TSIA-SD210 - Machine Learning

Lecture 1 - Introduction to Statistical Machine Learning

Florence d'Alché-Buc

Contact: florence.dalche@telecom-paristech.fr,
Télécom ParisTech France

Table of contents

Outline

Outline

AlphaGo Program Beats the European Human Go Champion

Last Jan 27 2016, for the first time, a machine learning program beat a human Go Champion in a real size grid. The machine learning program used Reinforcement Learning + deep learning (neural networks).

Go, a complex game popular in Asia, has frustrated the efforts of artificial-intelligence researchers for decades.

ARTIFICIAL INTELLIGENCE

Google masters Go

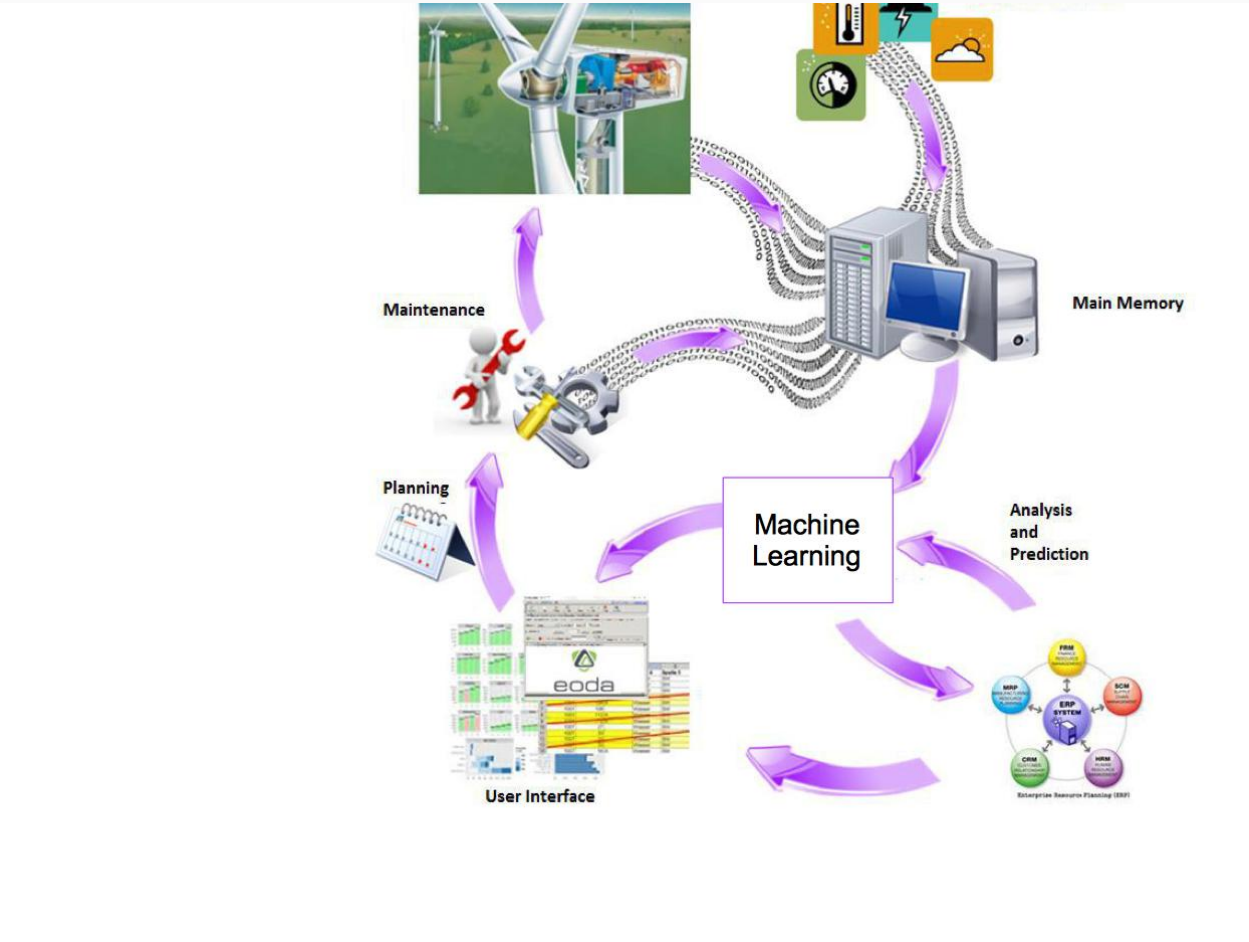
Deep-learning software excels at complex ancient board game.

AlphaGo: Ref: `http://www.nature.com/news/google-ai-algorithm-masters-ancient-game-of-go-1.19234`

▸ Read more

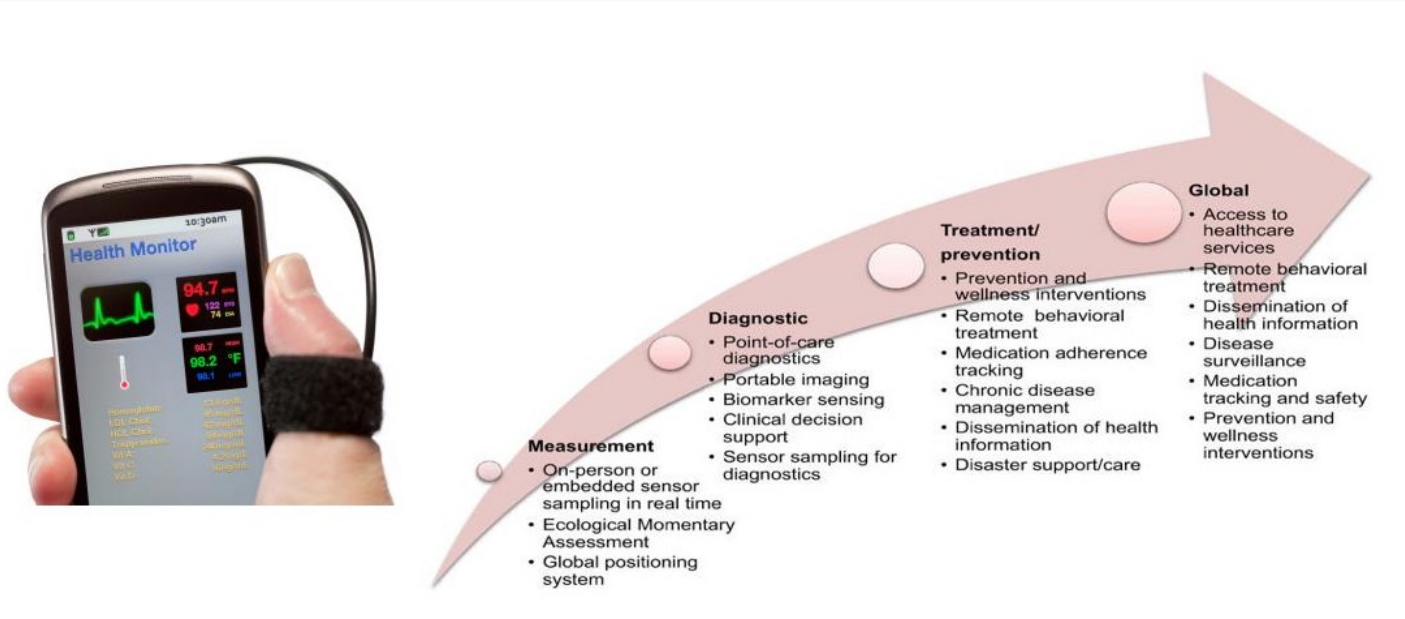
Predictive Maintenance

In manufacturing, data streaming from single components or entire pieces of equipment can used to predict the possibility of future failures, allowing the arrival of new components to be synchronised with that of the repair technician.



Read more: `http://www.itproportal.com/2015/08/01/predictive-analytics-will-shape-future-every-sector/#ixzz3zloPifwW`

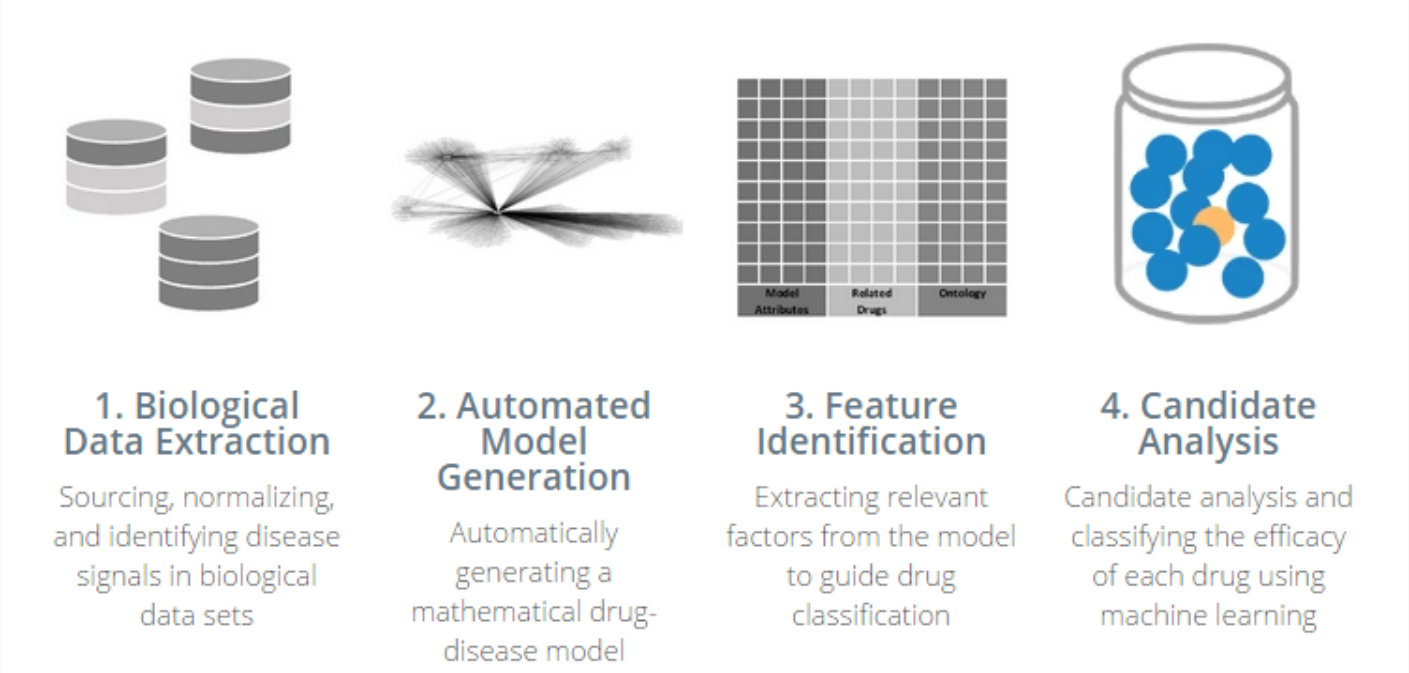
Mobile health monitoring



Read more: Figure Published in final edited form as: Am J Prev Med. 2013 August; 45(2) : 228 – 236..

Drug discovery

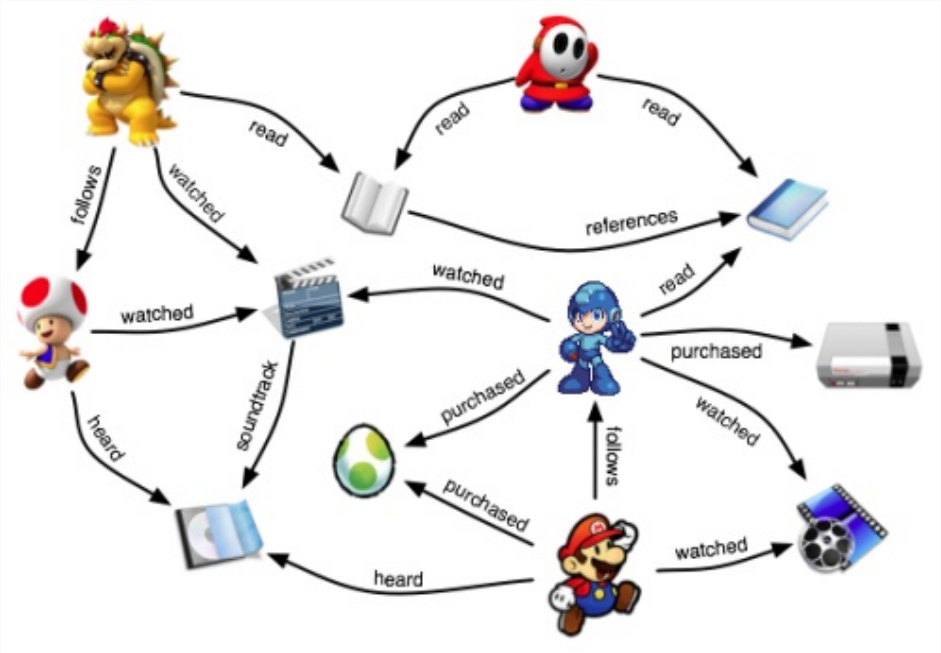
Drug-discovery has been revolutionized by Machine Learning.



