Advanced Computational Econometrics: Machine Learning

Chapter 1: Introduction

Thierry Denœux

July 2019

What is Machine Learning?

- "A field of study that gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).
- At the intersection of Computer Science (Artificial Intelligence) and Statistics (Computational Statistics, Statistical Learning).
- Part of the growing field of Data Science.
- ML exists since the appearance of the first computers in the 1950's, but it has recently gained considerable interest because of new applications such as
  - Search engines
  - Social networks
  - E-commerce (recommendation systems)
  - Big data analytics, etc.
- The biggest recruiters today in this field are big IT companies such as Google, Facebook, Microsoft, Amazon, Baidu, Alibaba, etc.

Overview

Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

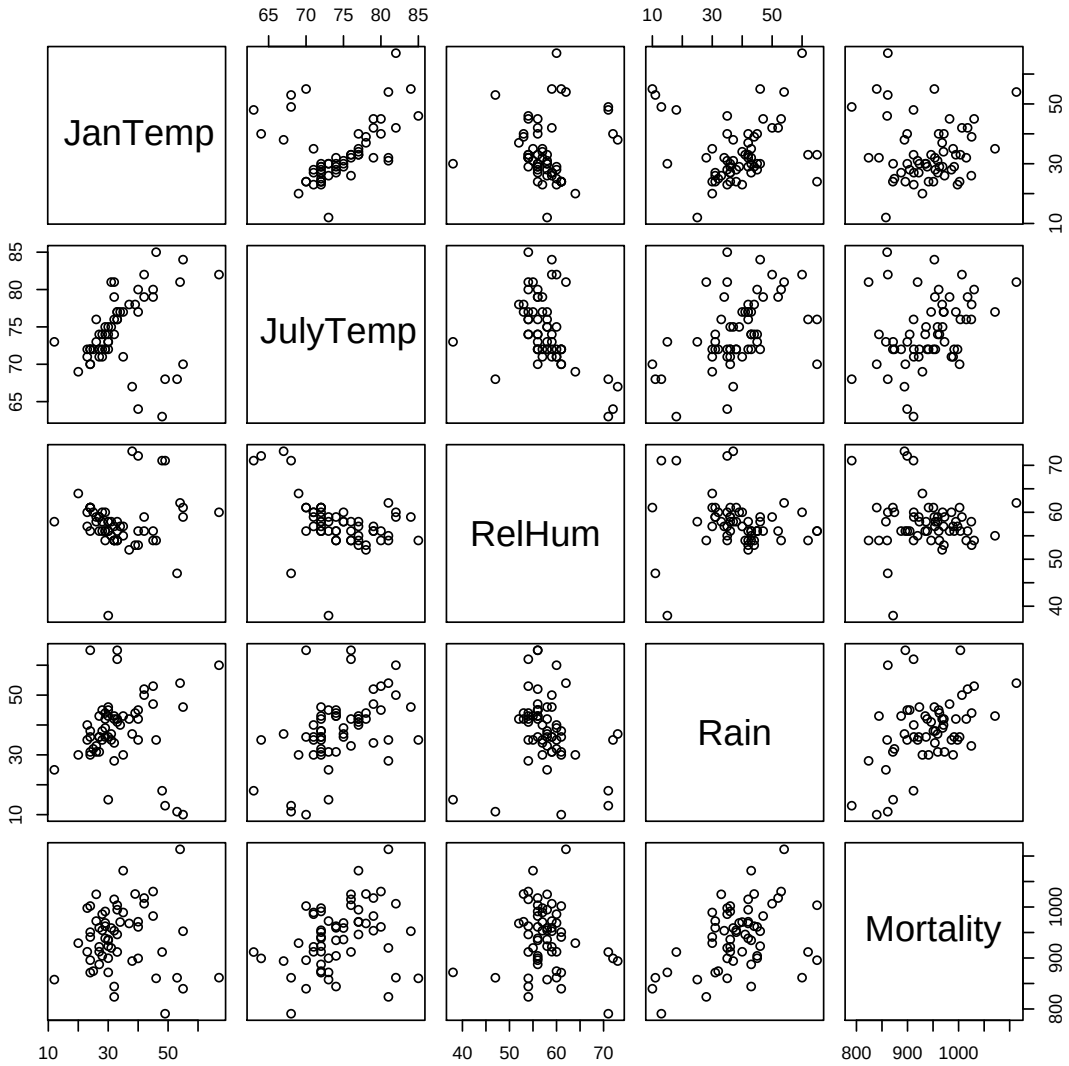
Air pollution and mortality

- Which sociological, climatic and pollution factors influence mortality?
- How do these variables interact?
- Can we predict the impact of a reduction of pollution on mortality?

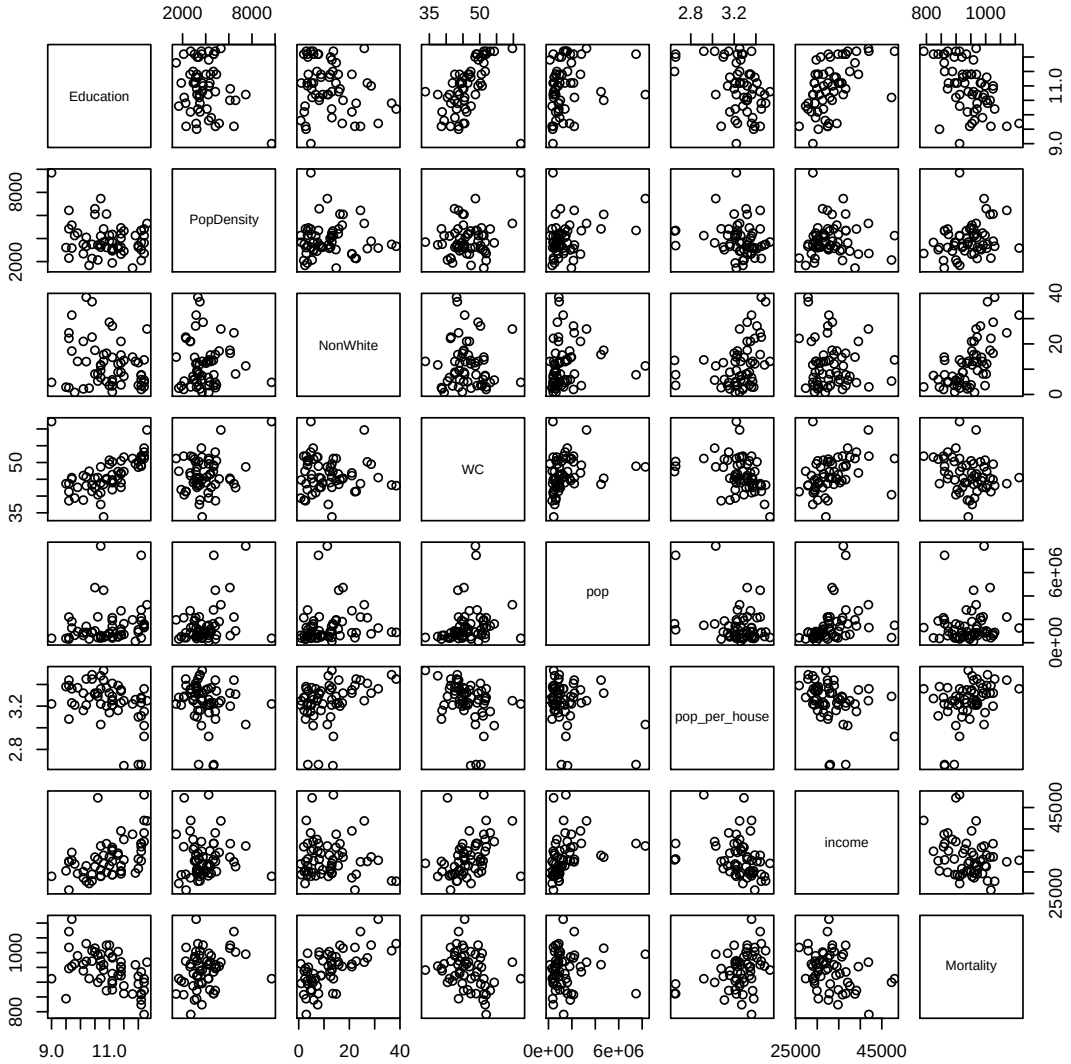
Pollution Data

- Data from McDonald and Schwing (1973) on 15 predictors and a measure of mortality in 60 US metropolitan areas in 1959-1961.
  1. Age Adjusted Mortality Rate (response)
  2. Mean annual precipitation in inches
  3. Mean January temperature in degrees Fahrenheit
  4. Mean July temperature in degrees Fahrenheit
  5. Percent of 1960 population that is 65 years of age or over
  6. Population per household, 1960
  7. Median school years completed for those over 25 in 1960
  8. Percent of housing units that are found with facilities
  9. Population per square mile in urbanized area in 1960
  10. % of 1960 urbanized area population that is non-white
  11. % employment in white-collar occupations in 1960 urbanized area
  12. % of families with income under 3,000 in 1960 urbanized area
  13. Relative population potential of hydrocarbons, HC
  14. Relative pollution potential of oxides of nitrogen, NOx
  15. Relative pollution potential of sulfur dioxide, SO2
  16. Percent relative humidity, annual average at 1 p.m.

