TSIA-SD210 - Machine Learning

Lecture 1 - Statistical Supervised Learning in a nutshell

Florence d'Alché-Buc

Contact: florence.dalche@telecom-paris.fr,
Télécom Paris, Institut Polytechnique de Paris, France

Table of contents

Outline

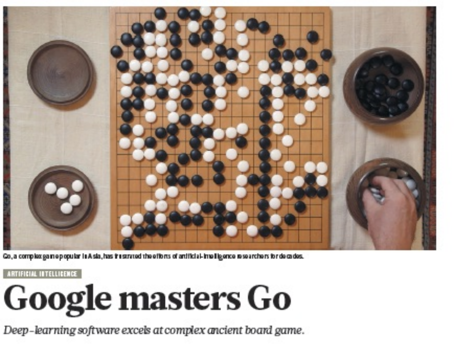
Introduction

Statistical Supervised Learning

References

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

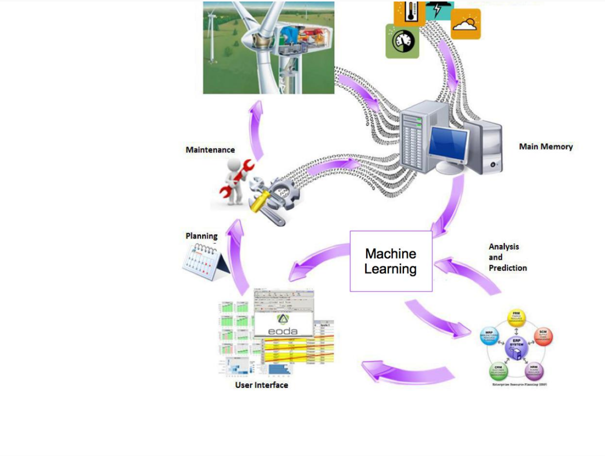
Deep-learning software excels at complex ancient board game.

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

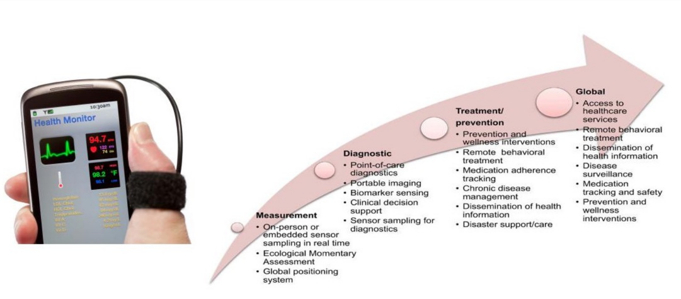
▸ Read more

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.



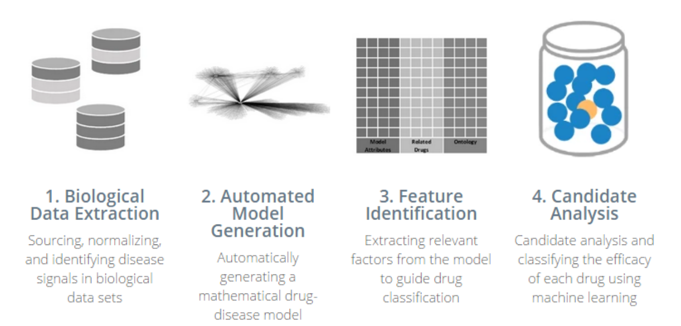
Mobile health monitoring


10:30am
Health Monitor
94.7
98.2 °F
Measurement
• On-person or embedded sensor sampling in real time
• Ecological Momentary Assessment
• Global positioning system
Diagnostic
• Point-of-care diagnostics
• Portable imaging
• Biomarker sensing
• Clinical decision support
• Sensor sampling for diagnostics
Treatment/ prevention
• Prevention and wellness interventions
• Remote behavioral treatment
• Medication adherence tracking
• Chronic disease management
• Dissemination of health information
• Disaster support/care
Global
• Access to healthcare services
• Remote behavioral treatment
• Dissemination of health information
• Disease surveillance
• Medication tracking and safety
• Prevention and wellness interventions

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.



