SY19 – Machine Learning
Chapter 1: Introduction

Thierry Denœux

Université de technologie de Compiègne

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

Automne 2024

Informations pratiques

- Support de cours: transparents (en anglais) mis sur la page Moodle du cours au plus tard la veille de chaque séance. (Premier cours: https://www.hds.utc.fr/~tdenoeux/dokuwiki/en/sy19).
- Poser les questions d'intérêt général (pratiques ou relatives au contenu du cours) sur le forum de discussion de Moodle.
- Equipe enseignante :
  - Thierry Denoeux (responsable) : cours
  - Cyprien Gilet: cours, TD
  - Sylvain Rousseau : TD
- Evaluation :
  - Deux projets en binôme : 25% + 25%
  - Examens median (20%) et final (30%) : questions de cours, note éliminatoire au final ≤ 6

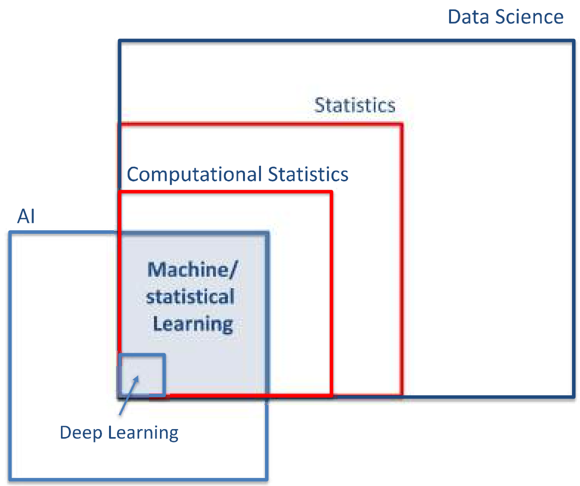
Overview

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

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

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
  - Trend analysis in social networks
  - E-commerce (recommendation systems)
  - Robotics, autonomous vehicles
  - Natural language recognition and generation
  - Finance (stock market forecasting, credit scoring, fraud detection,...)
  - Bioinformatics
  - Nondestructive testing, fault diagnosis
  - Mechanical engineering: design and optimization using surrogate models, etc.
- ML skills are in high demand by companies accross a wide range of areas.

Objectives of this course

- Understand the basic principles of ML
- Get working knowledge of the main ML techniques
  - Linear regression and classification (LDA, logistic regression)
  - Model selection: regularization (ridge regression, lasso), variable selection, linear feature extraction
  - Splines and additive models
  - Decision trees, random forests, bagging
  - Gaussian Mixture Models, EM algorithm
  - Kernel-based methods for classification (SVM), regression, novelty detection, clustering
  - Neural networks and deep learning
- Master the R software environment for data analysis and ML

Overview

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

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

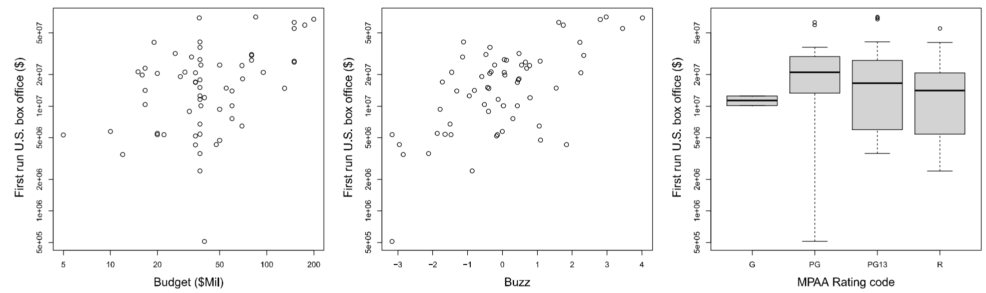
Examples of learning problems

- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.
- Analyze the contents of an image.

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



How to use these data to:

- Predict the BO receipt of a new movie?
- Quantify the uncertainty of the prediction?
- Understand what makes a movie commercially successful?

Examples of learning problems

- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.
- Analyze the contents of an image.

