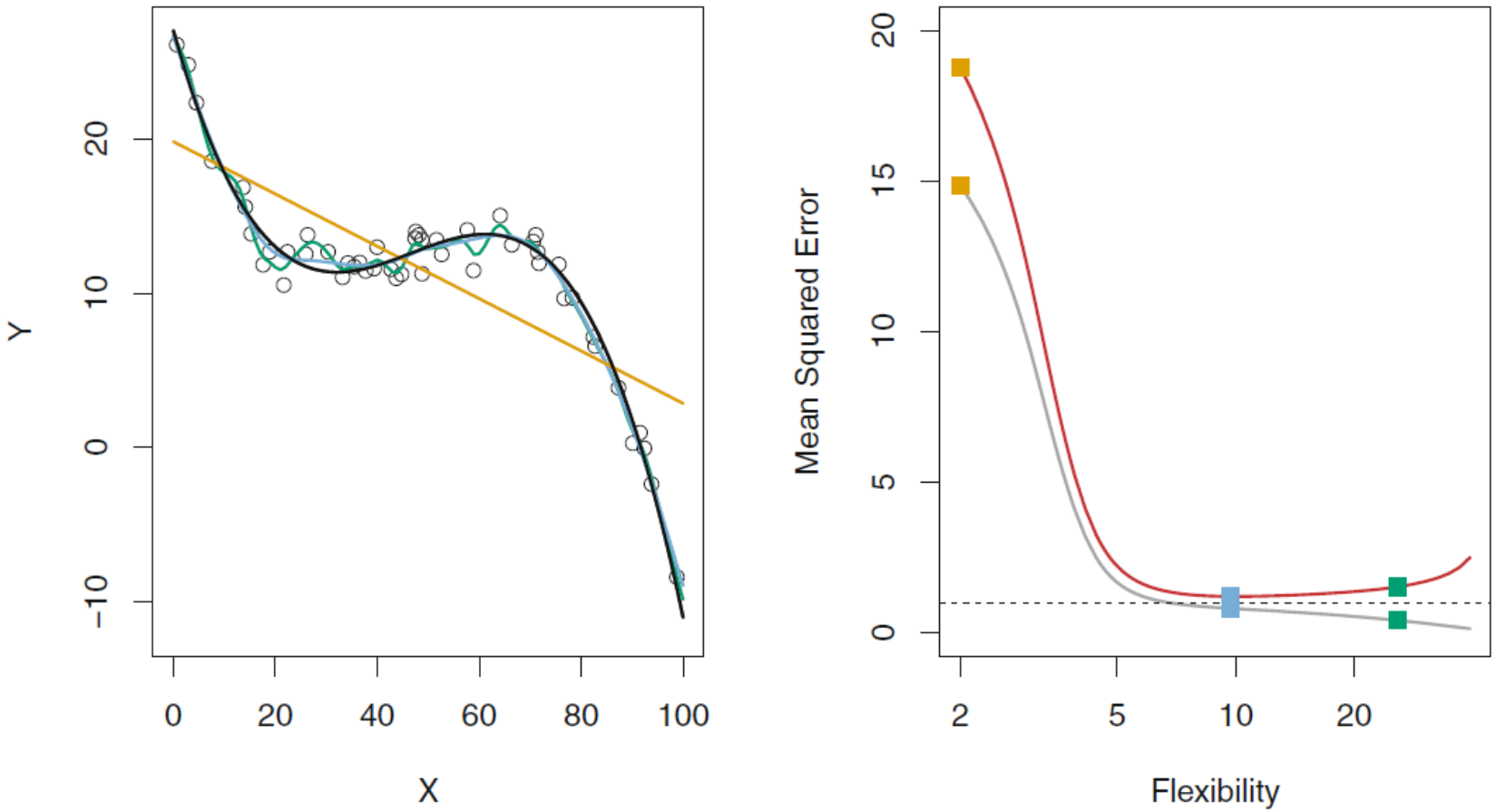

Figure 5: Bias-Variance Trade-off

Bias-Variance Trade-off IV

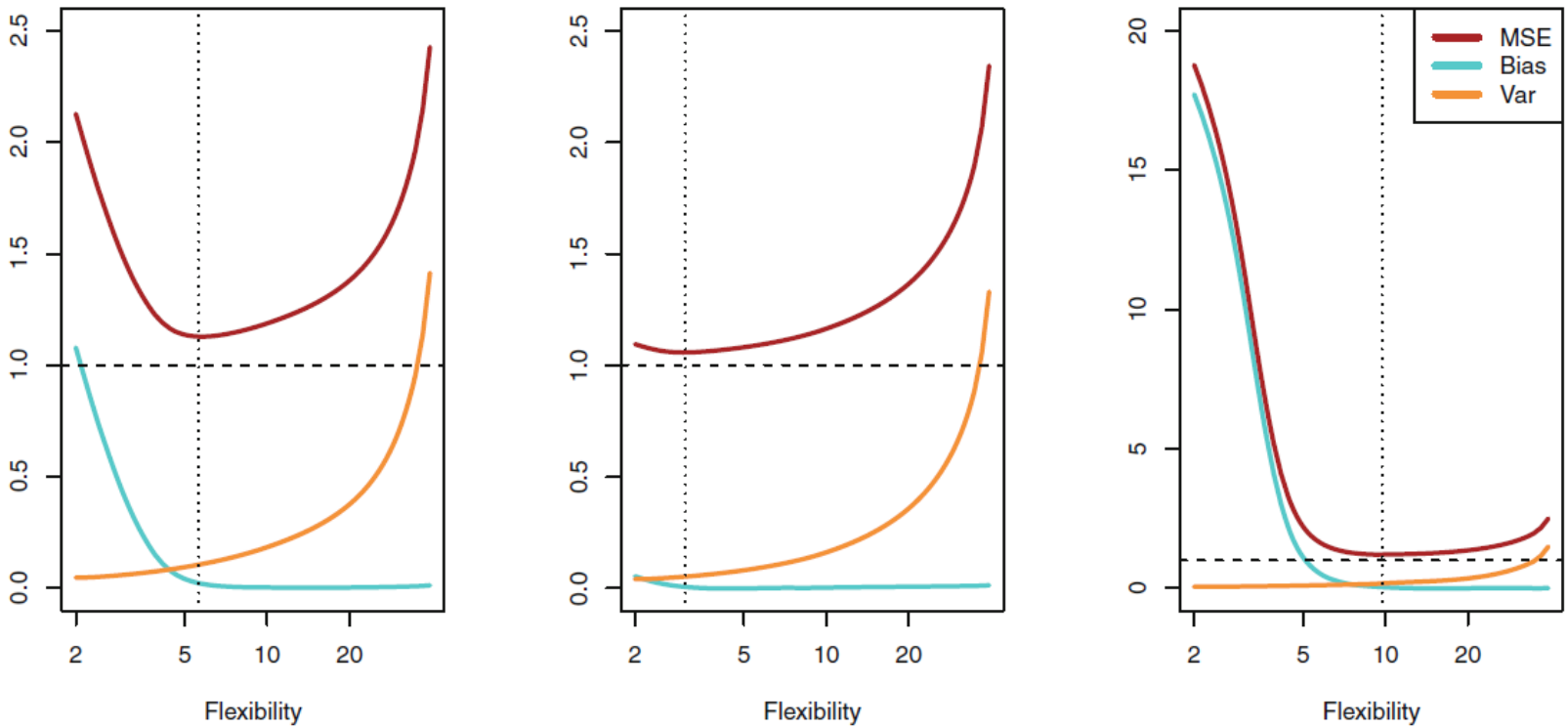

Figure 6: Bias-Variance Trade-off: Squared bias (blue curve), variance (orange curve), $\mathbb{V}\text{ar}\left[\epsilon\right]$ (dashed line), and test MSE (red curve) for the three data sets (previously shown). The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

Outline

1. Introduction to Statistical Learning
   - Modelling the Response
   - Assessing the Model Accuracy
   - Classification Setting
     - The Bayes Classifier
     - The K-Nearest Neighbours Classifier

Classification (& Clustering)