1. Biological Data Extraction

Sourcing, normalizing, and identifying disease signals in biological data sets

2. Automated Model Generation

Automatically generating a mathematical drug-disease model

3. Feature Identification

Extracting relevant factors from the model to guide drug classification

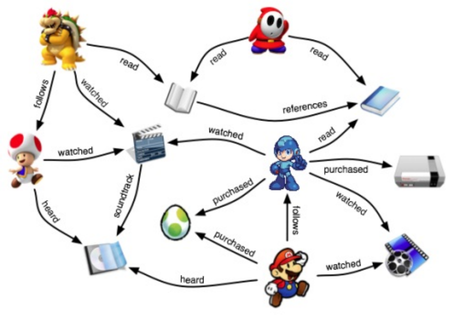
4. Candidate Analysis

Candidate analysis and classifying the efficacy of each drug using machine learning

Read more: ▸ Link

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

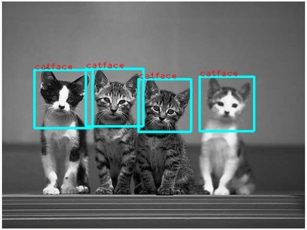
Recommendation system



- "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

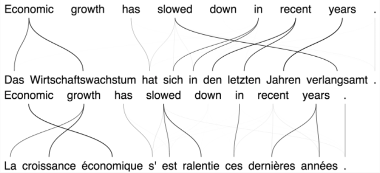


Read more: ▸ Link 1

Tuto Slides from Fei-Fei Li

and ▸ Link 2 for instance: website of Ivan Laptev

Machine Translation

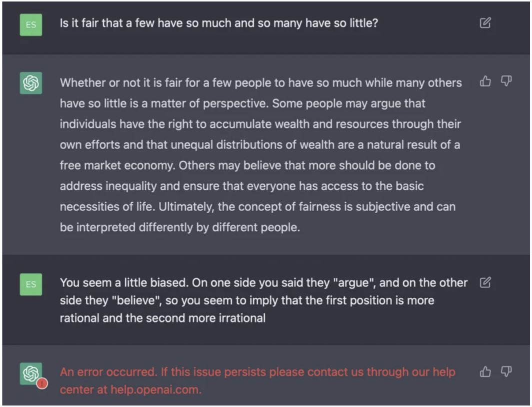

Economic growth has slowed down in recent years .
Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt .
Economic growth has slowed down in recent years .
La croissance économique s' est ralentie ces dernières années .

Read more: ▸ Link

Introduction to Neural Machine Translation with GPUs. Kyunghyun Cho.

Chatbot

Generative Pretrained Transformer 3 (175 billions of paremters)



Is it fair that a few have so much and so many have so little?

Whether or not it is fair for a few people to have so much while many others have so little is a matter of perspective. Some people may argue that individuals have the right to accumulate wealth and resources through their own efforts and that unequal distributions of wealth are a natural result of a free market economy. Others may believe that more should be done to address inequality and ensure that everyone has access to the basic necessities of life. Ultimately, the concept of fairness is subjective and can be interpreted differently by different people.

You seem a little biased. On one side you said they "argue", and on the other side they "believe", so you seem to imply that the first position is more rational and the second more irrational

An error occurred. If this issue persists please contact us through our help center at help.openai.com.

Read more: ▸ Link

ChatGPT:Optimizing Language Models for Dialogue

Machine Learning everywhere !

- Search engine, text-mining, automatic translation, chatbot
- Image recognition: face recognition, remote sensing ..
- Diagnosis, Fault detection
- Business analytics, Marketing, advertizing
- Prediction in Heath care, Personalized medecine
- Discovery tool in science
- Social networks, link prediction, recommendation
- Health Monitoring in industry, environment

With great power, it comes great responsibility

AI does not come for free

- Data acquisition, collection and annotation
- Carbon footprint
- Bias reproduction
- Black box
- Trustworthiness ?
- Arguable usage

Please keep in mind that the future of AI will require technical solutions to all these caveats !
Certainly a need for **onboarding** social and human sciences to the rescue*

Machine Learning

Definition

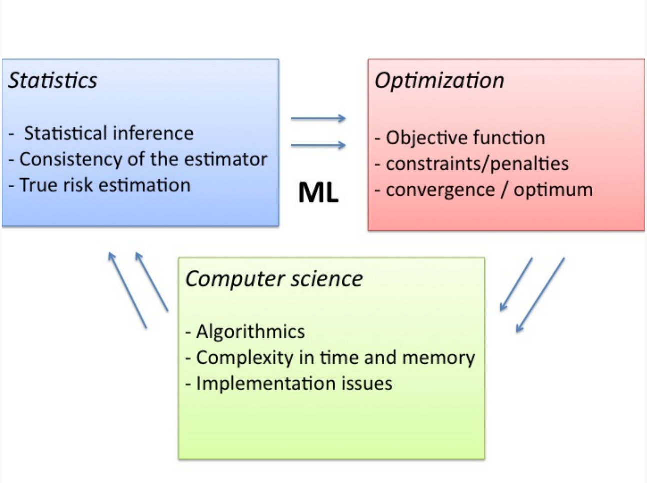
A type of artificial intelligence (AI) that provides computers with the ability to do certain tasks, such as recognition, diagnosis, planning, robot control, prediction, etc., without being explicitly programmed. It focuses on the development of algorithms that can teach themselves to grow and change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

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 modeling, game player, robot learning, time-series forecasting, recommendation...
- **Performance measure** : **today** accuracy on new data, ability to generalize -tomorrow: also transparency, fairness, privacy, frugality ...