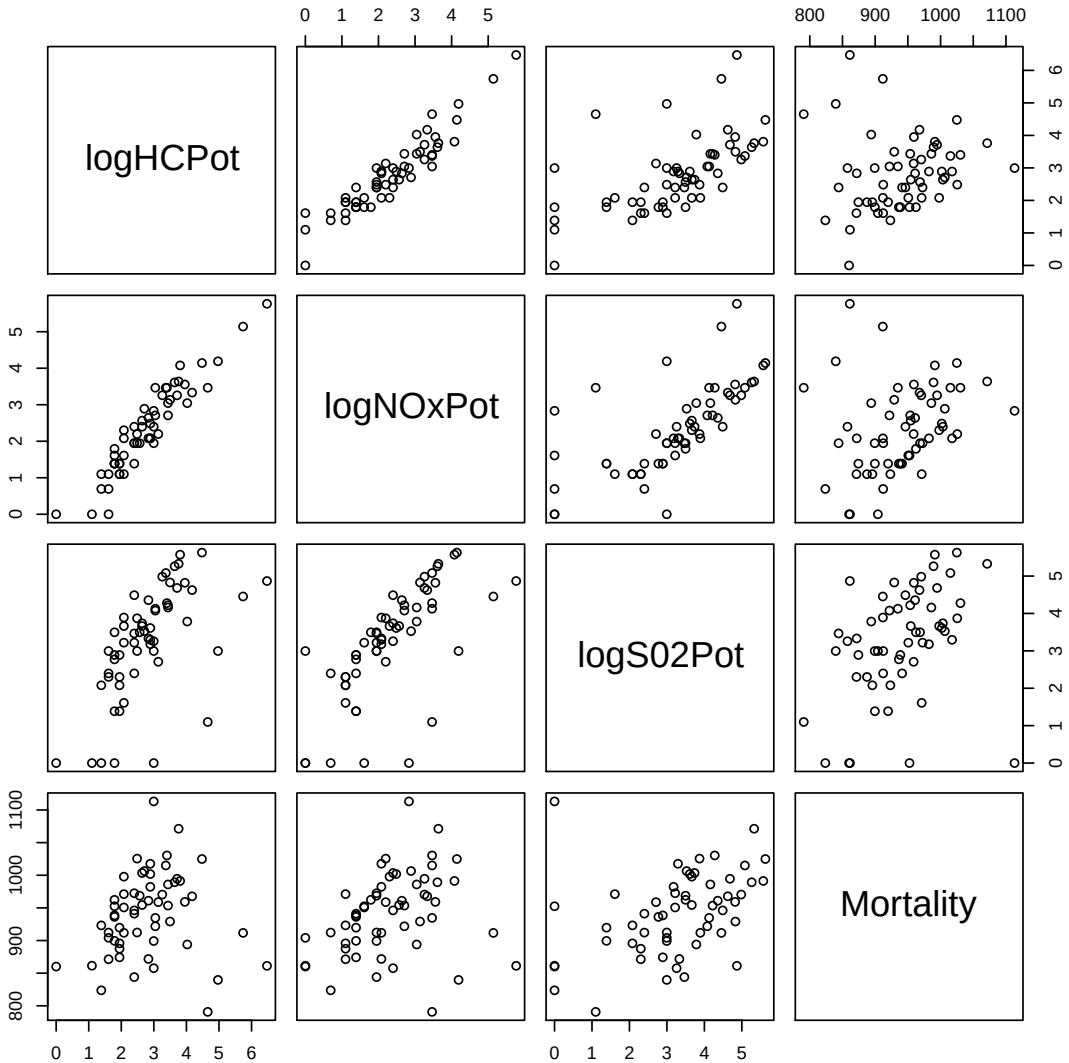
Weather variables vs. mortality



Sociological variables vs. mortality



Pollution variables vs. mortality



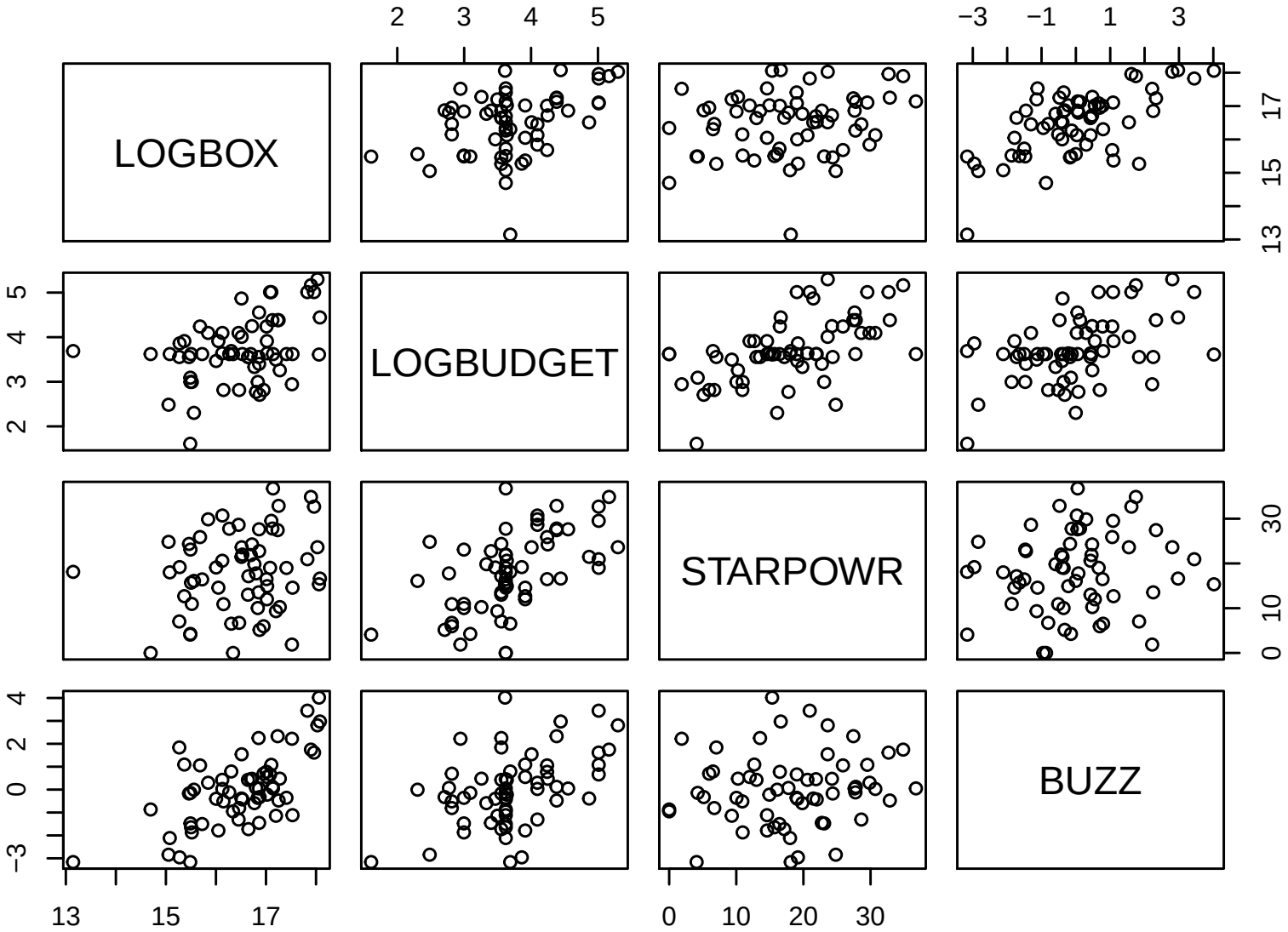
Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

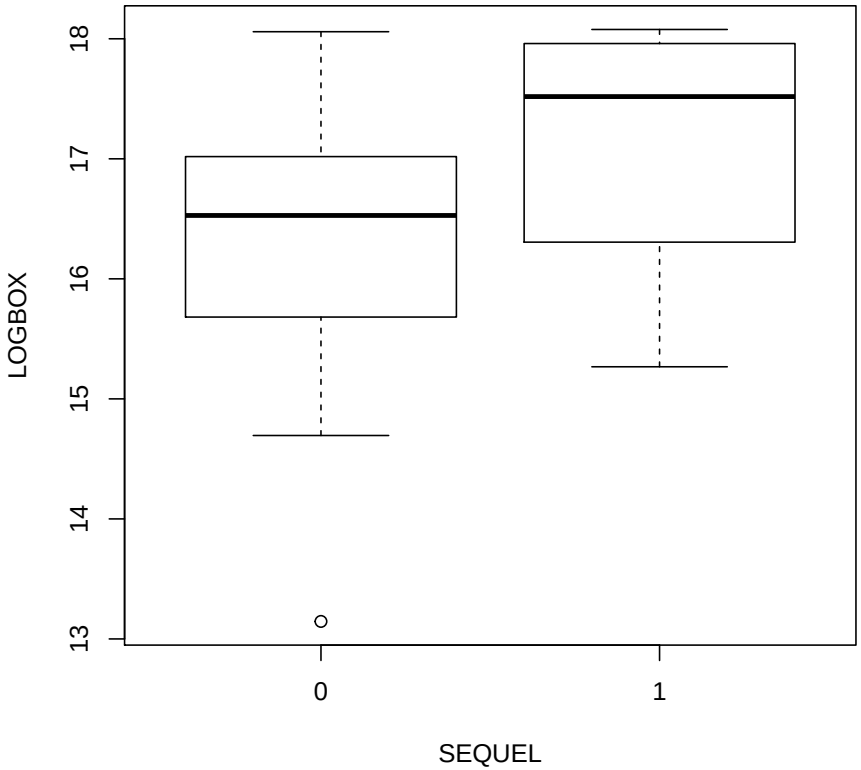
Movie Box Office data

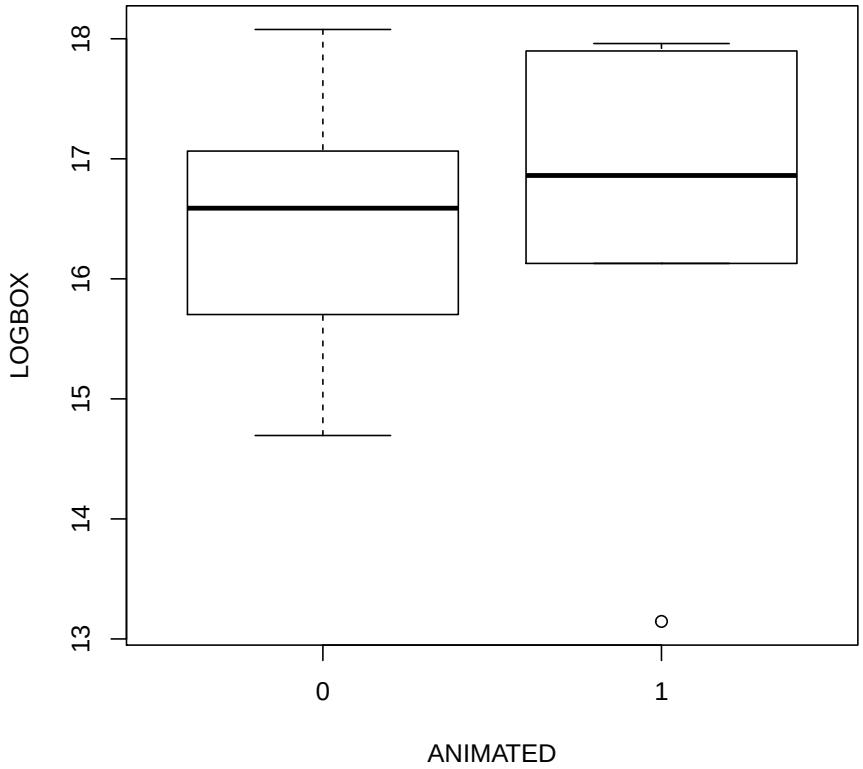
- Questions: Which factors influence the commercial success of a movie? Can we predict the box-office receipt before the movie has been released?
- Dataset about 62 movies released in 2009 (from Greene, 2012)
- Dependent variable: logarithm of Box Office receipts
- 11 input variables:
  - 3 binary variables (G, PG, PG13) to encode the MPAA (Motion Picture Association of America) rating, logarithm of budget (LOGBUDGET), star power (STARPOWR),
  - 1 binary variable to indicate if the movie is a sequel (SEQUEL),
  - 4 binary variables to describe the genre (ACTION, COMEDY, ANIMATED, HORROR)
  - 1 variable to represent internet buzz (BUZZ)

Box Office data



Box Office data



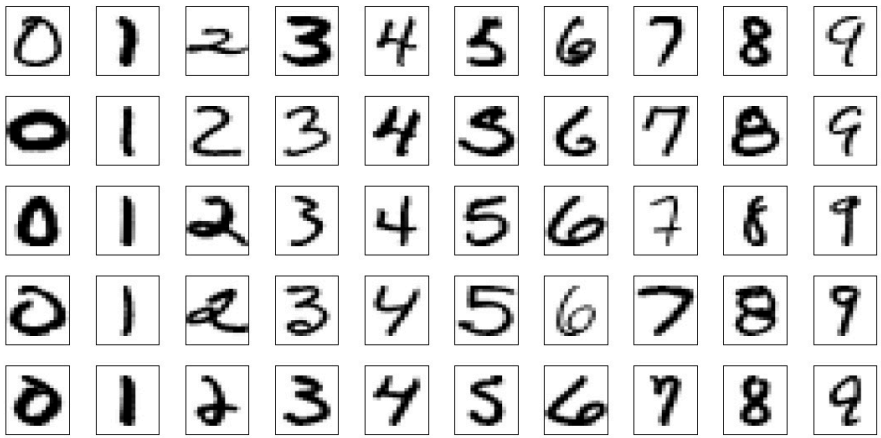


Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

Handwritten ZIP code

- Problem: read handwritten ZIP codes on envelopes from U.S. postal mail.



- The images are 16×16 eight-bit grayscale maps, with each pixel ranging in intensity from 0 to 255.
- The task is to predict, from the matrix of pixel intensities, the identity of each image $(0, 1, \cdots, 9)$ quickly and accurately.
- In this task, it is also import to assess the uncertainty, so as to defer decision-making when the uncertainty is too high.

Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

Spam detection

- Goal: build a customized spam filter.
- Data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as spam or email.
- Input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

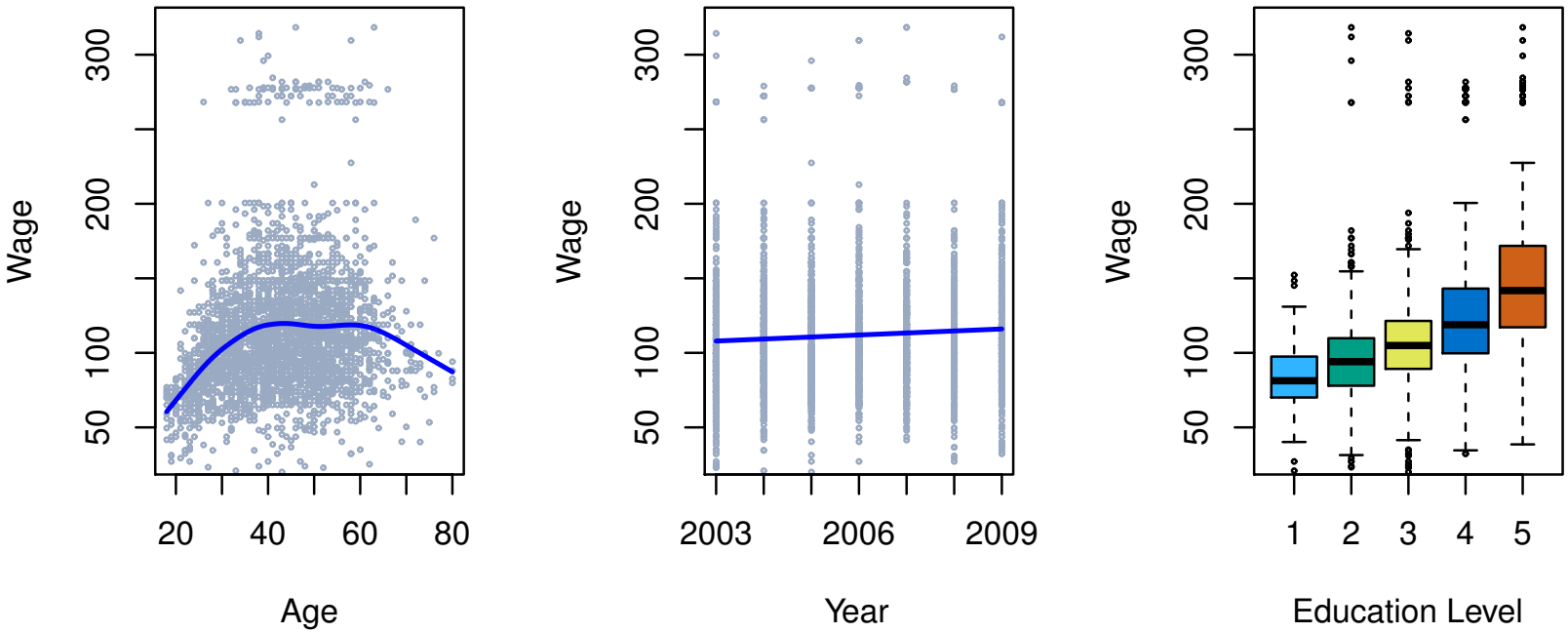
Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

Factors influencing wages

- Which factors influence wages? Are observations consistent with economic theories?



