Mathematical Foundations of Data Analysis (MFDA) - II

Boqiang Huang

huang@math.uni-koeln.de

Institute of Mathematics, University of Cologne, Germany

2019.04.16-18

Chapter 2. Linear Regression

- 1. A simple overview of statistical learning
  - 1.1. Concepts
  - 1.2. Statistical modeling
  - 1.3. Model accuracy
- 2. Linear regression
  - 2.1 Simple linear regresion
  - 2.2 Multiple linear regression
  - 2.3 Other considerations

Reference:

[1] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning: with applications in R, Springer, 2013.
[2] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics, 2016.
[3] A. Antoniou, W.-S. Lu, Practical optimization: algorithms and engineering applications, Springer, 2007.

1. A simple overview of statistical learning

1.1. Concepts

- Statistical learning vs machine learning

  Machine learning is a general concept in the artificial intelligence

  Statistical learning more emphasizes on statistical models and their interpretability, precision and uncertainty

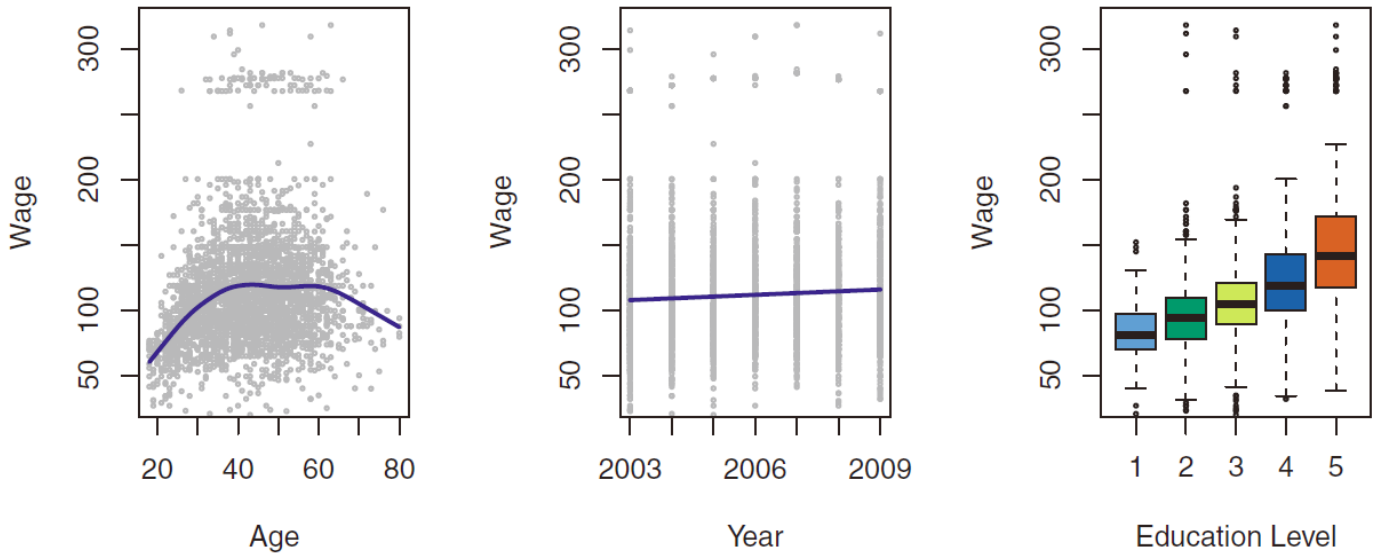
1. A simple overview of statistical learning

1.1. Concepts

- Statistical learning vs machine learning

  Machine learning is a general concept in the artificial intelligence

  Statistical learning more emphasizes on statistical models and their interpretability, precision and uncertainty

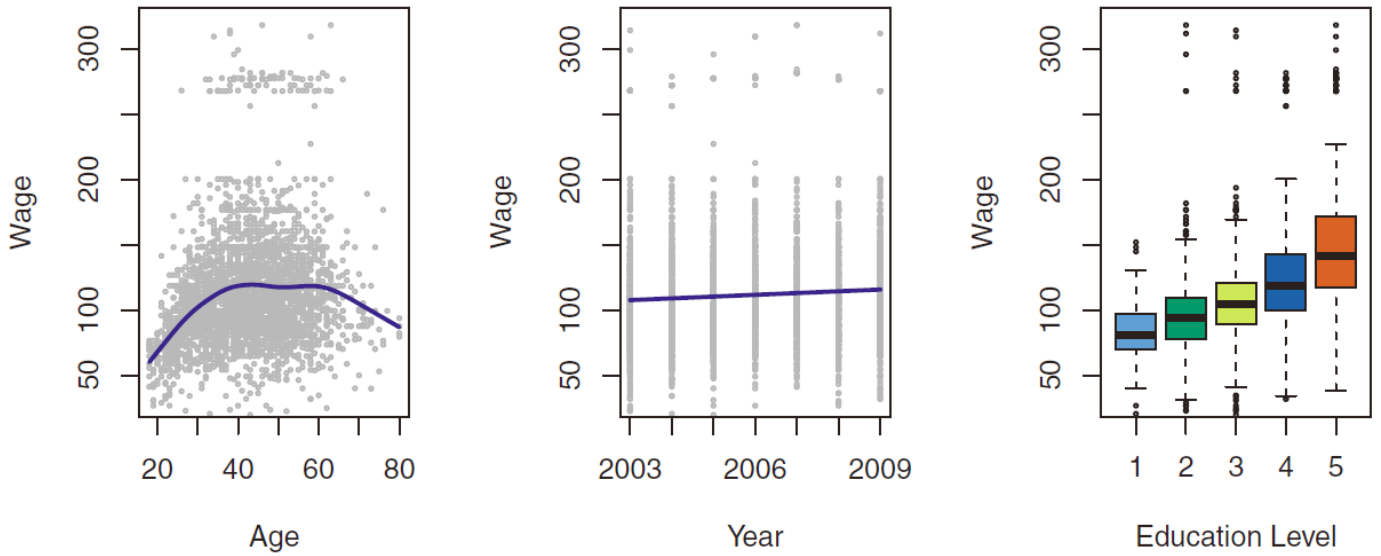


Income survey information for males from the central Atlantic region of the United States

1. A simple overview of statistical learning

1.1. Concepts

- Statistical learning vs machine learning
- Regression: predict a continuous or quantitative ouput value