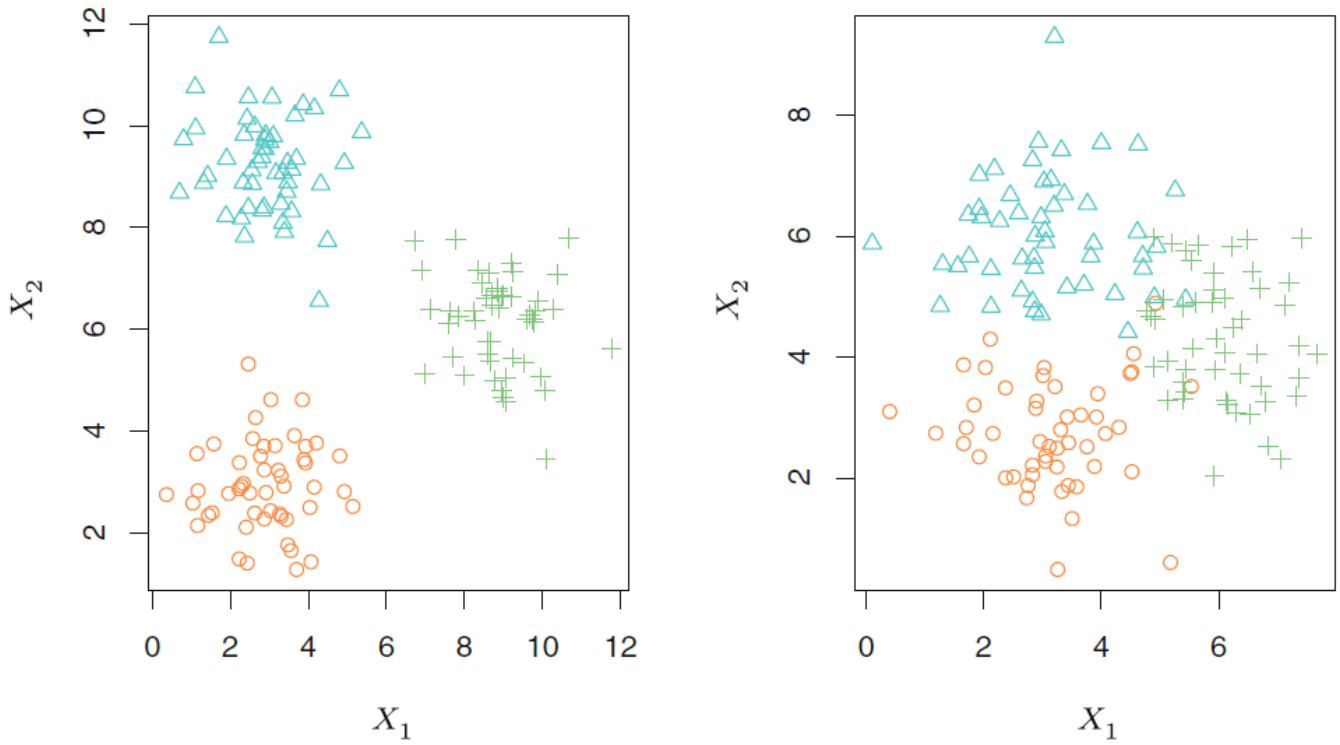

Figure 7: A clustering data set involving three groups. Each group is shown using a different coloured symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Training & Testing Error Rates

In the *classification* setting, we wish to estimate f on the basis of *training* observations $\{(x_i, y_i)\}_{i=1}^n$, where the y_i are *qualitative* and in a *class label*.

The *training error rate*, the proportion of mistakes that are made if we apply $\hat{f}$ to the training set:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \hat{y}_i\}},$$

Hence, the *test error rate*:

$$\text{Average}\left\{\mathbb{1}_{\{y_0 \neq \hat{y}_0\}}\right\}$$

for y_0 being in the test (data) set. For a "good" classifier the test error rate should be "small".

Bayes Classifier

The Bayes classifier assigns each observation to the most likely class, given its predictor values. In other words, it assigns a test observation with predictor (value) x_0 to the class j for which

$$\mathbb{P}\left\{Y = j \,|\, X = x_0\right\}$$

is "largest".
This classifier produces the lowest possible test error rate, called the *Bayes error rate*. Hence, at x_0, the error rate will be $1 - \mathbb{P}\left\{Y = j \,|\, X = x_0\right\}$. In general the overall Bayes error rate is given by

$$1 - \mathbb{E}\left[\max_j \mathbb{P}\left\{Y = j \,|\, X\right\}\right].$$

This is analogous to the *irreducible* error.