Read more: ▸ Link

Drug Discovery Today Volume 20, Number 3 March 2015. A. Lavecchia.

Recommendation system

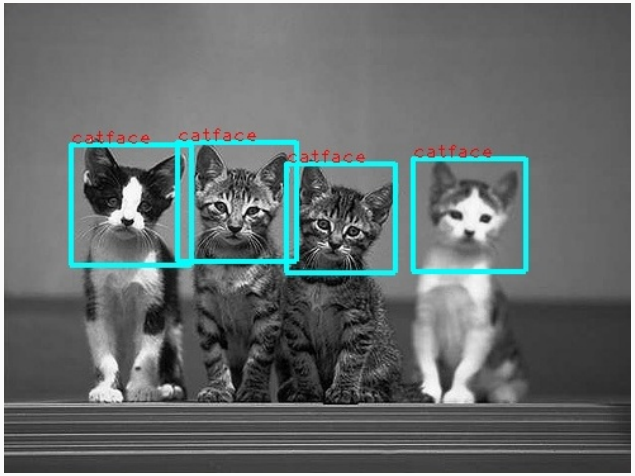


- "People read about 10 MB worth of material a day, hear 400MB a day and see 1MB of information every second"-The economist, Nov 2006.
- "We are leaving the age of information and entering the age of recommendation", Chris Anderson, Wired Magazine.

Read more: ▸ Link

Systems recommendation tutorial. X. Amatriain. RECSYS'14.

Object recognition

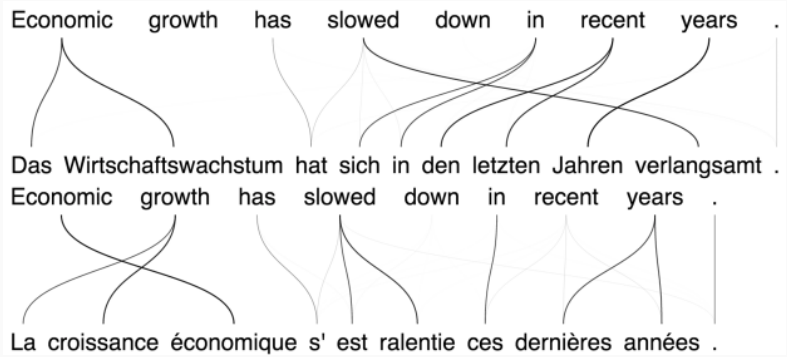


Read more: Link 1

Tuto Slides from Fei-Fei Li

and Link 2 for instance: website of Ivan Laptev

Machine Translation



Read more: ▸ Link

Introduction to Neural Machine Translation with GPUs. Kyunghyun Cho.

Machine Learning everywhere !

Use data to extract a prediction function

- Search engine, text-mining
- Diagnosis, Fault detection
- Business analytics
- Prediction in Heath care, Personalized medecine
- Social networks, link prediction, recommendation

Outline

A definition of Machine Learning

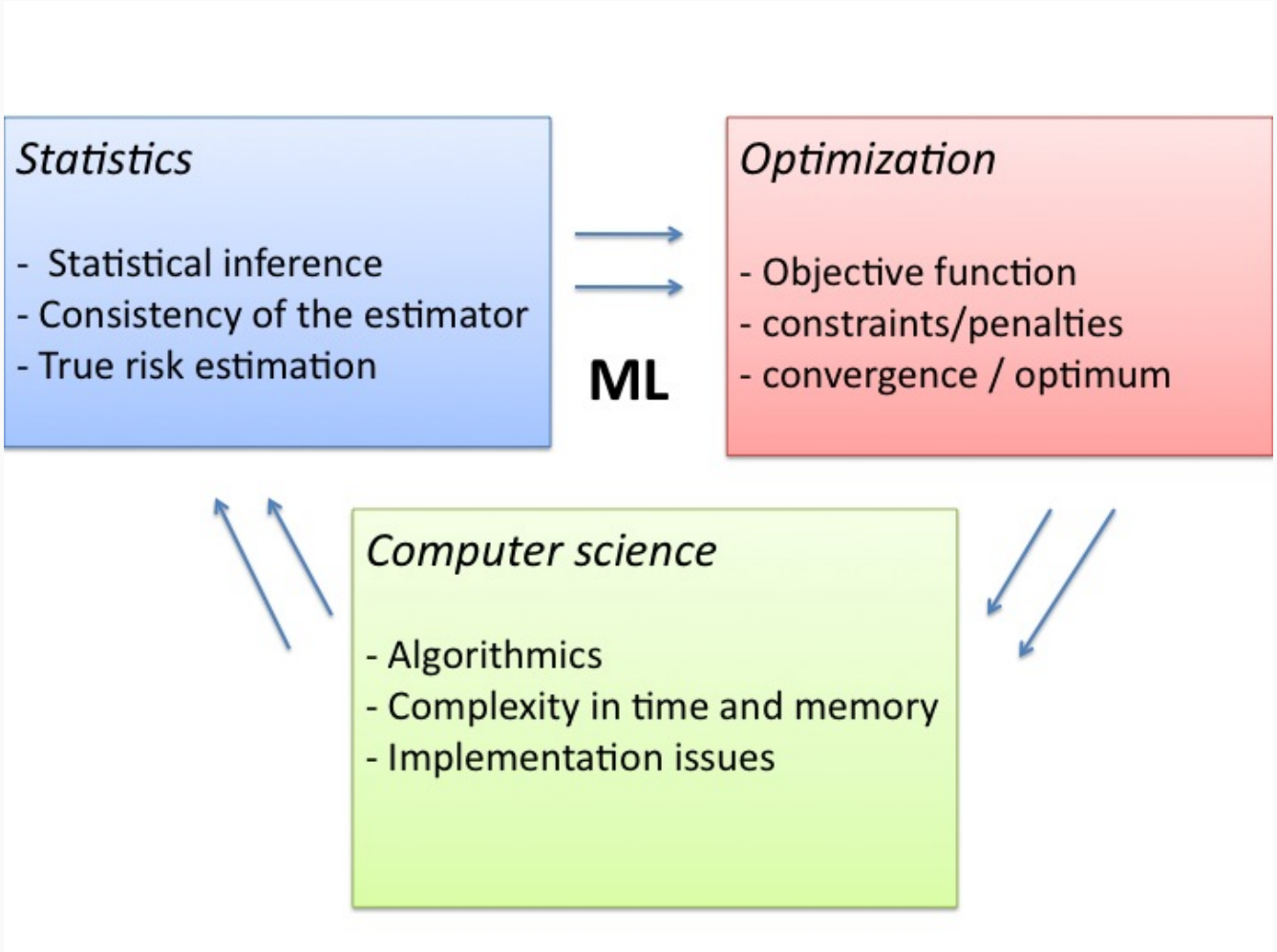
A type of artificial intelligence (AI) that provides computers with the ability to do certain tasks, such as recognition, diagnosis, planning, robot control, prediction, etc., without being explicitly programmed. It focuses on the development of algorithms that can teach themselves to grow and change when exposed to new data.

Experience, tasks and performance measure

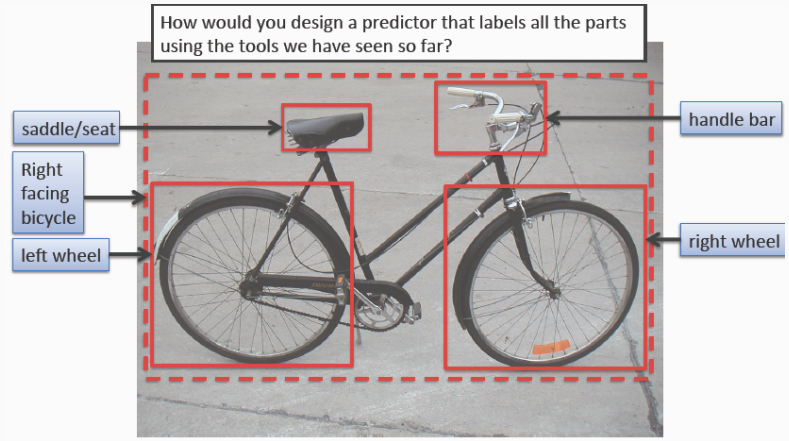
A definition by Tom Mitchell (`http://www.cs.cmu.edu/~tom/`
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T , as measured by P, improves with experience E.

- **Experience** : data provided off-line or on-line
- **Tasks** : pattern recognition, diagnostic, complex system modelling, game player, robot learning,...
- **Performance measure** : accuracy on new data, ability to generalize

Machine Learning



Example 1: object recognition in an image



- Read a data file
- Recognize if parts of the target object are present
- **Goal**: say if an object is present or not in the image.

First type of learning

Offline or batch learning: *the learning algorithm gets a datafile and outputs some function that can be used in turn on new data*

- pattern recognition (a wide panel of applications)
- diagnosis (health, plants)
- link prediction in networks
- data-mining
- social networks analytics

This course: **mainly batch learning.**

Example 2: a learning robot

Robot endowed with a set of sensors and a online learning algorithm:

- Sense the environment, act and measure the effect of action
- Goal: play football

Second type of learning

Online learning: *the learning algorithm keeps on interacting with the environment*

- robotics
- predictive maintenance
- security in cloud servers
- personalized advertising
- autonomous cars
- personalized healthcare
- security systems

Machine Learning

- Off-line learning
- Online learning

More and more, initialization with off-line learning and continuous update with online learning.
Important to understand well off-line learning before handling online learning

Outline

Machine learning : statistical or symbolic learning

In the 80's, there was a debate about how to address Machine Learning.