Machine Learning



Statistics

- Statistical inference
- Consistency of the estimator
- True risk estimation

ML

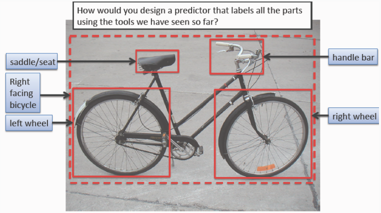
Optimization

- Objective function
- constraints/penalties
- convergence / optimum

Computer science

- Algorithmics
- Complexity in time and memory
- Implementation issues

Example 1: object recognition in an image



How would you design a predictor that labels all the parts using the tools we have seen so far?

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*

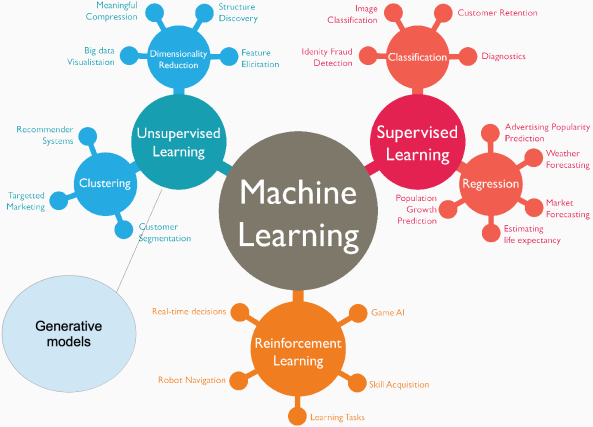
- robotics
- predictive maintenance
- security in cloud servers
- personalized advertising
- autonomous cars
- personalized healthcare
- security systems

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

Overview of Machine Learning



Machine Learning

Unsupervised Learning

Dimensionality Reduction

Meaningful Compression

Structure Discovery

Big data Visualistaion

Feature Elicitation

Clustering

Recommender Systems

Targetted Marketing

Customer Segmentation

Generative models

Supervised Learning

Classification

Image Classification

Customer Retention

Idenity Fraud Detection

Diagnostics

Regression

Advertising Popularity Prediction

Weather Forecasting

Population Growth Prediction

Market Forecasting

Estimating life expectancy

Reinforcement Learning

Real-time decisions

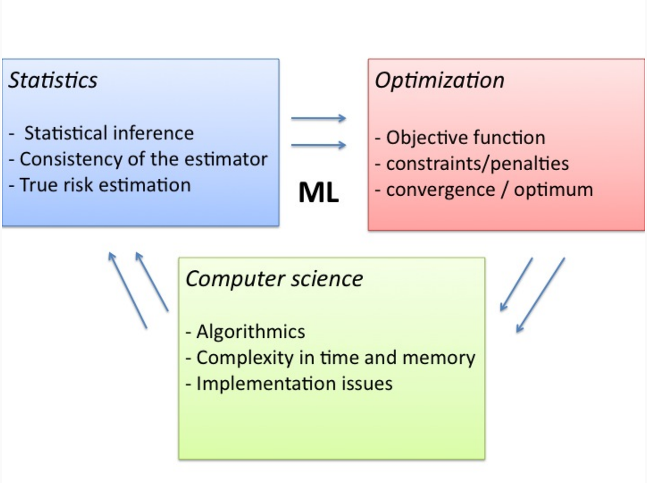
Game AI

Robot Navigation

Skill Acquisition

Learning Tasks

Machine Learning: what do you need ?



Statistics

- Statistical inference
- Consistency of the estimator
- True risk estimation

ML

Optimization

- Objective function
- constraints/penalties
- convergence / optimum

Computer science

- Algorithmics
- Complexity in time and memory
- Implementation issues

Teaching team in Machine Learning

Lecturers

- Matthieu Labeau, associate prof. (lecture + coordination)
- Ekhine Iruroski, associate prof. (lecture)
- Florence d'Alché, prof (lecture)
- Tamim El Ahmad, PhD student (practical session)
- Luc Brogat-Motte, PhD student (practical session)
- and many others... (10)

Evaluation of the course

- Three mandatory lab sessions, all graded, to submit **at the end of the session** (work in binomes)
- 2 best lab grades out of the 3: 5 pts each
- Exam, course questions, no notes: 10 pts