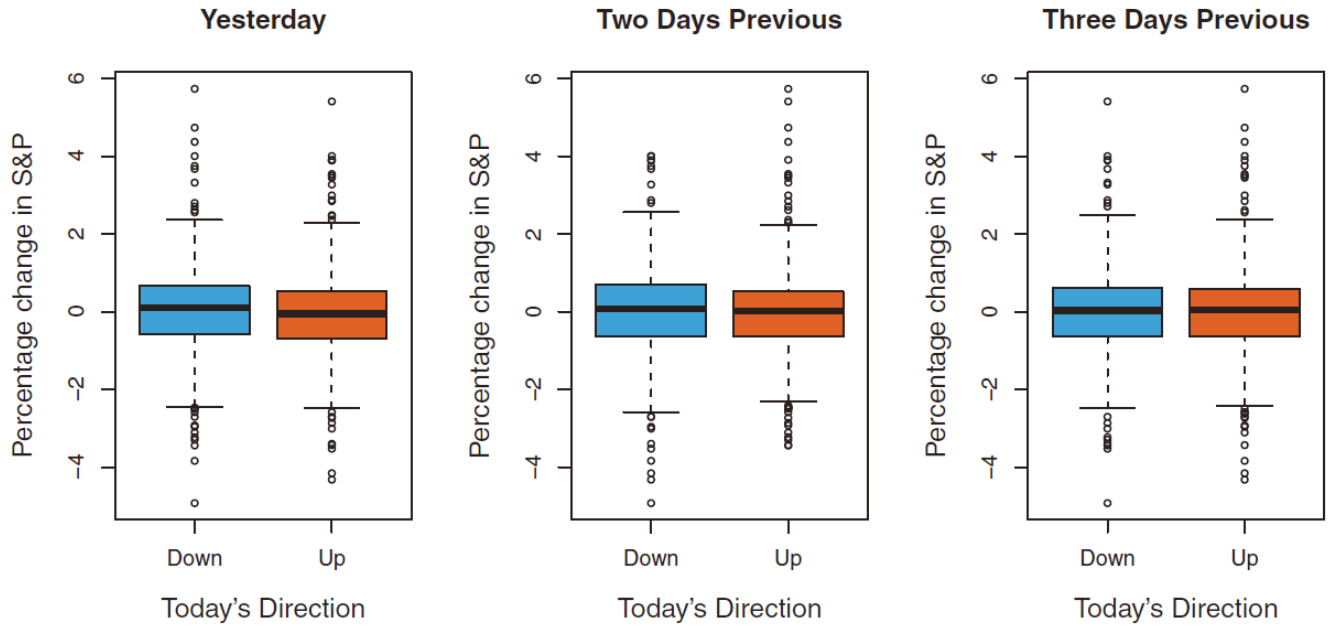


Income survey information for males from the central Atlantic region of the United States

1. A simple overview of statistical learning

1.1. Concepts

- Statistical learning vs machine learning
- Regression: predict a continuous or quantitative ouput value
- Classification: predict a non-numerical, categorical, or qualitative value

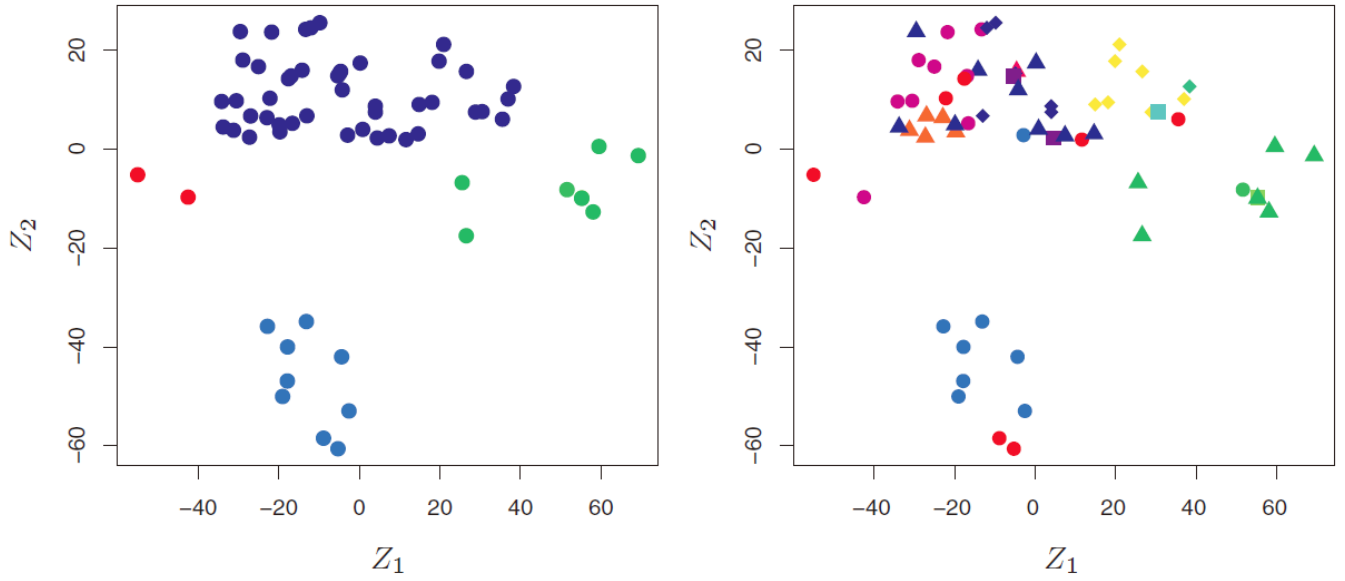


Previous day's percentage change in the Standard & Poor's 500 stock index

1. A simple overview of statistical learning

1.1. Concepts

- Statistical learning vs machine learning
- Regression: predict a continuous or quantitative ouput value
- Classification: predict a non-numerical, categorical, or qualitative value
- Clustering: observe input with no output, determine if there are groups or clusters



Representation of some gene expression data set in a 2-D space

1. A simple overview of statistical learning

1.2. Statistical modeling

Suppose that we observe a response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ such that

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

here ϵ is a random error term which is independent of X and has zero mean

1. A simple overview of statistical learning

1.2. Statistical modeling

Suppose that we observe a response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ such that

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

here ϵ is a random error term which is independent of X and has zero mean

- Prediction: a set of inputs X are readily available, but the output Y cannot be easily obtained

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

$\hat{f}$ represents the estimate of f, and $\hat{Y}$ represents the prediction for Y

1. A simple overview of statistical learning

1.2. Statistical modeling

Suppose that we observe a response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ such that

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

here ϵ is a random error term which is independent of X and has zero mean

- Prediction: a set of inputs X are readily available, but the output Y cannot be easily obtained

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

$\hat{f}$ represents the estimate of f, and $\hat{Y}$ represents the prediction for Y

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Remark: irreducible error provides an upper bound on the prediction accuracy for Y, which is unknown in practice!

1. A simple overview of statistical learning

1.2. Statistical modeling

Suppose that we observe a response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ such that

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

here ϵ is a random error term which is independent of X and has zero mean

- Inference: understand the way that Y is affected as $X_1, X_2, \ldots, X_p$ change

  for instance:

  which predictors are more important to the output Y ?

  how to quantify the relationship between Y and each predictor X_i?