- Symbolic learning
  - At that time, associated to Artificial Intelligence : about logical inference
  - **Goal of learning**: learn logical rules that are consistent (in a logical sense) with observed facts and given rules
  - Interest: interpretability $\neq$ a black box
- Numerical Learning / Connexionism (at this stage, statistical learning was not really born)
  - **Goal of learning**: learn weights (parameters) of neural networks to fit observed data
  - Interest: robustness to noise, efficiency of learning (stochastic gradient descent)
  - At the end of the 90's, connexionism has been replaced by **statistical learning**, giving a more general picture, conciliating machine learning with statistical inference

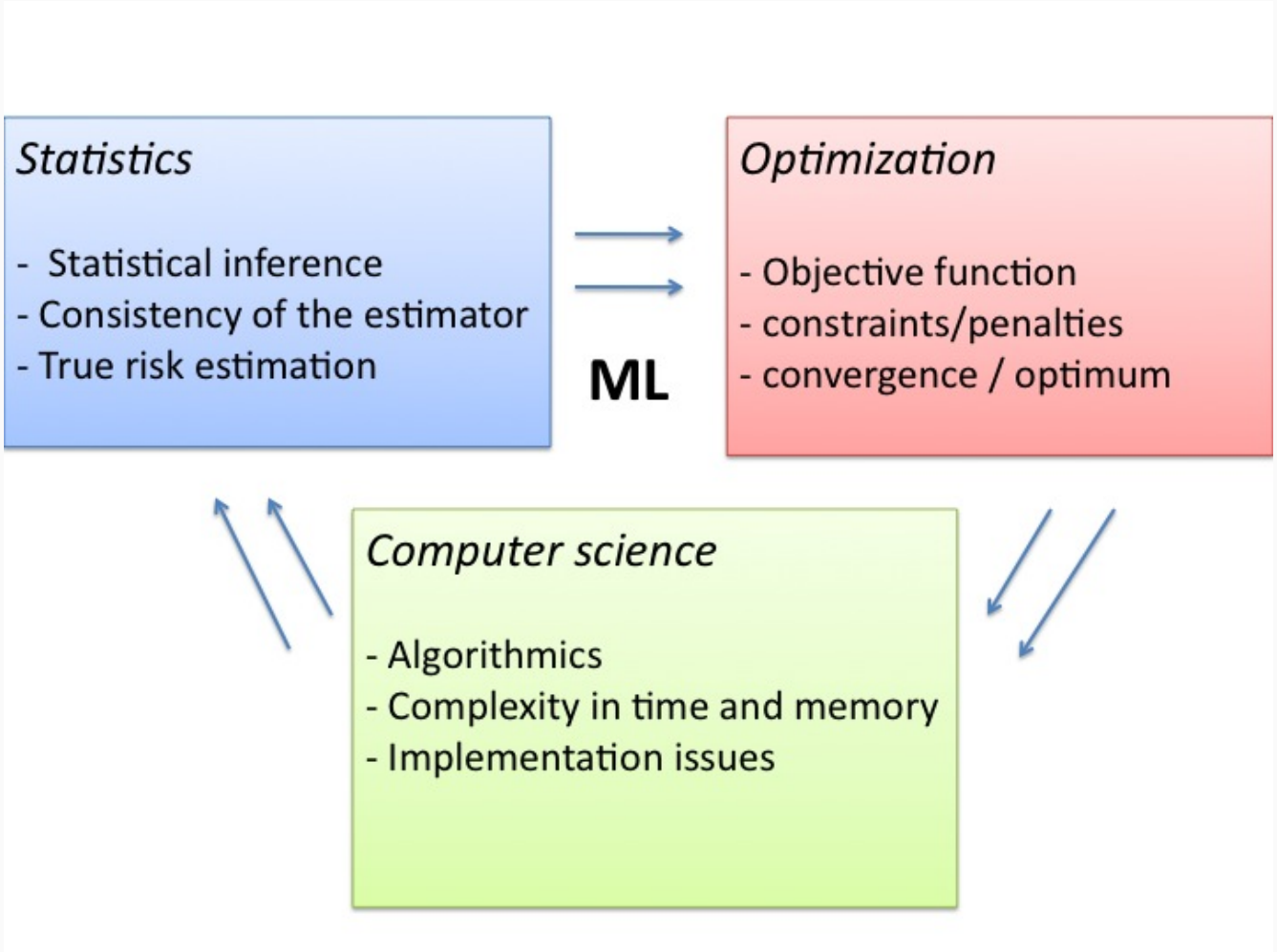
Slide form Rob Tibshirani (early 90's)

NEURAL NETS	STATISTICS
network	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation
optimal brain damage	model selection
large grant = $100,000	large grant= $10,000
nice place to have a meeting:	nice place to have a meeting:
Snowbird, Utah, French Alps	Las Vegas in August

Statistical Machine Learning

- we build learning algorithms: our algorithms provide estimators
- we are interested on some statistical properties like consistency of the estimators
- but also on the efficiency of the algorithms as optimization procedures.

Statistical Machine Learning

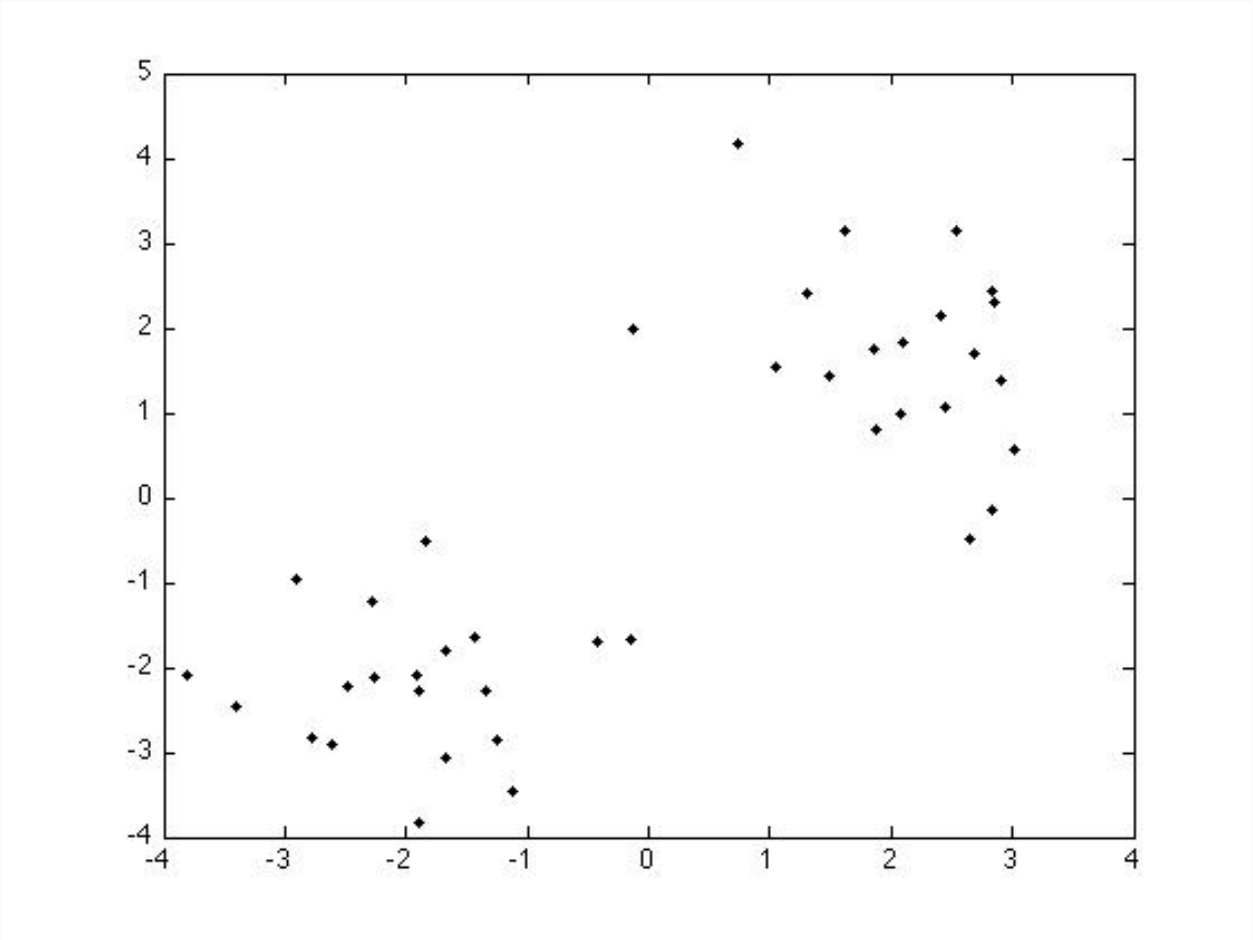


Supervised versus unsupervised learning

- **Supervised Learning (classification, regression)**:
  - Goal: Learn a function f to predict a variable y from an individual x.
  - Data: Learning set (x_i, y_i)
- **Unsupervised Learning (clustering, graphical model)**:
  - Goal: Discover a structure within a set of of individuals $\{x_i\}$.
  - Data: Training set $\{x_i\}$
- First case is better posed.
- Note: most of these algorithms can be implemented offline or online.

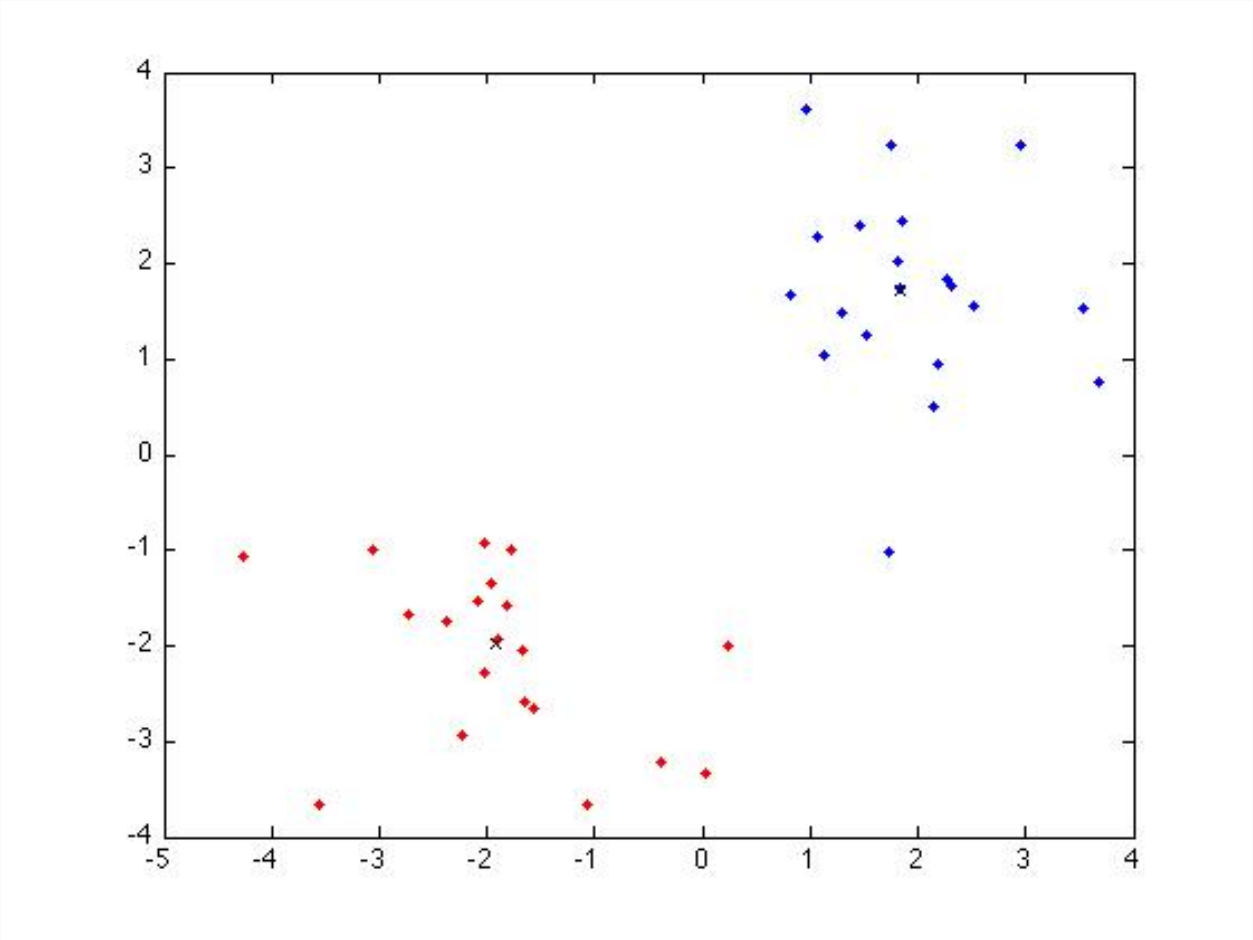
Example of clustering in 2D

Here are the data:

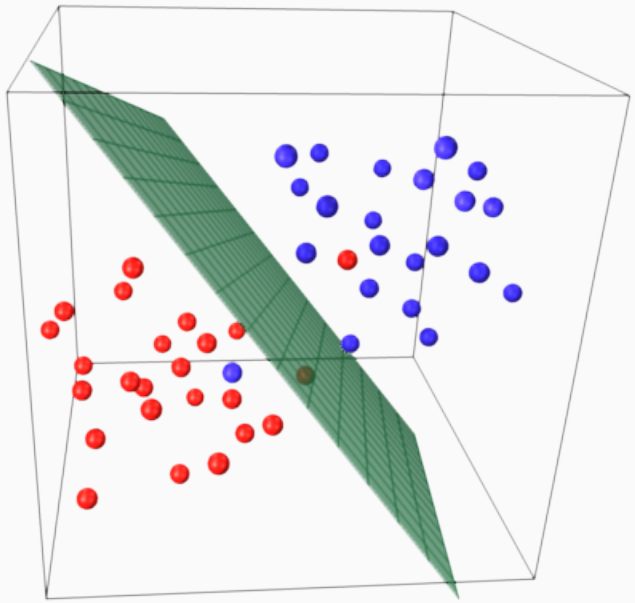


Example of clustering in 2D

Here are the data:



Example of supervised classification in 2D



Outline

Outline

Teaching team in Machine Learning

Lecturers

- Pietro Gori, associate prof. (image processing, machine learning)
- Umut Simsekli, associate prof. (computational Bayesian statistics, machine learning)
- Florence d'Alché, prof. (Machine learning)

Teaching Assistants

- Pietro Gori,associate prof. (Image processing, Machine learning)
- Giovanna Varni, associate prof. (social computing,Machine Learning)
- **Moussab Djerrab, PhD student, ML (Project/challenge)**
- Alexandre Garcia, PhD, Machine Learning
- Alexandre Lambert, PhD, Machine Learning

How to work for this course ?