Bayes Decision Boundary

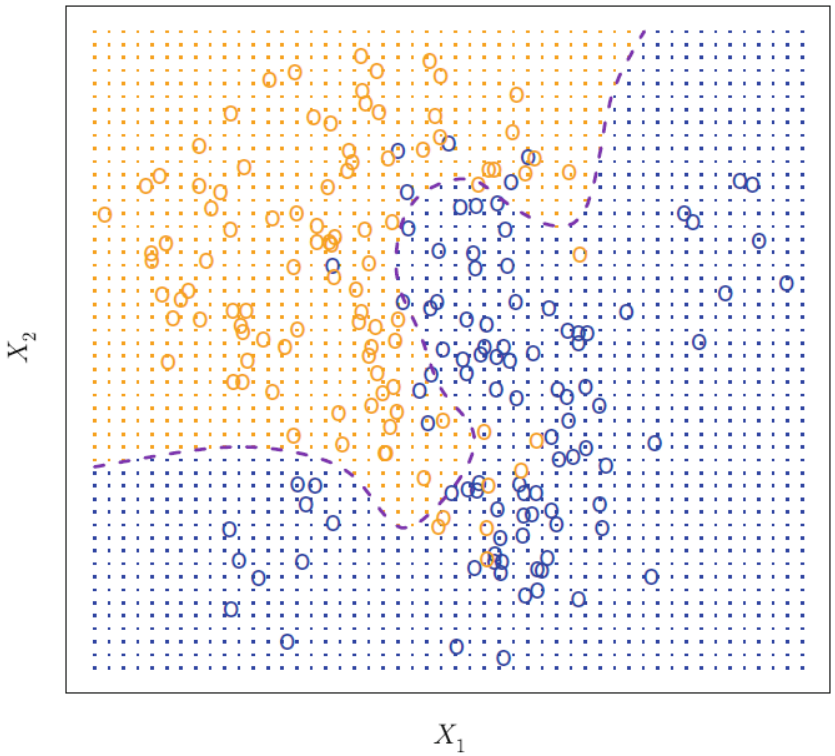

Figure 8: A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary $\mathbb{P}\{Y = \text{orange} \,|\, X = x_0\} = 0.5$. The orange background grid indicates the region in which a test observation will be assigned to the orange class.

K-Nearest Neighbours Classifier

Since we do not know (in general) the conditional distribution of Y given X for real data, the Bayes classifier stays in theoretical considerations. Thus, we classify observations based on the "estimated" conditional distribution of Y given X. Given a positive integer K and a test observation x_0:

- the K-nearest neighbours (KNN) classifier identifies K points in the *training* data that are closest to x_0, represented by $\mathcal{N}_0$;
- it estimates the conditional probability for class j as the fraction of the points in $\mathcal{N}_0$ whose response values equal j:

$$\mathbb{P}\{Y = j \,|\, X = x_0\} = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} \mathbb{1}_{\{y_i = j\}};$$

- KKN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

KNN Decision Boundary I

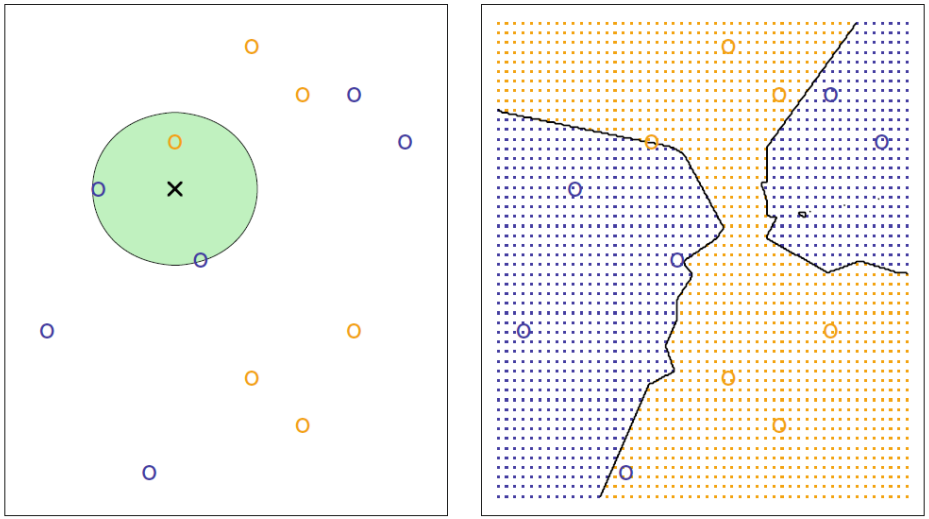

Figure 9: The KNN approach, using $K = 3$, is illustrated with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the 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.