Income survey data for males from the central Atlantic region of the USA

Examples of learning problems

- Study the relation between air pollution on mortality.
- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.

Expression recognition

joy

surprise

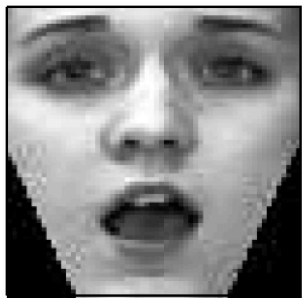
sadness

disgust

anger

fear

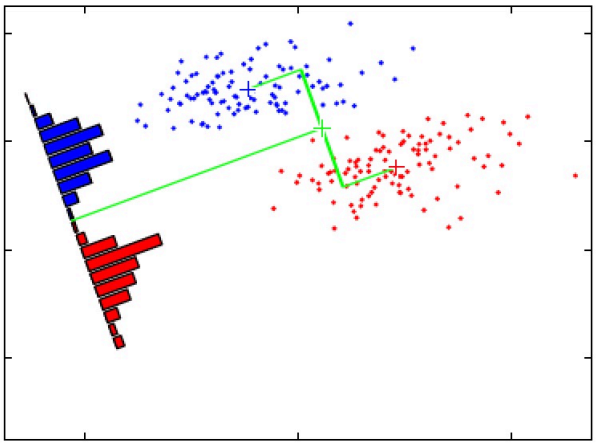
Learning



Projection in a 5D subspace (LDA)

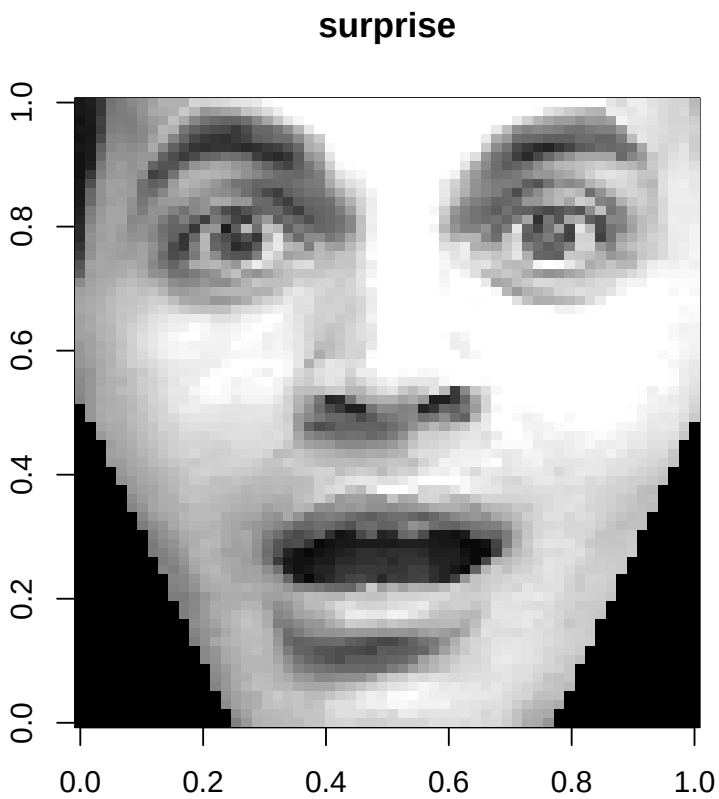
Logistic Regression

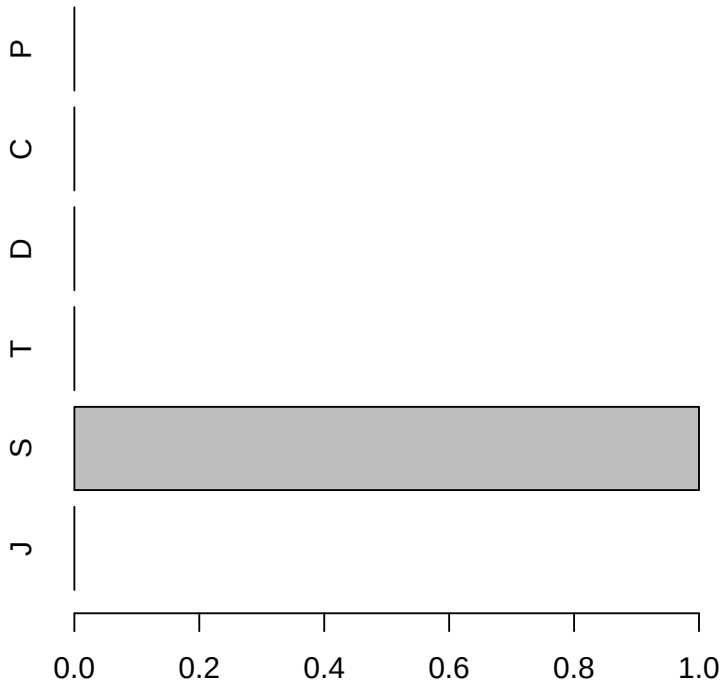
decision



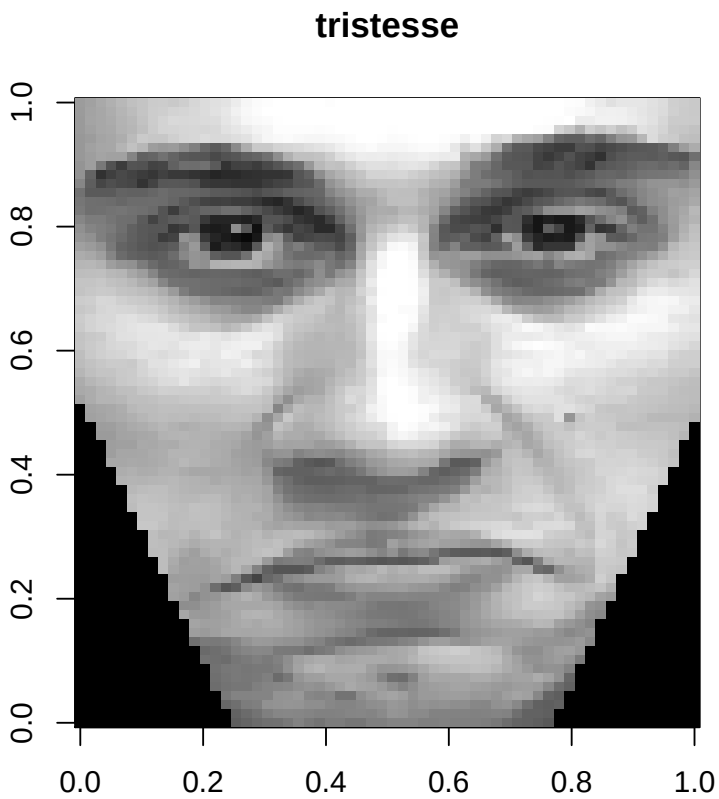
- 216 images 70×60 (36 per expression)
- 144 for learning, 72 for testing
- 5 features extracted by linear discriminant analysis
- test error rate: 23.6% (random: 83.3%)