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

Training data set: contains all necessary observations for teaching our method how to learn/estimate *f*

- Parametric methods
- Non-parametric methods

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

Training data set: contains all necessary observations for teaching our method how to learn/estimate *f*

- Parametric methods

  The mapping relationship $\hat{f}$ is clearly assumed, e.g. a linear model

$$\hat{f}(\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + + \beta_p \boldsymbol{X}_p$$

  Fit or train the model based on the observations from the training set

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{Y} - \hat{f}(\boldsymbol{X}) \right\|_{L_p}^{p}$$

  with or without contraints

  where $\boldsymbol{\beta}^* = [\beta_0^* \quad \cdots \quad \beta_p^*]^T \in \mathbb{R}^{p+1}$ will be the best parameters of the model

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

Training data set: contains all necessary observations for teaching our method how to learn/estimate *f*

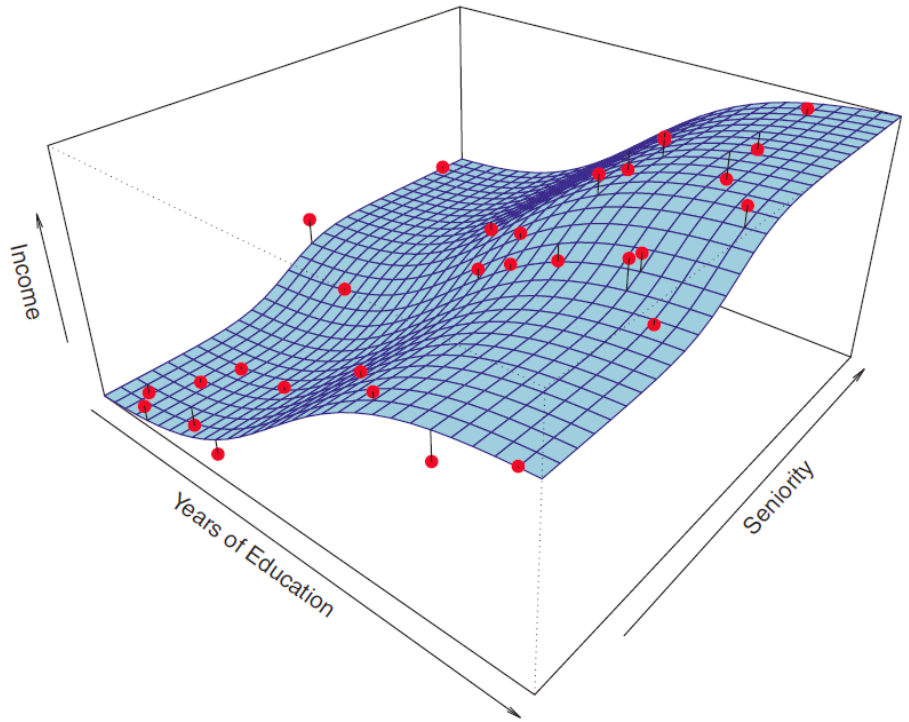
- Parametric methods

Advantage:	reduce the problem of estimating *f* down to estimating a set of parameters;
Disadvantage:	the estimation will be poor if the selected model $\hat{f}$ is far from the true *f*
Balancing:	more flexible model can fit more different functional forms for *f* overfit the data as the model follow the noise too much

1. A simple overview of statistical learning

1.2. Statistical modeling

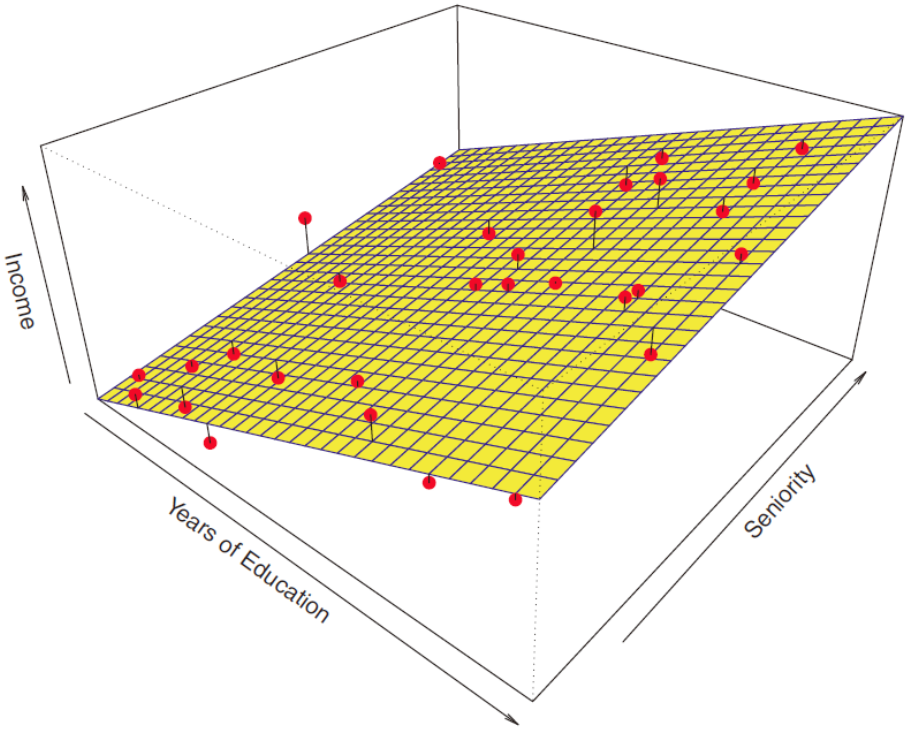
1.2.1 Estimating *f*



Simulated income data set based on the ideal blue surface

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

Red dots: 30 observed individuals



A linear model fit by least squares to the income data

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

Red dots: 30 observed individuals

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

Training data set: contains all necessary observations for teaching our method how to learn/estimate *f*

- Parametric methods

  Advantage: reduce the problem of estimating *f* down to estimating a set of parameters;

  Disadvantage: the estimation will be poor if the selected model $\hat{f}$ is far from the true *f*

  Balancing: more flexible model can fit more different functional forms for *f*
  overfit the data as the model follow the noise too much

- Non-parametric methods

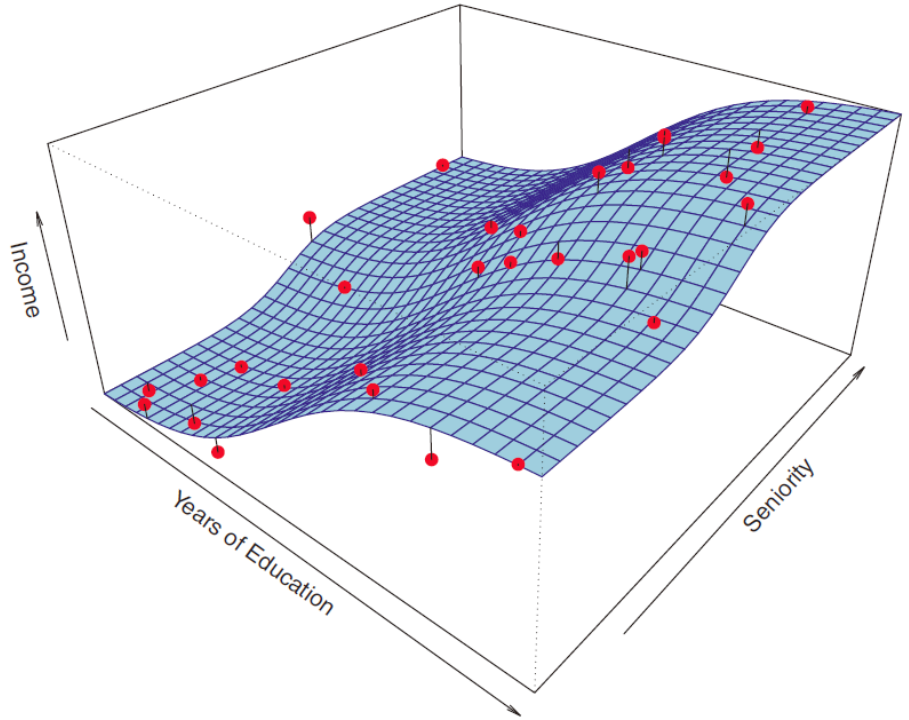
  Advantage: do not assume a particular functional form for *f*

