Machine Learning

MIRI Master

Lluís A. Belanche

belanche@cs.upc.edu

Soft Computing Research Group

Departament de Ciències de la Computació (Computer Science Department)

Universitat Politècnica de Catalunya - Barcelona Tech

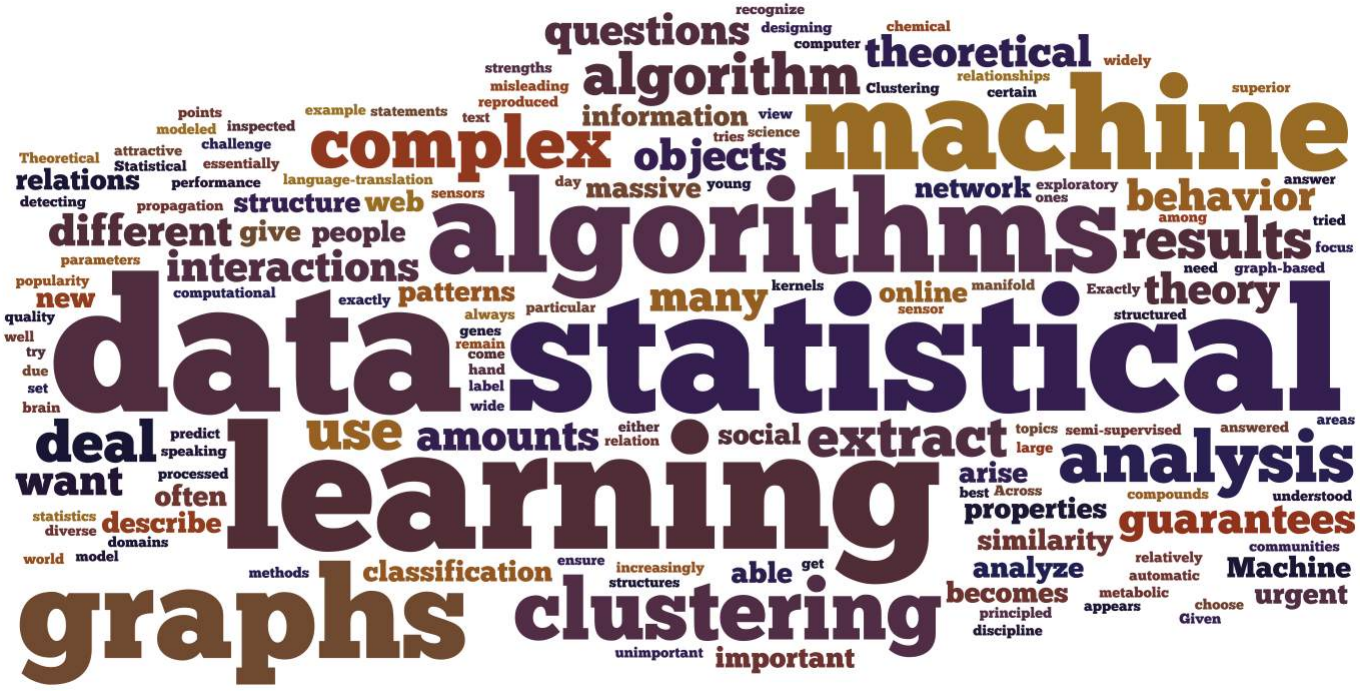
Spring Semester 2017-2018

LECTURE 1: Introduction to Machine Learning

Machine Learning

What is this course about?

Machine learning is a field of Computer Science that explores automatic methods for inferring models from data (e.g., for finding structure, for making predictions)



Machine Learning

Examples of learning tasks (ML subfields)

SUPERVISED LEARNING uses labeled data

Classification: predicting a class (or category) to each example (e.g., document classification); note multi-label, probabilistic generalizations

Regression: predicting a real value for each example (e.g., prediction of ph concentration); note multi-variable generalization

UNSUPERVISED LEARNING does not use (or have) data labels

Clustering: discovering homogeneous groups in data (clusters)

Dimensionality reduction: finding lower-dimensional data representations

Density estimation: estimating the probabilistic mechanism that generates data

Novelty detection: finding anomalous/novel/outlying data

SEMI-SUPERVISED LEARNING uses partly labeled data

Ranking: ordering examples according to some criterion (e.g., web pages returned by a search engine).

Reinforcement: delayed rewarding (e.g., finding the way out in a maze)

frontiers are increasingly vanishing ... (e.g. *Semi-supervised classification*)

Machine Learning

Course organization

- One **ML topic**: one pack of 4h
  - Theoretical lecture (2h) **I work!**
  - Practical lab session (2h) **You work!**
  - Each pack takes place in close-by sessions (hopefully)
- There will be special lab sessions (w/o direct link to theory):
  - one devoted to introduction to R and RStudio (non-presential)
  - one devoted to pre-processing
  - one devoted to resampling and reporting
- I will post the material (lecture slides, lab stuff) shortly **before** each class
- There is a **term project**, an oral **spotlight** presentation and a written **quiz**
- **Office** hours: Wednesdays 12-14h, office 326 (OMEGA building); previous e-mail appointment needed

Machine Learning

Course organization

- Labs use the R language (mandatory) along with the RStudio suite (recommended) as a front-end for R code programming
- You are welcome to bring your own laptop
- There is a **first part** (1h - 1h 30') in which you go through a prepared and complete R script on the current ML topic individually
- There is a **second part** (remainder) in which you can ask free questions or work towards your **term project**

Machine Learning

Course organization

The term project ...

- Must be done in groups of **3** (2?) people (singles are forbidden)
- **Topic** entirely of your choice (with some limits)
- Intended to start **early** and grow **mature** over time
- A final **written report** (along with the code) and an accompanying **spotlight** (short presentation) should be carefully prepared

Machine Learning

Course organization

The quiz is a written exam about generic knowledge

- It will take place on our last regular session
- Same place, same time (2h long)
- Collection of multiple-answer true/false questions

The spotlight is a short presentation (5-10') along with short Q/A (a summary of your project: goals and achievemnts)

Machine Learning

Course organization

- **Grading**: 35 % **quiz** + 50 % **project** + 15 % **spotlight**

- The project includes a skill (**reasoning**):

  "Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation."

- Several **rubrics** will be available to guide you, as well as detailed information on arrangements

Machine Learning

Course organization

Important dates:

- Lectures begin: **February 20**
- Lectures end: **May 31**
- **May 1, Easter Holydays**: NO CLASS
- **May 3** is Monday! (FIB only)
- **June 5** NO CLASS
- Quiz: **June 7** (usual classroom and hours)
- Term project
  1. unofficial start: **early March**
  2. final delivery: **June 25** (not to be extended)
  3. spotlights: **June 18** (hours and place TBA)

Machine Learning

Be warned ...

