ML01 – Introduction to Machine Learning

Lecture 1: Introduction

Thierry Denœux

tdenoeux@utc.fr
https://www.hds.utc.fr/~tdenoeux

Université de technologie de Compiègne

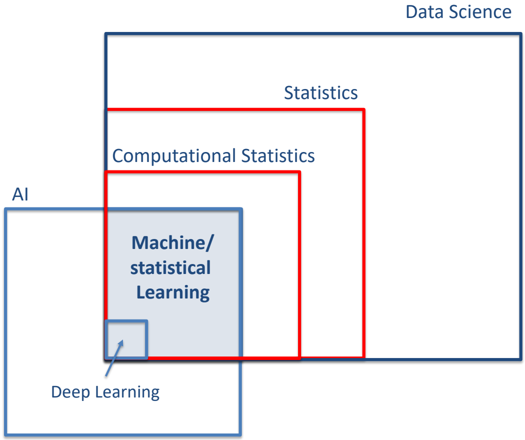
Spring 2021

Practical information

- Teachers:
  - Thierry Denœux (lectures)
  - Lunde Chen (practical classes)
- Time table:
  - Lectures on Tuesday (online via Voov, 14:10-16:45)
  - Practical classes (TD) on Thursday (C417, 14:10-16:45)
- Slides and other teaching material and course information will be posted each week.
- Evaluation:
  1. Course attendance: 10%
  2. Project: 30%
  3. Final exam (multiple-choice questionnaire): 60%

What is Machine Learning?

"A field of study that gives computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).



What is Machine Learning?

- Machine Learning (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)
  - Robotic perception, autonomous vehicles
  - Intelligent interfaces, etc.
- ML skills are in high demand by IT companies.

Objectives of this course

- Understand the basic principles of ML
- Get working knowledge of some elementary (but widely used) ML techniques:
  - Linear and quadratic classification
  - Linear regression
  - Model selection: variable selection, regularization
  - Decision trees
  - Neural networks
  - Clustering
  - Feature extraction and data visualization
- Master the R software environment for data analysis and ML

Overview

1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings

2. Some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - Regression: Bias-Variance trade-off

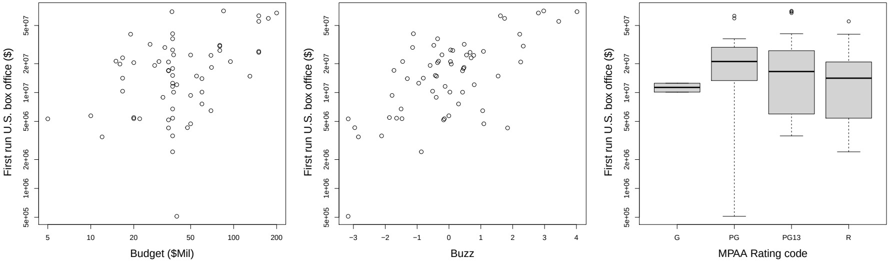
Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

Movie Box Office data

- Questions: Which factors influence the commercial success of a movie? Can we predict the box-office success before the movie has been released?
- Dataset about 62 movies released in 2009 (from *Econometric Analysis*, Greene, 2012)
- Response variable (to be predicted): Box Office receipts
- 11 predictors:
    - MPAA (Motion Picture Association of America) rating (G, PG, PG13)
    - Budget
    - Star power
    - Sequel (yes or no)
    - Genre (action, comedy, animated, horror)
    - Internet buzz

Box Office data

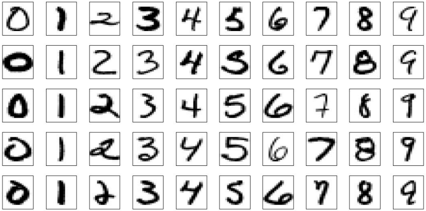


Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

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 recognize, from the matrix of pixel intensities, the digit in each image $(0, 1, \cdots, 9)$ quickly and accurately.

- Importance of uncertainty assessment (the task can be handed over to a human operator if the uncertainty is too high).

Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

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.
- Predictors: 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

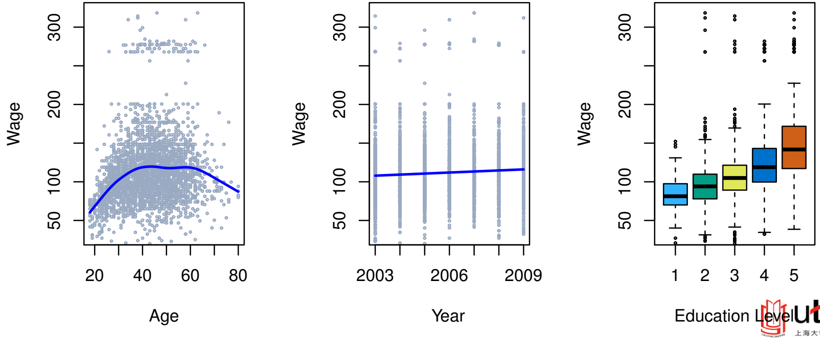
Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

Factors influencing wages

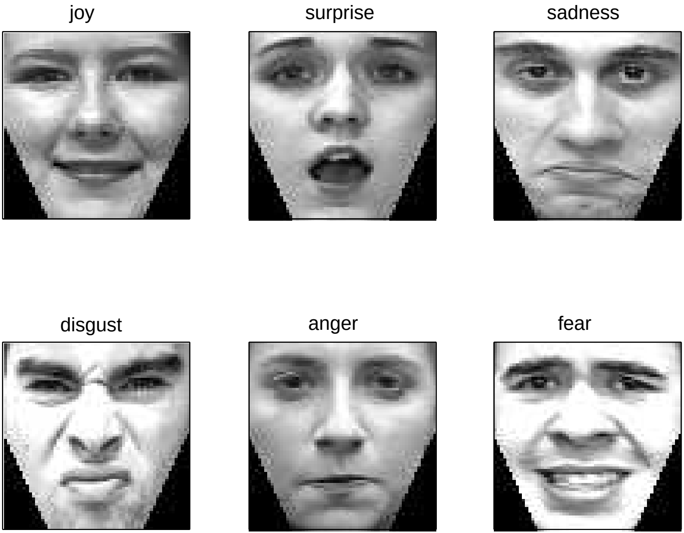
- Which factors influence wages? Are observations consistent with economic theories?
- Data: Income survey data for males from the central Atlantic region of the USA



Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

Expression recognition



joy

surprise

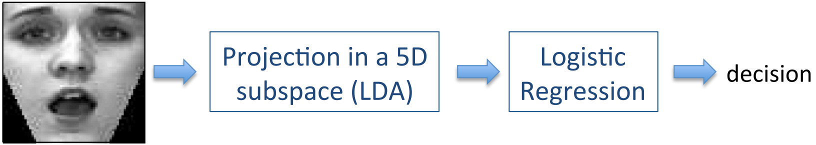
sadness

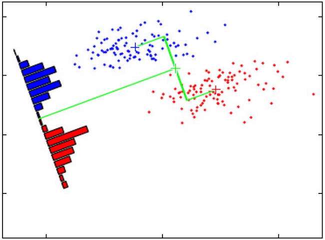
disgust

anger

fear

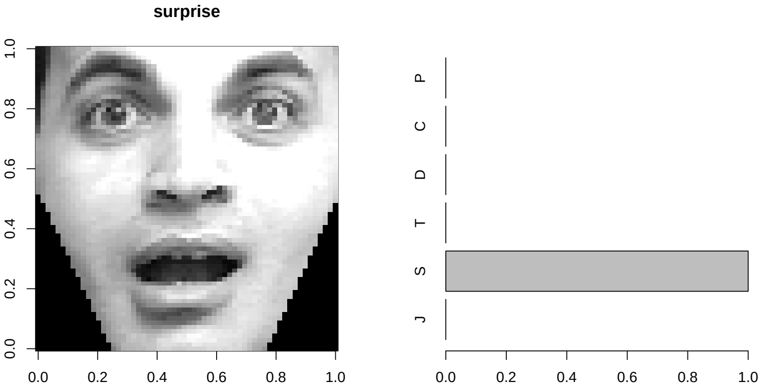
Learning



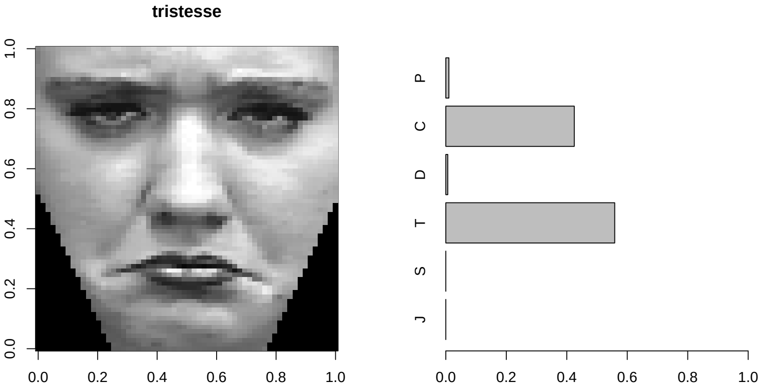


- 216 images 70 × 60 (36 per expression)
- 144 for learning, 72 for testing
- 5 features extracted by linear discriminant analysis
- Test error rate: 23.6% (random: 83.3%)

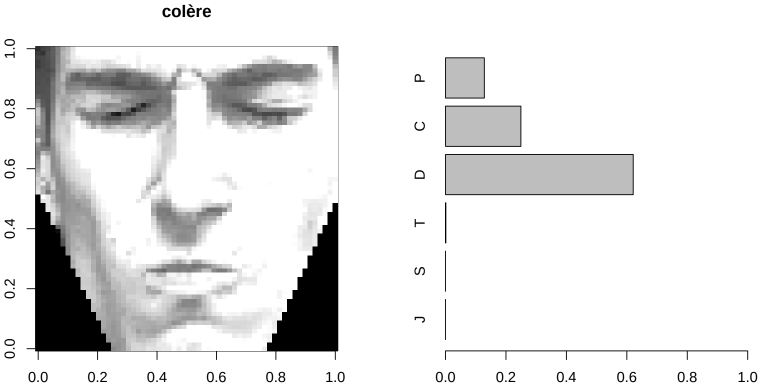
Results



Results



Results



Examples of learning problems

- 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.
- Detect and classify objects in a road scene.

Road scene analysis

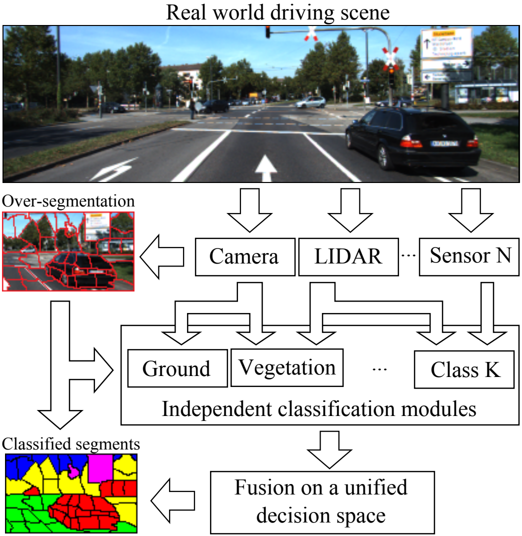
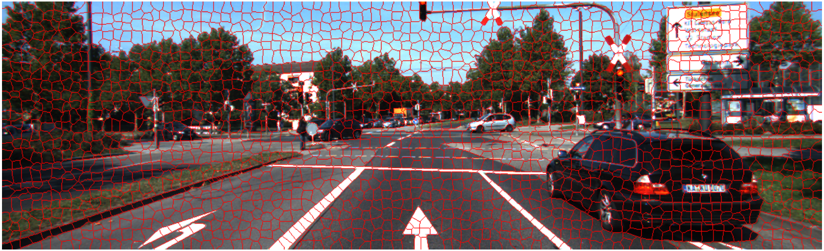
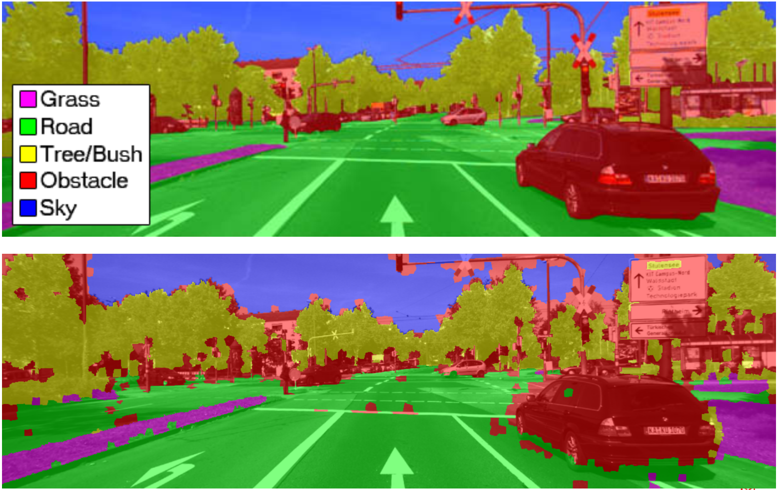