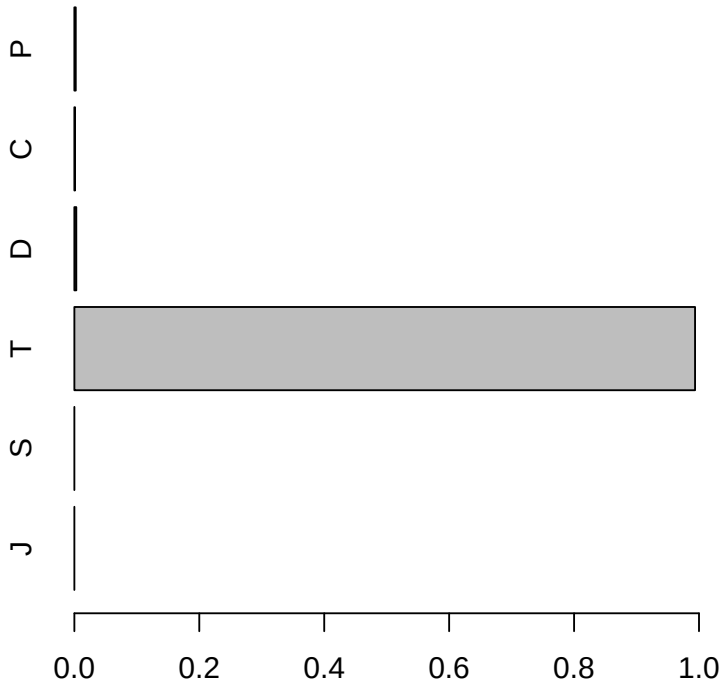
Results



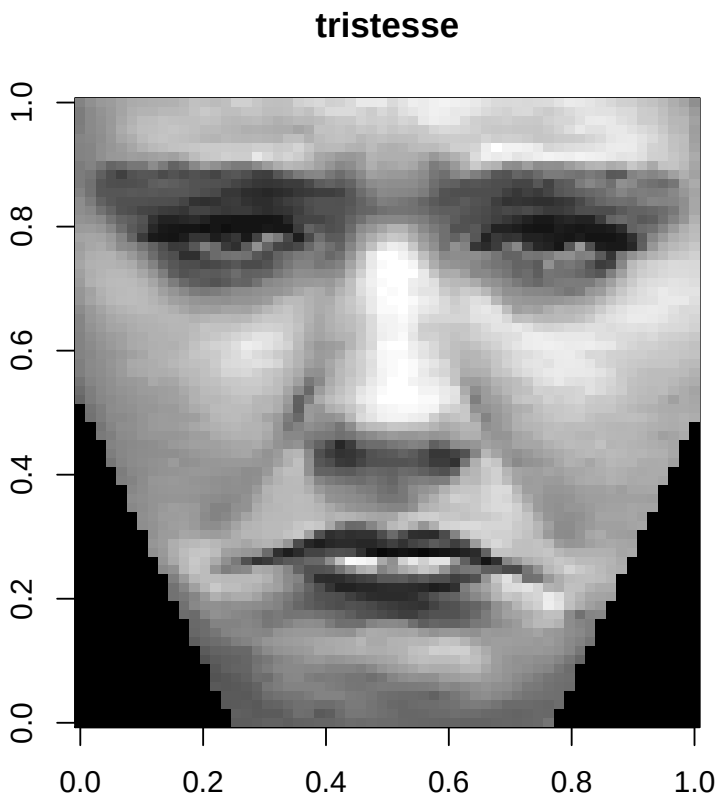


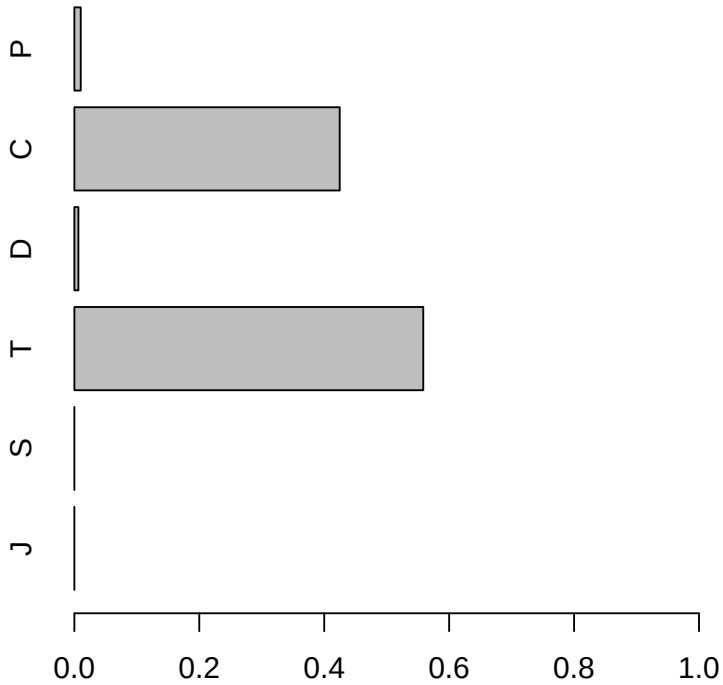
Results



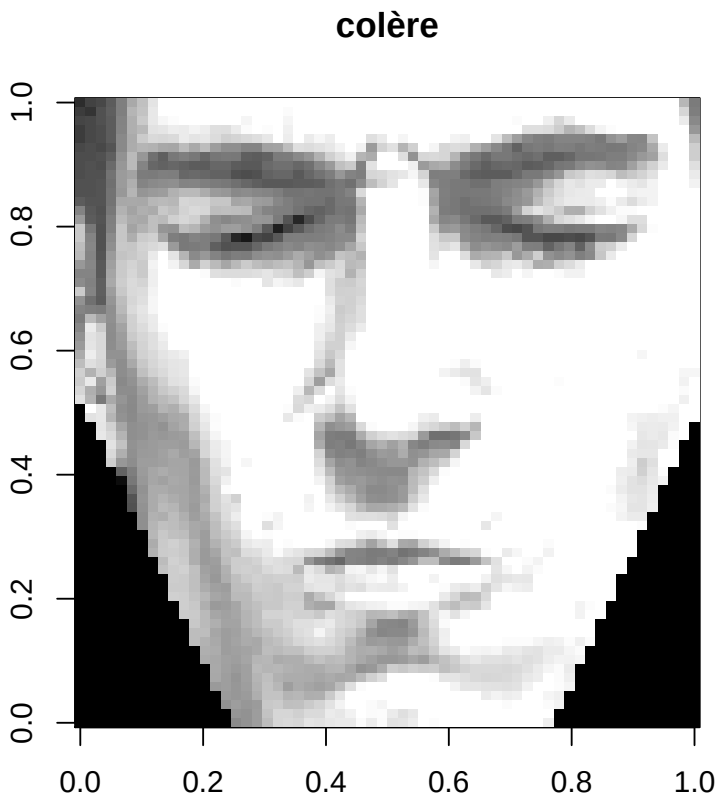


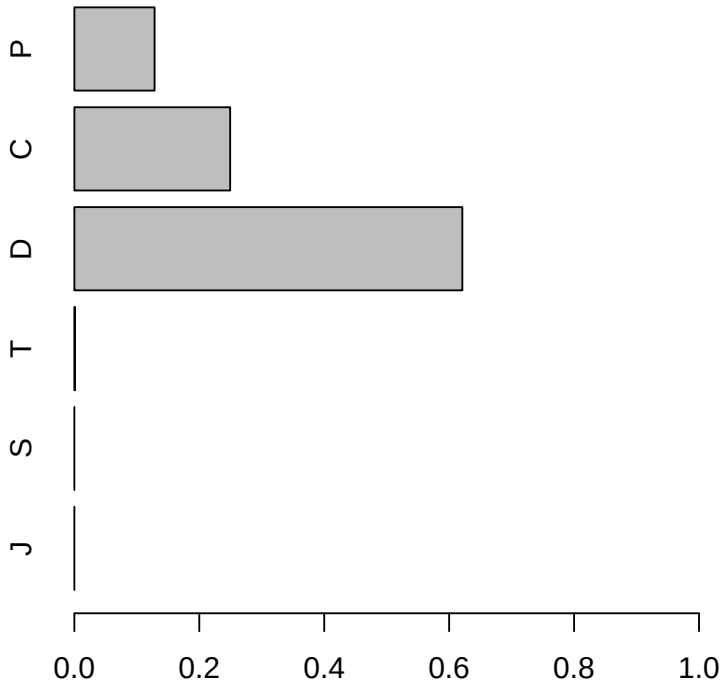
Results



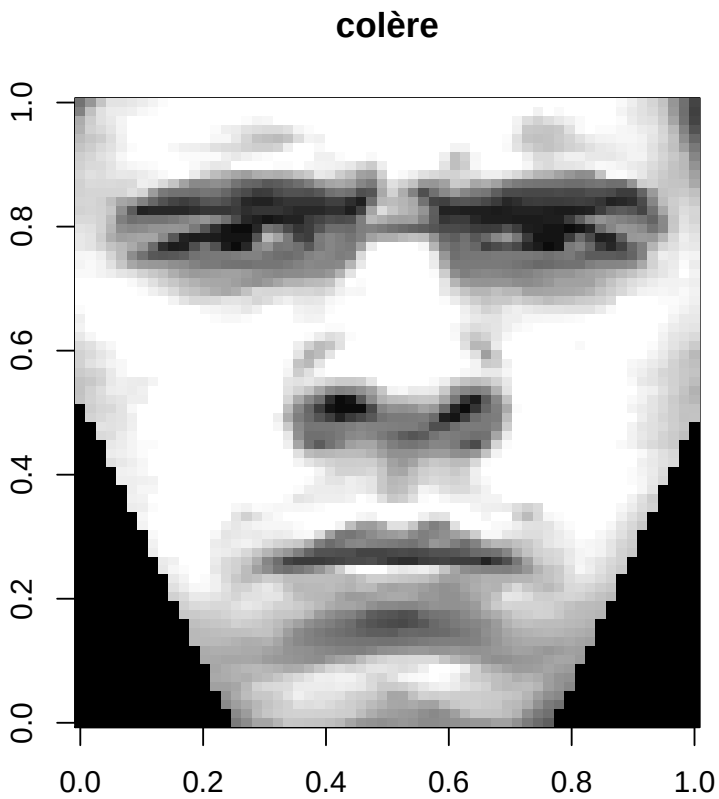


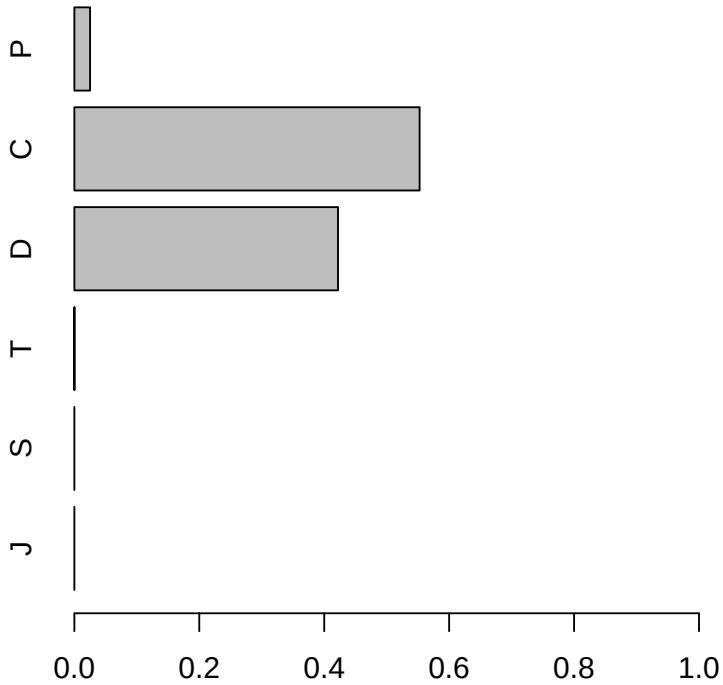
Results



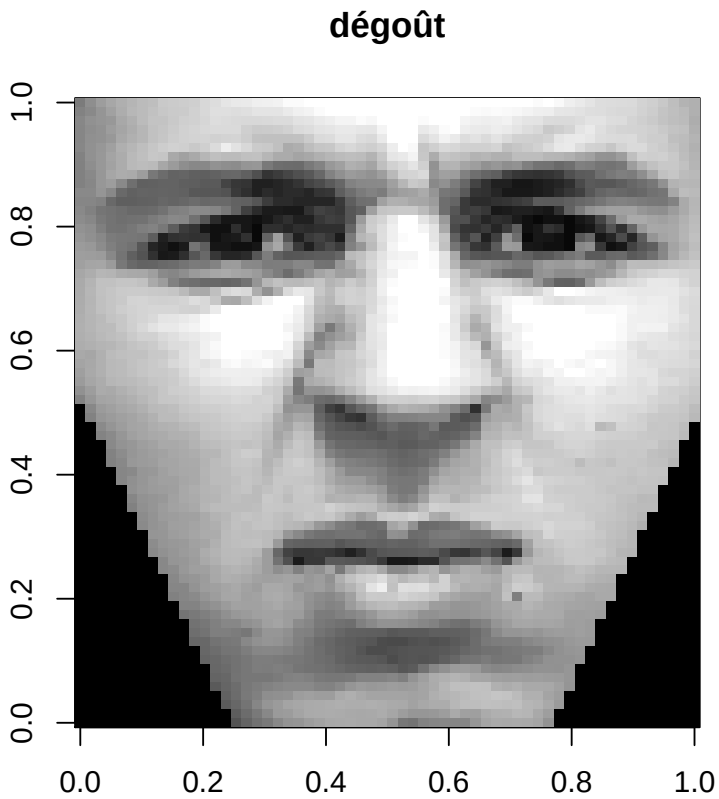


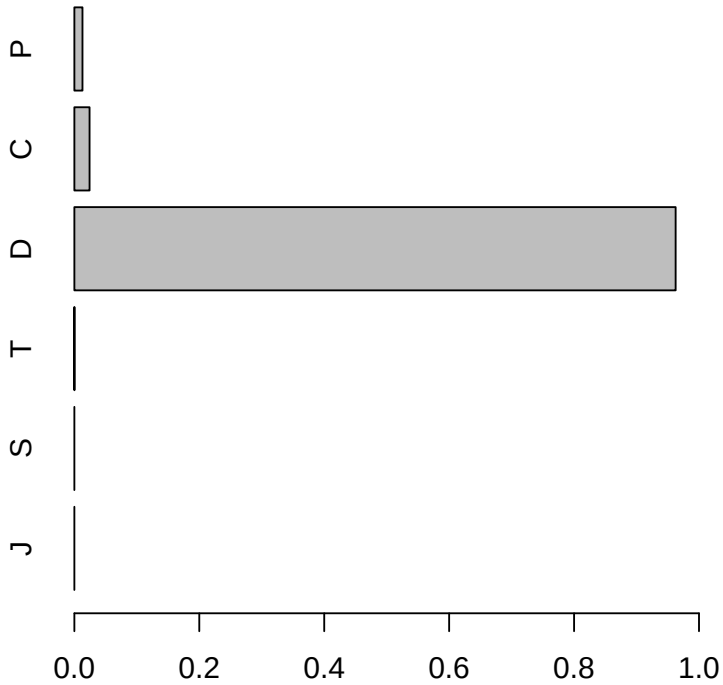
Results



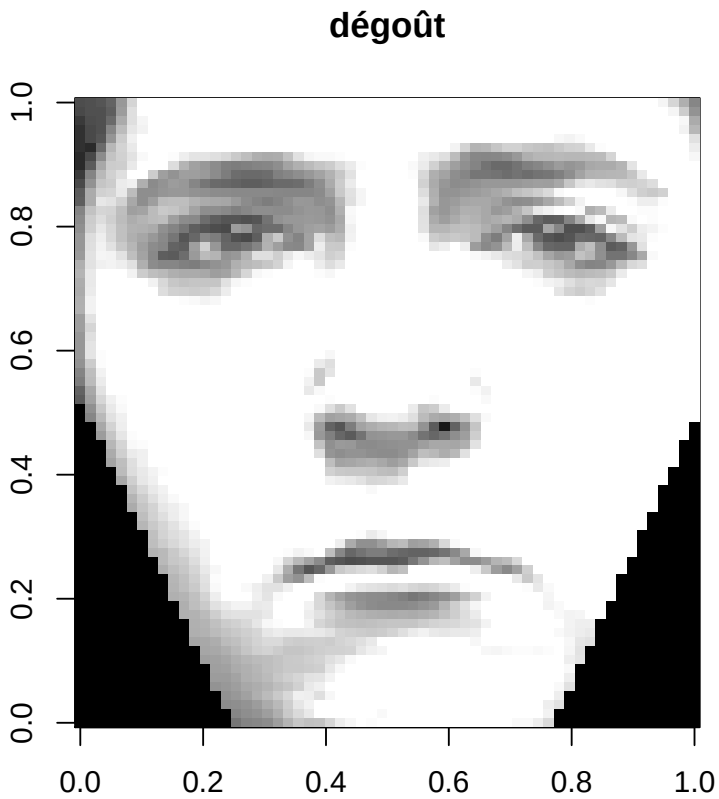


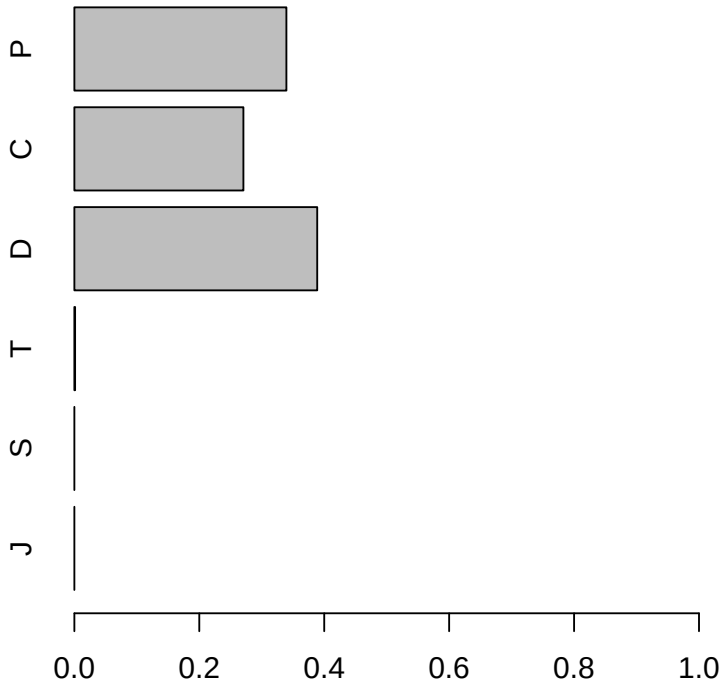
Results



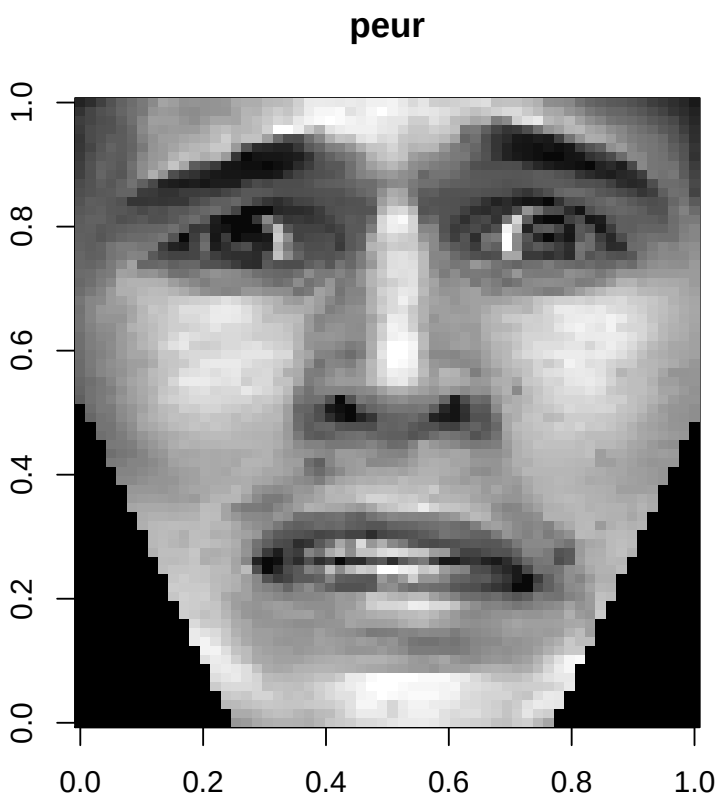


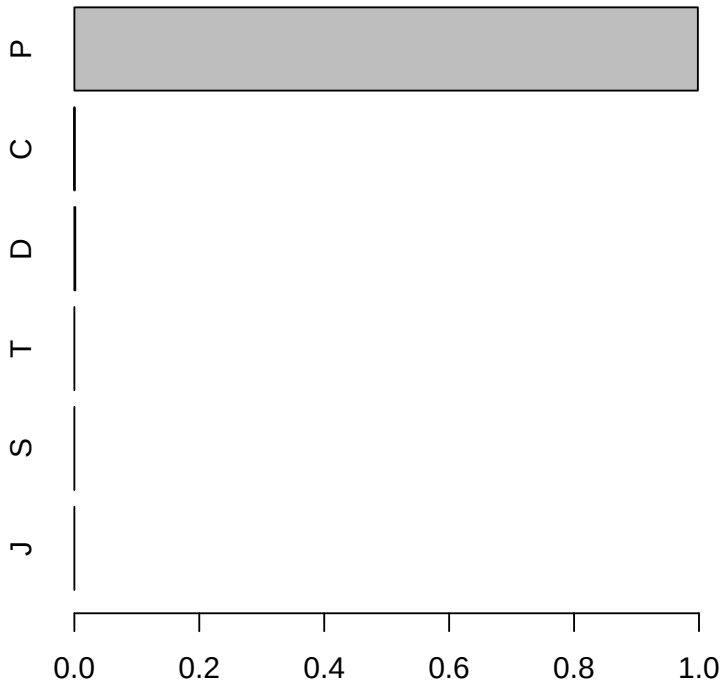
Results



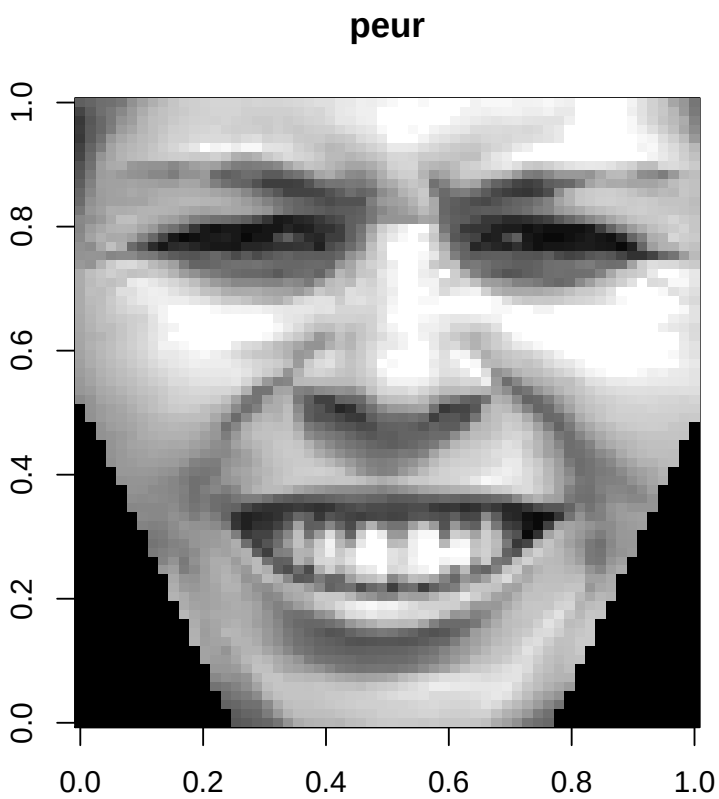


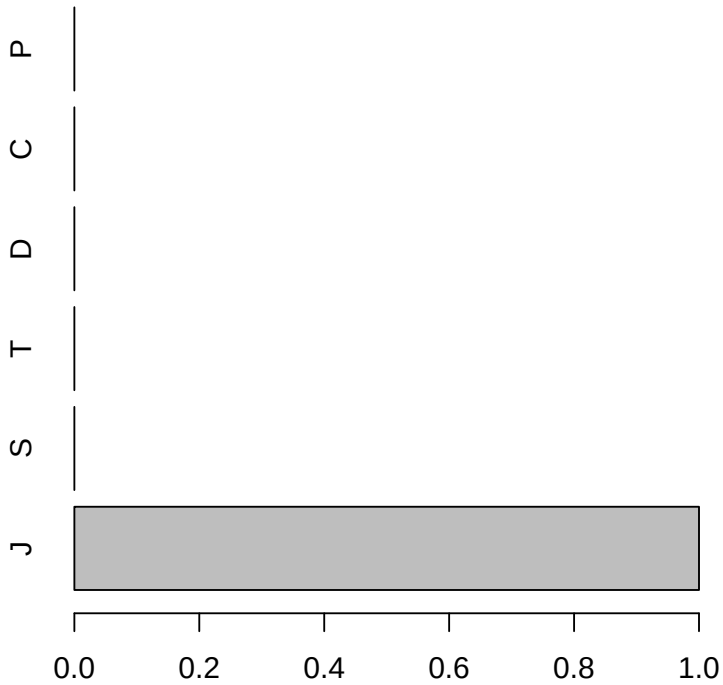
Results





Results





Overview

Supervised learning

- Starting point:
  - Outcome measurement Y (also called dependent variable, response, output, target).
  - Vector of p predictor measurements X (also called inputs, covariates, features, attributes, explanatory variables).
- In regression problems, Y is quantitative (e.g., price, blood pressure).
- In classification problems, Y is categorical: it takes values in a finite, unordered set $\mathcal{C}$ (survived/died, digit 0-9, facial expression, etc.).
- We have training data $(x_1, y_1), \ldots, (x_n, y_n)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

1. Accurately predict unseen test cases.