Planning of the course

- 1 Introduction to Statistical Machine Learning - Lecture
- 2 Trees and ensemble methods - Lecture
- 3 Practical session on Trees and Random forests
- 4 Support Vector Machines and Kernel Methods - Lecture
- 5 Practical session on SVM
- 6 Introduction to Neural Networks - Lecture
- 7 Practical session - Neural Networks
- 8 Exam

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

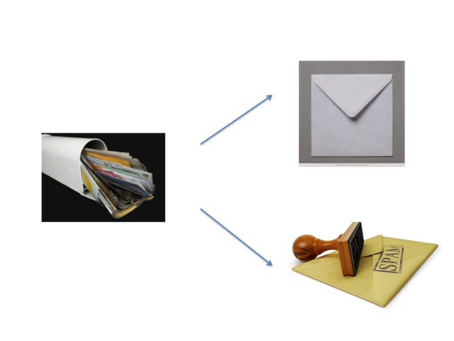
Introduction

Statistical Supervised Learning

- Focus on supervised classification
- Probabilistic and statistical setting of Supervised Learning
- Empirical risk minimization
- Relevance of Empirical Risk Minimization

References

Goal of Supervised classification



- Build a software that automatically classify data into two classes
- Two classes: relevant document / spams

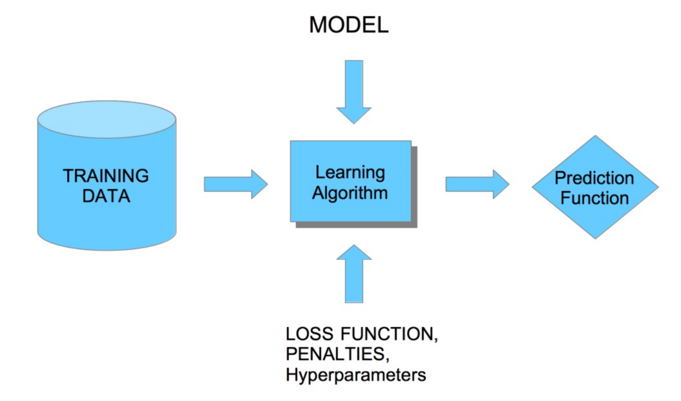
Use a training dataset to define the classifier

Computer science/algorithmics

- Training dataset: $\mathcal{S}_n = \{(document, label)\} = \{(x_i, y_i), i = 1, \ldots, n\}$
- 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 a document) : `print(clf.predict([[-0.8, -1]]))`

Read more about scikitlearn: `https://scikit-learn.org/stable/index.html`

Learning a classifier: applying a learning algorithm $\mathcal{A}$ to training data



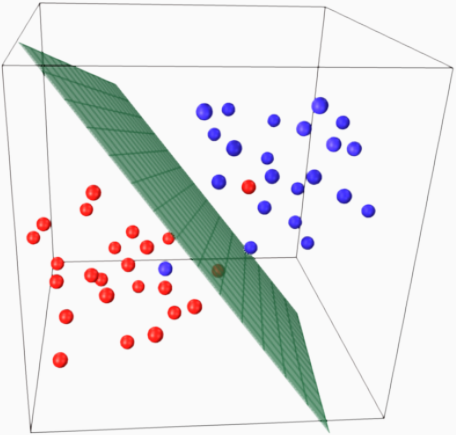
in *scikitlearn* : `clf.fit(Xtrain, ytrain)`

What do we need to determine a document classifier?

- Choose a way to represent a document(the input) : term-frequency inverse document frequency (tf idf), word2vec, ...
- output : y : 0 or 1, -1 or +1
- A classifier: linear or nonlinear ?
- Learning algorithm : minimizing some cost function
- Empirical measures: accuracy/ classification error, test error, Cross-validation

Read more: ▸ About TF-IDF, ▸ About word2vec.

A smple example: a linear classifier (formal neuron)



Building a document classifier?