After this course (and if you really exploit it), you should

1. **understand** many of the principles and foundations of the field;
2. **understand** the workings of many state-of-the-art methods;
3. **know how** to properly apply these methods to real data; and
4. **acquire** the necessary prerequisites to understand and apply more advanced methods that build on the covered topics

The course does **not** cover:

- all possible machine learning methods; or
- all possible applications of machine learning

Introduction to Machine Learning

Motivation

- We want human-like behaviour in **machines** (specially, in computers!): adaptable (flexible), fast, reliable and automatic

- Biological systems (like us) deal with imprecision, partial truth, uncertainties, noise, contradictions, ... mostly in a data-driven fashion (think of a baby learning to walk or talk)

- They also make predictions in intricate ways by own behavioral models (learnt by experience & almost impossible to verbalize)

- Very difficult to achieve by direct programming (brittleness)

Introduction to Machine Learning

A system (living or not) learns if it uses *past* experience to improve *future* performance:

1. **Acquiring** more knowledge (or more abilities) with time, and

2. **Reorganizing** this knowledge such that some problems are solved:

   a) in a more *efficient* way (using less resources) or

   b) in a more *effective* way (higher performance standards)

Introduction to Machine Learning

I have an idea ...

Let the **machine** ...

1. **learn** the information contained in a data sample (the experience);

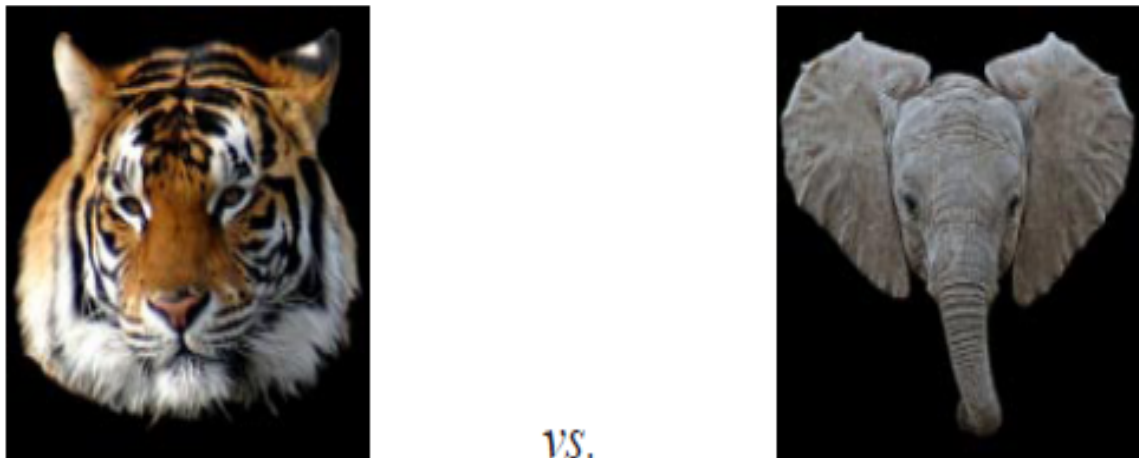
2. **build** a model with it and

3. **use** it to answer future queries

$\longrightarrow$ This is **Machine Learning**

Introduction to Machine Learning

Motivation



(example due to Jason Weston)

- Suppose we have a dataset $\mathcal{D}$ of 50 images of elephant faces and 50 of tiger faces, which we digitize into 100×100 pixel RGB images, so we have $\boldsymbol{x} \in \{0, \ldots, 255\}^d$ where $d = 3 \cdot 10^4$

- Given a *new* image, we want to answer the question: is it an elephant or a tiger? [we assume it is one or the other]

Introduction to Machine Learning

Motivation

Define a **classifier** as a function $f : \mathbb{R}^d \to \{-1, +1\}$

Key fact: Take a data sample $\mathcal{D} = \{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$; for any f there exists f^* s.t.

1. f and f^* coincide in all the N images in $\mathcal{D}$

2. f and f^* differ in at least one of all possible images (not in $\mathcal{D}$)

Moral: ML is about learning general structure from data (and $\mathcal{D}$ is just a sample!)

Introduction to Machine Learning

So what do we do?

Moral (more formal): Based on training data $\mathcal{D}$ *only*, there is no means of choosing which function f is better (generalization is not "guaranteed")

Consequence: we must add control to the "fitting ability" of our methods (complexity control)

training/empirical/apparent/resubstitution/in-sample error (in $\mathcal{D}$)
true/generalization/out-sample error (in all possible images)

$$\boxed{\text{true error } (f) \leq \text{training error } (f) + \text{complexity of } f}$$

Introduction to Machine Learning

Machine Learning in context

Machine Learning has strong bridges to other disciplines:

Statistics: inferential statistics, distribution and sampling theory, mathematical statistics

Mathematics: optimization, numerical methods, asymptotics, ...

Algorithmics: convergence, correctness, complexity (time, space), ...

Artificial Intelligence: general aims at "intelligent" behaviour

Introduction to Machine Learning

Applications of Machine Learning

Machine perception: image analysis, speech analysis, face and handwritten recognition, image and video captioning, ...

Natural language: translation, understanding, generation, ...

Business applications: fraud detection, credit concession, network intrusion, stock market analysis, ...

Scientific tasks: bioinformatics, chemoinformatics, microbiology, geology, astronomy, medical diagnosis, ...

Web analysis: hypertext, blogs, e-mail, social networks (recommender systems, sentiment analysis), ...

Introduction to Machine Learning

Examples

Need to identify **task**, **performance measure** and **experience (data)**

1. Playing Checkers
   - the task is to learn to win games without violating the rules
   - the performance measure is the fraction of games won
   - the experience is the set of games played against an opponent
2. Identification of handwritten digits
   - the task is to learn to recognize a member of the set $\{0, 1, ..., 9\}$
   - the performance measure is the fraction of correctly recognized elements
   - the experience could be a set of human labelled digits (*e.g.*, from the USPS)

Introduction to Machine Learning

Examples

Ill-posed examples (need a lot of refinement and specific information):

- classify my incoming e-mail
- predict the result of a basketball game
- predict who is going to win a video game
- predict whether this song is going to become a hit

Wrong examples (non-sensical for different reasons):

- sort an array of integers
- predict the lottery
- predict the next digit in the decimal expansion of π
- predict the terminal velocity of a falling object

Introduction to Machine Learning

Examples: your turn!

- predict whether a patient, already hospitalized for a heart attack, will develop a second attack, based on previous patient records (demographic, dietary, clinical ...).
- learn the land shape of Formentera island:

  1. knowing the boundary

  2. not knowing the boundary

Source: NASA World Wind

Introduction to Machine Learning

Inductive bias

Complete the series! 2, 4, 6, 8, ...