Spam detection

- Goal: build a customized spam filter.
- Data: 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

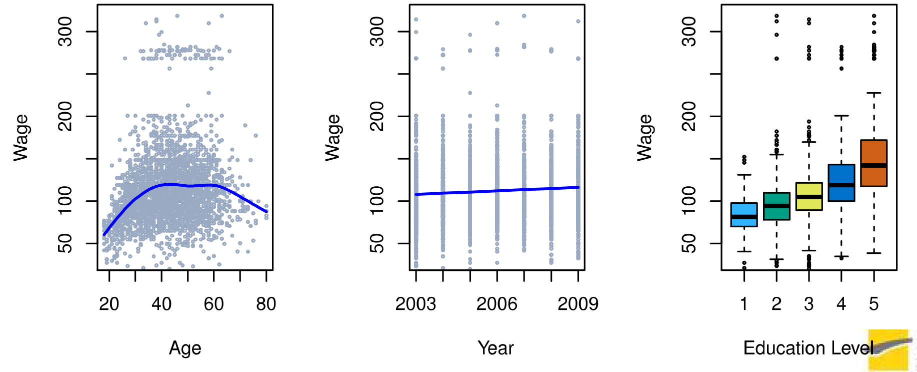
Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

Examples of learning problems

- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.
- Analyze the contents of an image.

Factors influencing wages

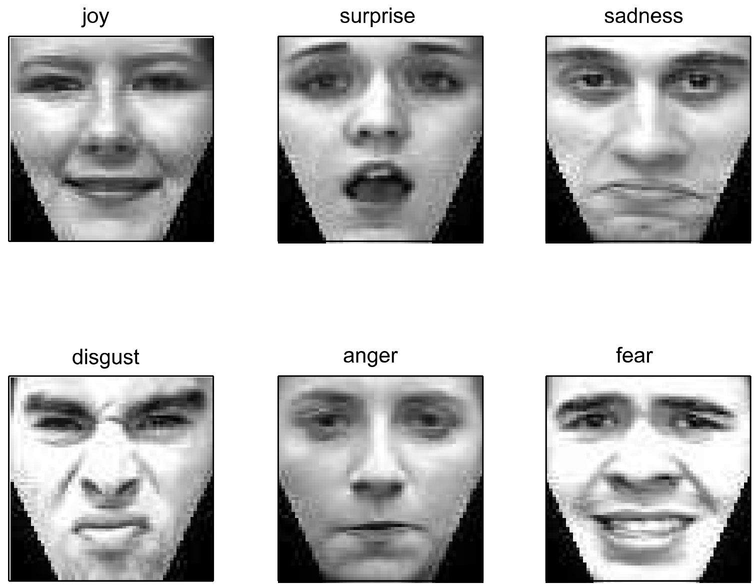
- Which factors influence wages? Are observations consistent with economic theories?
- Data: Income survey data for men from the central Atlantic region of the USA



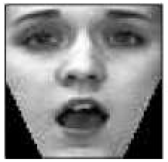
Examples of learning problems

- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.
- Analyze the contents of an image.