- n documents available at the "training phase"
- Document i → a vector $\mathbf{x}_i \in \mathbb{R}^p$,$i = 1, \dots, n$
- Label: $y_i \in \{0, 1\}$
- A linear classifier:$f(\mathbf{x}) = s(w_0 + w^T x)$
- with $s(z) = \frac{1}{1+exp(-\frac{1}{2}z)}$, $z \in \mathbb{R}$
- Simple example: minimization of $\mathcal{L}(w; \mathbf{x_1}, \dots, \mathbf{x_n}) = \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2$
- Find w such that $\mathcal{L}(w; \mathbf{x_1}, \dots, \mathbf{x_n})$ be minimal

Essential questions to be answered to learn a predictive model

Once we have acquired and labeled the dataset

- Representation
    - Choose a way to represent (input) data
    - Choose a family of predictive models, i.e. a hypothesis space
- Optimization
    - Define a loss function, possibly constraints or penalties
    - Express the learning problem as an optimization one
    - Develop an optimization algorithm or take it off-the-shelves
- **Validation / Evaluation**
    - Define Evaluation metrics
    - Model selection procedure

Outline

Introduction

Statistical Supervised Learning

References

A probabilistic setting for the learning problem

- 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
- Let us call $p(X, Y)$ the joint probability distribution of the random pair (X, Y)
- $\mathcal{Y} = \mathbb{R}$ in case of regression
- $\mathcal{Y} = \{1, -1\}$ in case of binary supervised classification

Statistical view of the learning problem: notations

First, we need further notations. We denote:

- $\mathcal{D}$, the class of measurable functions from $\mathcal{X}$ to $\mathcal{Y} \subset \mathbb{R}$
- Hypothesis space:= $\mathcal{H} \subset \mathcal{D}$, the space of classification (regression) models
- (local) loss function:= $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$: for instance, the zero-one prediction loss $\ell(y, h(x)) := 1_{y \neq h(x)}$
- **True risk** of $h \in \mathcal{H}$ ñoted R(h):= $\mathbb{E}_{(X,Y)\sim\rho}[\ell(Y, h(X))]$

Probabilistic and Statistical view of the learning problem: definition

Supervised learning

Supervised learning consists in searching for the solution of the following optimization problem:

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_{(X,Y)}[\ell(h(X), Y)]$$

with the help of a training sample: $S_n := \{(x_i, y_i)_{i=1}^n\}$ containing n identical independent realizations of (X, Y) AND without knowledge of P.

What is the best solution when we know p

Suppose we know $p(x, y)$. No training data.
Now solve the following problem:

$$\arg \min_{h:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)}[\ell(h(X), Y)]$$

Binary Supervised Classification

Let us focus on binary classification and on the zero-one loss $\ell_{0,1}$. Imagine now that $h(x) \in \{-1, +1\}$.

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

$$
\begin{aligned}
R(h) &= \sum_{y=-1,1} P(Y=y) \int_{\mathbb{R}^p} \ell_{0,1}(h(x), y) p(x|Y=y) dx \\
&= \sum_{y=-1,1} P(Y=y) \int_{\mathbb{R}^p} 1_{h(x) \neq y} p(x|Y=y) dx \\
&= P(Y=-1) \int_{\mathbb{R}^p} 1_{h(x) \neq -1} p(x|Y=-1) dx + P(Y=+1) \int_{\mathbb{R}^p} 1_{h(x) \neq +1} p(x|Y=+1) dx
\end{aligned}
$$

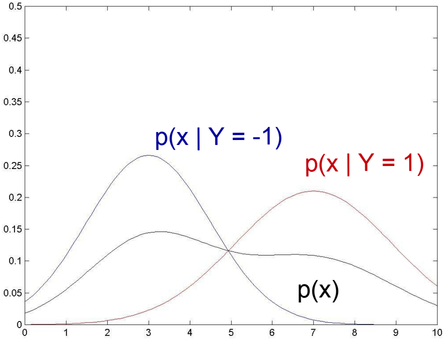
Bayes Rule

Bayes rule

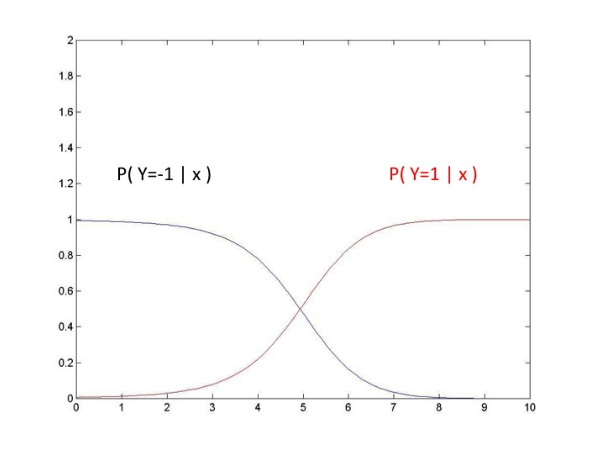
$$P(Y = k|x) = \frac{p(x|Y = k)P(Y = k)}{p(x|Y = -1).P(Y = -1) + p(x|Y = 1).P(Y = 1)}$$