Answer 1: 132 (model 1: $f(n) = n^4 - 10n^3 + 35n^2 - 48n + 24$)

Answer 2: 10 (model 2: $f(n) = 2n$)

How can we rule out the more complex one? (and many others)

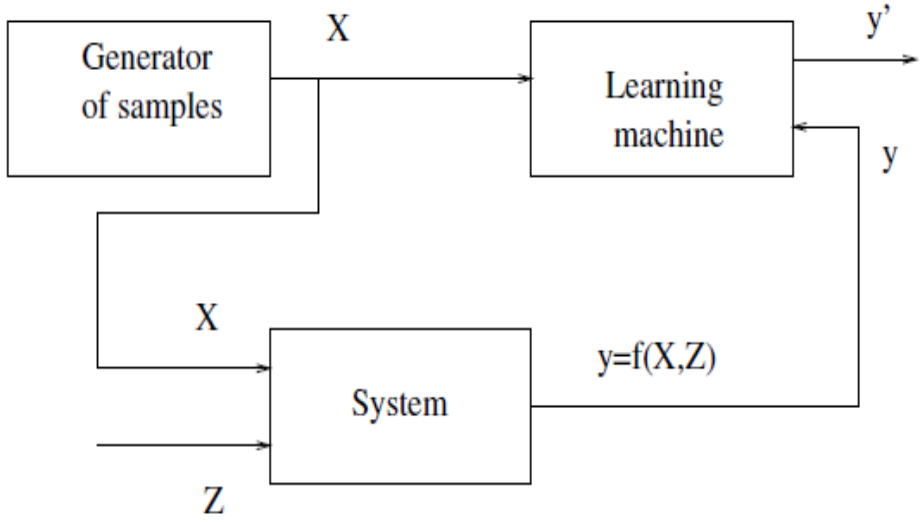
1. Supply more "training" data: 2, 4, 6, 8, 10, 12, 14, ...

2. Regularize: add a penalty to higher-order terms

3. Reduce the hypothesis space (e.g. restrict to quadratic models)

Introduction to Machine Learning

Formulation



X are the measured variables

Z are the non-measured variables

y is the true function

y' is the modeled function

Introduction to Machine Learning

Example 1a: classification of images [easy]

- Predict the class (a category) subject to little probabilistic uncertainty
- Little chance to "hand-craft" a solution, without learning
- Note heavy pre-processing

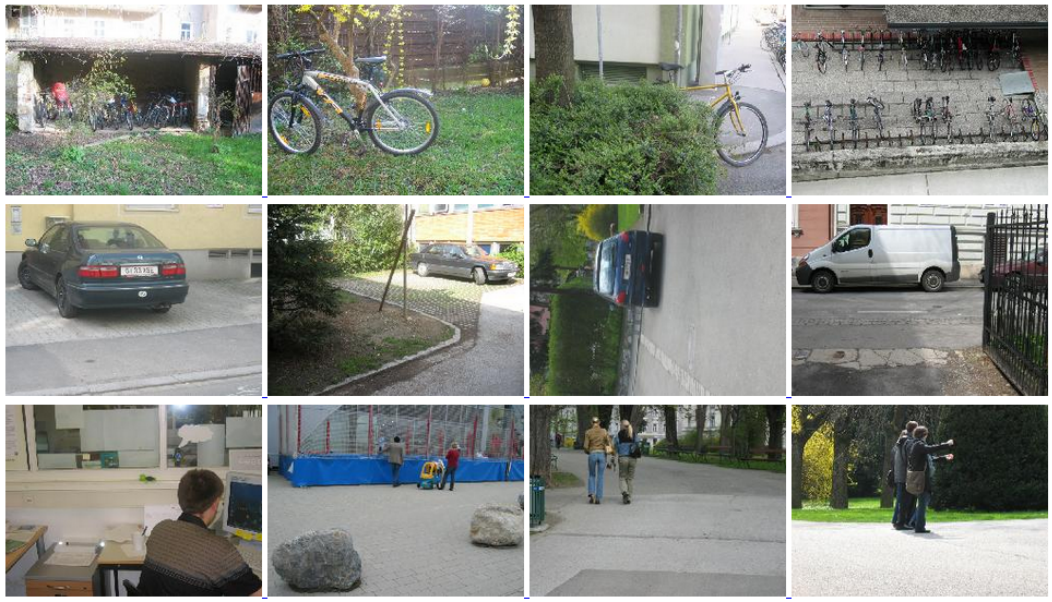
Introduction to Machine Learning

Example 1b: classification of images [medium]

- Predict the class (a category) subject to little probabilistic uncertainty, but larger variety
- Negligible chance to "hand-craft" a solution, without learning
- Note some pre-processing, unique label

Introduction to Machine Learning

Example 1c: classification of images [hard]

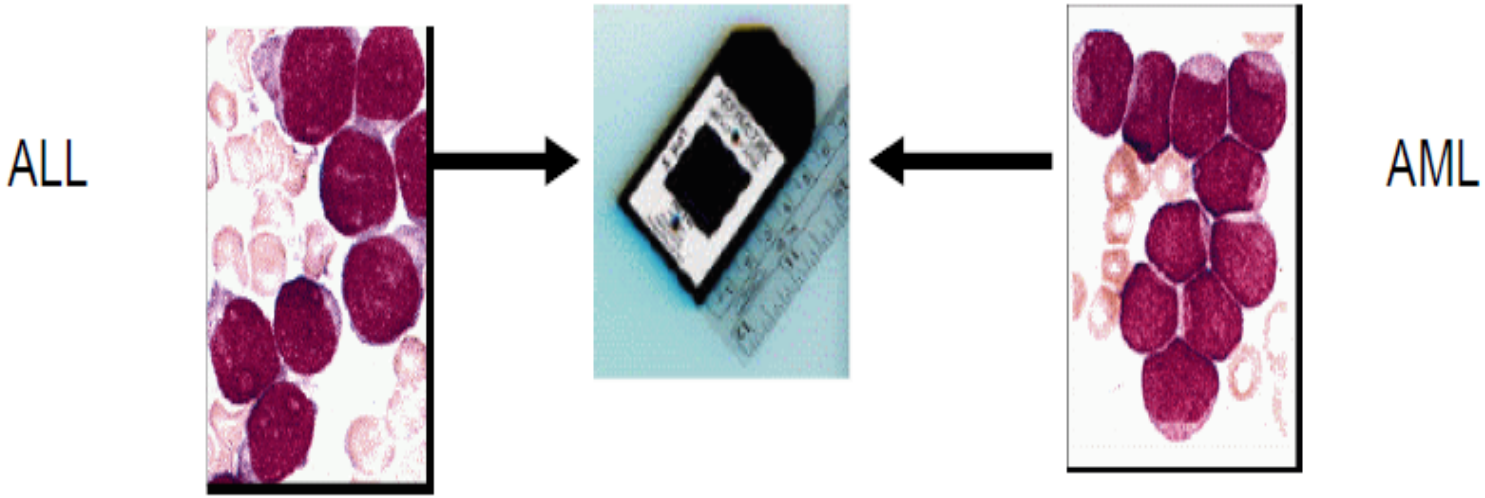


- Predict the class (a category) subject to some probabilistic uncertainty, general background, position, sixze, occlusion, illumination, ...
- Null chance to "hand-craft" a solution, without learning
- Note no pre-processing, multi-label