Several books/sources can help you

1. The elements of statistical learning, Hastie, Tibshirani, Friedman, Springer (free pdf).
2. Pattern recognition and machine learning, Chris Bishop
3. Several video-lectures and online courses
4. Machine Learning Meetup in Paris every month ▸ ML Meetup

Evaluation of the course

- 2 practical session graded (binomes)
- 1 real project/challenge proposed this year by **BearingPoint** (4-student team)
- 1 exam (questions about the course)

Planning of the course

1. Feb 2, 2018: Introduction to Statistical Learning and presentation of the project
2. Feb 9, 2018: Statistical Learning methodology - Model selection
3. Feb 16, 2018: Kernel methods + Practical session 1
4. March 2, 2018: Decision Trees + Practical session 2
5. March 9, 2018: Ensemble methods + Practical session 3
6. March 16, 2018: Neural networks (towards Deep learning)
7. March 30, 2018: Practical session 4
8. April 6, 2018: exam + Project session
9. Later in April : de-brieffing, best projects

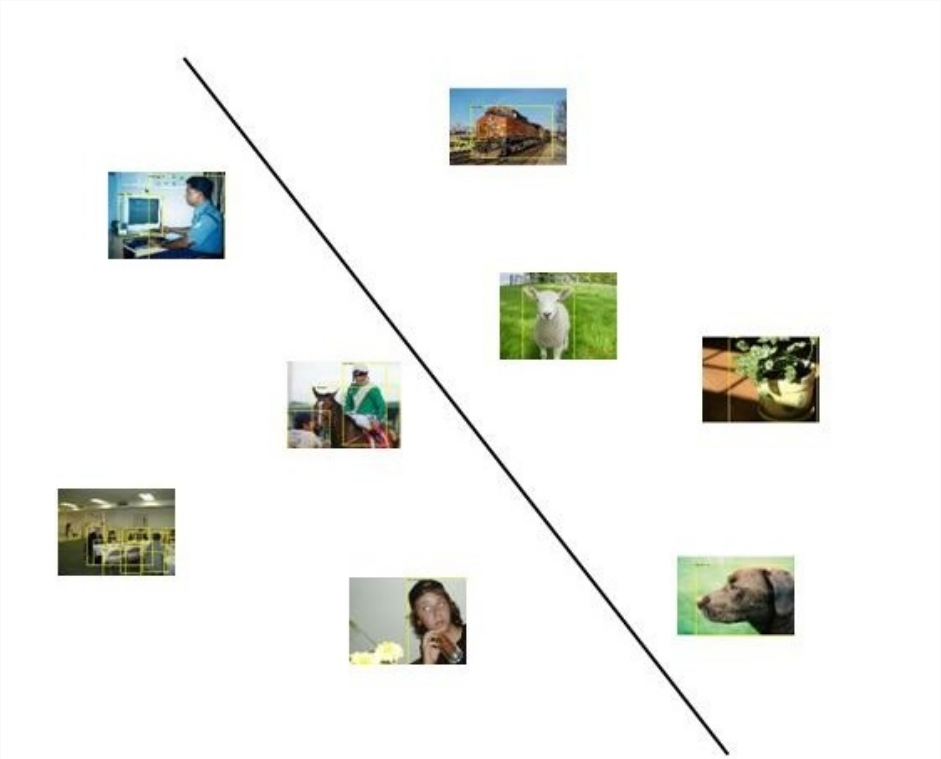
Outline

Outline

Outline

Outline

Goal of Supervised classification



- Build a software that automatically classify data into two classes
- Images with human versus images with no human

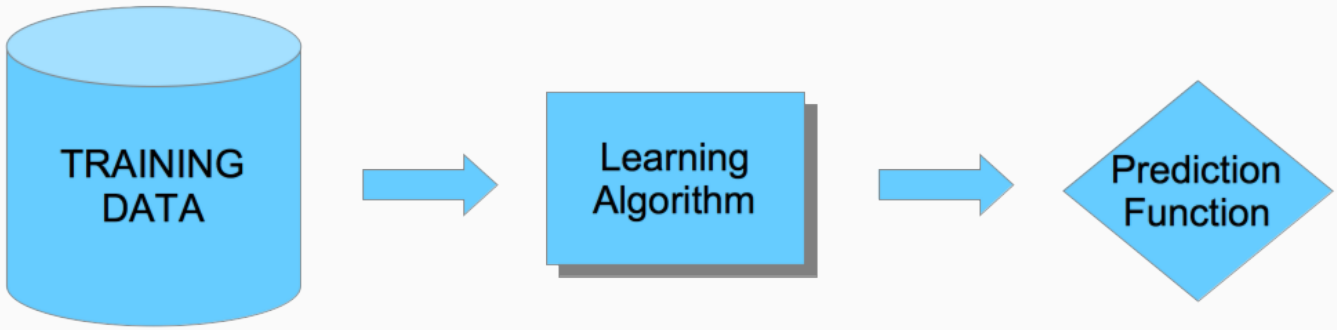
Use training dataset to define the classifier

Computer science/algorithmics

- Training dataset: $\mathcal{S}_n = \{(image, label)\} = \{(x_i, y_i), i = 1, \ldots, n\}$
- Define an algorithm $\mathcal{A}$ that takes the training dataset and provide a function that classifies the data
- At the end, two pieces of code:
  - a program that implements $\mathcal{A}$: in *scikitlearn* : `clf.fit(Xtrain, ytrain)`
  - a program that makes a prediction given some input (here an image) : `print(clf.predict([[-0.8, -1]]))`

Learning a classifier

a program that implements the learning algorithm $\mathcal{A}$



Prediction with a classifier

a program that makes a prediction given some input (here an image)

in *scikitlearn* : `print(clf.predict(newx))`

Learning: a statistical inference view

- Training dataset: $\mathcal{S}_n = \{(image, label)\} = \{(x_i, y_i), i = 1, \ldots, n\}$
- Define an estimation procedure that takes the training dataset and provide a function that classifies the data
- At the end,
  - What are the properties of the estimator ? Consistency ....
  - What error does make the estimated function ?

Learning: a mathematical programming / optimization view

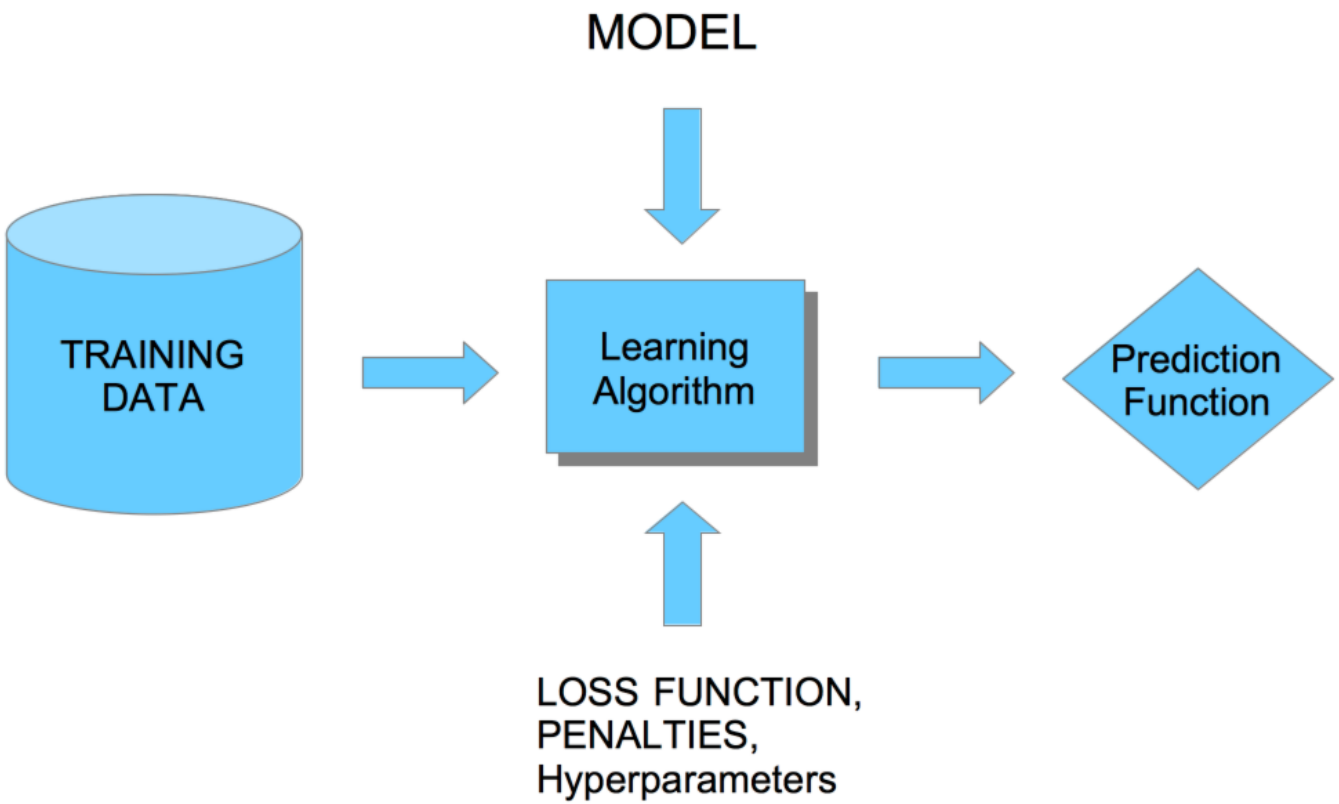
- Training dataset: $\mathcal{S}_n = \{(image, label)\} = \{(x_i, y_i), i = 1, \ldots, n\}$
- Define an optimization problem and then an optimization algorithm that takes the training dataset and provide a function that classifies the data
- At the end,
  - What are the properties of the optimization algorithm ? convergence towards a global/local minimum...
  - Complexity in time ... Complexity in memory...

What do we need to determine an image classifier?