  Disadvantage: more observation is required comparing to parametric methods, overfit still exsits

1. A simple overview of statistical learning

1.2. Statistical modeling

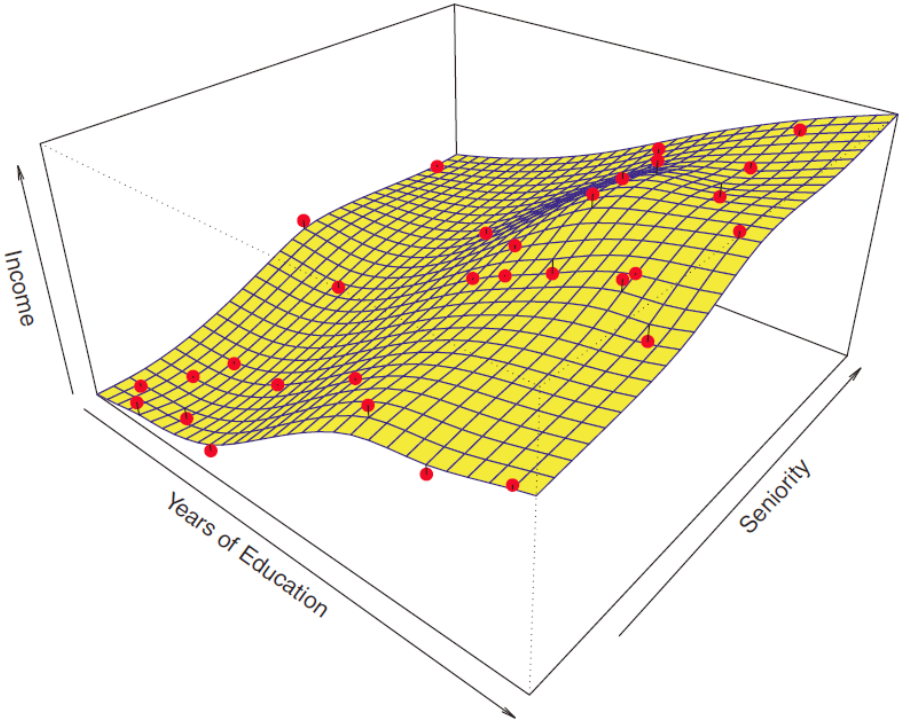
1.2.1 Estimating *f*



Simulated income data set based on the ideal blue surface

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

Red dots: 30 observed individuals



A smooth thin-plate spline fit to the income data

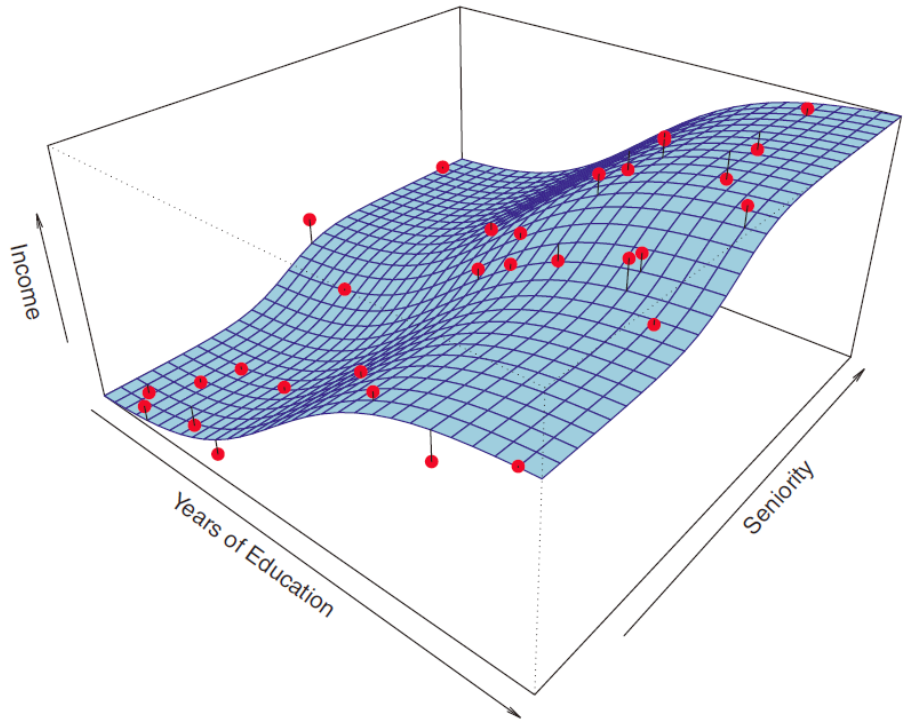
$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{Y} - \hat{f}(\boldsymbol{X}) \right\|_{L_p}^{p}$$

Red dots: 30 observed individuals

1. A simple overview of statistical learning

1.2. Statistical modeling

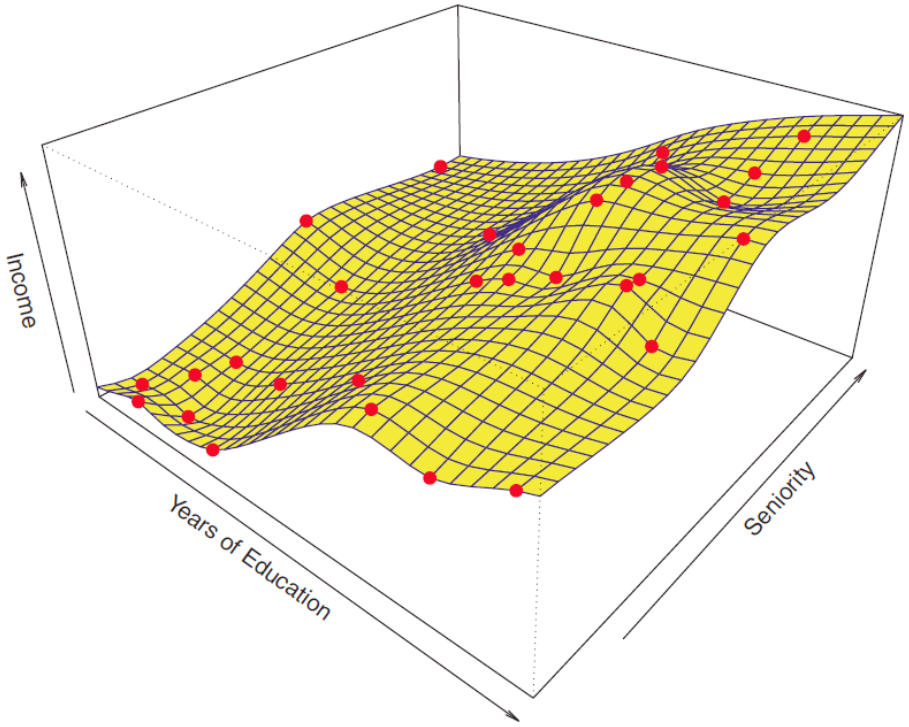
1.2.1 Estimating *f*



Simulated income data set based on the ideal blue surface

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

Red dots: 30 observed individuals



Another thin-plate spline overfit to the income data

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{Y} - \hat{f}(\boldsymbol{X}) \right\|_{L_p}^{p}$$