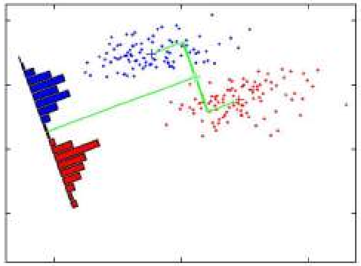
Expression recognition



Learning

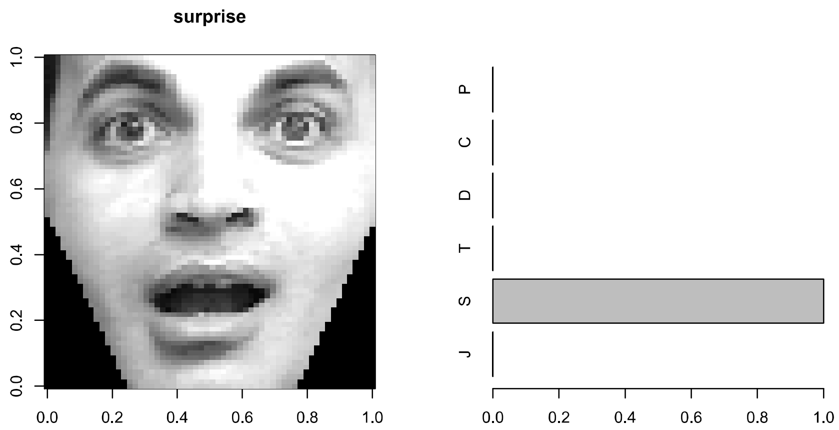


Projection in a 5D subspace (LDA) → Logistic Regression → decision

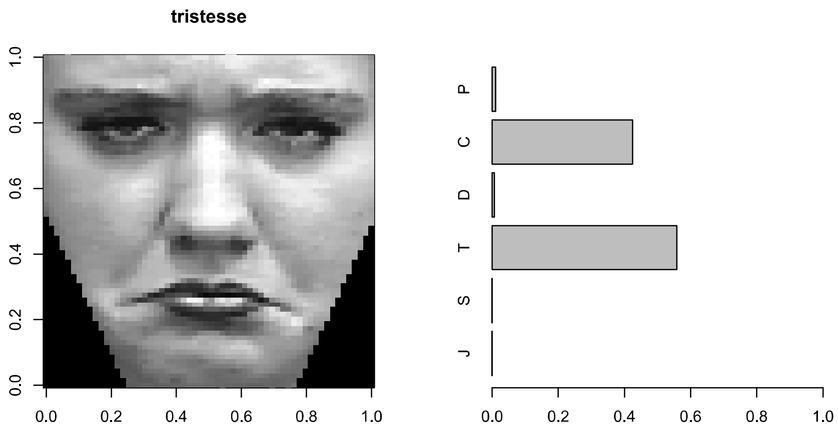


- 216 images 70 × 60 (36 per expression)
- 144 for learning, 72 for testing
- 5 features extracted by linear discriminant analysis
- Test error rate: 23.6% (random: 83.3%)