2. Understand which inputs affect the outcome, and how.
3. Assess the quality of our predictions and inferences.

Unsupervised learning

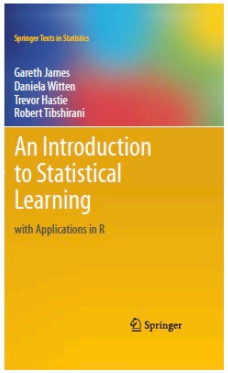
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy:
  - Find groups of samples that behave similarly (clustering)
  - Find linear combinations of features with the most variation (principal component analysis, PCA), etc.
- Sometimes difficult to know how well your are doing.
- Can be useful as a pre-processing step for supervised learning (feature extraction, etc.)

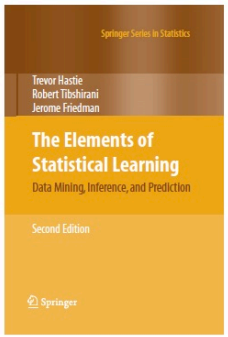
Course outline

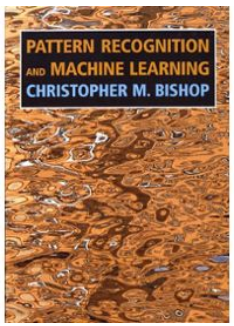
Topics to be covered (or a subset thereof):

1. Linear classification
2. Model selection
3. Splines
4. Gaussian mixture models
5. Tree-based methods
6. Support Vector Machines
7. Support Vector Regression and KPCA
8. Relevance Vector Machines
9. Neural networks

Course texts



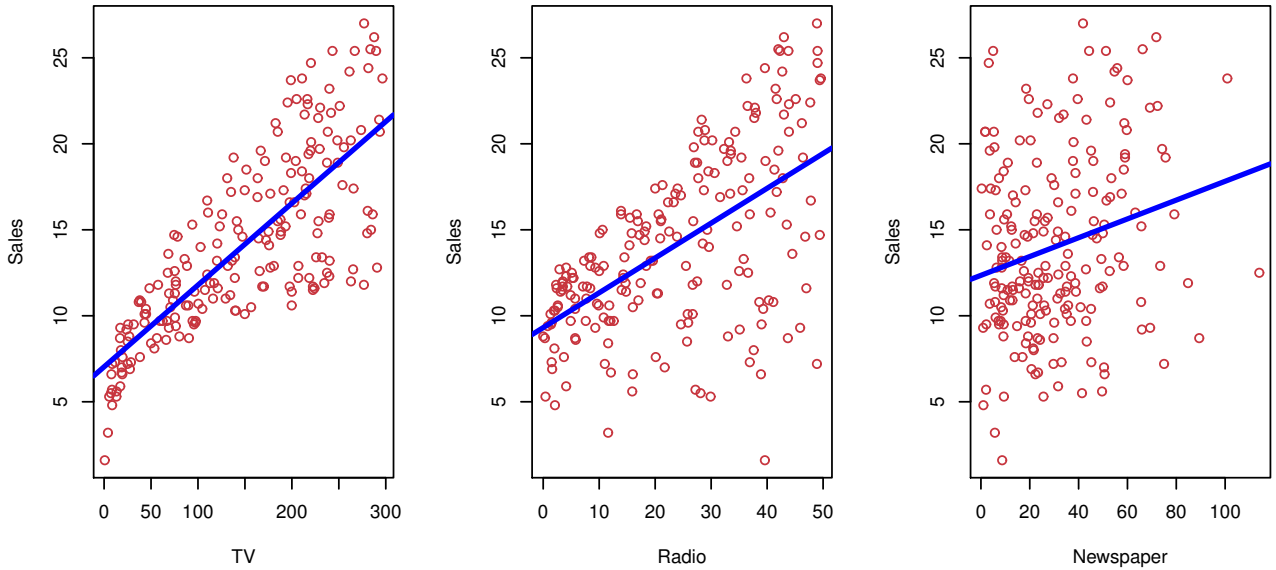




1. "An Introduction to Statistical Learning" (ISLR): emphasis on basic principles and application, no mathematical details. Available at `http://www-bcf.usc.edu/~gareth/ISL`
2. "The Elements of Statistical Learning" (ESL): more mathematically advanced and theoretical. Available at `http://statweb.stanford.edu/~tibs/ElemStatLearn`
3. "Pattern Recognition and Machine Learning" (PRML): same level as ESL, covers some other topics. Available at BUTC.