Red dots: 30 observed individuals

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

1.2.2 Trade-off between prediction accuracy and model interpretability

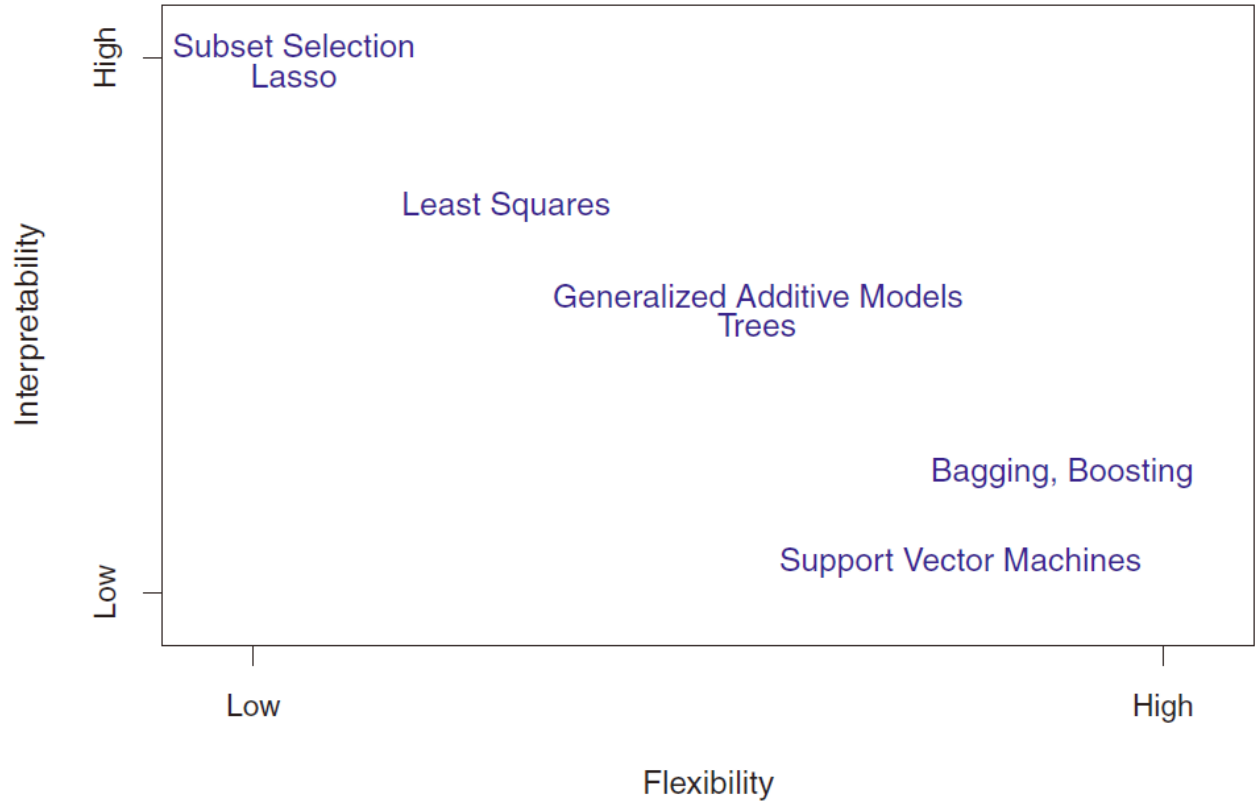
less flexiable	small range of shapes to estimate *f*	easy to explain
more flexiable	large range of shapes to estimate *f*	hard to explain

1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

1.2.2 Trade-off between prediction accuracy and model interpretability



1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

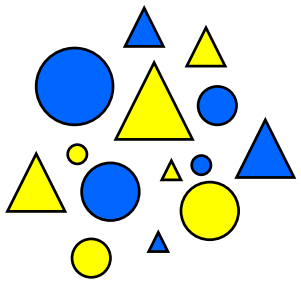
1.2.2 Trade-off between prediction accuracy and model interpretability

1.2.3 Supervised vs unsupervised learning

- Supervised learning

  for each observation of the predictor $\boldsymbol{x}_i$, there is an associated response $\boldsymbol{y}_i$

  wish to understand the relationship between the response and the predictor



the whole data set

training set

1. A simple overview of statistical learning

1.2. Statistical modeling

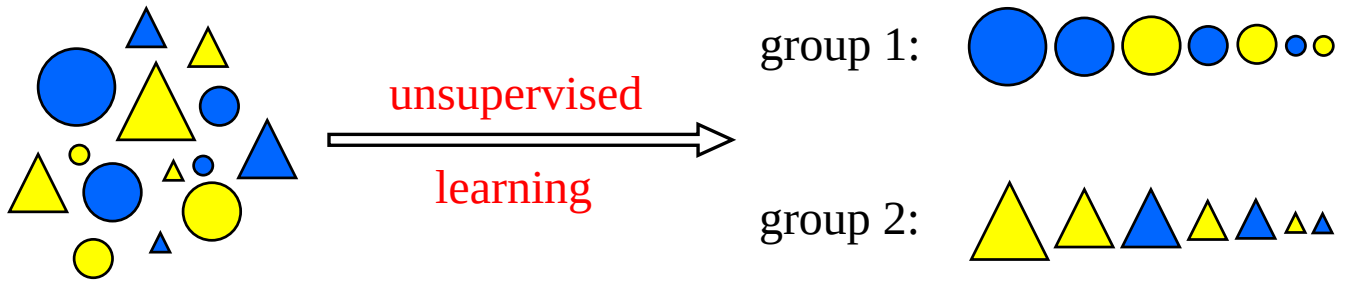
1.2.1 Estimating *f*

1.2.2 Trade-off between prediction accuracy and model interpretability

1.2.3 Supervised vs unsupervised learning

- Supervised learning
- Unsupervised learning

each observation only has predictor $\boldsymbol{x}_i$, there is no associated response $\boldsymbol{y}_i$



1. A simple overview of statistical learning

1.2. Statistical modeling

1.2.1 Estimating *f*

1.2.2 Trade-off between prediction accuracy and model interpretability

1.2.3 Supervised vs unsupervised learning

1.2.4 Regression vs classification

- Regression: quantitative variables take on numerical values
- Classification: qualitative variables take on values in one of *K* different classes or categories

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

Regression: Mean Squared Error (MSE)

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

Training data set: Tr $\coloneqq \{x_i, y_i\}_1^N$ $\quad$ $\mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i\in\mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$

Test data set: Te $\coloneqq \{x_i, y_i\}_1^M$ $\quad$ $\mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i\in\mathsf{Te}}[y_i - \hat{f}(x_i)]^2$

Remark: choose a method that gives the lowest test MSE, as opposed to the lowest training MSE