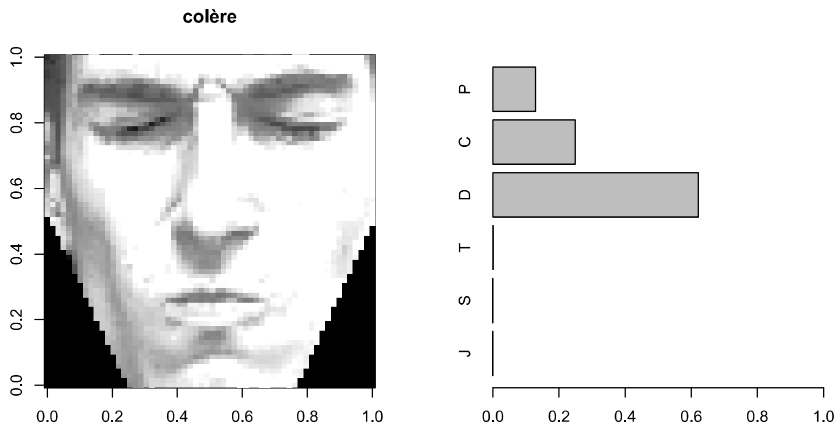
Results



Results



Results



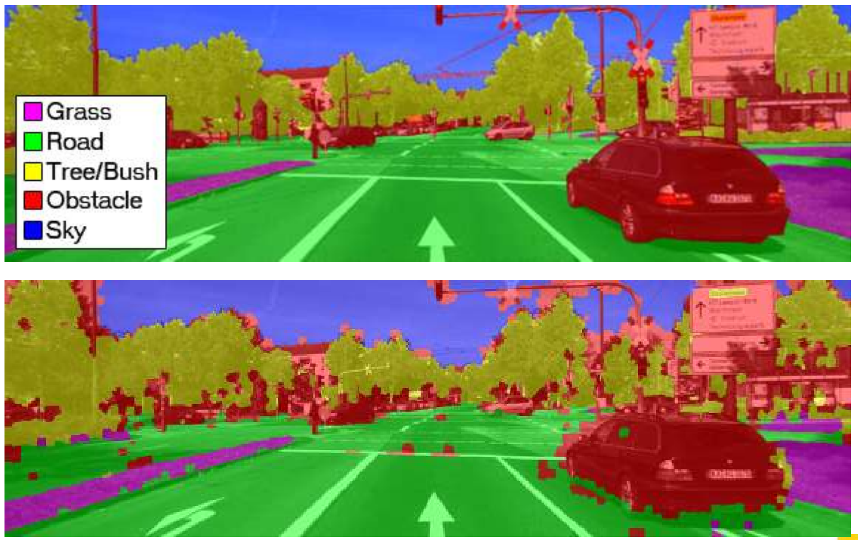
Examples of learning problems

- Predict the box office receipt of a movie from the genre, budget, star power, buzz, etc.
- Customize an email spam detection system.
- Establish the relationship between salary and demographic variables in population survey data.
- Recognize the expression on a face.
- Analyze the contents of an image.

Semantic segmentation

The semantic segmentation tasks consists in classifying each pixel to segment the image into regions corresponding to different kinds of objects.

Road scene analysis



Overview

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

Supervised learning

- We have a training/learning set $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$ of n observations (examples, instances) of
  - A response variable Y (also called output, target, outcome)
  - A vector of p predictors X (also called inputs, features, attributes, explanatory variables).
- The task is to predict Y given X for new data.
- Different cases:

  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/classification: Y is ordinal, i.e., it takes values in a finite, ordered set $\mathcal{C}$ (example: "small", "medium", "large")

Objectives of supervised learning

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

1. Accurately predict unseen test cases
2. Understand which predictors affect the response, and how
3. Quantify the uncertainty of the predictions
4. Assess the quality of our predictions and inferences

Unsupervised learning

- No response variable, just a collection $\{x_i\}_{i=1}^n$ of feature/attribute vectors 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:

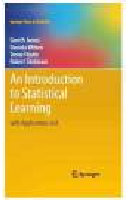
$$\mathcal{L} = \underbrace{\{(x_i, y_i)\}_{i=1}^{n_s}}_{\text{labeled data}} \cup \underbrace{\{x_i\}_{i=n_s+1}^{n}}_{\text{unlabeled data}}.$$

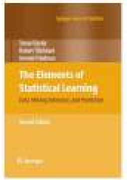
- A common situation, as data labeling is usually very costly.

Overview

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

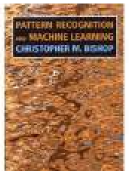
Course texts

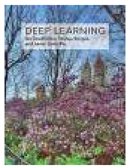




- "An Introduction to Statistical Learning" (ISLR): emphasis on basic principles and application, no mathematical details. Second edition available at `https://www.statlearning.com`
- "The Elements of Statistical Learning" (ESL): more mathematically advanced and theoretical. Available at `http://statweb.stanford.edu/~tibs/ElemStatLearn`

Course texts (continued)



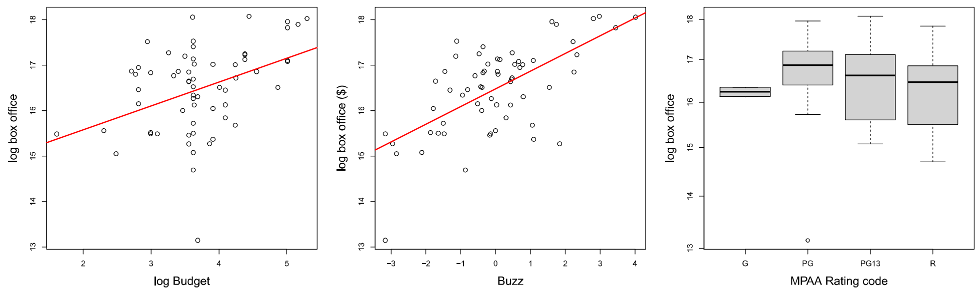


- "Pattern Recognition and Machine Learning" (PRML): same level as ESL, covers some other topics. Available at BUTC.
- "Deep Learning": recent textbook on neural networks. Available at `http://www.deeplearningbook.org`