Image over-segmentation



Result



Overview

1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings

2. Some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - Regression: Bias-Variance trade-off

Supervised learning

Definitions

- Each observation consists of
  - A response variable Y (also called output, target, outcome)
  - A vector of p predictors X (also called inputs, features, attributes, explanatory variables).
- Supervised learning tasks:

  Regression: Y is quantitative (e.g., price, blood pressure).

  Classification: Y is nominal/categorical, i.e., it takes values in a finite, unordered set $\mathcal{C}$ (survived/died, digit 0-9, facial expression, etc.).

  Ordinal regression: Y is ordinal, i.e., it takes values in a finite, ordered set $\mathcal{C}$ (example: "small", "medium", "large")
- We have training/learning data $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. These are observations (examples, instances) of these variables.

Supervised learning

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 response variable, just a collection of variables (features) observed for a set of instances.
- Unsupervised learning tasks:

  Clustering: Find groups of observations that behave similarly

  Feature extraction: Find a small number of new features that contain as much relevant information as possible

  Novelty detection: Learn a rule to detect data from a previously unseen distribution (outliers, new states, etc.)
- Unsupervised learning is sometimes useful as a pre-processing step prior to supervised learning.

Semi-supervised learning

- Same task as supervised learning, but the response variable is only observed for a subset of the learning data.
- The learning set has the following form:

$$\mathcal{L} = \underbrace{\{(x_i, y_i)\}_{i=1}^{n_s}}_{\text{labeled data}} \cup \underbrace{\{x_i\}_{i=n_s+1}^{n}}_{\text{unlabeled data}}.$$

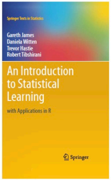
- A common situation, as data labeling is usually very costly.

Overview

1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings

2. Some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - Regression: Bias-Variance trade-off

Course texts



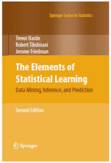
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

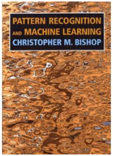
Data Mining, Inference, and Prediction

Second Edition

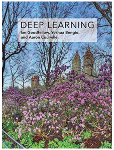
Springer

- "An Introduction to Statistical Learning" (ISLR): emphasis on basic principles and application, no mathematical details. Available at http://faculty.marshall.usc.edu/gareth-james/ISL
- "The Elements of Statistical Learning" (ESL): more mathematically advanced and theoretical. Available at http://statweb.stanford.edu/~tibs/ElemStatLearn

Course texts (continued)

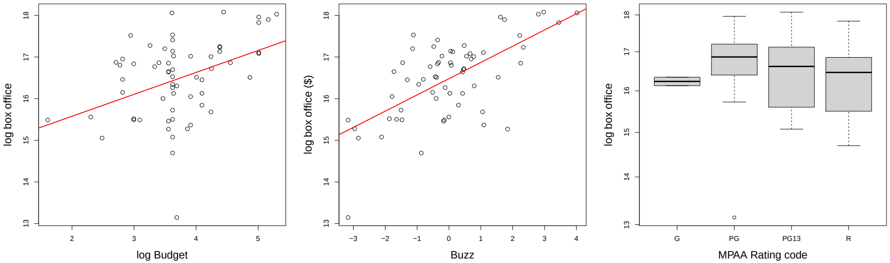


- "Pattern Recognition and Machine Learning" (PRML): same level as ESL, covers some other topics. Available at BUTC.



- "Deep Learning": recent textbook on neural networks. Available at `http://www.deeplearningbook.org`

A regression problem



- Shown are the log of box office receipt vs log of budget, rating and buzz index for 62 2009 movies, with red linear-regression line fits.
- Can we predict box office receipt using any single predictor?
- Perhaps we can do better using a model

$$\text{Box office} \approx f(\text{Budget}, \text{Buzz}, \text{Rating})$$

Notation

- Here 'Box office' is a response that we wish to predict. It is denoted as Y (variables are usually denoted by capital letters).
- 'Budget' is a predictor; we name it X_1. Likewise name 'Buzz' 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 (sources of variation of Y not explained by X).