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

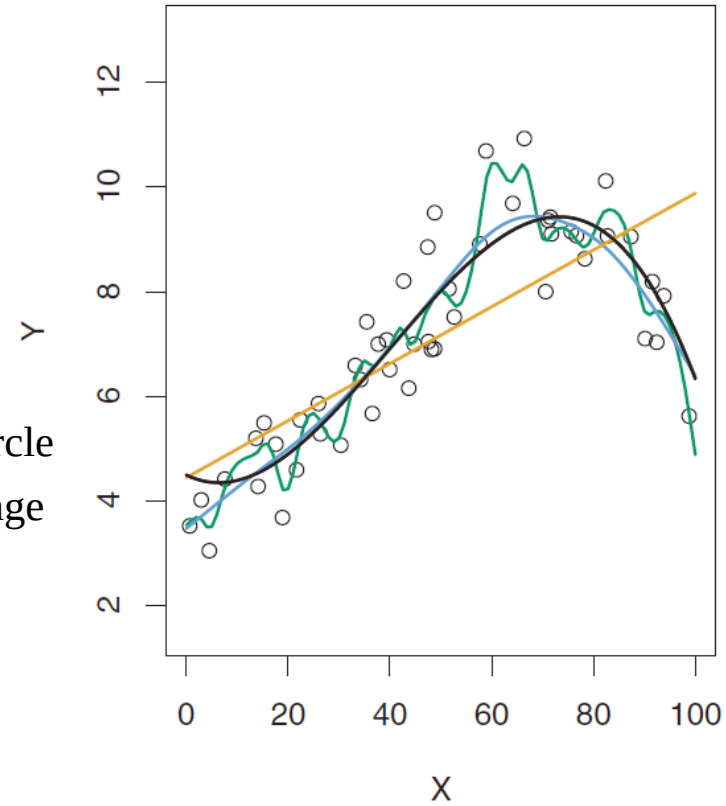
Ideal *f*, black curve

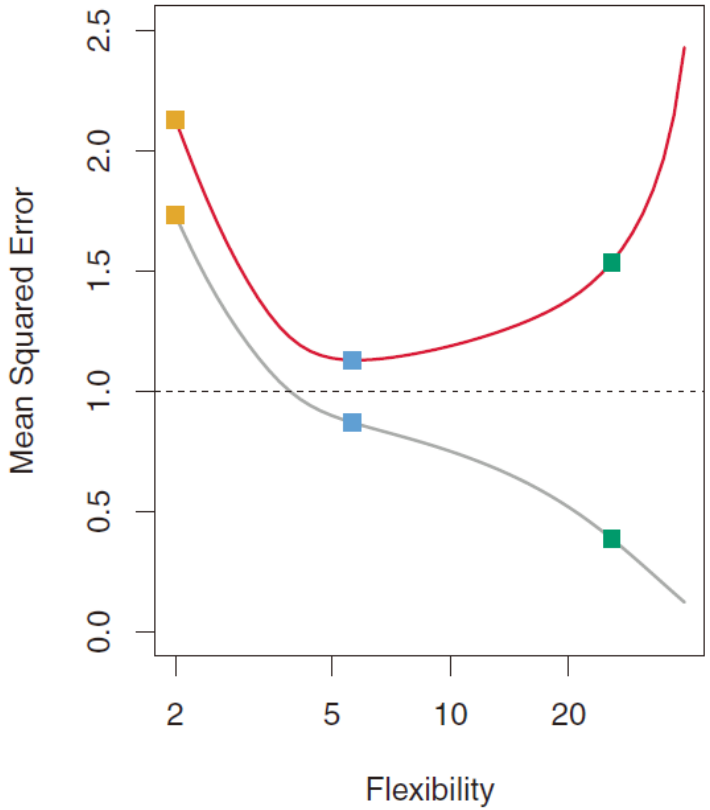
Observations, black circle

Linear regression, orange

Spline fit 1, blue

Spline fit 2, gree





Training MSE, grey

Test MSE, red

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

Regression: Mean Squared Error (MSE)

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

Training data set: Tr $\coloneqq \{x_i, y_i\}_1^N$ $\quad \mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i\in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$

Test data set: Te $\coloneqq \{x_i, y_i\}_1^M$ $\quad \mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i\in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$

Remark: choose a method that gives the lowest test MSE, as opposed to the lowest training MSE

Real world: easy to compute the training MSE based on the training data set

difficult to estimate the test MSE because usaually no test data are availiable

cross-validation !!!

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a test observation

if the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$)

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

Expected test MSE

$$\mathrm{Bias}(\hat{f}(x_0))] = E[\hat{f}(x_0)] - f(x_0)$$

to minimize the expected test error, we need a statistical learning method with low variance and low bias

the more flexibility the higher variance, but the lower bias!!!

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

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

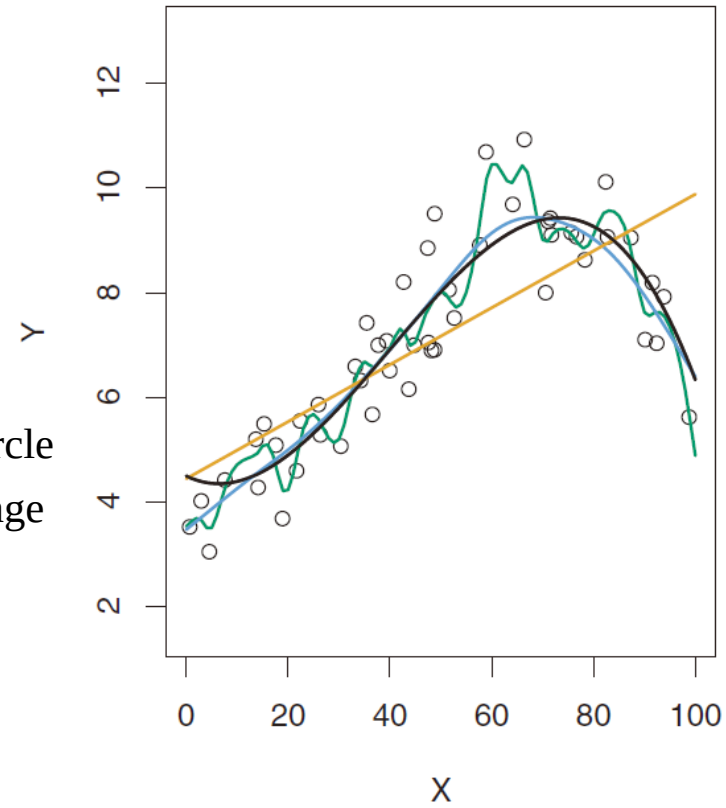
Ideal *f*, black curve

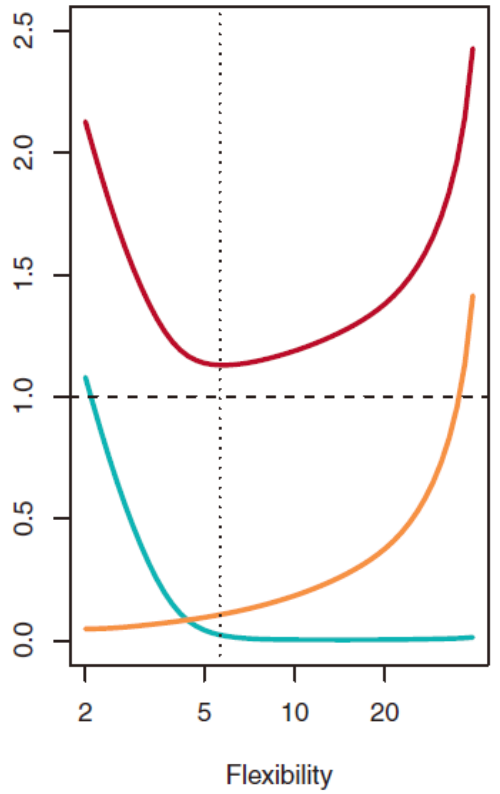
Observations, black circle

Linear regression, orange

Spline fit 1, blue

Spline fit 2, gree





Test MSE, red

$\mathrm{Var}(\epsilon)$, dashed line

Variance, orange

Squared bias, blue

1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

1.3.3 The classification setting

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

$I(y_i \neq \hat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$

Test rate $\text{Ave}\,(I(y_0 \neq \hat{y}_0))$

The Bayes clasifier: assigns each observation to the most likely class, given its predictor values