Overview

Introductory example



- Shown are Sales (in thousands of units) vs TV, Radio and Newspaper advertising expenses (in thousands of \$) for 200 markets, with a blue linear-regression line fit separately to each.
- Can we predict Sales using any single predictor?
- Perhaps we can do better using a model

$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Notation

- Here 'Sales' is a response that we wish to predict. We generically refer to the response as Y.
- 'TV' is an input, or predictor; we name it X_1. Likewise name 'Radio' as X_2, and so on.
- We can refer to the input vector collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

- Now we write our model as

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

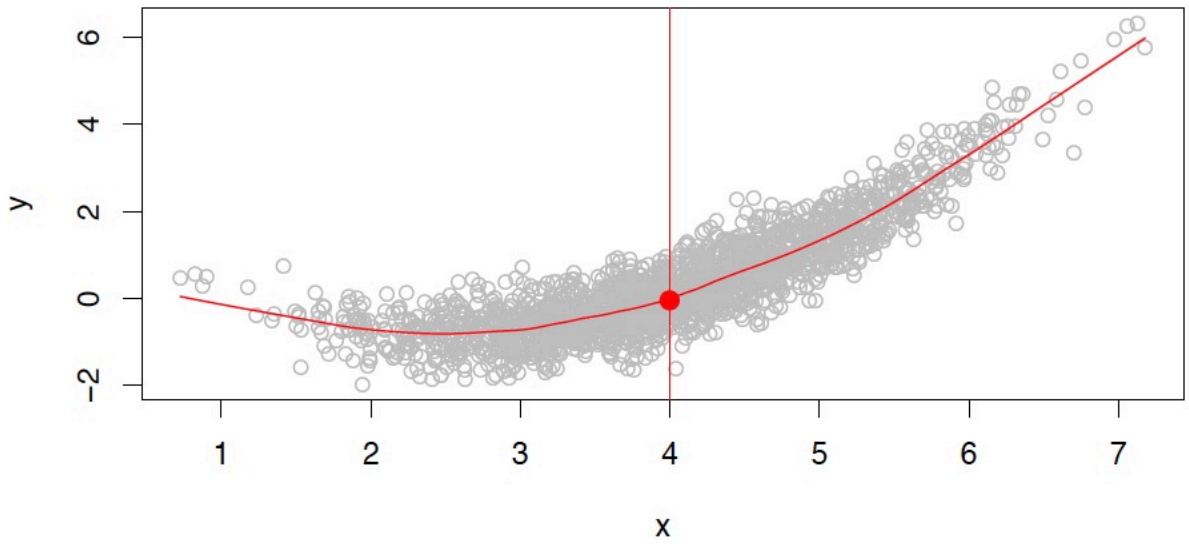
where ϵ captures measurement errors and other discrepancies.

What is $f(X)$ good for?

- With a good f we can make predictions of Y at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.

Overview

Regression function



- Is there an ideal $f(X)$? In particular, what is a good value for $f(X)$ at any selected value of X, say $X = 4$?
- There can be many Y values at $X = 4$. A good value is

$$f(4) = \mathbb{E}(Y|X = 4)$$

  where $\mathbb{E}(Y|X = 4)$ is the expected value of Y given $X = 4$.
- Function $f(x) = \mathbb{E}(Y|X = x)$ is called the regression function.

Justification of the regression function

- Assume that we predict Y at $X = x$ by some value $g(x)$, and the quality of the prediction is measured by the squared error $(y - g(x))^2$.
- We want to find the best function g, i.e., the function g that minimizes the mean squared error

$$MSE(g) = \mathbb{E}[(Y - g(X))^2].$$

- We can write

$$\mathbb{E}\left[(Y - g(X))^2 \mid X = x\right] = \mathrm{Var}(Y|X = x) + (f(x) - g(x))^2$$

- The regression function f minimizes $\mathbb{E}[(Y - g(X))^2 \mid X = x]$ for all x: consequently, it minimizes $MSE(g)$. It is the best possible prediction function (according to the MSE).

Reducible vs. irreducible error

- In practice, we never know the true f, but we can estimate it by some function $\widehat{f}$.
- The MSE at $X = x$ is then

$$\mathbb{E}\left[(Y - \widehat{f}(X))^2 \mid X = x\right] = \underbrace{(f(x) - \widehat{f}(x))^2}_{\text{reducible}} + \underbrace{\text{Var}(\epsilon \mid X = x)}_{\text{irreducible}}$$

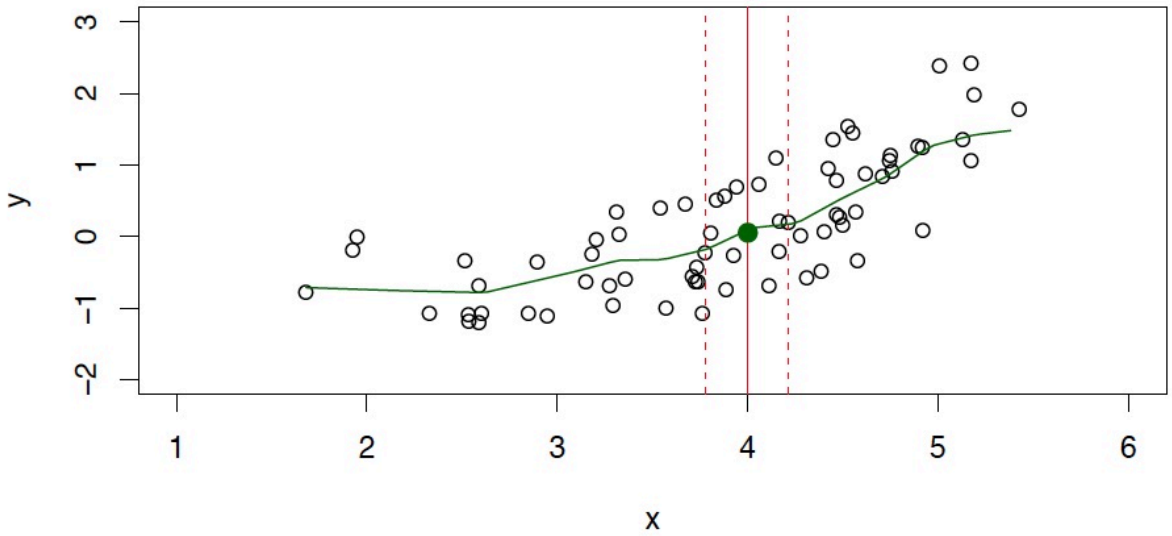
- Even if we knew $f(x)$, we would still make prediction errors, because of the second term $\text{Var}(\epsilon|X = x)$, which cannot be reduced.
- A learning method will try to minimize the reducible component $(f(x) - \widehat{f}(x))^2$ of the error.

Overview

How to estimate f

- Typically we have few if any data points with $X = 4$ exactly. So, we cannot compute $\mathbb{E}(Y|X = x)$!
- However, we can compute the mean value of Y in a neighborhood $\mathcal{N}(x)$ of x:

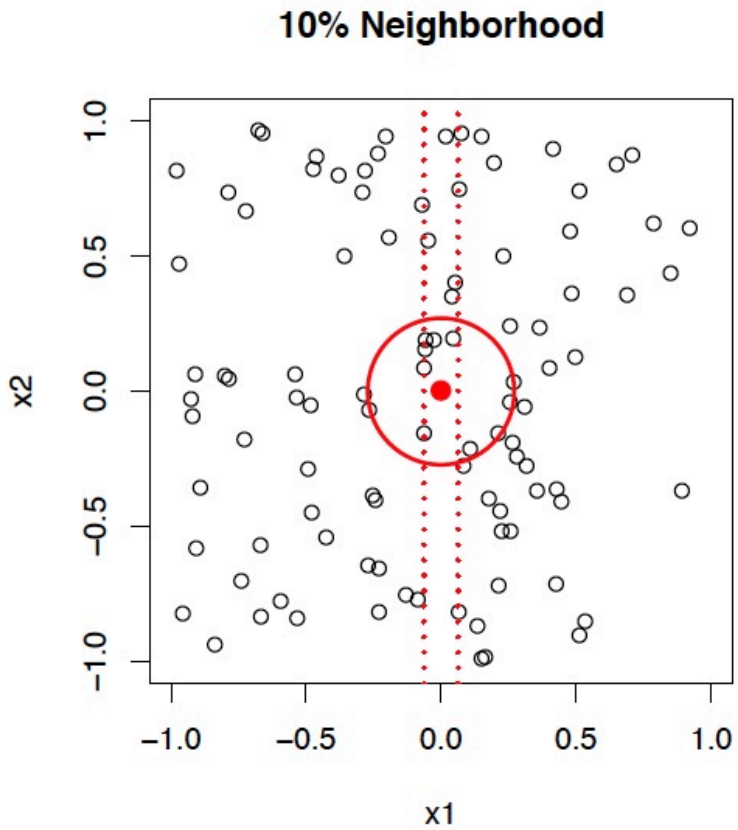
$$\widehat{f}(x) = \text{Ave}\,(Y \mid X \in \mathcal{N}(x))$$

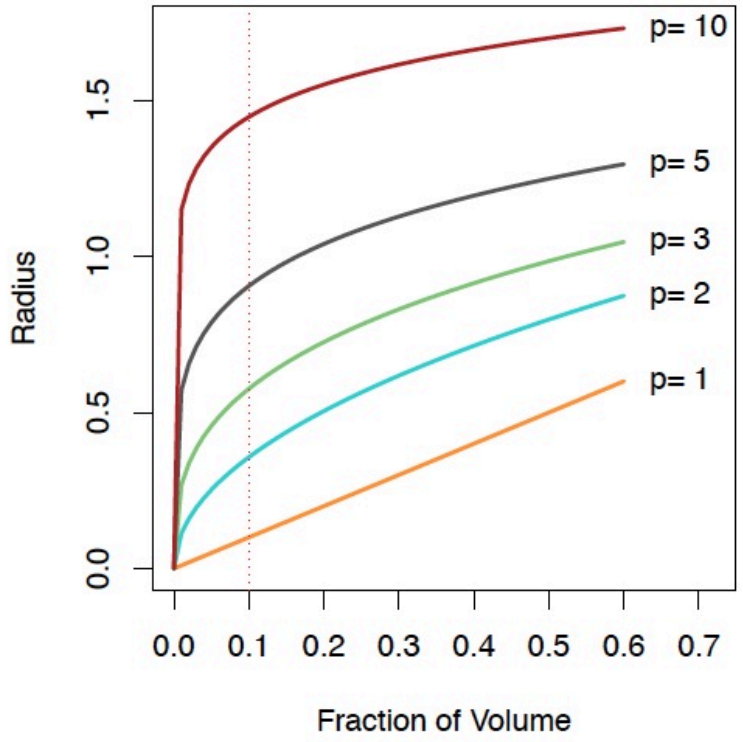


Nearest neighbor averaging

- The neighborhood $\mathcal{N}(x)$ can be defined as the region containing the K nearest neighbors of x in the training data.
- Nearest neighbor averaging can be pretty good for small p – i.e., $p \leq 4$ and n not too small.
- We will discuss smoother versions, such as spline smoothing later in the course.
- Nearest neighbor methods can perform badly when p is large. Reason: the curse of dimensionality. Nearest neighbors tend to be far away in high dimensions:
  - We need to get a reasonable fraction of the n values of y_i to average to bring the variance down – e.g. 10%.
  - A 10% neighborhood in high dimensions may no longer be local, so we lose the spirit of estimating $\mathbb{E}(Y|X = x)$ by local averaging.