(all images from `http://pascallin.ecs.soton.ac.uk/challenges/VOC/databases.html`)

Introduction to Machine Learning

Example 2: classification of Leukemia types

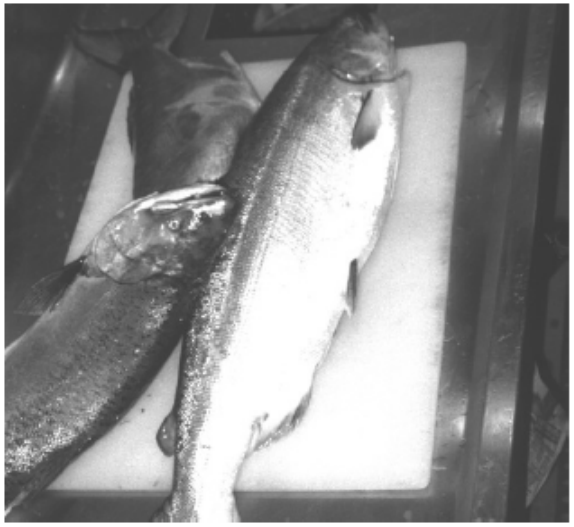


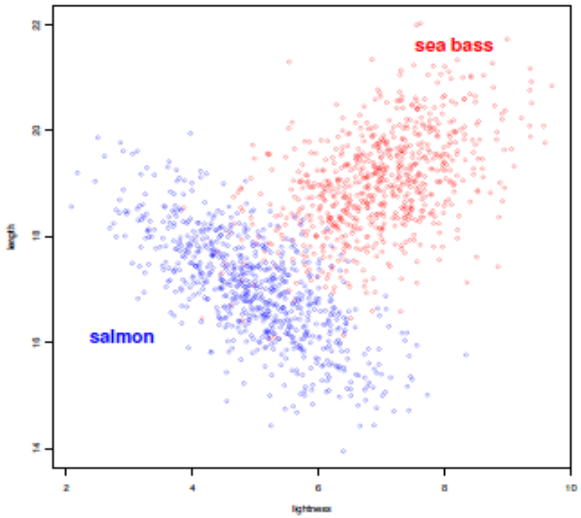
- 38 training instances, 34 test instances, $\sim 7,000$ genes
- ALL: Acute Lymphoblastic Leukemia; AML: Acute Myeloid Leukemia
- Results on test data: 33/34 correct (the 1 error may be misslabeled)
- Very small dataset, high-dimensional, large irrelevance and redundancy, probably subject to some probabilistic uncertainty

Introduction to Machine Learning

Example 3: fish classification

- A fish processing plant wants to automate the process of sorting incoming fish according to species (salmon or sea bass)
- The system consists of a conveyor belt, a robotic arm, a vision system with an overhead CCD camera and a computer.
- After some preprocessing, each fish is characterized by two features: average lightness and length: subject to heavy probabilistic uncertainty



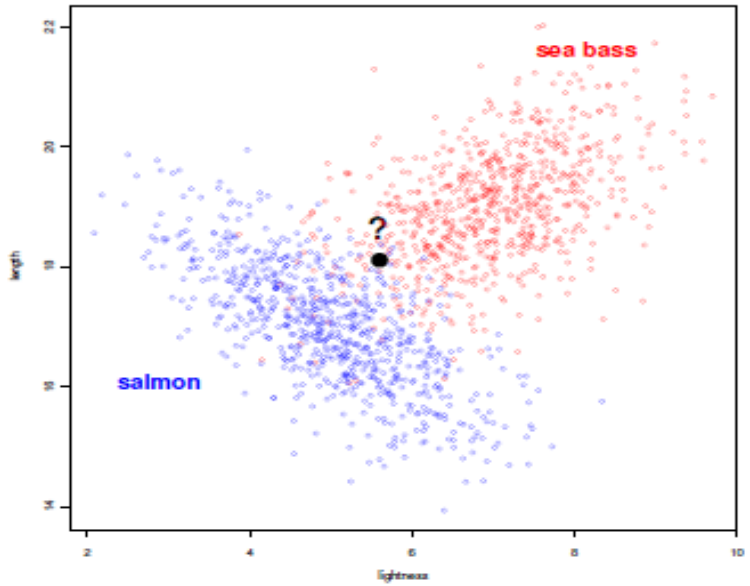


(from *Pattern Classification*)

Introduction to Machine Learning

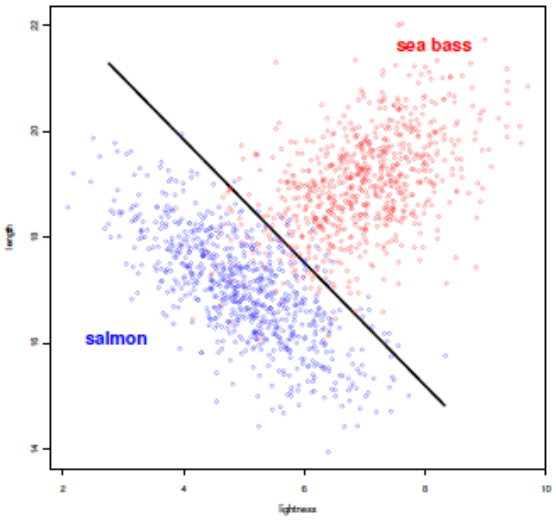
Example 3: fish classification

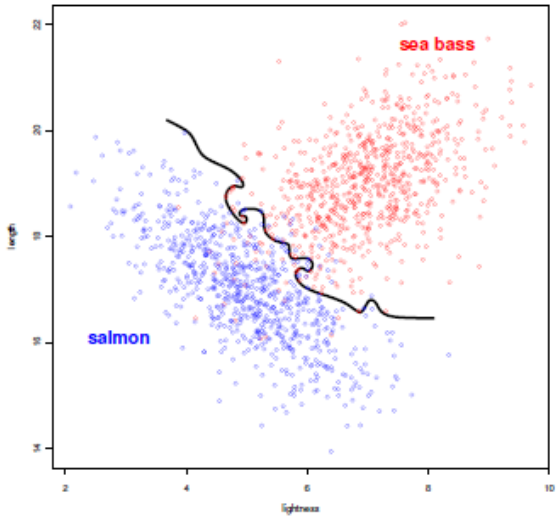
Given labeled training data coming from some unknown joint probability distribution, should we predict the new point as salmon or sea bass?