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

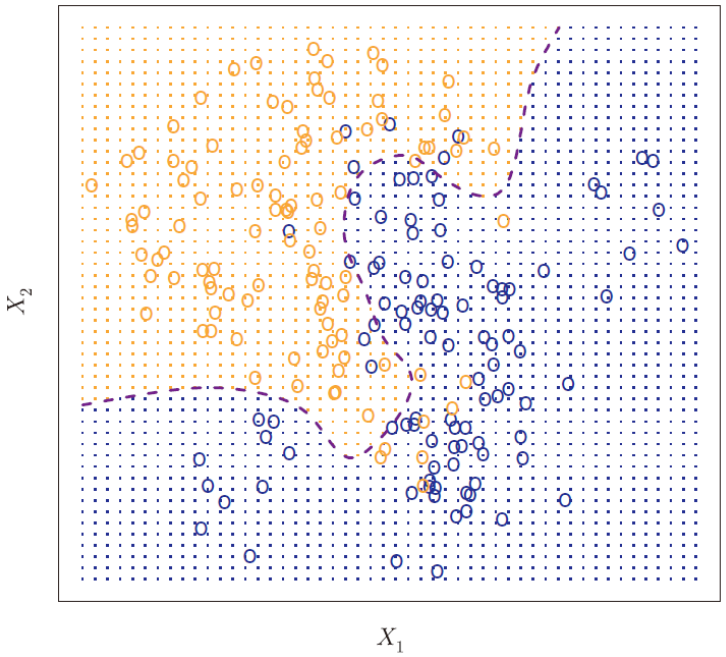
1. A simple overview of statistical learning

1.3. Model accuracy

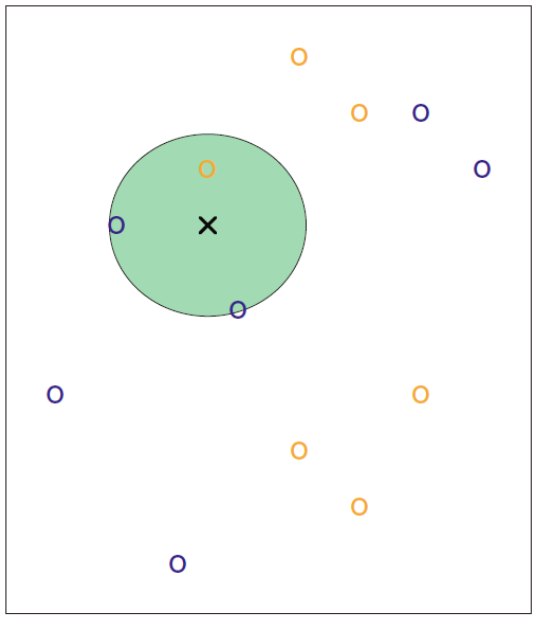
1.3.1 Measuring the quality of fit

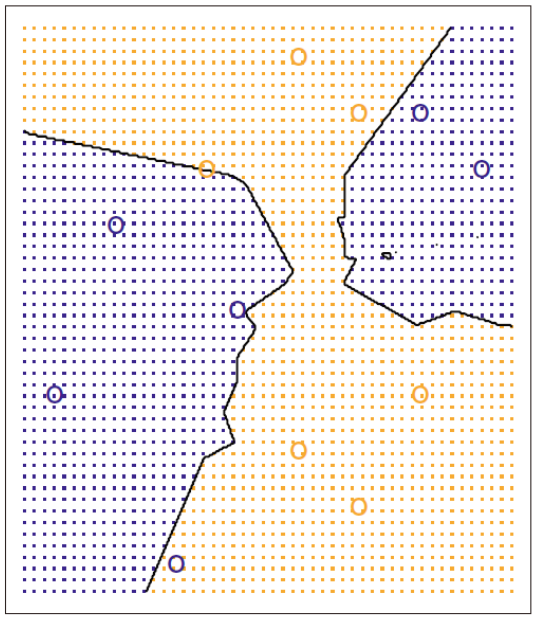
1.3.2 The bias-variance trade-off

1.3.3 The classification setting



Simulated data set consisting of 100 observations in two groups





Idea of KNN

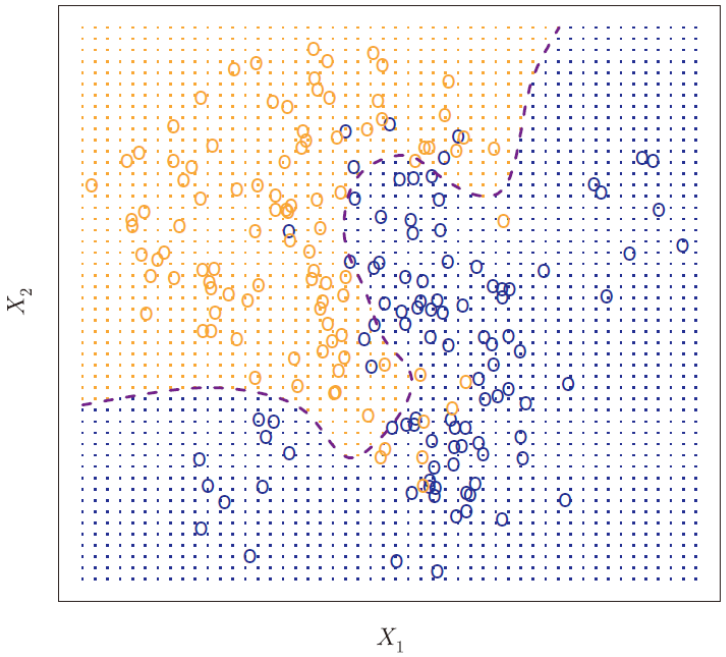
1. A simple overview of statistical learning