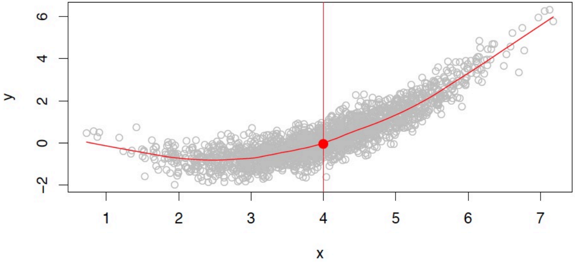
What is $f(X)$ good for?

- 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.
- Is there an optimal function f?

Overview

1. Introduction
    - Examples
    - Supervised vs. unsupervised learning
    - Recommended readings
2. Some basic concepts
    - The regression function
    - Nonparametric vs. parametric estimation
    - Regression: Bias-Variance trade-off

Regression function



- 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 typical value is

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

where $\mathbb{E}(Y|X = 4)$ is the expected value of Y given $X = 4$.

Optimality of the regression function

Definition (Regression function)

Function $f(x) = \mathbb{E}(Y \mid X = x)$ is called the regression function.

- We will show that the regression function is, in some sense, optimal.
- 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):

$$\begin{aligned} \text{MSE}(g) &= \mathbb{E}_{X,Y}\left[(Y - g(X))^2\right] \\ &= \mathbb{E}_X\left\{\mathbb{E}_Y\left[(Y - g(X))^2 \mid X = x\right]\right\} \end{aligned}$$

Optimality of the regression function (continued)

- We can write

$$\begin{aligned}\mathbb{E}_Y[(Y-g(X))^2 \mid X=x] &= \mathrm{Var}(Y \mid X=x) + (f(x)-g(x))^2 \quad (1a)\\ &= \mathrm{Var}(\epsilon \mid X=x) + (f(x)-g(x))^2 \quad (1b)\end{aligned}$$

Proof.

- The regression function f minimizes $\mathbb{E}[(Y-g(X))^2|X=x]$ for all x: consequently, it minimizes $\mathrm{MSE}(g)$. We write

$$f = \arg\min_g \mathrm{MSE}(g)$$

- The regression function 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

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

- 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

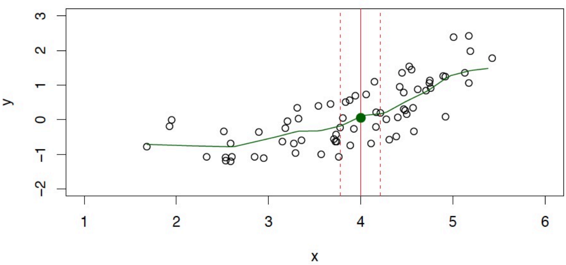
1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings

2. Some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - Regression: Bias-Variance trade-off

How to estimate f?

- Learning set: $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Typically we have few if any data points with $x_i = 4$ exactly. So, how can we estimate $\mathbb{E}(Y \mid X = x)$?
- Solution: we can compute the mean value of Y in a neighborhood $\mathcal{N}(x)$ of x:

$$\widehat{f}(x) = \text{Ave}\{y_i : x_i \in \mathcal{N}(x)\}$$



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.
- Formally, let $x_{(1)}, \ldots, x_{(n)}$ be the training vectors ordered by increasing distance to x. We often use the Euclidean distance

$$d(x, x_i) = \|x - x_i\| = \left(\sum_{j=1}^{p} (x_j - x_{ij})^2 \right)^{1/2}$$

- We then have

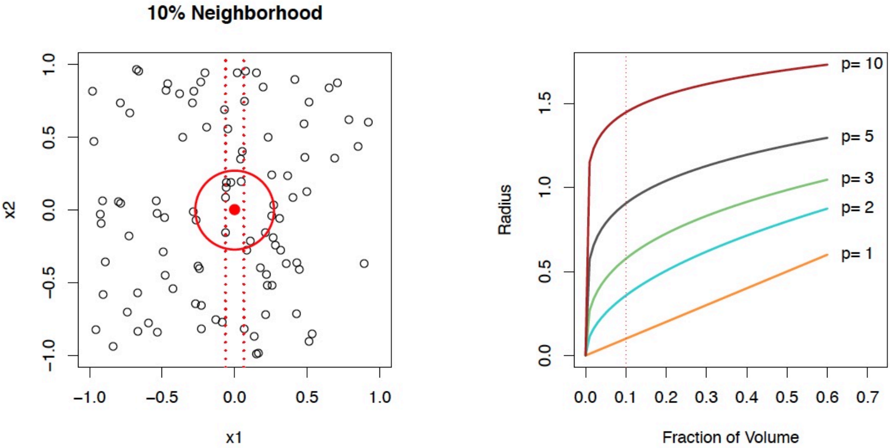
$$\widehat{f}(x) = \frac{1}{K} \sum_{i=1}^{K} y_{(i)}.$$