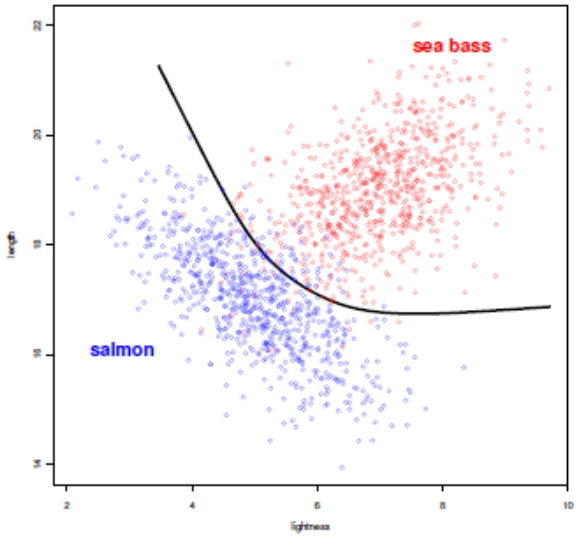


Introduction to Machine Learning

Example 3: fish classification



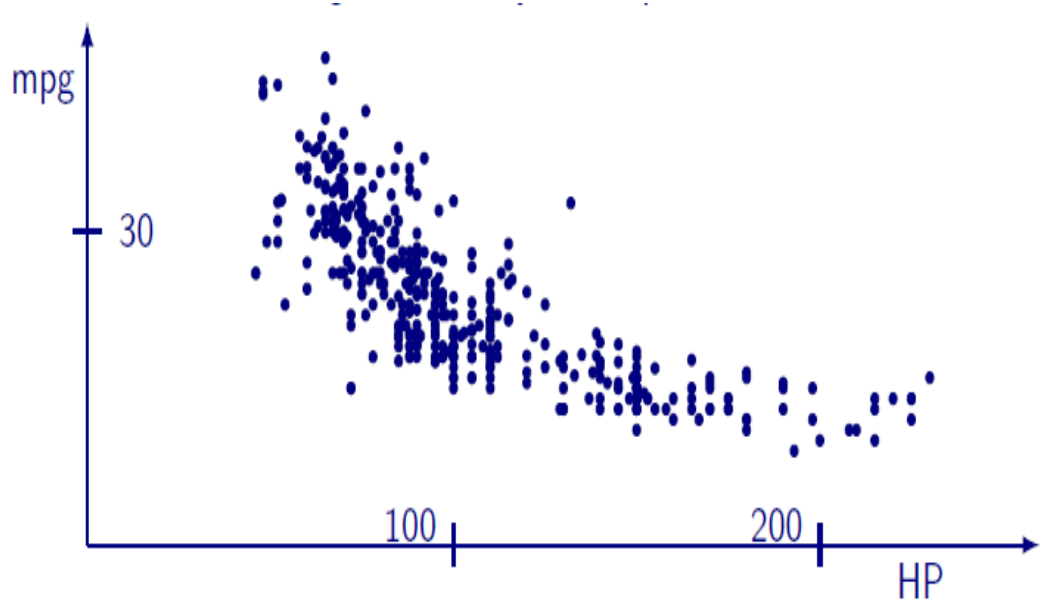




The **goal** is to obtain a model based on available **training data** (*known* examples) with high classification accuracy on unseen *unknown* examples (**test data**), i.e. achieving good **generalization**

Introduction to Machine Learning

Example 4: regression

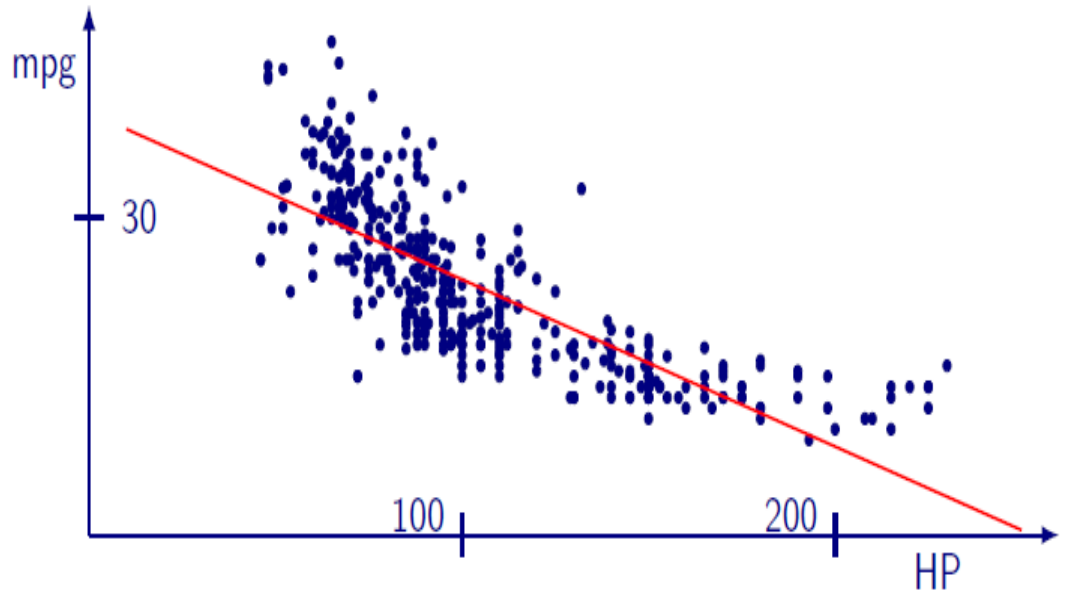


- Predict some quantitative outcome subject to probabilistic uncertainty
- Example: predict gas mileage (`mpg`) of a car as a function of horsepower (`HP`)

(`auto-mpg` data set from UCI Machine Learning Repository)