- Choose a way to represent image
- Choose a family of classification functions
- Formulate the learning problem as an optimization one
- Define an optimization algorithm
- Evaluate the quality of the classifier learnt from data

Learning a classifier

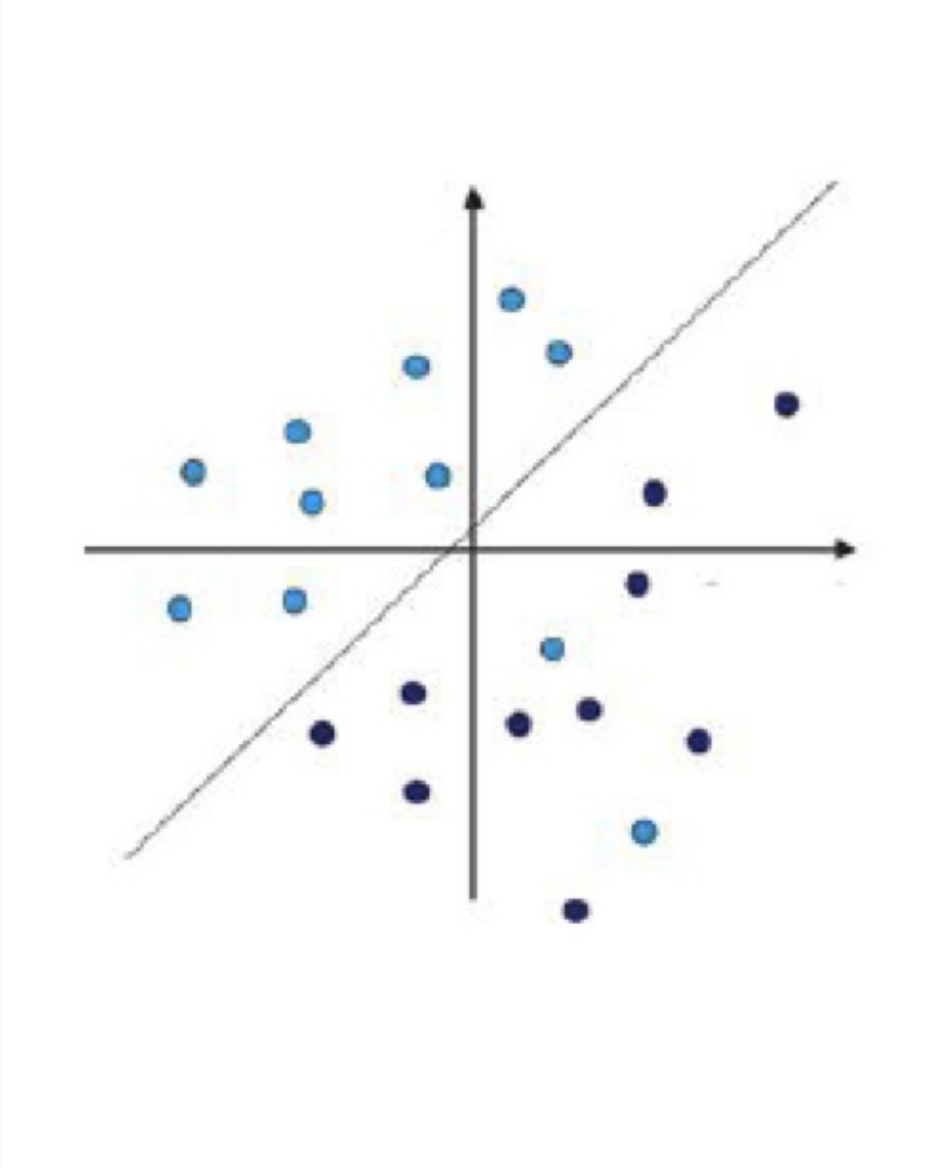
a program that implements the learning algorithm $\mathcal{A}$



What do we need to determine an image classifier?

- Choose a way to represent image (the input) : x: a grey-level matrix as a huge vector
- output : y : 0 or 1
- A classifier: linear or nonlinear ?
- Learning algorithm : minimizes some cost function
- Empirical measures: accuracy/ classification error, test error, Cross-validation

Linear classifier



Building an image classifier?

- n images
- Image i $\rightarrow$ a vector $\mathbf{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$
- Label: $y_i \in \{0, 1\}$
- A linear classifier: $h(\mathbf{x}) = s(w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p)$
- with $s(z) = \frac{1}{1+exp(-\frac{1}{2}z)}$, $z \in \mathbb{R}$
- Simple example: minimization of $\mathcal{L}(w; \mathbf{x_1}, \ldots, \mathbf{x_n}) = \frac{1}{n} \sum_{i=1}^{n} (h(\mathbf{x}_i) - y_i)^2$
- Find w such that $\mathcal{L}(w; \mathbf{x_1}, \ldots, \mathbf{x_n})$ be minimal

Outline

A probabilistic view of the learning problem (1): no data !

- Let's call X a random vector that takes its value in $\mathcal{X} = \mathbb{R}^p$
- X describes the properties (we say , features) of the objects
- Y a random variable that takes its value in $\mathcal{Y}$: Y encodes some output property
- $\mathcal{Y} = \mathbb{R}$ in case of regression
- $\mathcal{Y} = \{1, -1\}$ in case of binary supervised classification

A probabilistic view of the learning problem (2)

- Let's note $\mathcal{D}$, the class of measurable functions from $\mathcal{X}$ to $\mathcal{Y} \cup \mathbb{R}$.
- Given $\mathcal{H} \subset \mathcal{D}$ and a local loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the problem of supervised learning consists in solving the following optimization problem:
  - $\hat{h} = \arg min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]$
- Zero-One Loss: $\ell(h(x), y) = 1$ if $y \neq \text{sign}(h(x)))$, 0 otherwise.
- Margin (a criterion to be maximized) : $m(h(x), y) = y\text{sign}(h(x))$
- Equivalently : a loss to be minimized:
  $\ell(h(x), y) = \max(0, 1 - yh(x))$
- Prediction for x: take $\text{sign}(h(x))$

Binary Supervised Classification

- True risk (also called *generalization error*):$R(h) = \mathbb{E}_P[\ell(h(x), y)]$
- Find h that minimizes :

$$R(h) = \sum_{y=-1,1} P(Y = y) \int_{\mathbb{R}^p} \ell(h(x), y) p(x|Y = y) dx$$

Bayes Rule

Bayes rule

$$P(Y = k|x) = \frac{p(x|Y = k)P(Y = k)}{p(x|Y = -1).P(Y = -1) + p(x|Y = 1).P(Y = 1)}$$

$P(Y = k)$: prior probability
$P(Y = k|x)$: posterior probability of $Y = k$ given x
$p(x|Y = k)$: likelihood or probability density of x conditionnally to $Y = k$
Note that $P(Y = 1) + P(Y = -1) = 1$

Bayes classifier: a proposal for classification in an ideal world

Definition

$$h_{bay}(x) = argmax_{k=1,-1} P(Y = k|x)$$

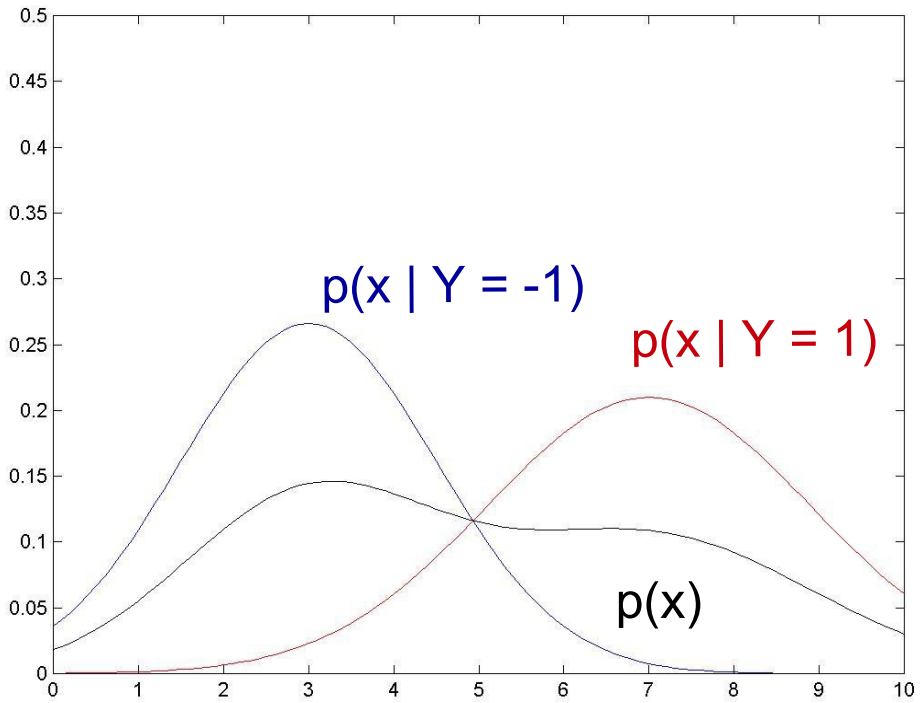
.

Bayes risk

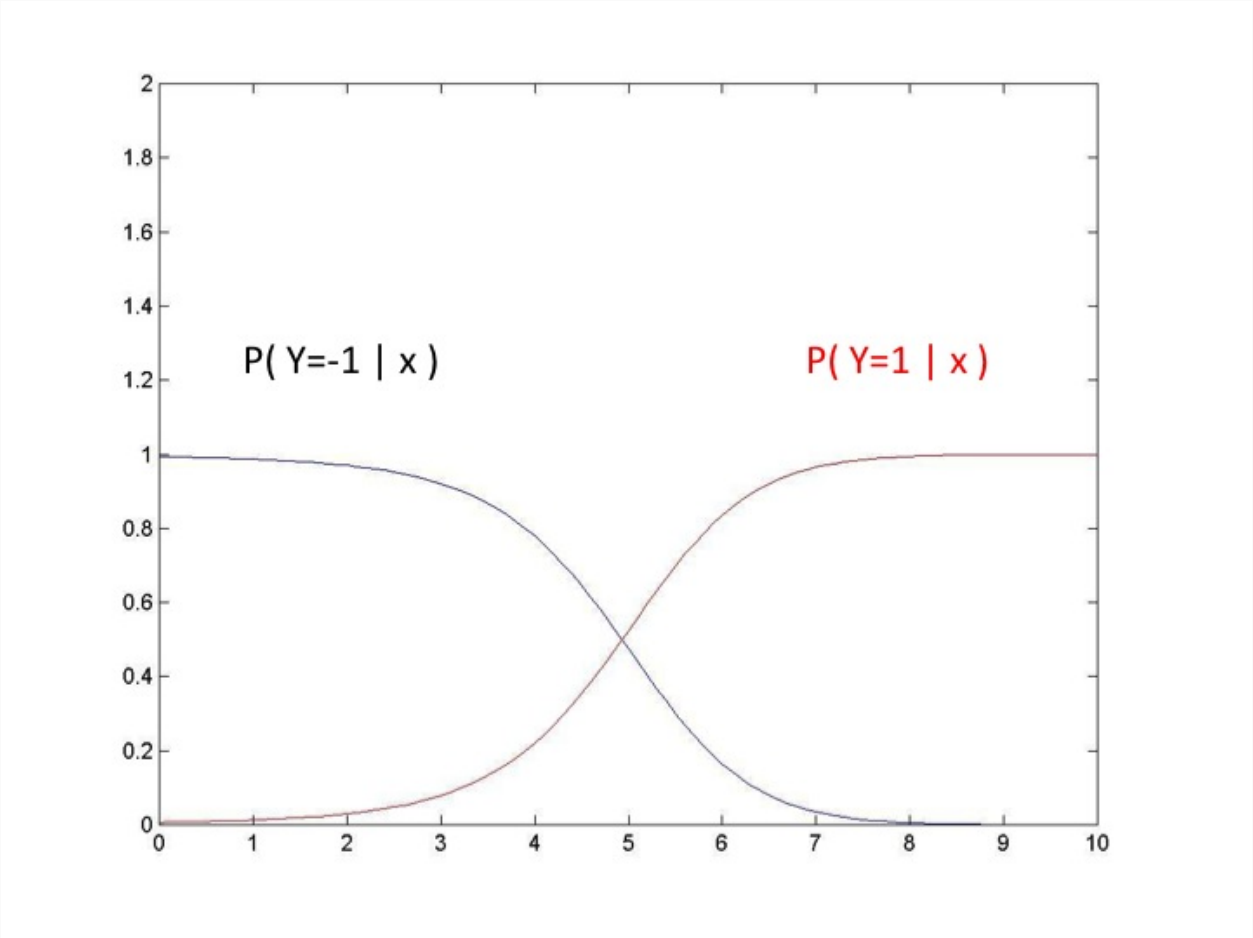
$$\begin{aligned} R(h_{bay}) &= \int_{R_1} P(h_{bay}(x) \neq 1)p(x)dx + \int_{R_{-1}} P(h_{bay}(x) \neq -1)p(x)dx \quad (1) \\ &= \int_{R_1} P(y = -1|x)p(x)dx + \int_{R_{-1}} P(y = 1|x)p(x)dx \quad (2) \end{aligned}$$

It can be shown that $R_{Bayes} = R(h_{Bayes})$ is minimal.