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

A 1D example with Gaussian probability distribution

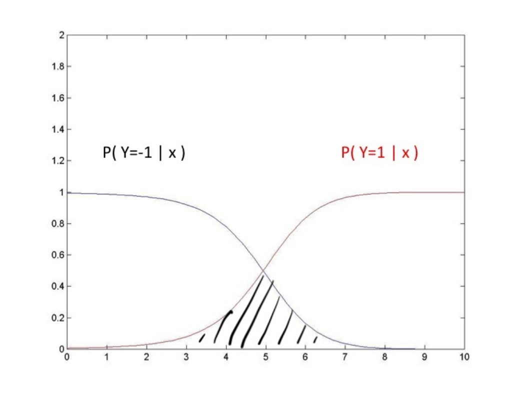


Bayesian classifier: Gaussian probability distribution



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

Bayesian classifier: Gaussian probability distribution



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

What is the best classifier h_{target} for the zero-one loss ?

Let $\eta(x) = P(Y = 1|x)$ for all x in $\mathcal{X}$. Then the Bayes classifier defined by:

$$h_{bayes}(x) = 1_{\eta(x) \geq 1/2}$$

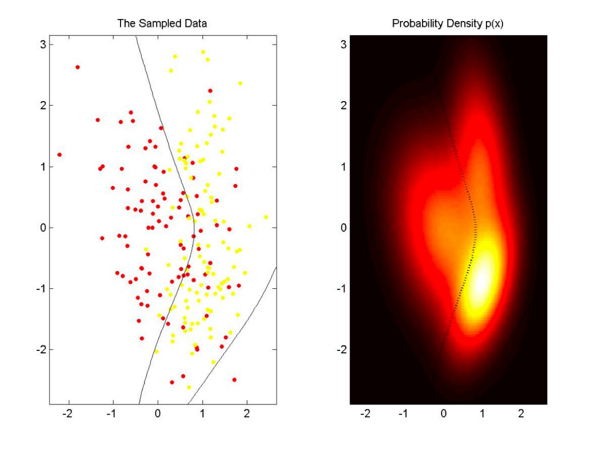
It can be shown that $h_{target} := h_{Bayes}$ and $R_{Bayes} = R(h_{Bayes})$ is the minimal risk associated to the zero-one loss.

First take-home message

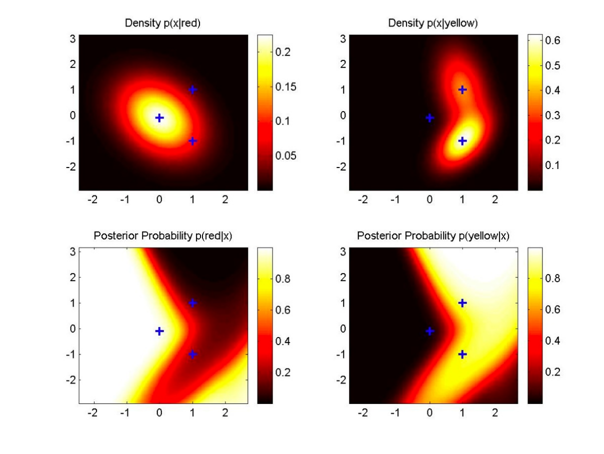
- The target function in supervised classification is the Bayes classifier for the $0-1$loss
- The target function in regression is $h(x) = \mathbb{E}[Y|x]$ for the square loss
- More generally, the nature of the target function depends heavily on the loss

Goal of learning:
Find a proxy of the target function h_{target} using an i.i.d. training sample $\mathcal{S}_n$ without the entire knowledge of p. Go further: see examples of other losses `https://arxiv.org/pdf/1612.03663.pdf`

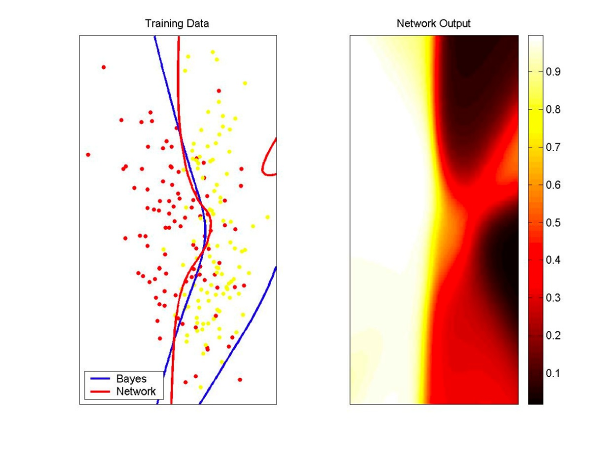
Example in 2D



Example in 2D



Using training set



Novel definition of statistical learning

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 can be 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.
- $\mathcal{S}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$.
- Empirical risk: $R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$