Introduction to Machine Learning

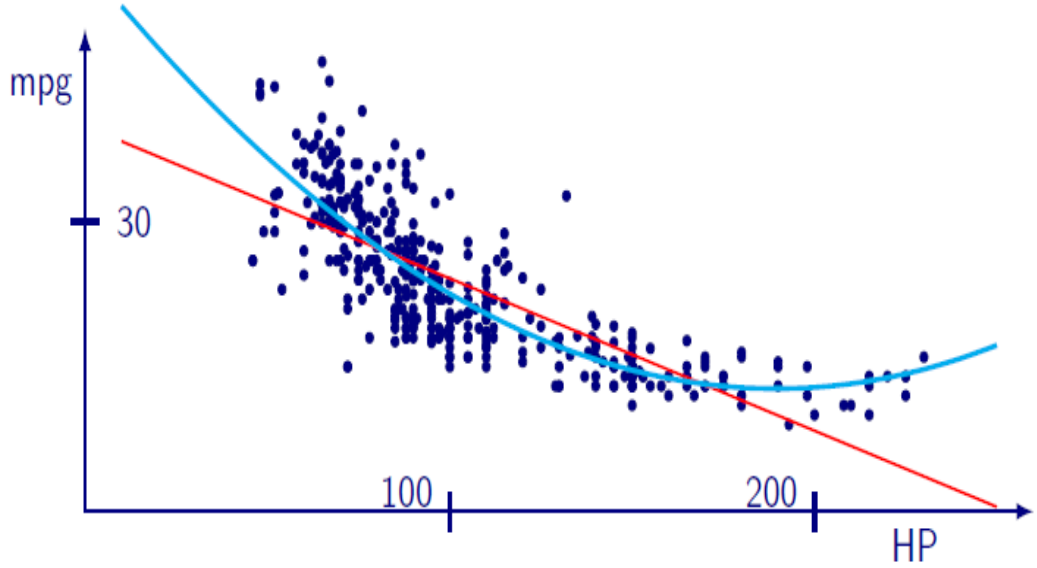
Example 4: regression



We can start by fitting a straight line to explain the relationship ...

Introduction to Machine Learning

Example 4: regression

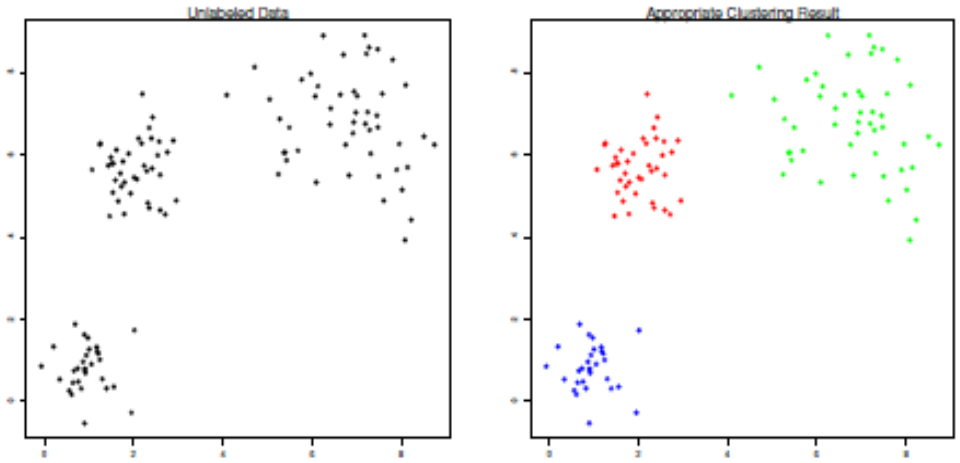


We can then fit a quadratic function ...

Is it a better fit? Will it lead to better predictions?

Introduction to Machine Learning

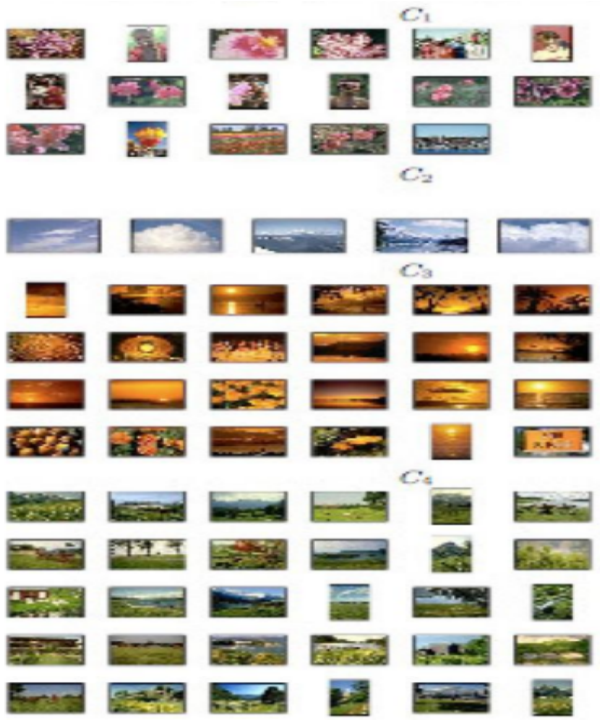
Example 5a: clustering [easy]



- Group unlabeled data into (non?)-overlapping subsets (**clusters**), according to a similarity measure
- Large intra-cluster (**within**) similarity and small inter-cluster (**between**) similarity
- Many times similarity is just the inverse of a (metric) distance

Introduction to Machine Learning

Example 5b: clustering [medium]