A 1D example with Gaussian probability distribution



Bayesian classifier: Gaussian probability distribution



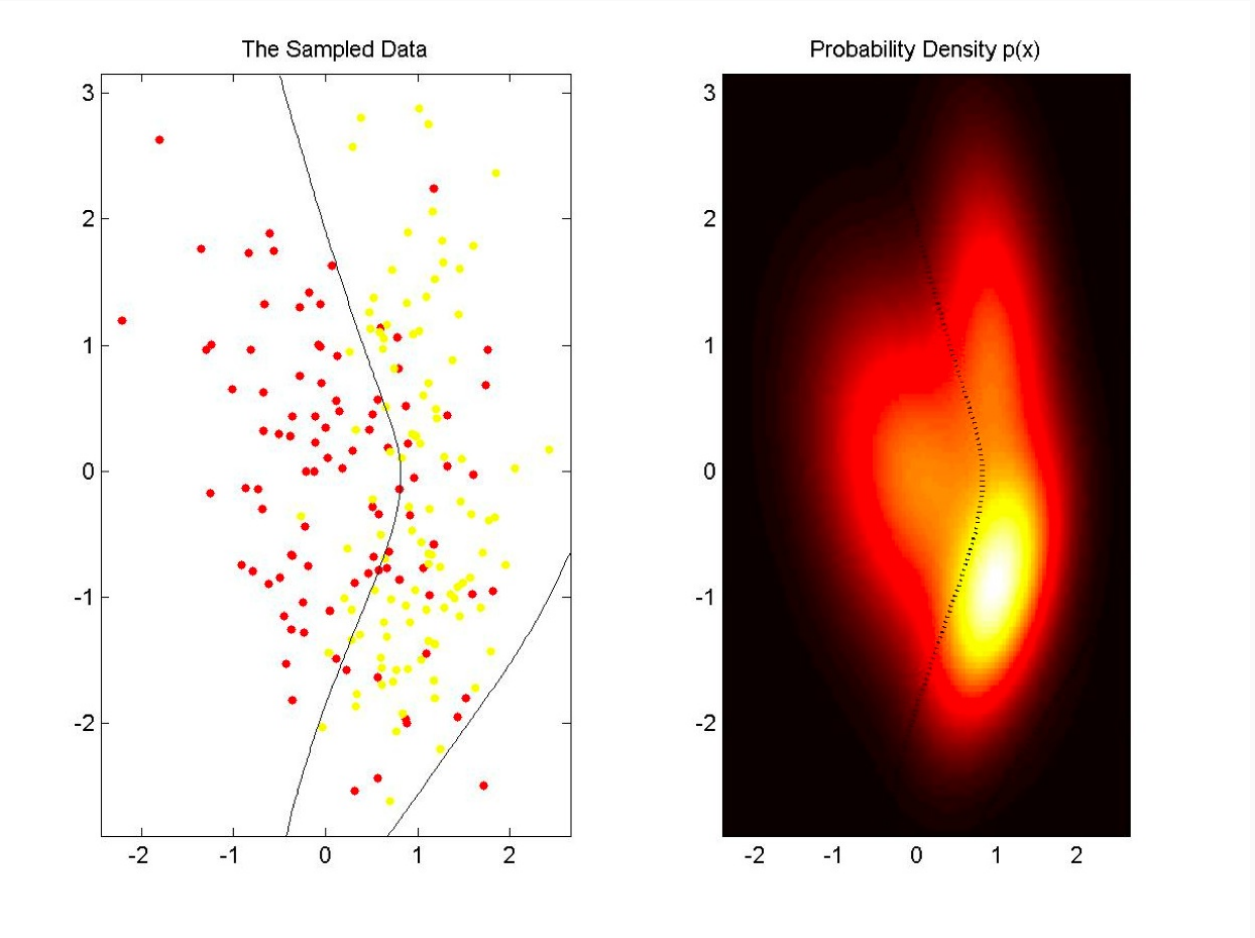
Exercise: what is the true risk of the Bayes Classifier ?

First take-home message

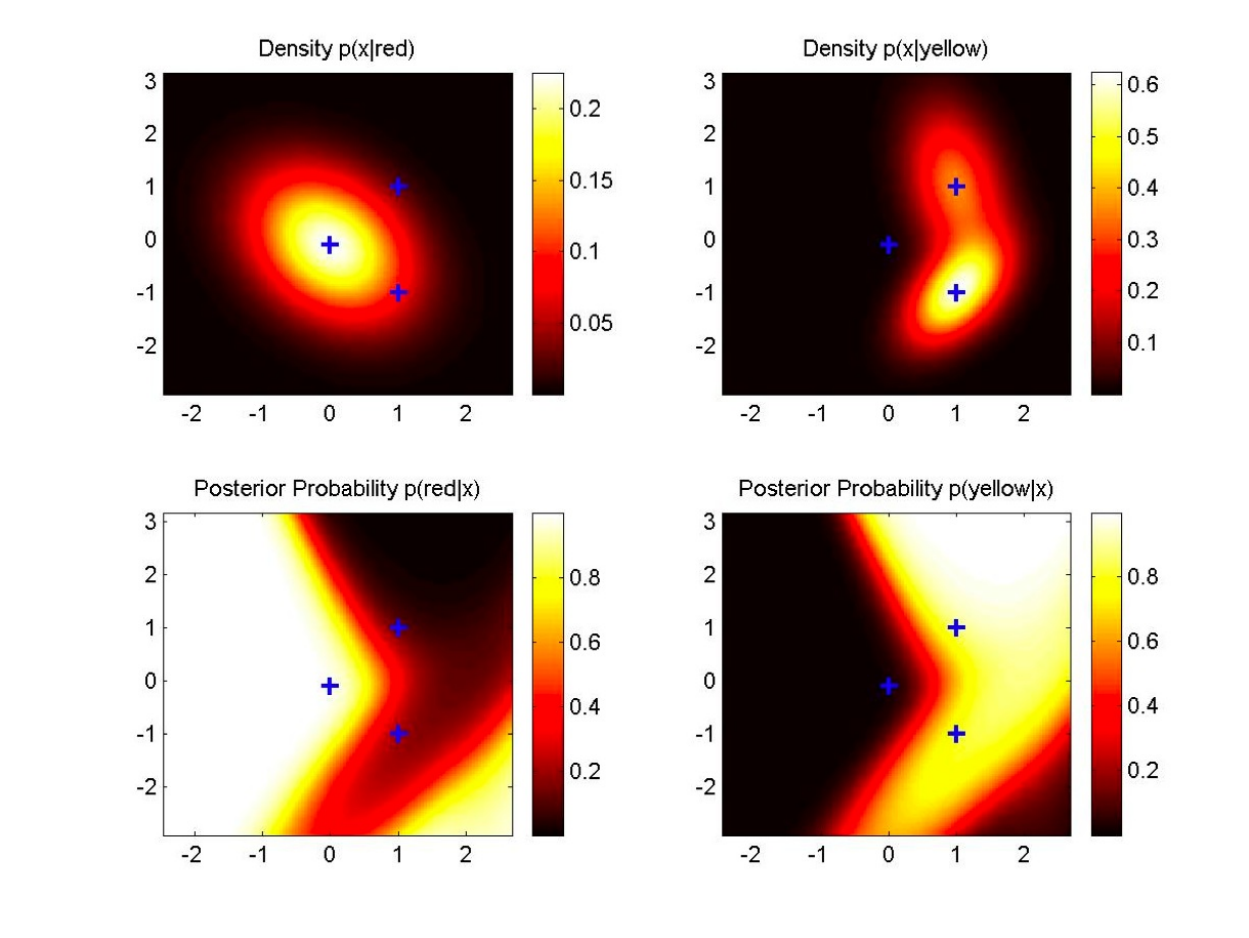
- The target function in supervised classification is the Bayes classifier for the $0-1$loss
- The target function in regression is $h(x) = \mathbb{E}[Y|x]$ for the square loss
- Now we call h_{target} the true target function

Exercise: *prove that* $\mathbb{E}[Y|x]$ *is the target function for the square loss.*

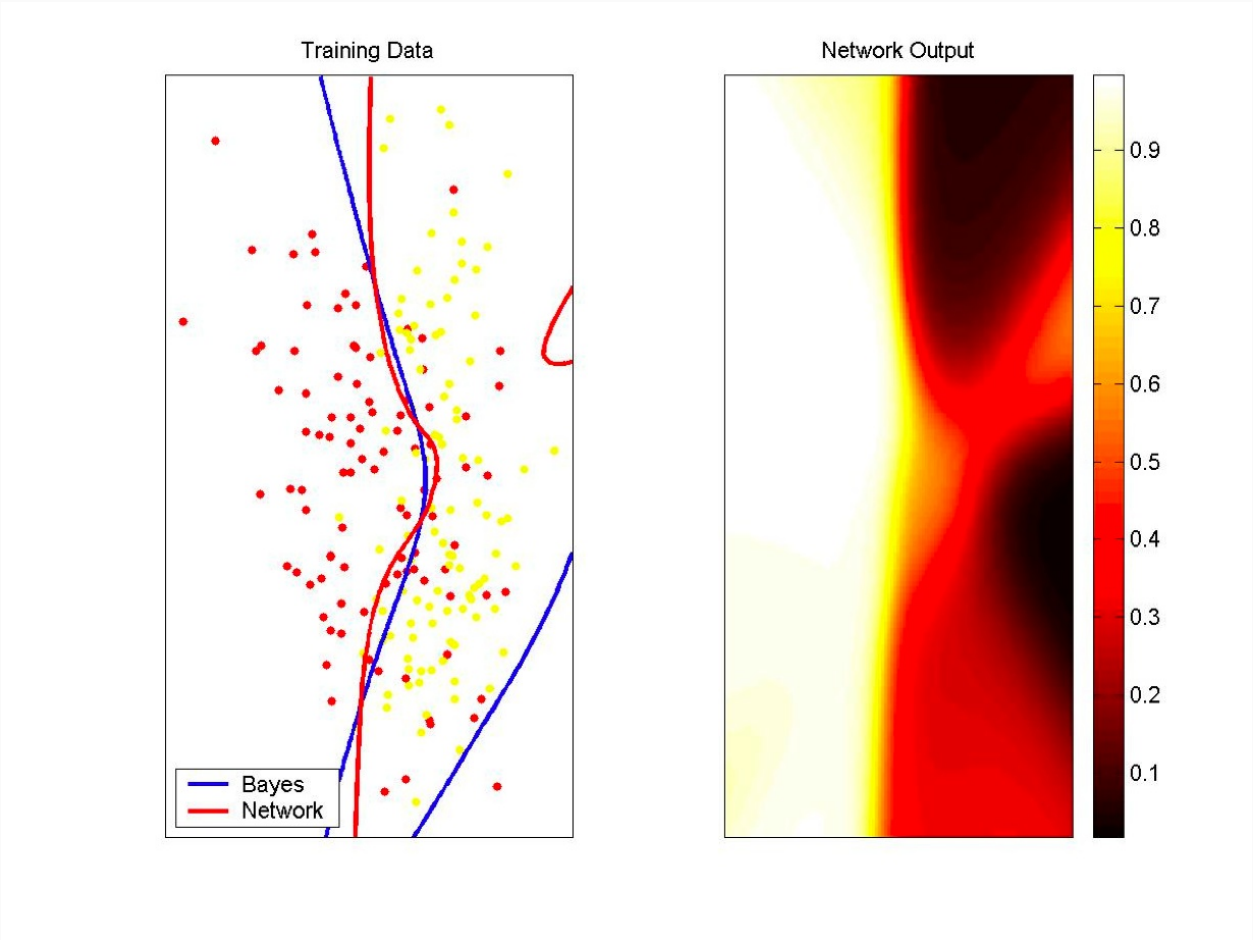
Example in 2D



Example in 2D



Using training set



Outline

Here comes the data: a statistical definition of the learning problem

Definition

- $\mathcal{S}_n$ is an i.i.d sample of size n, drawn from the joint probability law P(X,Y) fixed but unknown.
- $\mathcal{S}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$.
- Statistical learning is defined by:
  - Define a learning algorithm $\mathcal{A} : S_n \rightarrow \mathcal{A}(S_n) \in \mathcal{H}$ such that $\forall P$, S_n drawn from P, $R(\mathcal{A}(S_n))$ converges towards $R(h_{target})$ in probability

Statistical Machine Learning by ERM

Definition

- $\mathcal{S}_n$ is an i.i.d sample of size n, drawn from the joint probability law P(X,Y) fixed but unknown.
- $\mathcal{S}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$.