Overview

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

A regression problem



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

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

Formalization

- We can write

$$\boxed{Y = g(X) + \epsilon}$$

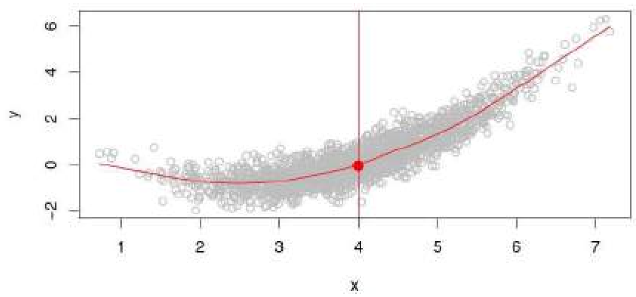
  where
  - X is the vector of predictors
  - g is a linear or nonlinear prediction function
  - ϵ is a random error term
- With a good g we can
  - Make predictions of Y at new points $X = x$.
  - Understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y and, sometimes, how each component X_j of X affects Y.
- Is there an optimal function g?

Overview

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

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

Regression function



- What is a good value for $g(X)$ at any selected value of X, say $X = 4$?
- There can be many Y values at $X = 4$. A typical value is the conditional expectation

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

Definition (Regression function)

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

Loss function

- Assume we predict Y given $X = x$ by $g(x)$. A “good” function g should be such that $g(x)$ is often “close” to Y.
- A common error measure (or loss function) is the squared error $(y - g(x))^2$.
- A good prediction function should have the lowest possible squared error $(y - g(x))^2$, on average.

Definition (Mean squared error)

The mean squared error (MSE) of g is

$$\text{MSE}(g) = \mathbb{E}_{X,Y}\left[(Y - g(X))^2\right]$$

Optimality of the regression function

Theorem

The regression function minimizes the MSE, i.e.,

$$f = \arg\min_g MSE(g)$$

Proof:

1. $\text{MSE}(g) = \mathbb{E}_{X,Y}\left[(Y - g(X))^2\right] = \mathbb{E}_X\left\{\mathbb{E}_Y\left[(Y - g(X))^2 \mid X\right]\right\}$
2. We can write

$$\mathbb{E}_Y[(Y - g(X))^2 \mid X = x] = (f(x) - g(x))^2 + \underbrace{\text{Var}(Y \mid X = x)}_{\text{Var}(\epsilon|X=x)} \quad (1)$$

Proof.

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

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.

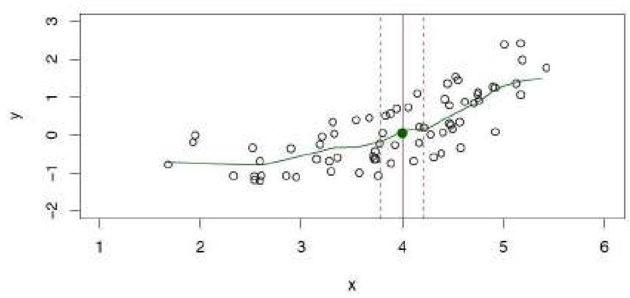
Overview

1. Introduction
   - Examples
   - Supervised vs. unsupervised learning
   - Recommended readings
2. Regression: some basic concepts
   - The regression function
   - Nonparametric vs. parametric estimation
   - 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 regression

- The neighborhood $\mathcal{N}(x)$ can be defined as the region containing the K nearest neighbors (NN) of x in the training data.
- To define the neighbors, 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

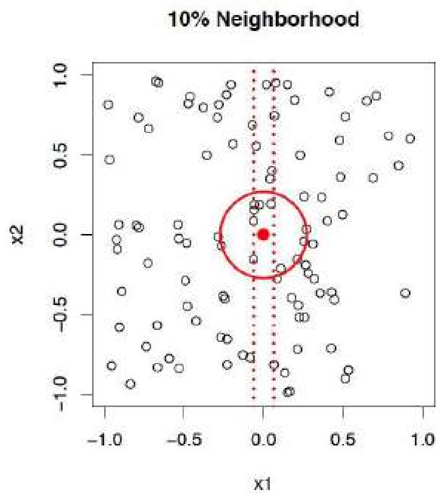
$$\widehat{f}(x) = \frac{1}{K}\sum_{i=1}^{K} y_{(i)},$$