- 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.

Curse of dimensionality

- Nearest neighbor methods can perform badly when p is large.
- Reason: nearest neighbors tend to be far away in high dimensions. This is called the curse of dimensionality.
- We need to use a reasonable fraction of the n values of Y in the 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 $\mathbb{E}(Y \mid X = x)$ by local averaging.

Curse of dimensionality (continued)



Parametric models

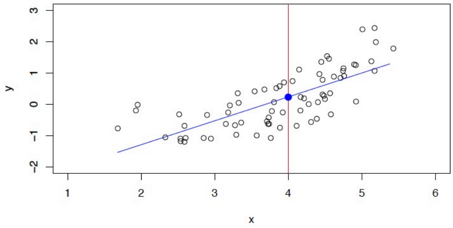
- A parametric model assumes that f belongs to a parametrized family of functions with a simple form.
- The linear model assumes the following form for f:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

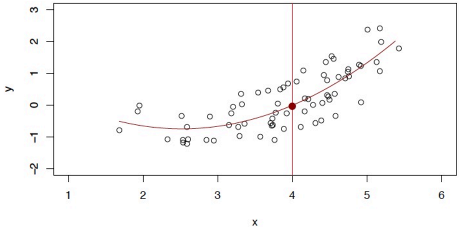
  It is specified in terms of a vector of $p+1$ parameters $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$.
- We estimate the parameters by fitting the model to training data.
- Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown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

Linear vs. quadratic

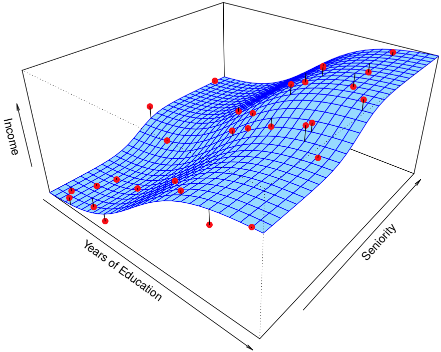
A linear model $\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ gives a reasonable fit here



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



Simulated example

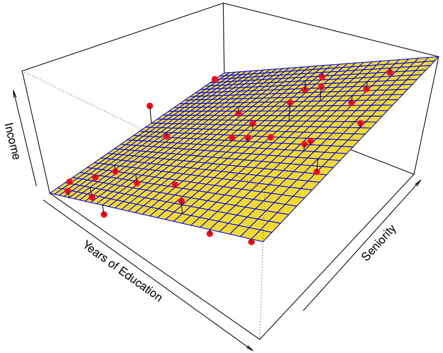


Red points are simulated values for income from the model

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

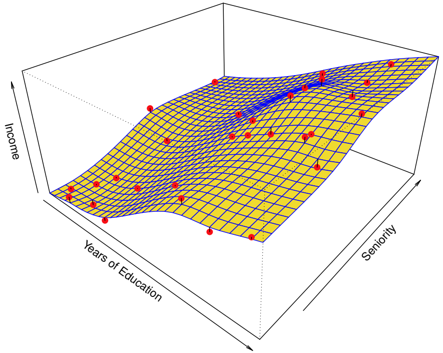
f is the blue surface.

Linear regression model fit



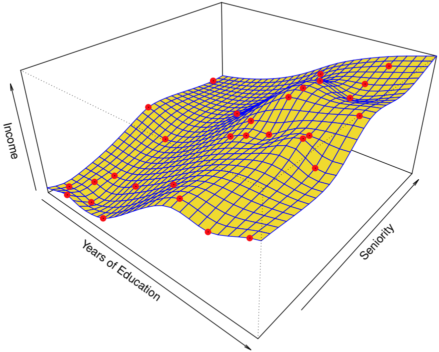
A linear model does not fit the data very well, but it provides a simple description of the effect of the two predictors on the response.