Statistical learning by Empirical Risk Minimization

- $\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$

instead of $min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]$

Empirical risk minimization

Definition

- Empirical risk: $R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i))$
- $\mathcal{A}(S_n) = \arg\, min_{h \in \mathcal{H}} R_n(h)$
- Where $\mathcal{H}$ is a tractable hypothesis set

Outline

Excess risk, approximation error and estimation error

Let us consider the 0/1 loss : Let R_{Bayes} be the Bayes Risk and $R_{\mathcal{H}} = \inf_{h \in \mathcal{H}} R(h)$ the smallest risk you can achieved in the function space $\mathcal{H}$.
Let $h_n \in \mathcal{H}$ be the classifier learnt from dataset S_n by minimization of the empirical risk or any method based on the dataset S_n

Excess risk, approximation error and estimation error

$$R(h_n) - R_{Bayes} = R(h_n) - R_{\mathcal{H}} + R_{\mathcal{H}} - R_{Bayes}$$

The excess risk of h_n compared to Bayes risk is equal to the sum of the two terms:

- $R(h_n) - R_{\mathcal{H}}$: an *estimation error* that measures to which point h_n is close to the best solution in $\mathcal{H}$
- $R_{\mathcal{H}} - R_{Bayes}$: an *approximation error* , inherent to the chosen class of functions, for instance, the approximation error is large if the true separation is nonlinear whereas I have chosen a linear classifier.

How to choose $\mathcal{H}$?

A compromise bias/variance

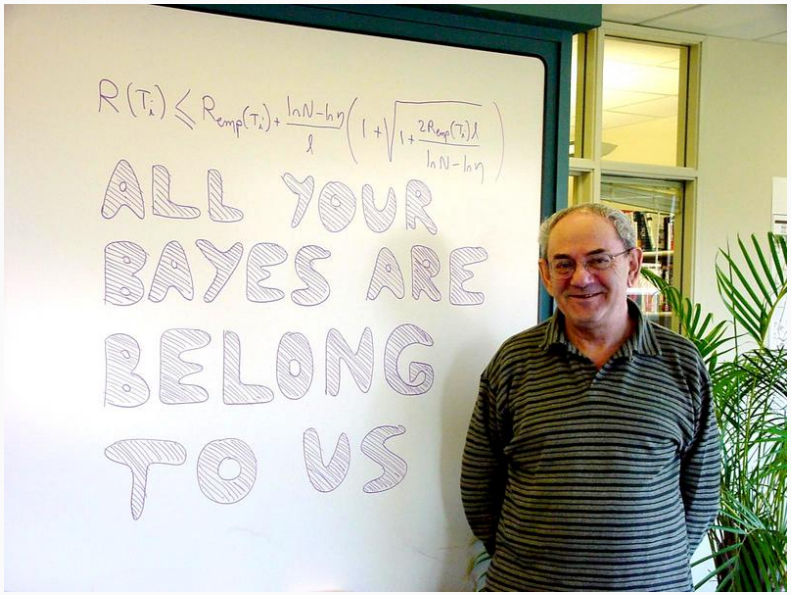
- If $\mathcal{H}$ is too small, you cannot reach the target (large bias, no universality) : risk of UNDERFITTING
- If $\mathcal{H}$ is too big, you cannot reduce variance (large variance, no consistency) : risk of OVERFITTING (we'll come back to that)

Is empirical risk minimization meaningful ?

Vapnik and Chervonenkis's results

- $\forall \mathbb{P}, \mathcal{S}_n$ drawn from P, $\forall h \in \mathcal{H}, R(h) \leq R_n(h) + \mathcal{B}(d, n)$
- where d is a measure of complexity of $\mathcal{H}$

Generalization bounds



Vladimir Vapnik in front of a white board, claiming for statistical learning against Bayesian inference (frequentist against bayesian stat.

Question: learning guarantee
If we measure the empirical risk $R_S(h)$ associated to a classifier h, what can we say about its true risk $R(h)$?

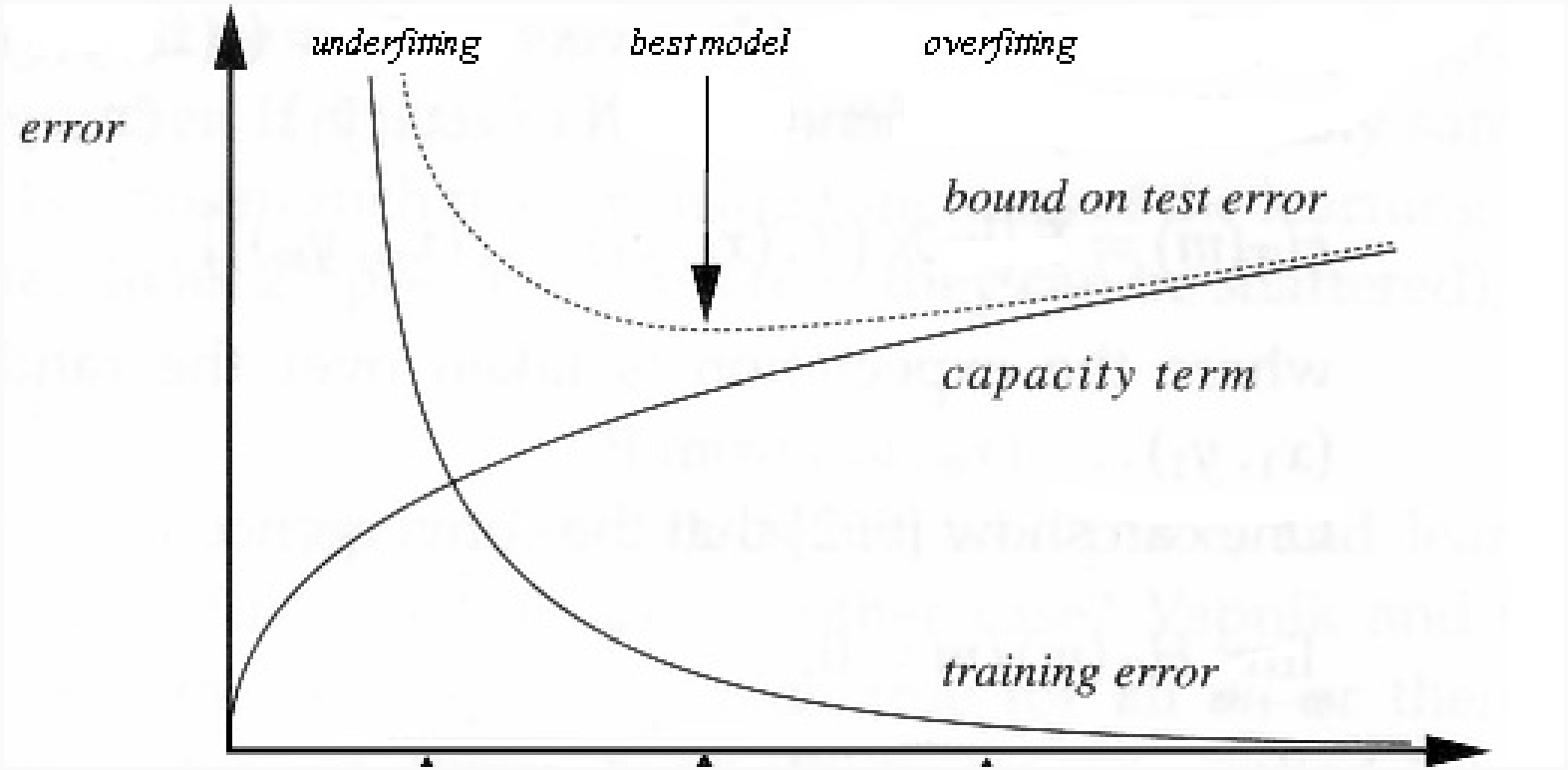
Read more: ▸ Link towards a small tutorial with proof

VC-dimension generalization bounds

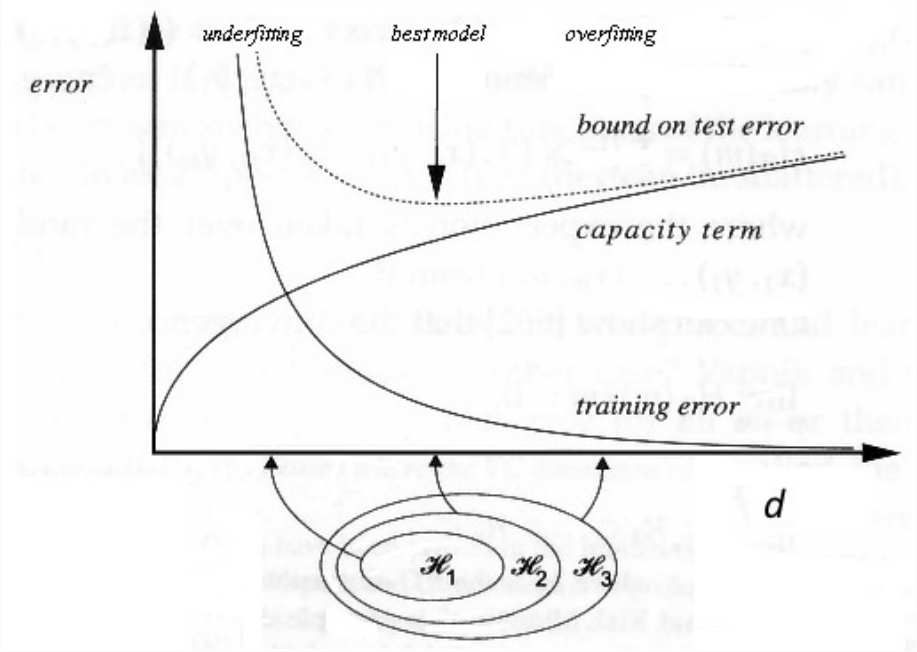
Theorem:
Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d_{VC}. Then, for any $\delta > 0$, the following holds for all $h \in \mathcal{H}$ with probability greater than 1 - δ

$$R(h) \leq R_n(h) + \sqrt{\frac{8 d_{VC}(\ln \frac{2n}{d_{VC}} + 1) + 8 \log(\frac{4}{\delta})}{n}}$$

Error (risk) versus h



Principle of Structural Risk Minimization



Vapnik proposed to replace empirical minimization principle by structural risk minimization, the underlying idea is to control the complexity of family $\mathcal{H}$ while reducing the empirical error.

Shattering

Definition: **Shattering**

$\mathcal{H}$ is said to shatter a set of data points $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ if, for all the 2^n possible assignments of binary labels to those points, there exists a function $h \in \mathcal{H}$ such that the model h makes no errors when predicting that set of data points.