KNN Decision Boundary II

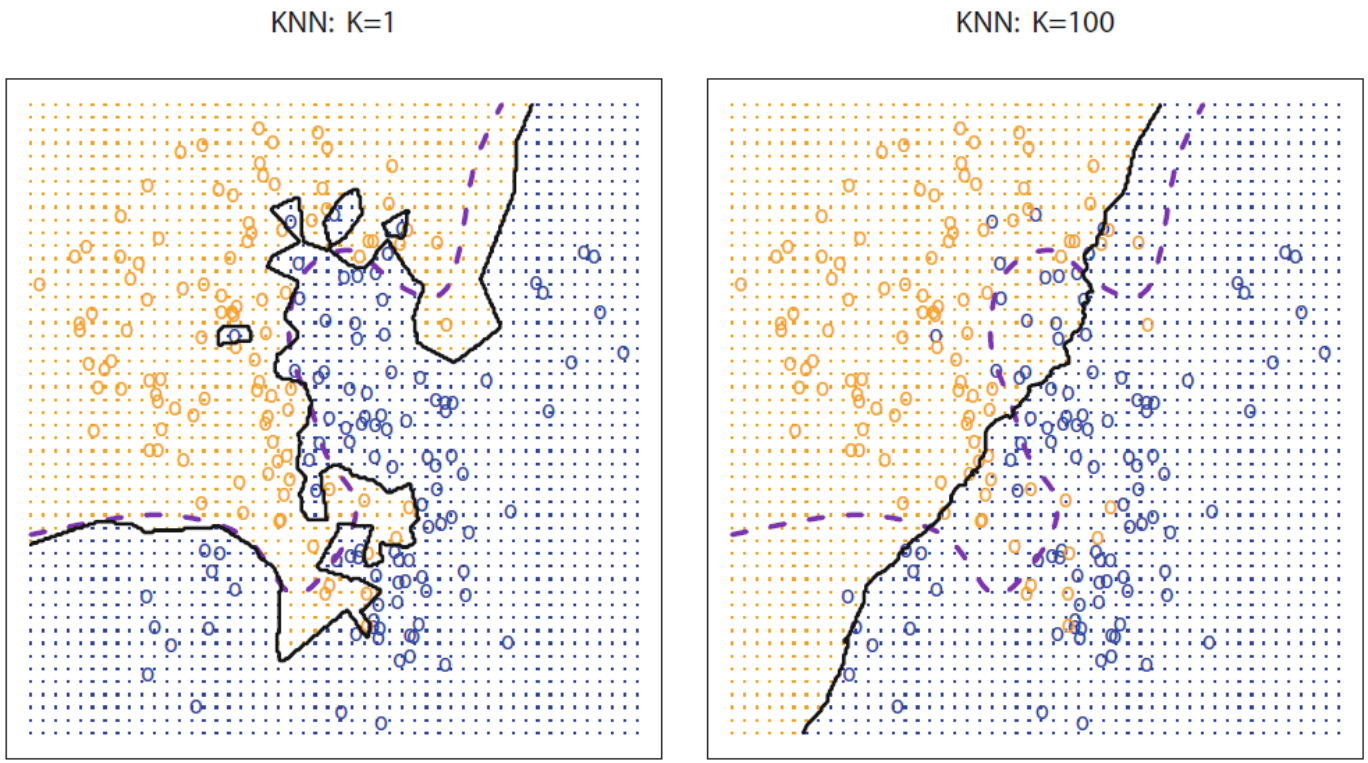

Figure 10: A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.