More flexible regression model



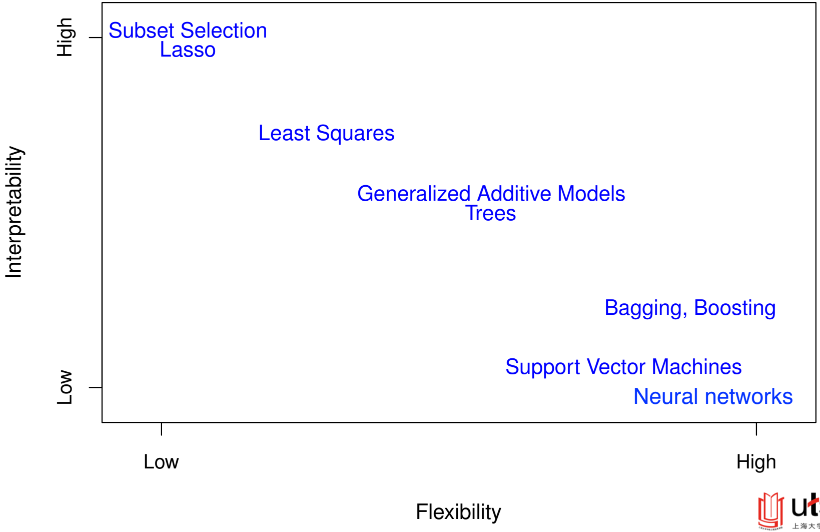
More flexible regression model fit to the simulated data. Here we used a technique called a thin-plate spline to fit a flexible surface.

Even more flexible spline regression model



Here an even more flexible spline regression model interpolates the data points (it makes no errors on the training data)! Also known as overfitting.

Interpretability/flexibility trade-off



Overview

1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings
2. Some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - Regression: Bias-Variance trade-off

Assessing Model Accuracy

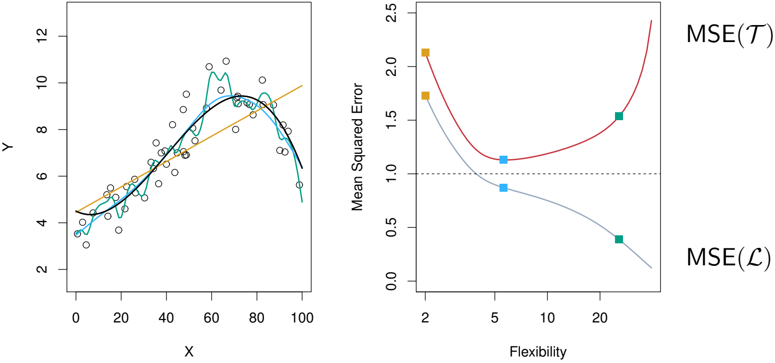
- Suppose we have a regression problem. We fit a model $f(x)$ to some learning data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$ and we wish to see how well it performs.
- We could compute the average squared prediction error over $\mathcal{L}$:

$$\text{MSE}(\mathcal{L}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{f}(x_i)\right]^2$$

  This may be biased toward more overfit models.
- Instead we should, if possible, compute it using fresh test data $\mathcal{T} = \{(x_i', y_i')\}_{i=1}^m$:

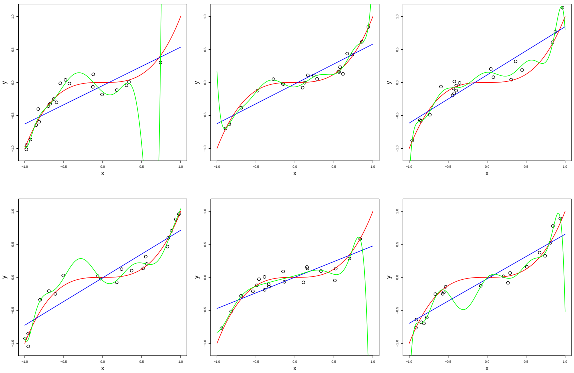
$$\text{MSE}(\mathcal{T}) = \frac{1}{m}\sum_{i=1}^{m}\left[y_i' - \widehat{f}(x_i')\right]^2$$

Learning and test errors for 3 models



The most flexible model (with more parameters) does not perform best.
Why?