Vapnik-Chervonenkis dimension

Definition: **VC-dimension**

The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$:

$$d_{VC}(\mathcal{H}) = max\{m : \exists(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathcal{X}^m \text{that are shattered by } \mathcal{H}\}$$

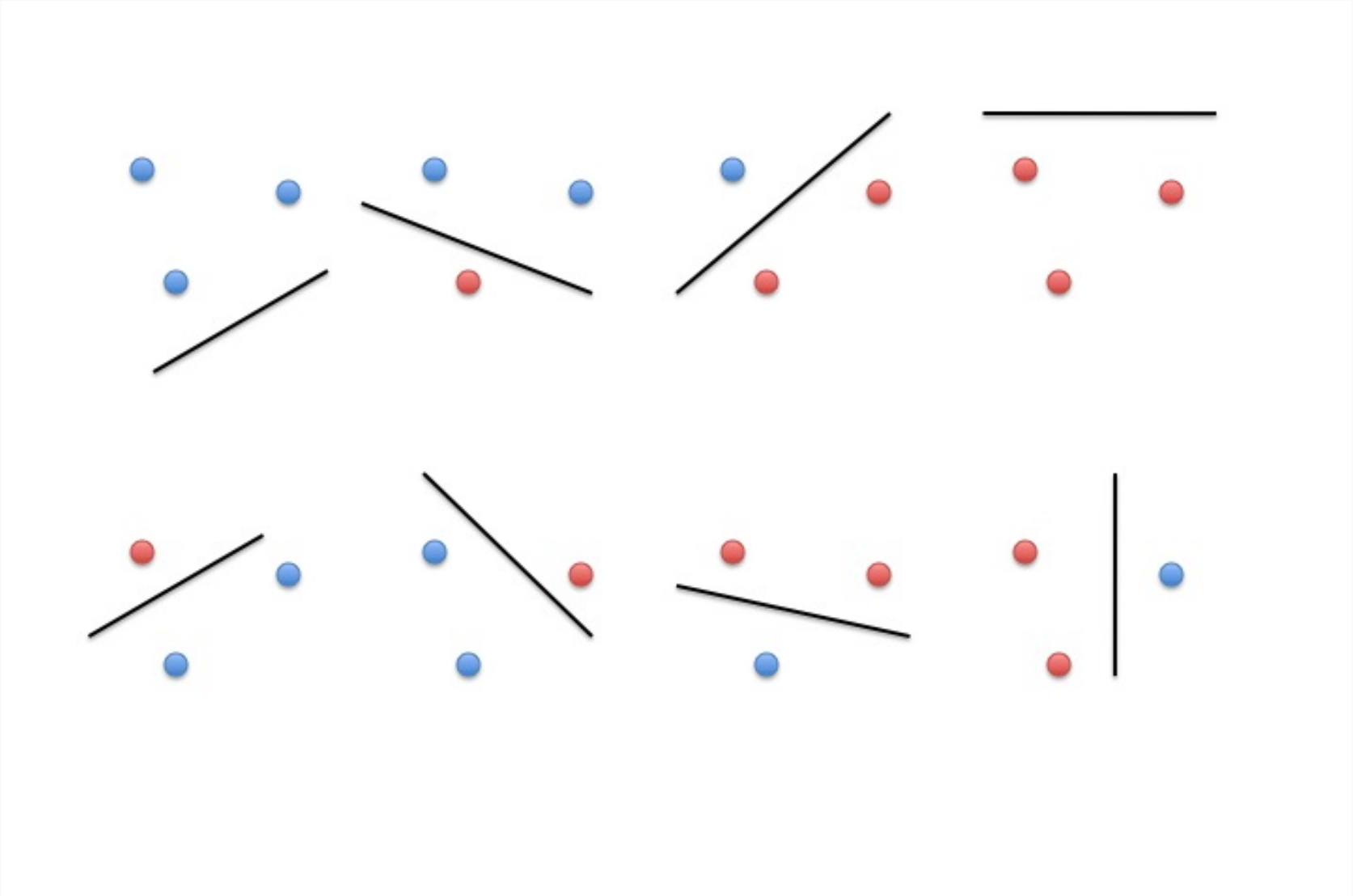
N.B.: if $d_{VC}(\mathcal{H}) = d$, then there exists a set of d points that is fully shattered by $\mathcal{H}$, but this DOES NOT imply that all sets of dimension d or less are fully shattered !

VC-dimension of Hyperplanes

What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?
Obviously $d_{VC}(\mathcal{H}_2) \geq 2$
Let us try with 3 points :

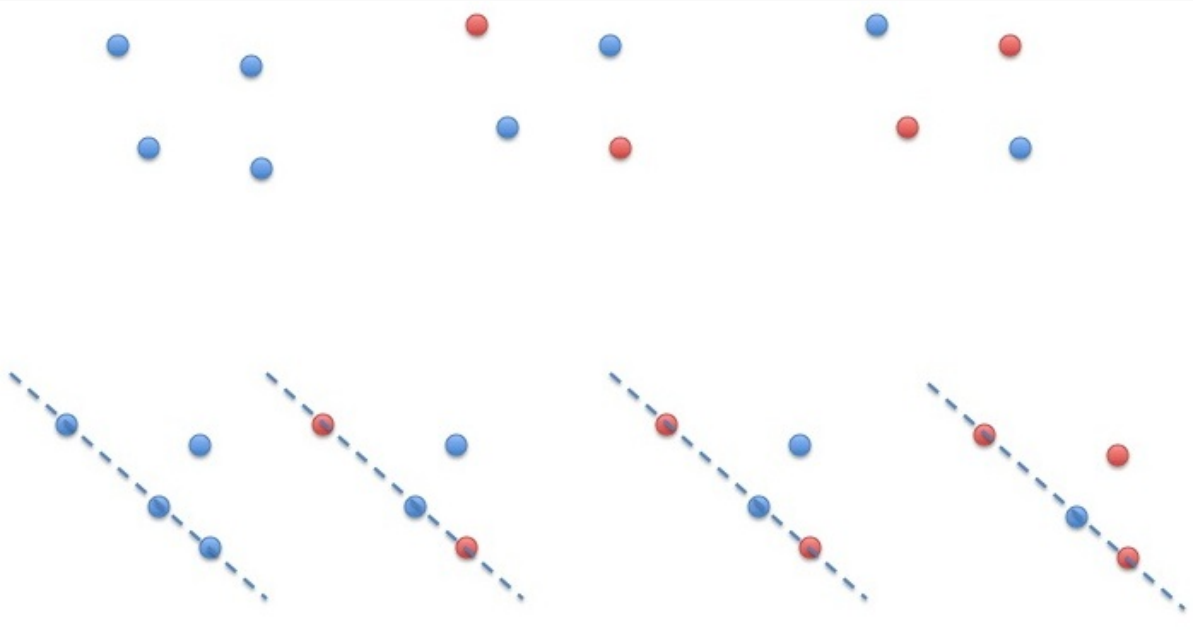
VC-dimension of Hyperplanes

What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?
Let us consider the following triplet of points



VC-dimension of Hyperplanes

What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?
For any set of 4 points, either 3 of them (at least) are aligned or no triplet of points is aligned.



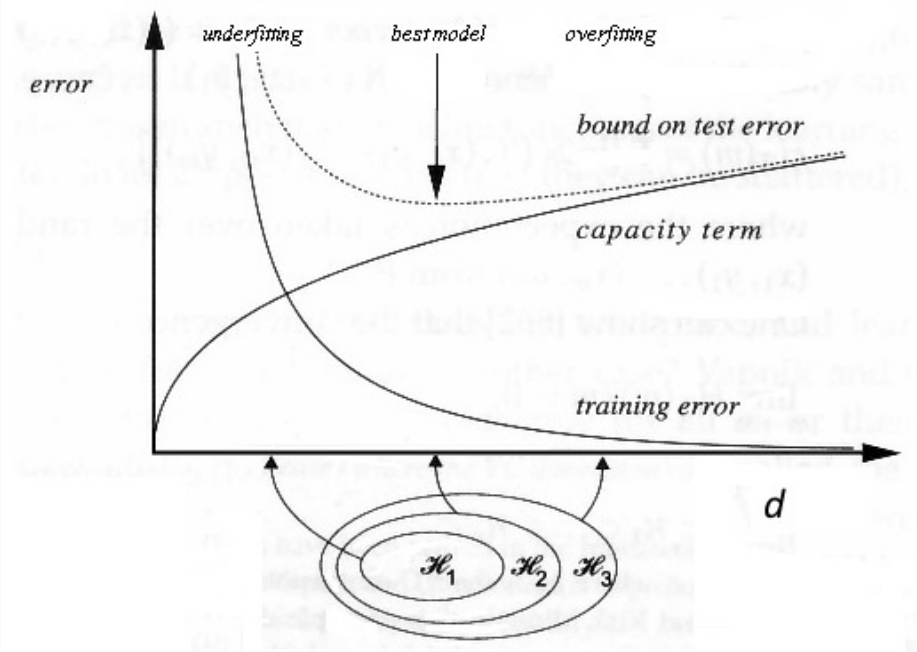
We can show that it is not possible for $\mathcal{H}_2$ to shatter 4 points.
Then $d_{VC}(\mathcal{H}_2) = 3$.

VC-dimension of Hyperplanes

More generally, one can prove :

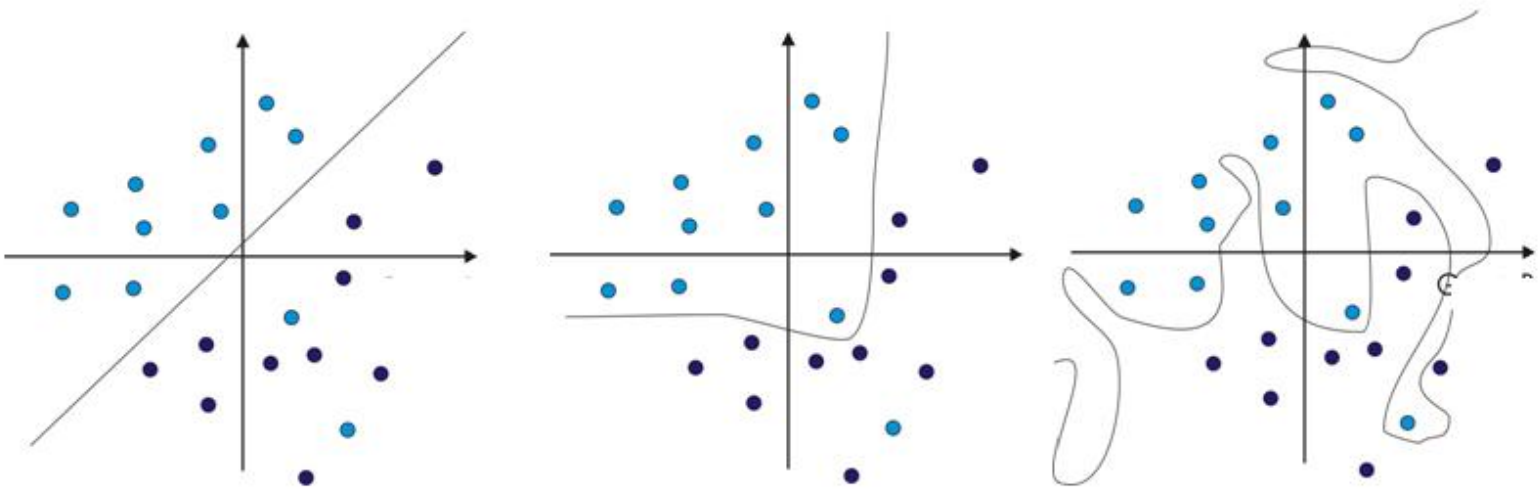
$$d_{VC}(\mathcal{H}_d) = d + 1$$

Principle of Structural Risk Minimization



Vapnik proposed to replace empirical minimization principle by structural risk minimization, the underlying idea is to control the complexity of family $\mathcal{H}$ while reducing the empirical error.

In practice, how to avoid overfitting



Optimization problem in practice: regularization

Pb1
$Min_h R_n(h)$ s.c $\Omega(h) \leq C$

Pb2
$Min_h \Omega(h)$ s.c $R_n(h) \leq C$

Pb3
$Min_h R_n(h) + \lambda\Omega(h)$

- Ω(h): measures the complexity of a single function h

A practical methodology of machine learning

- Three main problems to be solved :
  - **Representation**: determine in which representation space the data will be encoded and determine which family of mathematical functions will be used
  - **Optimization**: using statistical criteria, formulate the learning problem as an optimization problem, develop an optimization algorithm
  - **Evaluation**: provide a performance estimate

Statistical learning for supervised classification

Two main families of approaches:

1. Discriminant approaches : just find a classifier which does not estimate the Bayes classifier
2. Generative probabilistic approaches that are built to model $h(x) = \hat{P}(Y = 1|x)$ using $p(x|Y = 1)$, $p(x|Y = -1)$ and prior probabilities.

Outline

Bibliography

- The elements of Statistical Learning, Hastie, Tibshirani and Friedman, Springer, 2001.
- Chris Bishop, Pattern recognition and Neural networks, Springer, 1999.
- James, Gareth, et al. An introduction to statistical learning. Vol. 6. New York: springer, 2013.
- Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2012. (more 3A/M2 level)
- Abu-Mostafa, Y. S., Magdon-Ismail, M., Lin, H. (2012). Learning from data: a short course.

Outline