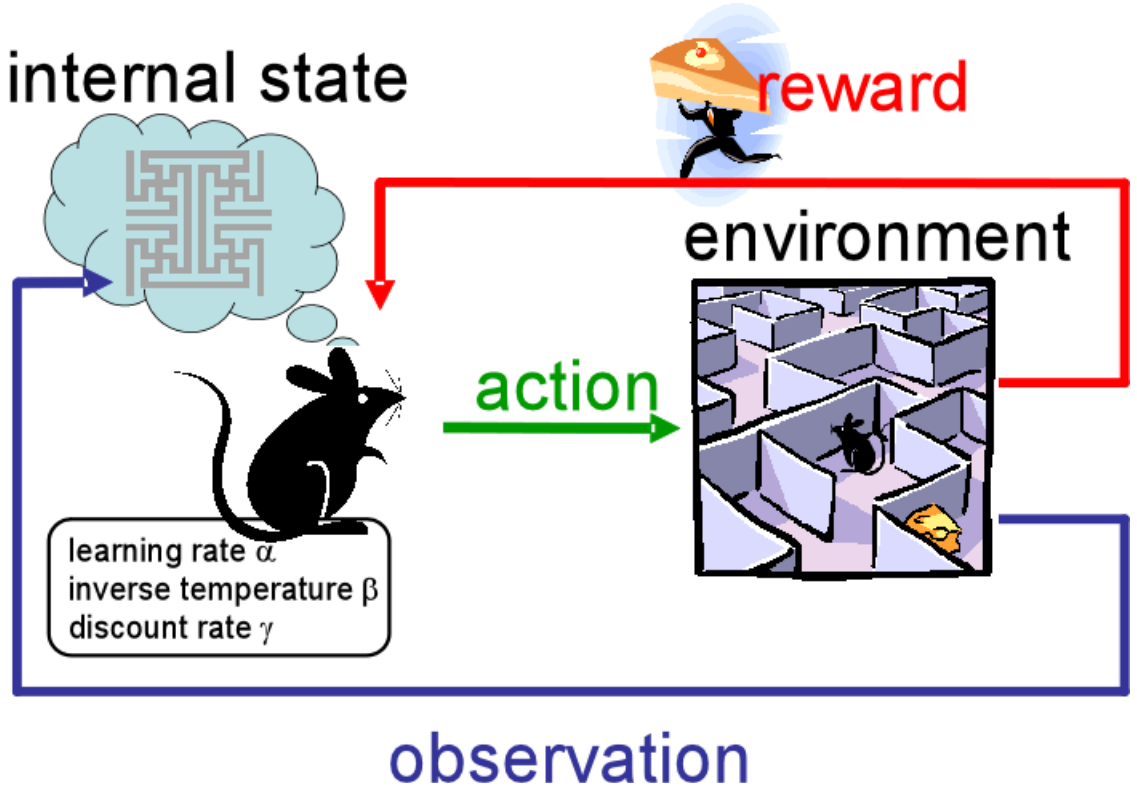
Introduction to Machine Learning

Example 6: clustering

Difficulty with clustering: large subjectivity ... how should we group them?

Introduction to Machine Learning

Example 7: reinforcement learning



- No supervised output but sequence of (delayed) rewards
- Example: robot in a maze

Introduction to Machine Learning

Example 8: Dimensionality reduction

- Each image has thousands or millions of pixels

Fig. 3. Images of faces (*11*) mapped into the embedding space described by the first two coordinates of LLE. Representative faces are shown next to circled points in different parts of the space. The bottom images correspond to points along the top-right path (linked by solid line), illustrating one particular mode of variability in pose and expression.

- Can we give each image a coordinate, such that (only) similar images are nearby?

The LLE algorithm

Introduction to Machine Learning

The Rosetta stone

Machine Learning	Statistics
model	model
parameter/weight	parameter/coefficient
training	fitting
learning	modelling
regression	regression
classification	discrimination
clustering	clustering/classification
inputs/features/variables	independent variables explanatory variables predictors
outputs/targets	dependent variables response variables
instances/examples	individuals/observations
error/loss function	fit criterion

Introduction to Machine Learning

Prediction vs. Inference

Prediction: produce a good estimate for the predicted variable

Inference:

1. Which predictors actually affect the predicted variable?

2. How strong are these dependencies?

3. Are these relationships positive or negative?

Introduction to Machine Learning

Example: Direct mailing

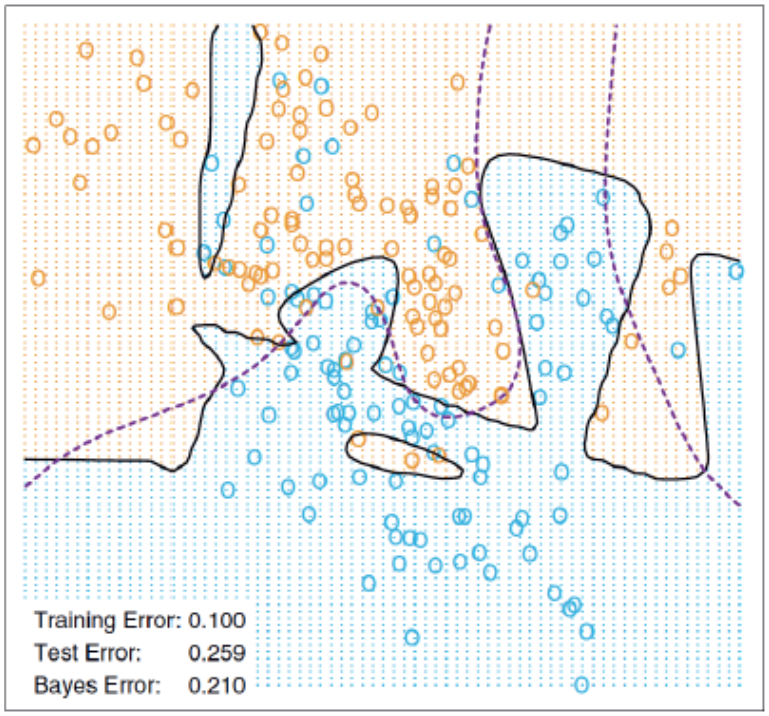
Predicting how much money an individual will donate (the response) based on observations from $90,000$ people on which we have recorded over 400 different characteristics (the predictors)

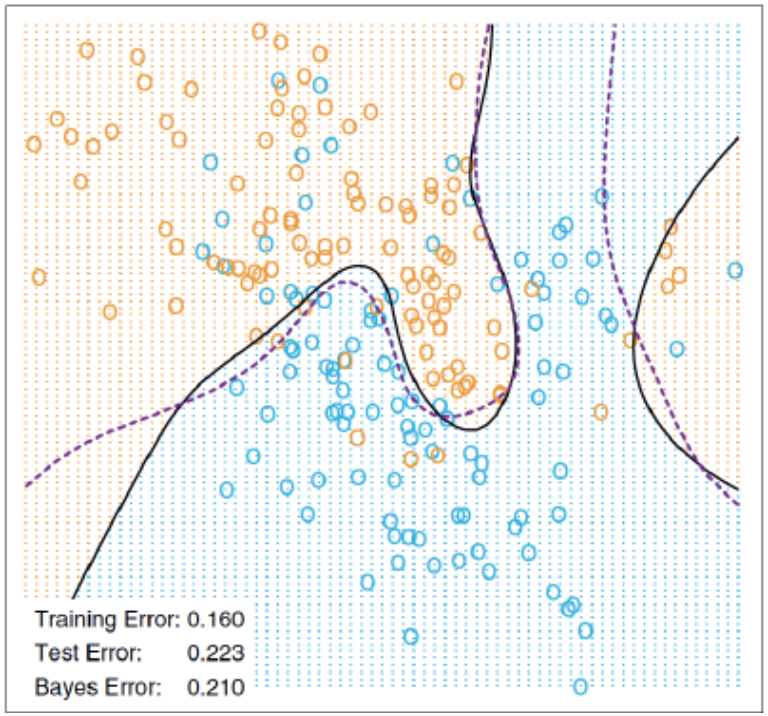
What do we pretend?

1. For a given individual should I send out an e-mail (yes/no)?
2. What is the probability that a specific individual will donate?
3. What is the expected donation for a specific individual?
4. What are the characteristics more strongly linked to donation?
5. How much increase in donation is associated with a given increase in a specific predictor?