Another example



Red line is truth. Blue and green lines correspond, respectively, to a linear model and a polynomial of degree 10.

The linear model is stable but biased. The polynomial model is more flexible, so it is less biased, but it is unstable. Bias and variance both account for prediction error.

Bias-variance decomposition

- Let $\widehat{f}$ be the estimated regression function learnt from data set $\mathcal{L}$.
- If the true model is $Y = f(X) + \epsilon$, with $f(x) = \mathbb{E}(Y|X = x)$, then the MSE averaged over all learning sets $\mathcal{L}$ conditionally on $X = x$ is

$$\mathbb{E}_{\mathcal{L},Y}\left[\left(Y - \widehat{f}(X)\right)^2 \mid X = x\right] =$$

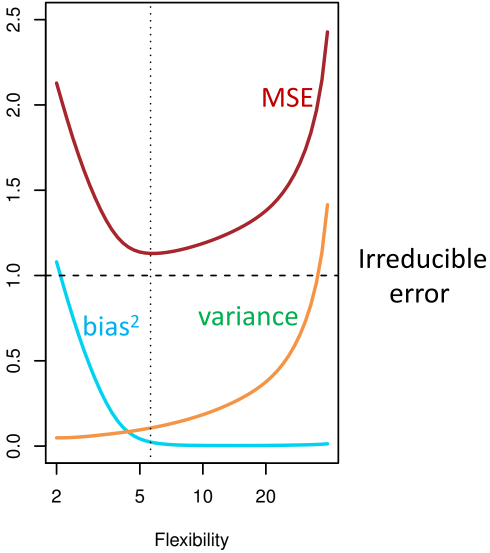
$$\underbrace{\text{Var}_{\mathcal{L}}(\widehat{f}(x))}_{\text{variance}} + \underbrace{\left[\mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - f(x)\right]^2}_{\text{bias}^2} + \underbrace{\text{Var}_Y(\epsilon \mid X = x)}_{\text{irreducible error}} \quad (2)$$

Proof.

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.

Graphical illustration



Proof of Equation (1)

$$
\begin{aligned}
\mathbb{E}_Y[(Y-g(X))^2 \mid X=x] &= \mathbb{E}_Y[(Y-f(x)+f(x)-g(x))^2 \mid X=x] \\
&= \underbrace{\mathbb{E}_Y[(Y-f(x))^2 \mid X=x]}_{\mathrm{Var}(Y|X=x)} + (f(x)-g(x))^2 \\
&\quad + 2(f(x)-g(x))\underbrace{\mathbb{E}_Y[Y-f(x) \mid X=x]}_{\mathbb{E}[Y|X=x]-f(x)=0}
\end{aligned}
$$

Given $X=x$,

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

so

$$\mathrm{Var}(Y \mid X=x) = \mathrm{Var}(\epsilon \mid X=x)$$

Back

Proof of Equation (2) I

First, we insert $\mathbb{E}_{\mathcal{L}}[\widehat{f}(X) \mid X = x] = \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)]$:

$$\mathbb{E}_{\mathcal{L},Y}\left[\left(Y - \widehat{f}(X)\right)^2 \mid X = x\right] =$$

$$\mathbb{E}_{\mathcal{L},Y}\left[\left(Y - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] + \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - \widehat{f}(X)\right)^2 \mid X = x\right] =$$

$$\underbrace{\mathbb{E}_{Y}\left[\left(Y - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)]\right)^2 \mid X = x\right]}_{A} +$$

$$\underbrace{\mathbb{E}_{\mathcal{L}}\left[\left(\widehat{f}(x) - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)]\right)^2\right]}_{B = \mathrm{Var}_{\mathcal{L}}[\widehat{f}(x)]} +$$

$$\underbrace{2\mathbb{E}_{\mathcal{L},Y}\left[(Y - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)])(\mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - \widehat{f}(X)) \mid X = x\right]}_{C}$$

Proof of Equation (2) II

- We have already seen from Eq. (1) that A can be written as

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

- In C, the first term in the product depends only on Y and the second term depends only on $\mathcal{L}$. As Y and $\mathcal{L}$ are independent, we can write

$$C = 2\mathbb{E}_Y\left[Y - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] \mid X = x\right] \underbrace{\mathbb{E}_{\mathcal{L}}\left[\mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - \widehat{f}(X) \mid X = x\right]}_{=\mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - \mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] = 0}$$

QED

Back