The curse of dimensionality





Overview

Parametric and structured models

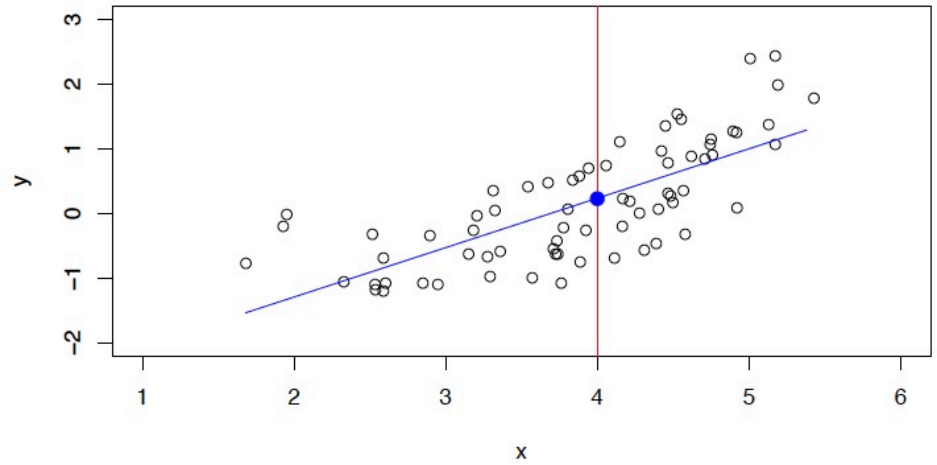
The linear model is an important example of a parametric model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

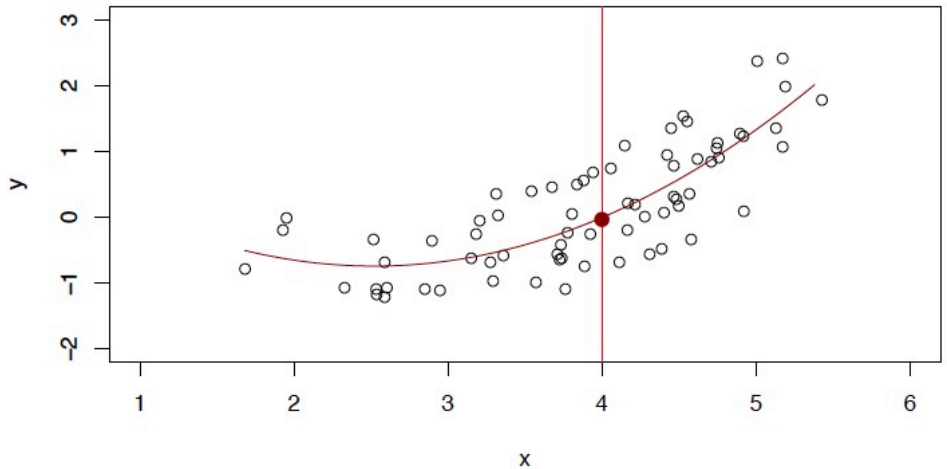
- A linear model is specified in terms of a vector of $p + 1$ parameters $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$.
- We estimate the parameters by fitting the model to training data.
- Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

Linear vs. quadratic

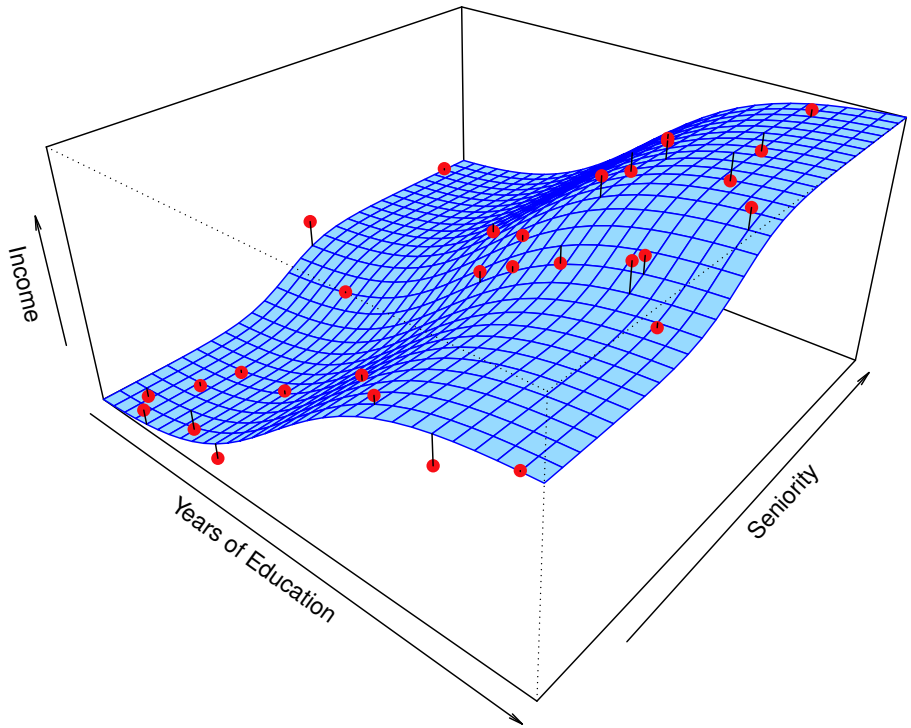
A linear model $\widehat{f}_L(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$ gives a reasonable fit here:



A quadratic model $\widehat{f}_Q(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X + \widehat{\beta}_2 X^2$ fits slightly better:



Simulated example

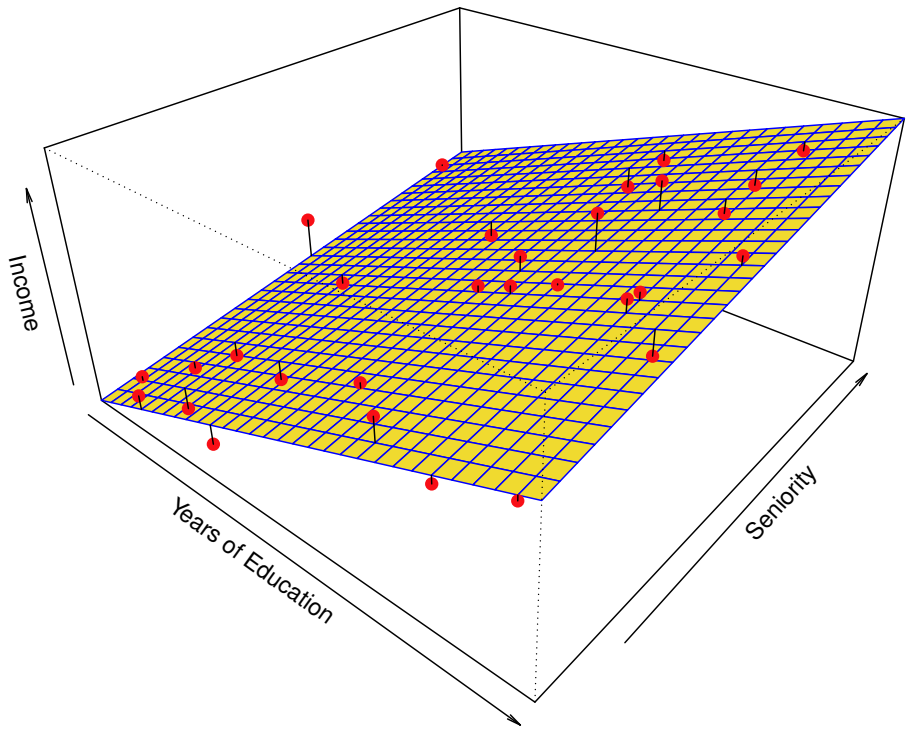


Red points are simulated values for income from the model

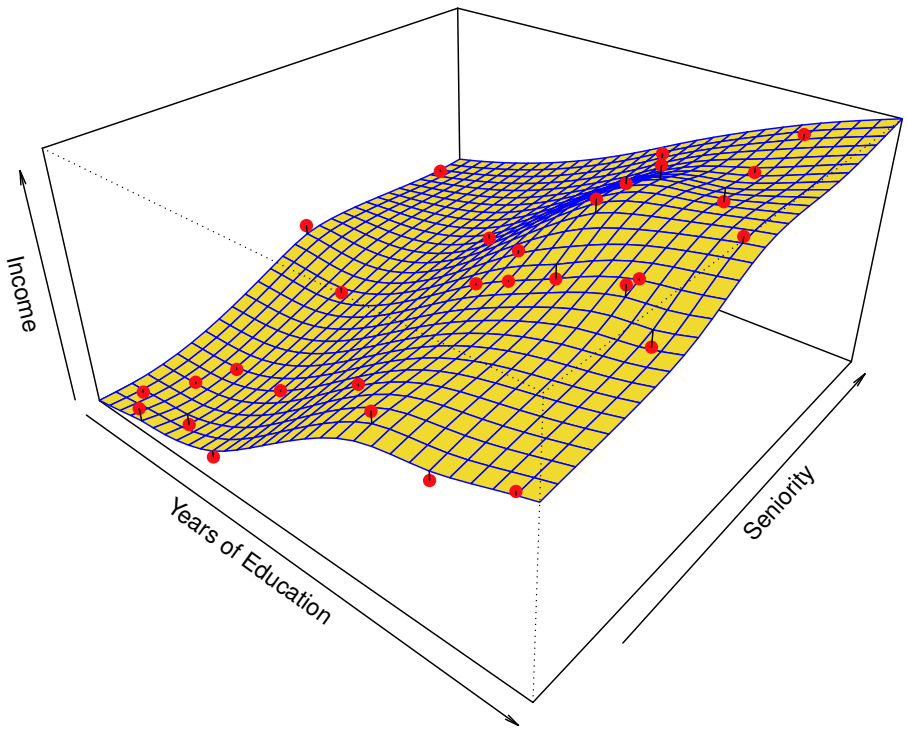
$$\text{income} = f(\text{education}, \text{seniority}) + \epsilon$$

f is the blue surface.

Linear regression model fit

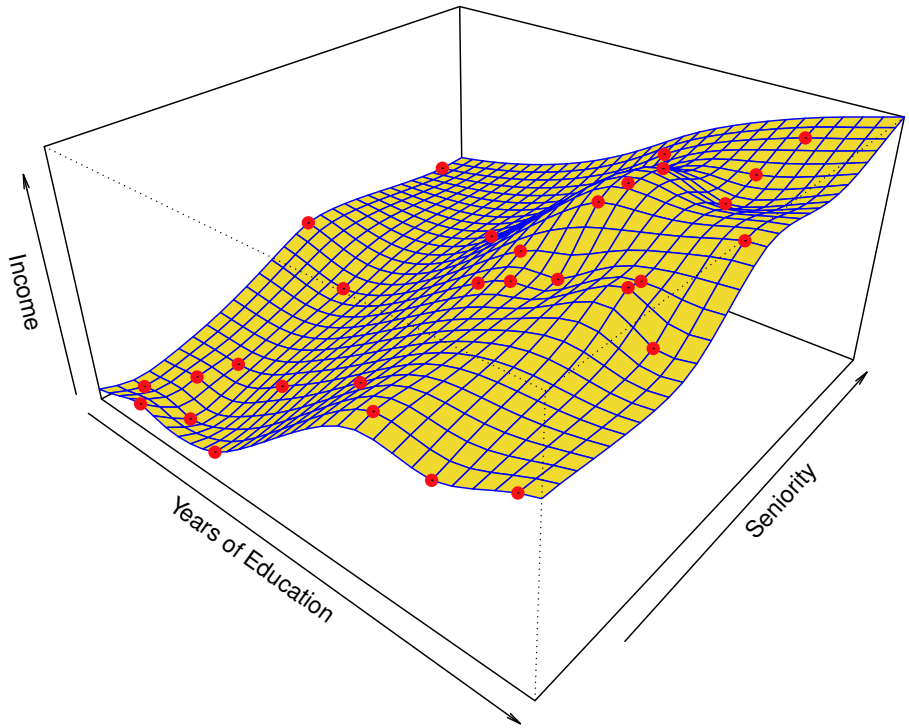


More flexible regression model



More flexible regression model $\widehat{f}_S(\text{education}, \text{seniority})$ fit to the simulated data. Here we used a technique called a thin-plate spline to fit a flexible surface.

Even more flexible spline regression model

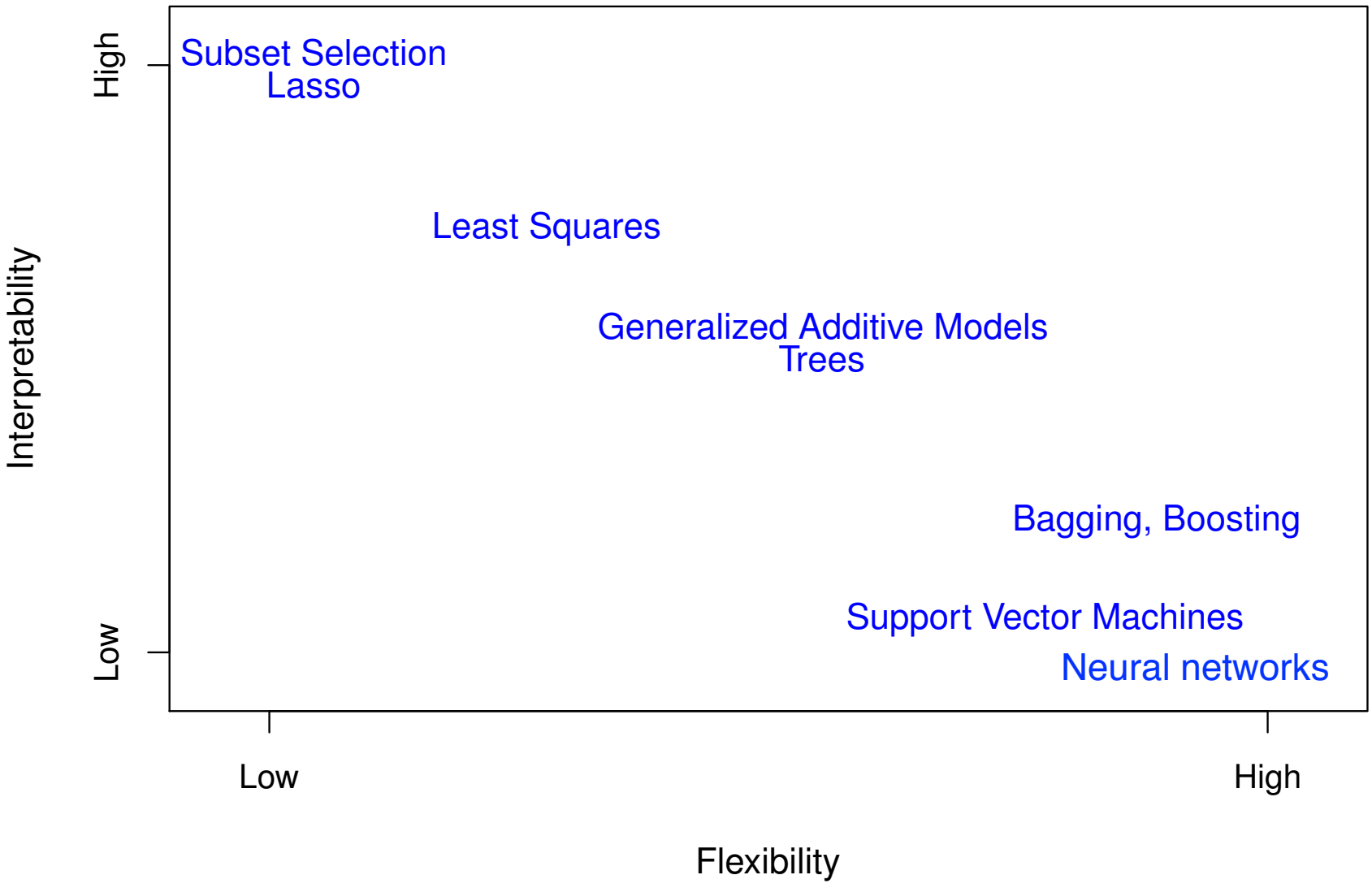


Even more flexible spline regression model fit to the simulated data. Here the fitted model makes no errors on the training data! Also known as overfitting.