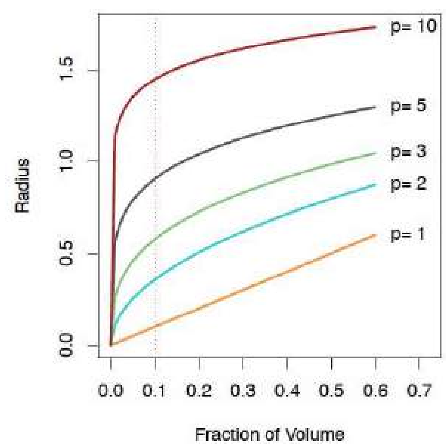
where $y_{(1)}, \ldots, y_{(K)}$ are the values of Y for the K NN of x.
- This method is called nearest neighbor regression. It is a nonparametric method. (We do not assume any functional form for f a priori). This method can be pretty good for small p – i.e., $p \leq 4$ and n not too small.

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: example





Parametric models

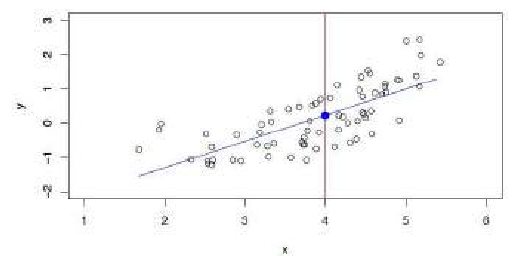
- A parametric model assumes that f belongs to a parametrized family of functions with a simple form.
- The simplest parametric model is the linear model, which assumes the following form for f:

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

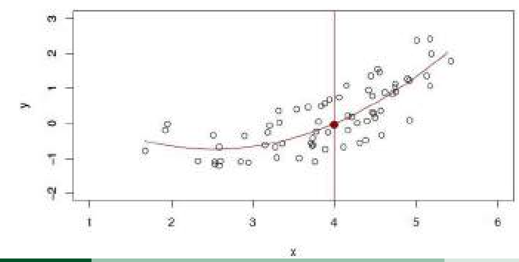
  It is specified in terms of a vector of $p+1$ parameters $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$.
- We estimate the parameters by fitting the model to training data.
- Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown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

Linear vs. quadratic

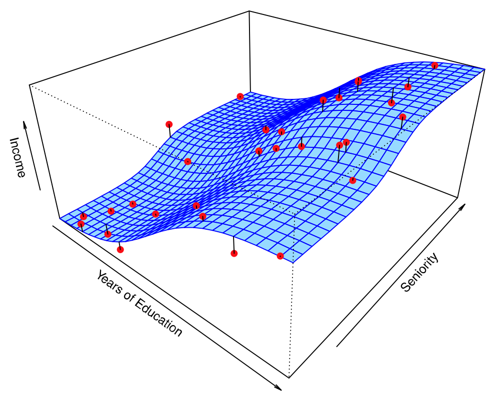
A linear model $\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ gives a reasonable fit here:



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



Simulated example

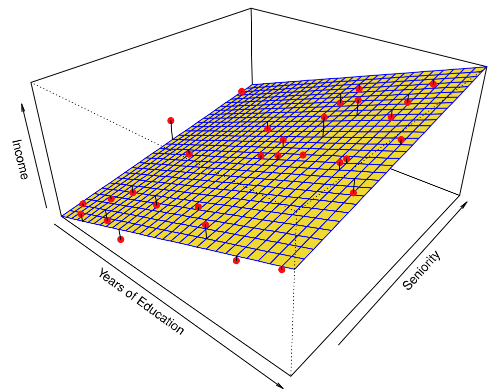


Red points are simulated values for income from the model

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

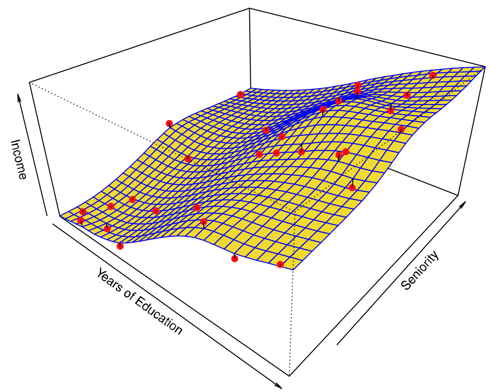
f is the blue surface.