When h is fixed, Law of large numbers : $R_n(h)$ tends towards $R(h)$ almost surely. $(P(lim_n R_n(h) = R(h)) = 1)$

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

Outline

Introduction

Statistical Supervised Learning

References

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

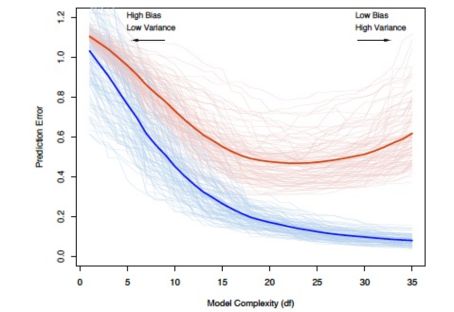
$$R(h_n) - R_{Bayes} = R(h_n) - R_{\mathcal{H}} + R_{\mathcal{H}} - R_{Bayes}$$

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.

Bias-variance dilemna

Experimental study



How to choose $\mathcal{H}$?

A compromise bias/variance

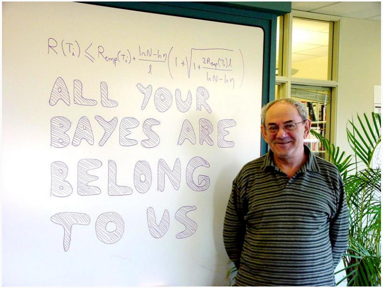
- If $\mathcal{H}$ is too small, you cannot reach the target (large bias, no universality) : risk of UNDERFITTING
- If $\mathcal{H}$ is too big, you cannot reduce variance (large variance, no consistency) : risk of OVERFITTING (we'll come back to that)

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


$R(T_i) \leqslant R_{emp}(T_i) + \frac{\ln N - \ln\eta}{\ell}\left(1 + \sqrt{1 + \frac{2R_{emp}(T_i)\ell}{\ln N - \ln \eta}}\right)$
ALL YOUR BAYES ARE BELONG TO US

Vladimir Vapnik in front of a white board, claiming for statistical learning against Bayesian inference

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

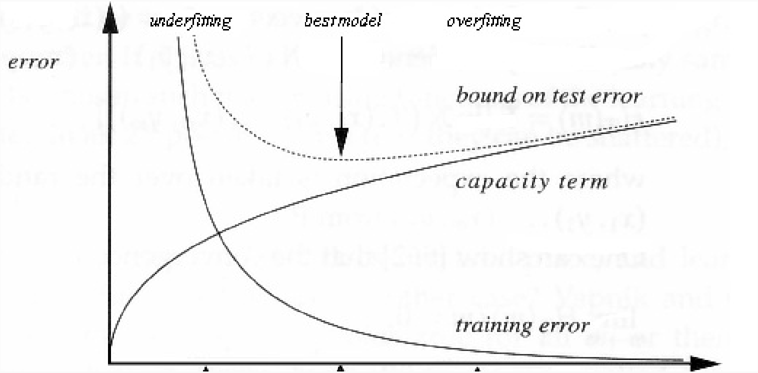
VC-dimension generalization bounds

Theorem:

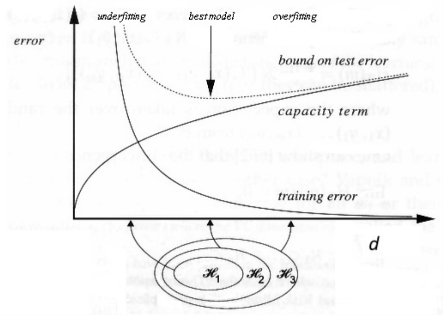
Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d_{VC}. Then, for any $\delta > 0$, the following holds for all $h \in \mathcal{H}$ with probability greater than 1 - δ

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

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 !

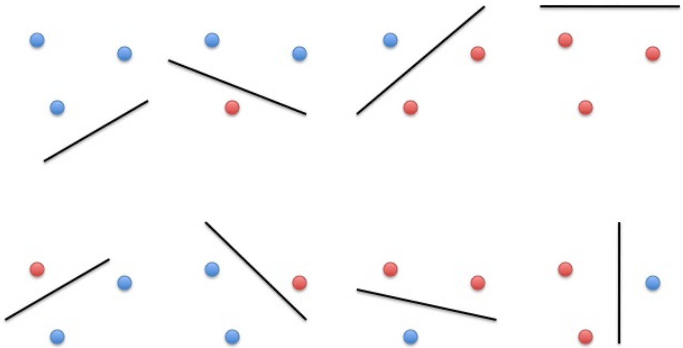
VC-dimension of Hyperplanes

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

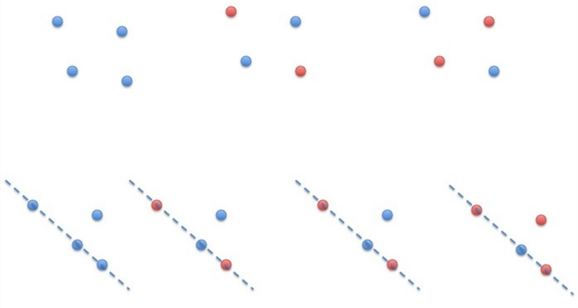
What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathcal{H}_2$) ?