Some trade-offs

- Prediction accuracy versus interpretability: linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit: how do we know when the fit is just right?
- Parsimony versus black-box: we often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Interpretability vs. flexibility



Overview

Assessing Model Accuracy

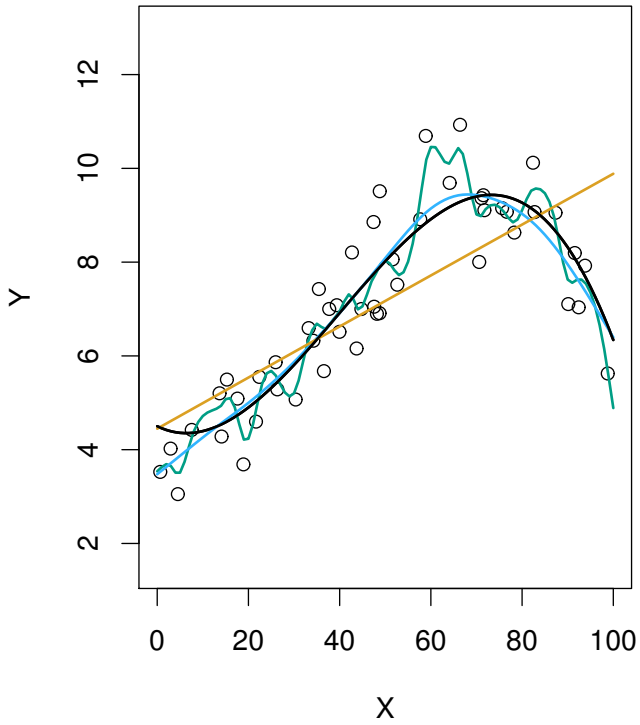
- Suppose we fit a model $f(x)$ to some learning data $\mathcal{L} = \{x_i, y_i\}_{i=1}^n$, and we wish to see how well it performs.
- We could compute the average squared prediction error over $\mathcal{L}$:

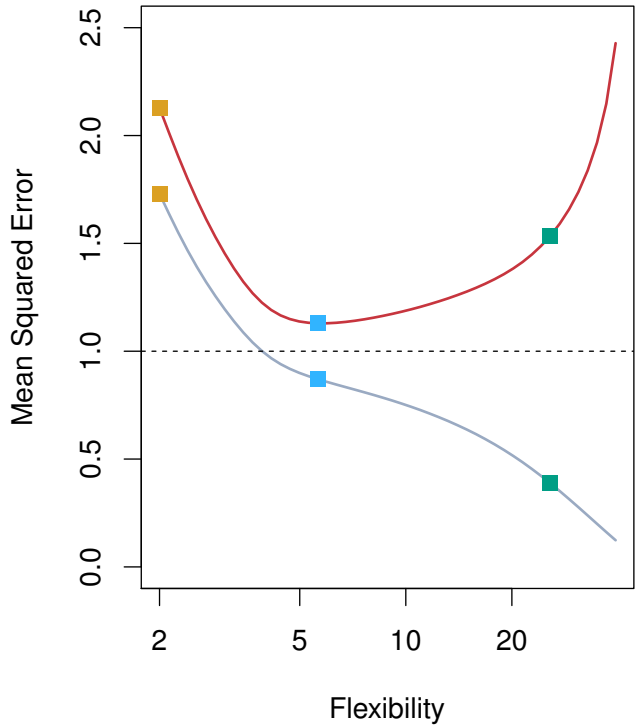
$$\text{MSE}_{\mathcal{L}} = \text{Ave}_{i \in \mathcal{L}} \left[y_i - \widehat{f}(x_i) \right]^2$$

- This may be biased toward more overfit models.
- Instead we should, if possible, compute it using fresh test data $\mathcal{T} = \{x_i, y_i\}_{i=1}^m$:

$$\text{MSE}_{\mathcal{T}} = \text{Ave}_{i \in \mathcal{T}} \left[y_i - \widehat{f}(x_i) \right]^2$$

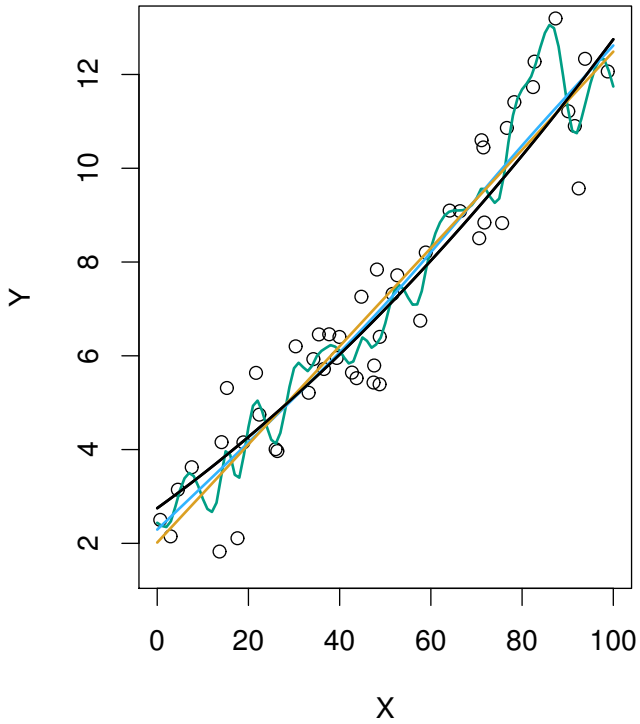
Case 1

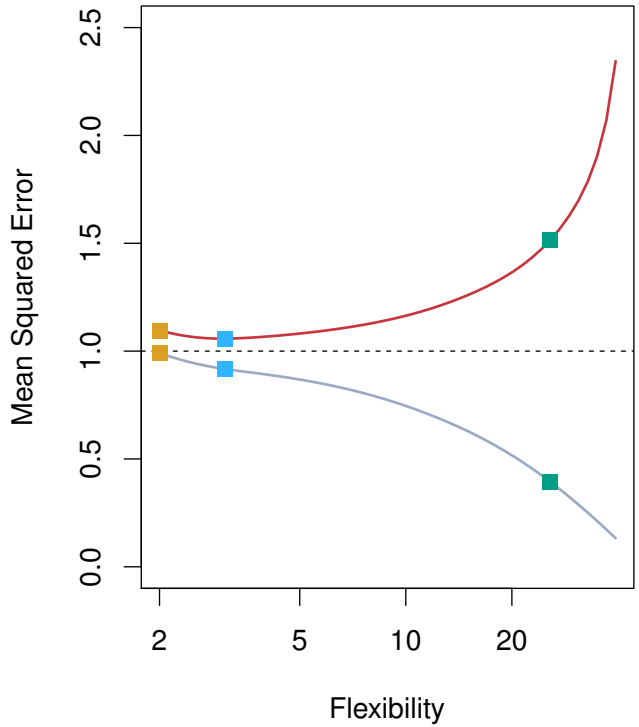




Black curve is truth. Red curve on right is $MSE_{\mathcal{T}}$, grey curve is $MSE_{\mathcal{L}}$. Orange, blue and green curves/squares correspond to fits of different flexibility.

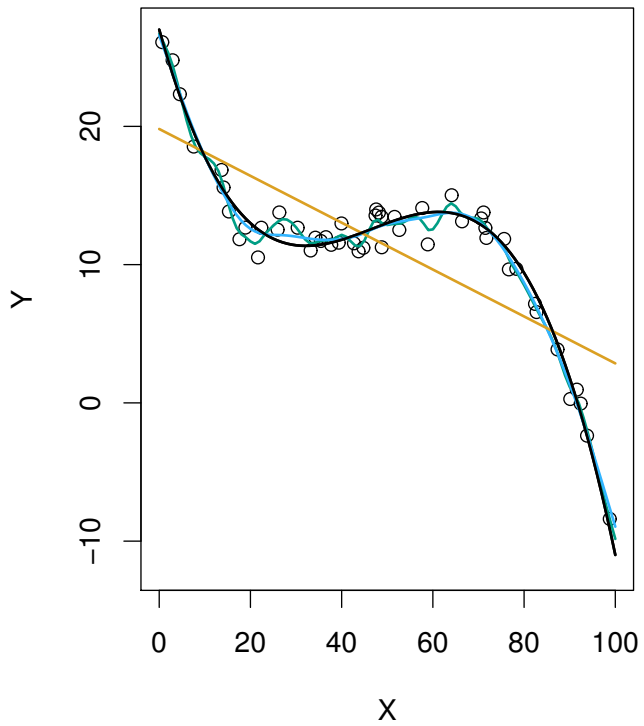
Case 2

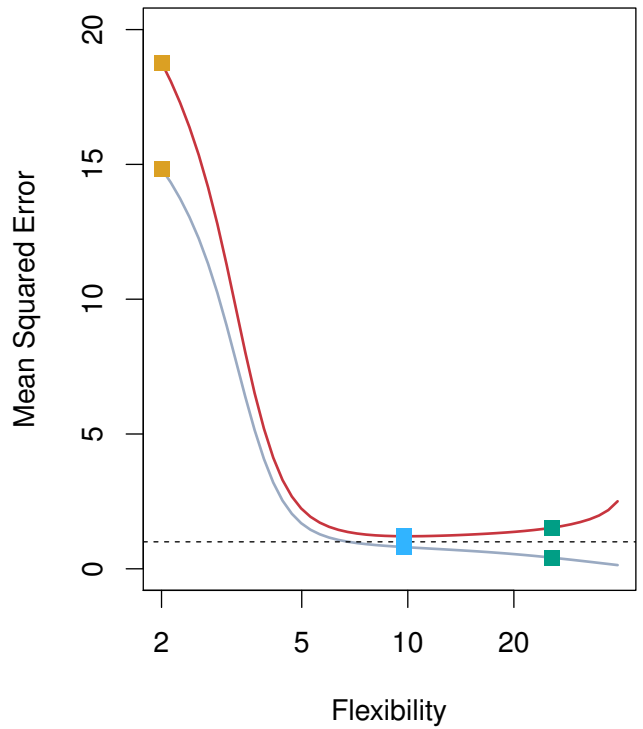




Here the truth is smoother, so the smoother fit and linear model do really well.

Case 3





Here the truth is wiggly and the noise is low, so the more flexible fits do the best.

Bias-variance decomposition

- Suppose we have fit a model $\widehat{f}(x)$ to some learning data $\mathcal{L}$, and let (x_0, y_0) be a test observation drawn from the population.
- If the true model is $Y = f(X) + \epsilon$, with $f(x) = \mathbb{E}(Y|X = x)$, then the MSE for fixed x_0

$$\mathbb{E}\left[\left(Y - \widehat{f}(X)\right)^2 \mid X = x_0\right] =$$

$$\text{Var}(\widehat{f}(x_0)) + [\text{Bias}(\widehat{f}(x_0))]^2 + \text{Var}(\epsilon \mid X = x_0)$$

where $\text{Bias}(\widehat{f}(x_0)) = \mathbb{E}[\widehat{f}(x_0)] - f(x_0)$, and the expectation averages over the variability of Y as well as the variability in $\mathcal{L}$.

- Prove it! (Insert $\mathbb{E}[\widehat{f}(x_0)]$, then insert $f(x_0)$ in $\mathbb{E}[(Y - \mathbb{E}[\widehat{f}(x_0)])^2]$.)

Bias-variance trade-off

- Typically as the flexibility of $\widehat{f}$ increases, its variance increases, and its bias decreases.
- So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

Bias-variance trade-off for the three previous examples