Linear regression model fit



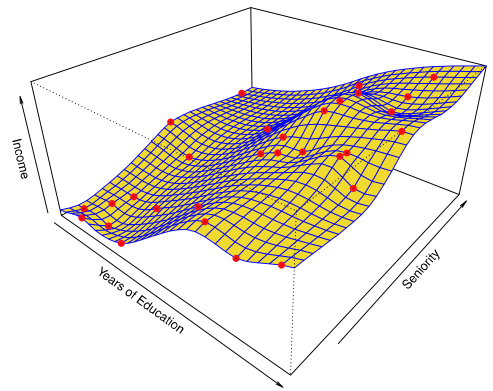
A linear model does not fit the data very well, but it provides a simple description of the effect of the two predictors on the response.

More flexible regression model



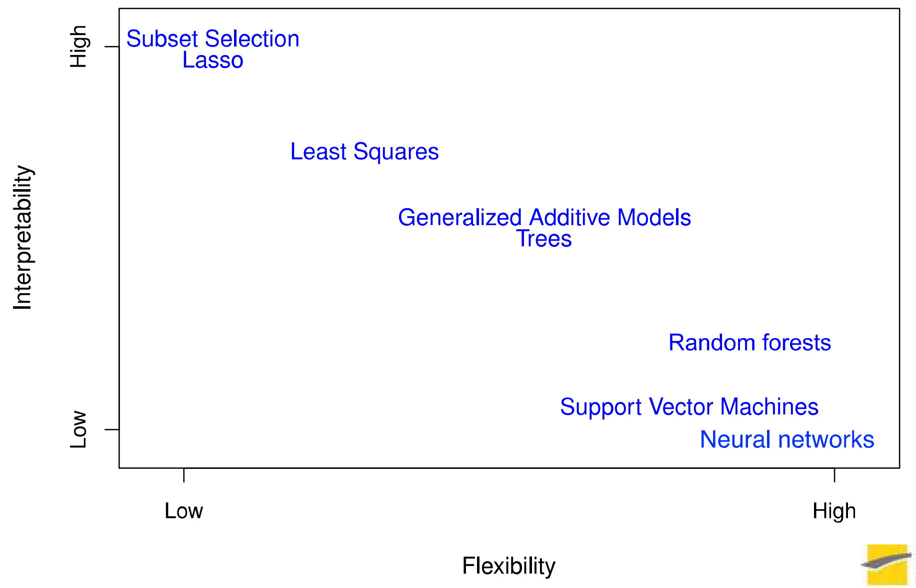
More flexible regression model fit to the simulated data. Here we used a model called a thin-plate spline to fit a flexible surface.

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 Regression: some basic concepts
- The regression function
- Nonparametric vs. parametric estimation
- 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 $\mathcal{L}$:

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

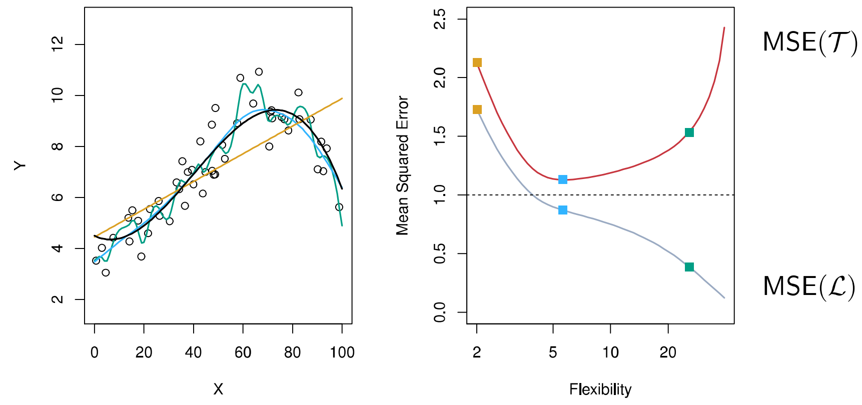
This is called the learning error. It can be severely biased toward more overfit models.