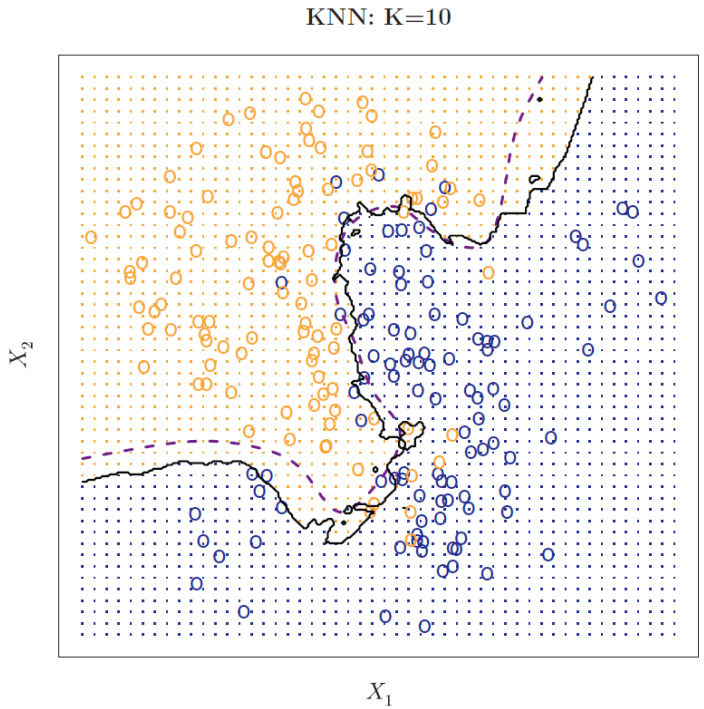
1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

1.3.3 The classification setting





Simulated data set consisting of 100 observations in two groups

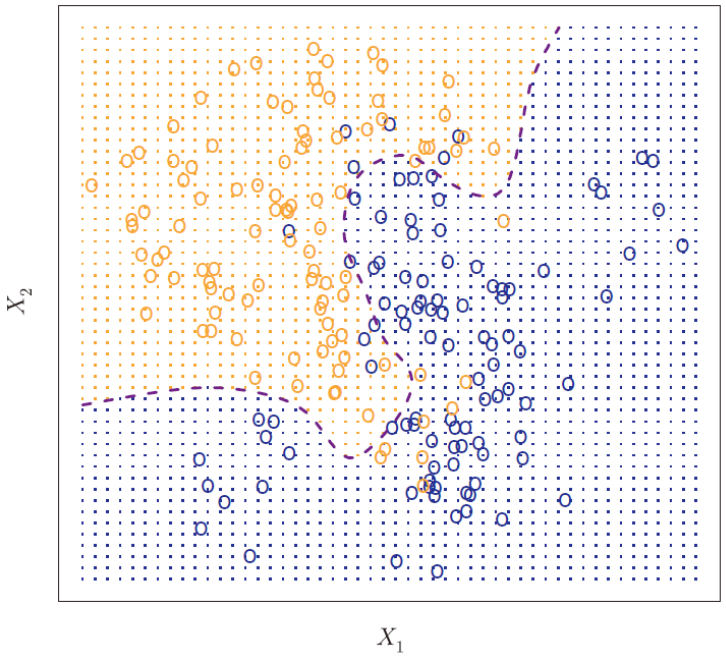
1. A simple overview of statistical learning

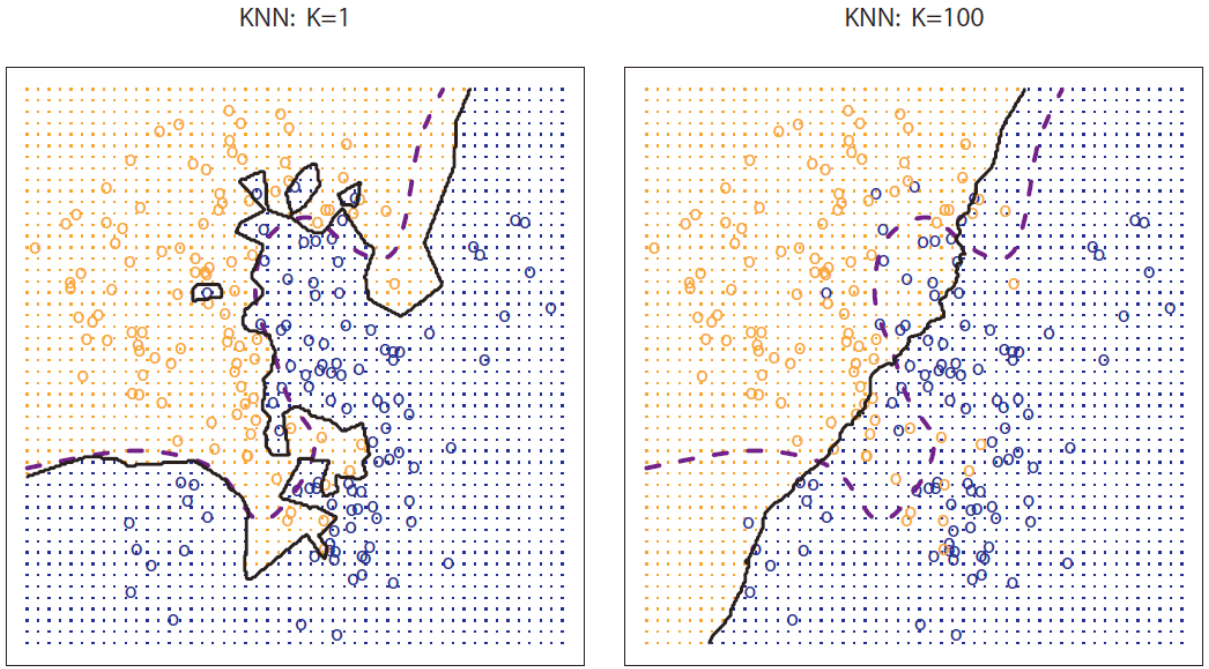
1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

1.3.3 The classification setting





Simulated data set consisting of 100 observations in two groups

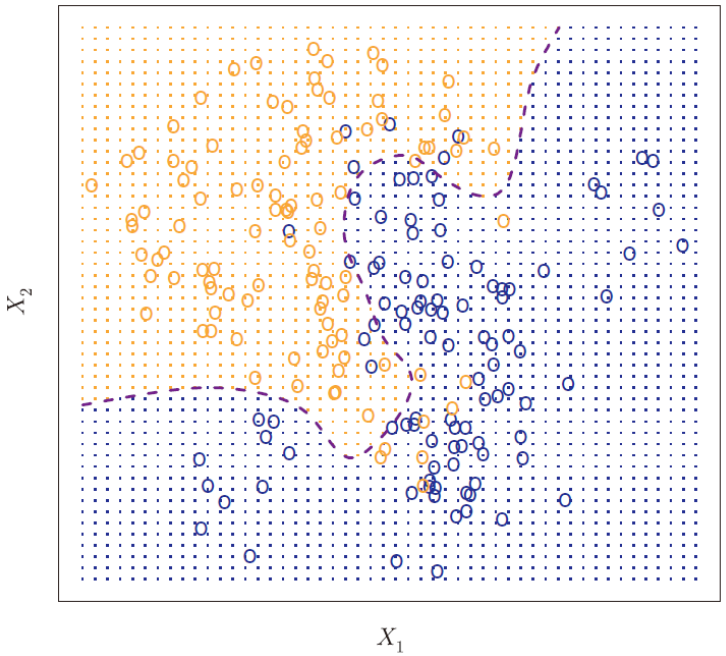
1. A simple overview of statistical learning

1.3. Model accuracy

1.3.1 Measuring the quality of fit

1.3.2 The bias-variance trade-off

1.3.3 The classification setting



Simulated data set consisting of 100 observations in two groups