Introduction to Machine Learning

Model selection



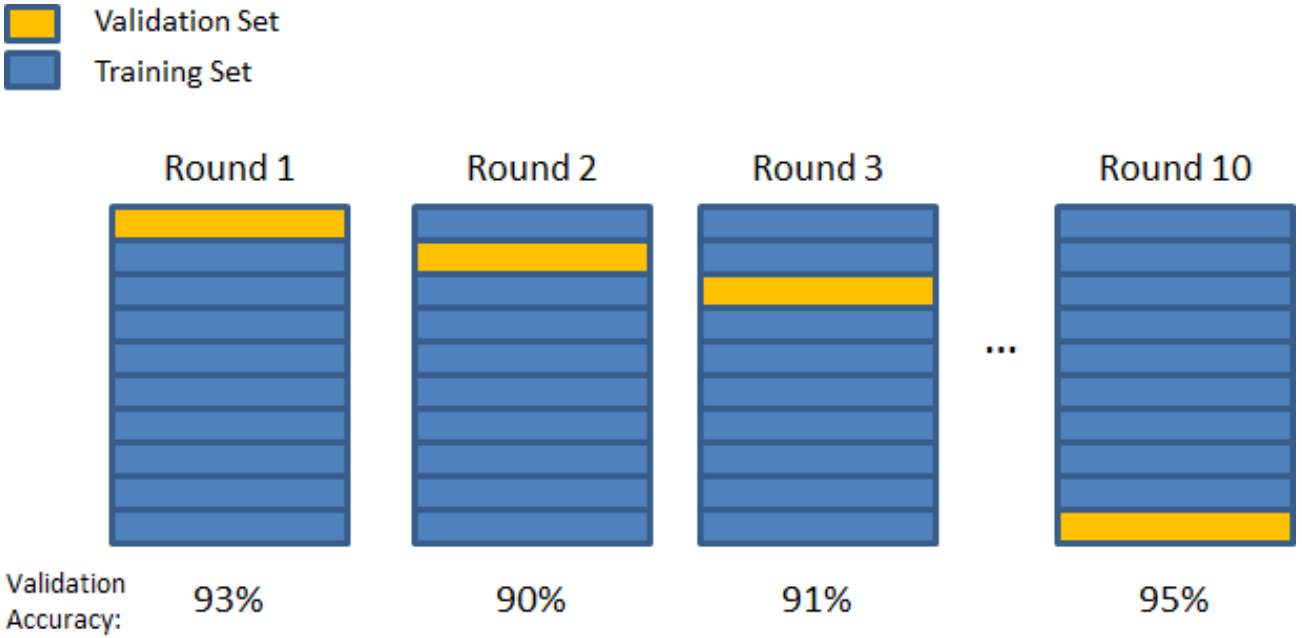


We need to perform three different tasks:

1. **Fit** models to data (estimate coefficients)
2. **Choose** one of these models (based on prediction error)
3. **Estimate** the true performance of the chosen model

Introduction to Machine Learning

Resampling methods



(from `https://chrisjmccormick.files.wordpress.com/2013/07/10_fold_cv.png`)

Introduction to Machine Learning

Resampling methods

Method:

- **Training data** is used to fit models
- **Validation data** is used to average prediction errors and choose the model with the lowest prediction error
- The chosen model is **refit** using the full training data
- **Test data** is used to estimate true performance of the chosen model

Introduction to Machine Learning

On data pre-processing

Each problem requires a different approach in what concerns data cleaning and preparation. This pre-process is very important because it can have a deep impact on performance; it can easily take you a significant part of the time.

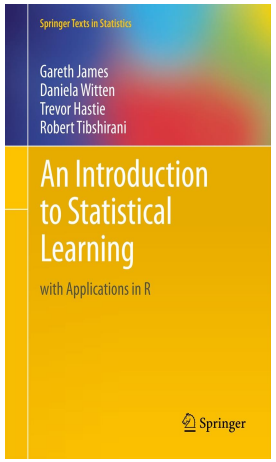
1. treatment of lost values (missing values)
2. treatment of anomalous values (outliers)
3. treatment of incoherent or incorrect values
4. coding of non-continuous or non-ordered variables
5. possible elimination of irrelevant or redundant variables (feature selection)
6. creation of new variables that can be useful (feature extraction)
7. normalization of the variables (*e.g.* standardization)
8. transformation of the variables (*e.g.* correction of serious skewness and/or kurtosis)

Introduction to Machine Learning

Recommended reading

- A free online version of *An Introduction to Statistical Learning, with Applications in R* by James, Witten, Hastie and Tibshirani (Springer, 2013) is available from January 2014.

- Springer has agreed to this, so no need to worry about copyright. However, you may not distribute printed versions of this pdf file.



`http://www-bcf.usc.edu/~gareth/ISL/`

Machine Learning

Some very nice books

Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer, 2006
`http://research.microsoft.com/~cmbishop/PRML`

Pattern Classification, Richard 0. Duda and Peter E. Hart and David G. Stork, Wiley-Interscience, 2001, sec. ed.
`http://rii.ricoh.com/~stork/DHS.html`

The Elements of Statistical Learning (10th edition) Hastie, Tibshirani and Friedman (2009). Springer-Verlag.
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Introduction to Machine Learning, Ethem Alpaydin (3rd Ed.), The MIT Press, 2015
`https://www.cmpe.boun.edu.tr/~ethem/i2ml3e/`

Introduction to Machine Learning

Required software

R and RStudio are freely available:

- R is an open source software for statistical computing and publication quality graphics and a very usable programming language (mix of imperative, OO and functional)

  Get R from `http://cran.r-project.org/`

- RStudio is a friendly IDE for R (Windows, Mac, and Linux)

  Get RStudio from `http://www.rstudio.com/`

Introduction to Machine Learning

Recommended background

1. **Probability theory:** fundamental in dealing with the uncertainty inherent in real problems (Bayes formula, multivariate Gaussian distribution, conditioning, ...)

2. **Basic statistics:** Maximum Likelihood, bias & variance, ...

3. Basic **real analysis** and **linear algebra**

4. Basic **matrix theory**

5. Fluency in some high-level **programming** language

Machine Learning

Syllabus

1. Introduction to Machine Learning

2. Theoretical issues (I): regression

3. Linear regression and beyond

4. Theoretical issues (II): classification

5. Generative classifiers

6. Discriminative classifiers

7. Clustering: k-means and E-M

8. Learning with kernels (I): The SVM

9. Learning with kernels (II): Kernel functions

10. Artificial neural networks (I): Delta rule, MLP-1

11. Artificial neural networks (II): MLP-2, RBF

12. Artificial neural networks (III): DL and CNNs

13. Ensemble methods: Random Forests

14. Advanced topics and frontiers