KNN Decision Boundary III

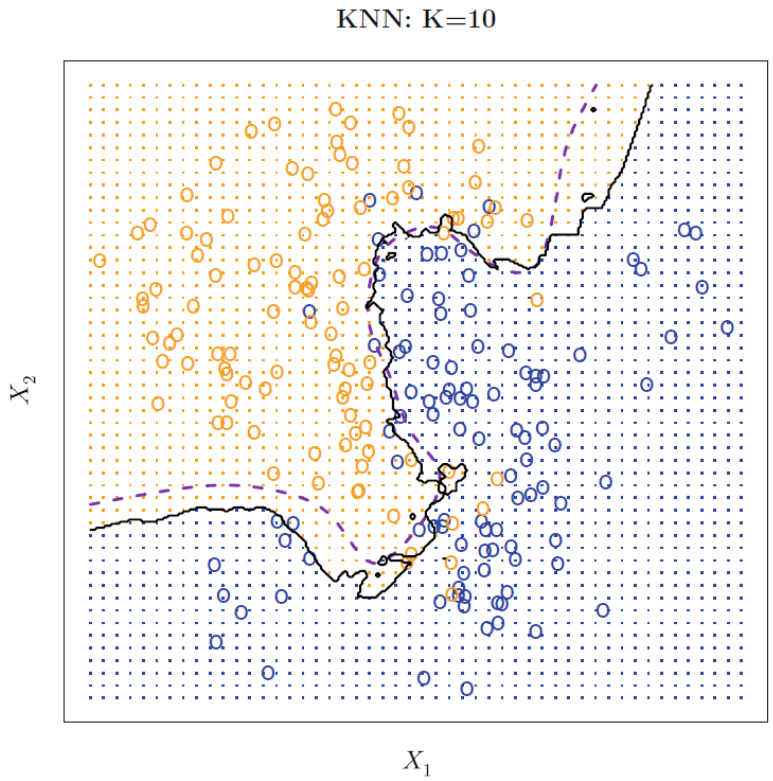

Figure 11: The black curve indicates the KNN decision boundary on the data, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.

KNN Trade-off

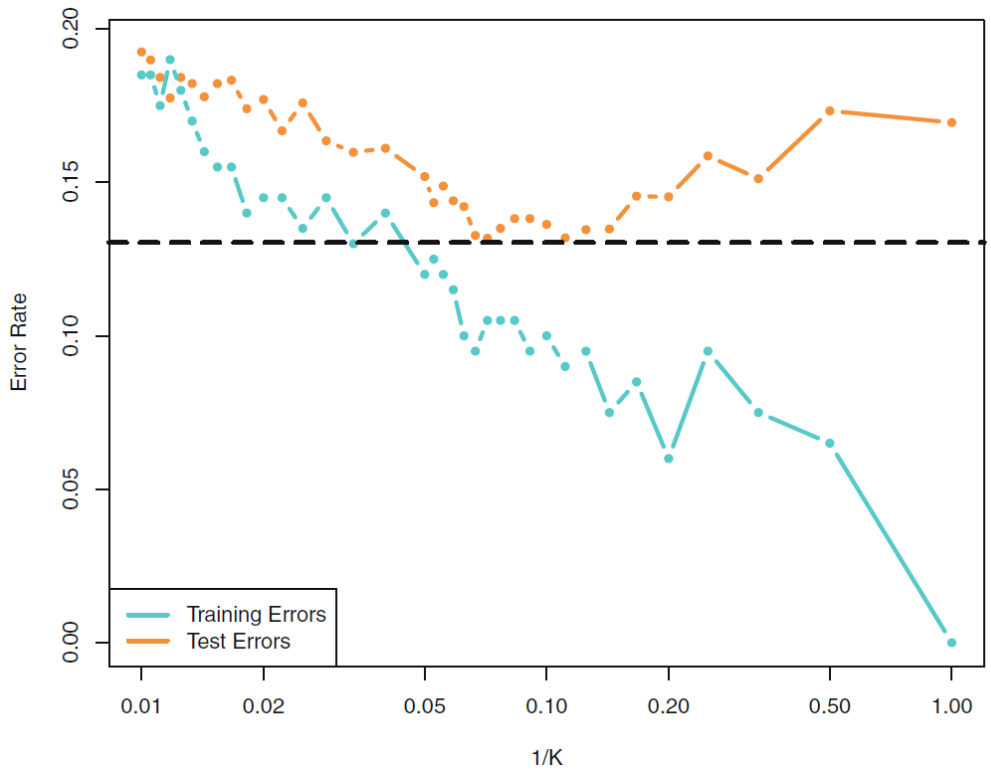

Figure 12: The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbours K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.

Thanks!...

Thanks!...