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.



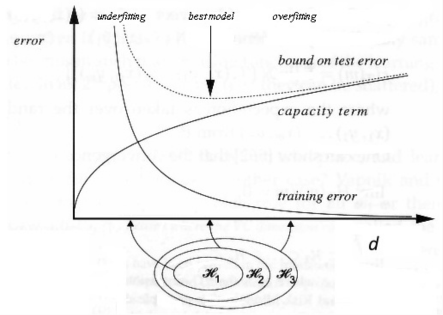
We can show that it is not possible for $\mathcal{H}_2$ to shatter 4 points.
Then $d_{VC}(\mathcal{H}_2) = 3$.

VC-dimension of Hyperplanes

More generally, one can prove :

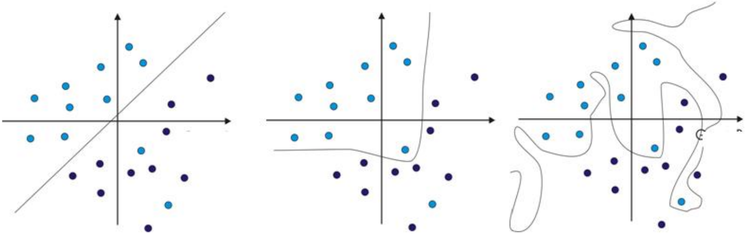
$$d_{VC}(\mathcal{H}_d) = d + 1$$

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*

Machine Learning with regularization

Supervised Learning

Let $\mathcal{S}_n = \{(x_i, y_i)\}_{i=1}^n$, a i.i.d. sample drawn from p a joint probability distribution defined on (X, Y): X takes its values in $\mathbb{R}^d$ and Y is real-valued.

Regularized empirical risk minimization

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, $\Omega : \mathcal{H} \to \mathbb{R}^+$, the goal is now to find a solution of:

$$\arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) + \lambda \Omega(h) \tag{1}$$

Role of $\Omega(h)$: control of the model complexity, more generally imposition of some prior knowledge

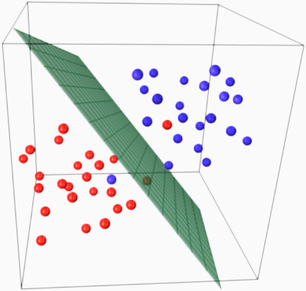
Machine Learning: two tasks

Let $\mathcal{S}_n = \{(x_i, y_i)\}_{i=1}^n$, a i.i.d. sample drawn from μ a joint probability distribution defined on (X, Y): X takes its values in $\mathbb{R}^d$ and Y is real-valued.

- **Learning**: get $h_n = \mathcal{A}(\mathcal{S}_n, \mathcal{H}, \ell, \lambda, \Omega)$ with
  - $\mathcal{S}_n$: training data
  - $\mathcal{H}$: class of functions
  - λ: some hyperparameter
  - ℓ: Local loss function
  - Ω: regularizing function
  - $\mathcal{A}$: learning algorithm
- **Prediction**: given x, and compute $h_n(x)$

Machine Learning: key components

- Data representation
- Hypothesis space
- Loss function
- Learning algorithm
- Evaluation metrics
- Model selection



Example of supervised learning

Statistical learning for supervised classification

Two main families of approaches:

1. Discriminant approaches : just find a classifier which discriminates
2. Generative probabilistic approaches: build a plug-in estimator of $\hat{P}(Y = 1|x)$ using $p(x|Y = 1)$, $p(x|Y = -1)$ and prior probabilities.

General principles to build a model

- Local average (k-neighbours, decision tree)
- Agregation/ committee (random forest, boosting)
- Hierarchy (=decision tree)
- Layer composition (shallow versus deep)
- Working in the whole space / subspaces

Parametric/non-parametric modeling

Parametric modeling

- Linear models
- Neural networks

Non-parametric modeling

- Local average models: k-neighbors, trees
- Kernel models

Outline

Introduction

Statistical Supervised Learning

References

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.