- Instead we should, if possible, estimate the error using fresh test data $\mathcal{T} = \{(x_i', y_i')\}_{i=1}^m$:

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

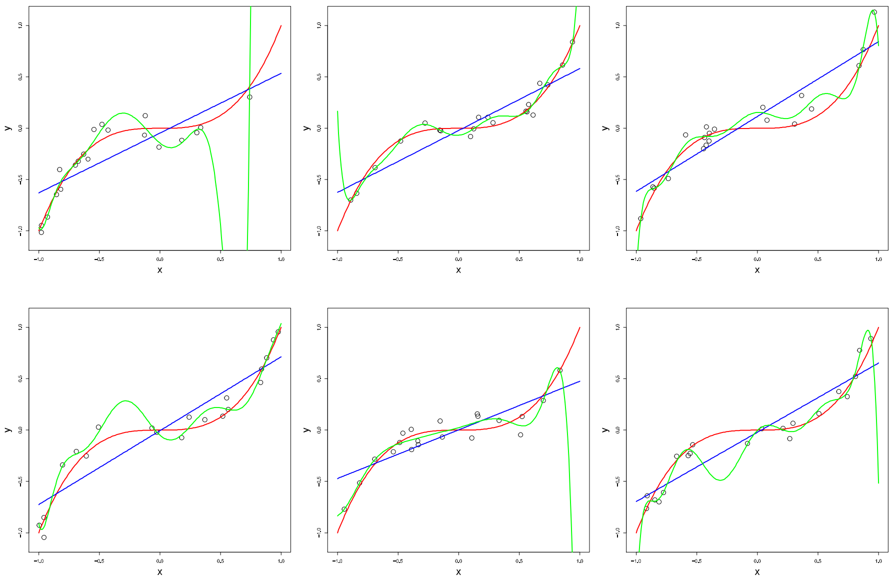
This is the test error.

Learning and test errors for 3 models



- Black curve is truth. Orange, blue and green curves/squares correspond to fits of different flexibility.
- The most flexible model (with more parameters) does not perform best. Why?

Another example (see next slide)



Example (continued)

- Red curve is truth. Blue and green curves 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.

Formalization

Theorem (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

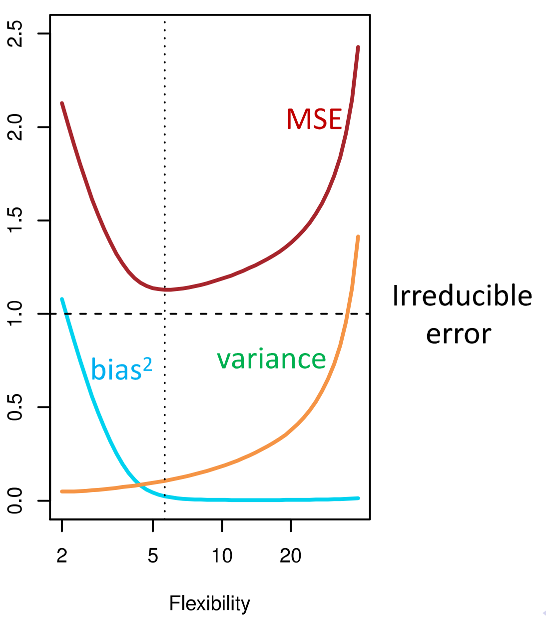
$$\mathbb{E}_{\mathcal{L},Y}\left[\left(Y - \widehat{f}(X)\right)^2 \mid X = x\right] = \underbrace{\left[\mathbb{E}_{\mathcal{L}}[\widehat{f}(x)] - f(x)\right]^2}_{bias^2} + \underbrace{\mathrm{Var}_{\mathcal{L}}(\widehat{f}(x))}_{variance} + \underbrace{\mathrm{Var}_Y(\epsilon \mid X = x)}_{irreducible\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

Proof.

Bias-variance trade-off

- When the flexibility of $\widehat{f}$ increases, $\widehat{f}(x)$ becomes closer to Y: its bias decreases, and as its variance increases.
- So choosing the right degree of flexibility based on average test error amounts to a bias-variance trade-off.
- We will come back to the very important issue of model selection in a later chapter.

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

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

- 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