

TSIA-SD210 - Machine Learning

Lecture 1 - Introduction to Statistical Supervised Learning

Matthieu Labeau

(Slides par Florence d'Alché-Buc)

Contact: matthieu.labeau@telecom-paris.fr,
Télécom Paris, Institut Polytechnique de Paris, France

Table of contents

1. Motivation
2. About this course
3. A practical and computational introduction to Supervised Learning
4. A probabilistic and statistical view of Supervised Learning
5. References

Motivation

- A definition of Machine Learning

- Statistical learning

About this course

- A practical and computational introduction to Supervised Learning

- A probabilistic and statistical view of 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

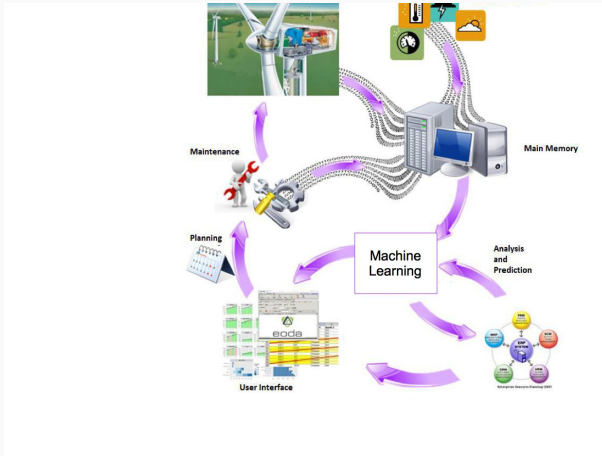
Deep-learning software excels at complex ancient board game.

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

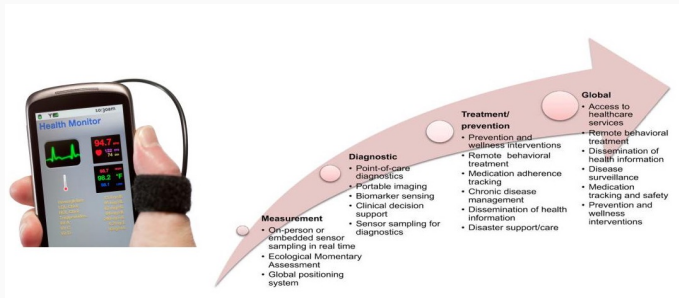
► Read more

Predictive Maintenance

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



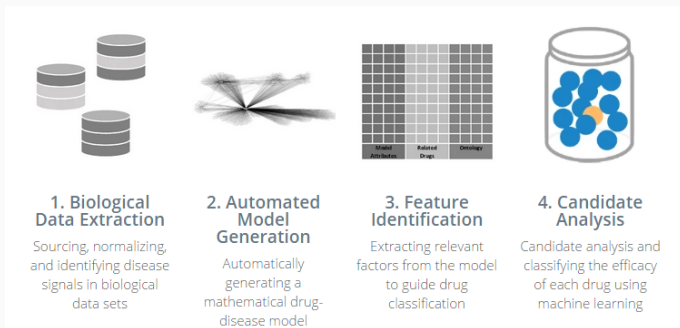
Mobile health monitoring



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

Drug discovery

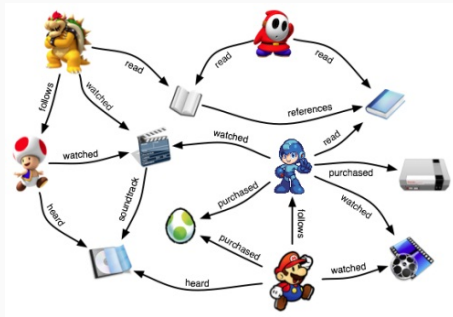
Drug-discovery has been revolutionized by Machine Learning.



Read more: [▶ Link](#)

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

Recommendation system

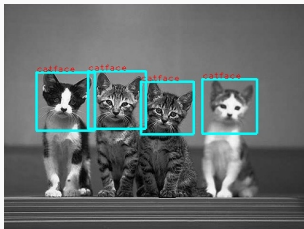


- "People read about 10 MB worth of material a day, hear 400MB a day and see 1MB of information every second"-The economist, Nov 2006.
- "We are leaving the age of information and entering the age of recommendation", Chris Anderson, Wired Magazine.

Read more: [▶ Link](#)

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

Object recognition

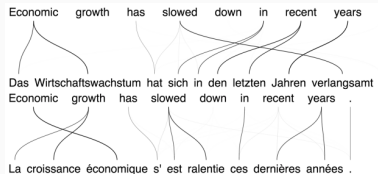


Read more: [▶ Link 1](#)

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 Health care, Personalized medicine
- Social networks, link prediction, recommendation

Motivation

- A definition of Machine Learning

- Statistical learning

About this course

- A practical and computational introduction to Supervised Learning

- A probabilistic and statistical view of Supervised Learning

- Formal setting of Supervised Learning

- Empirical risk minimization

- Relevance of Empirical Risk Minimization

References

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.

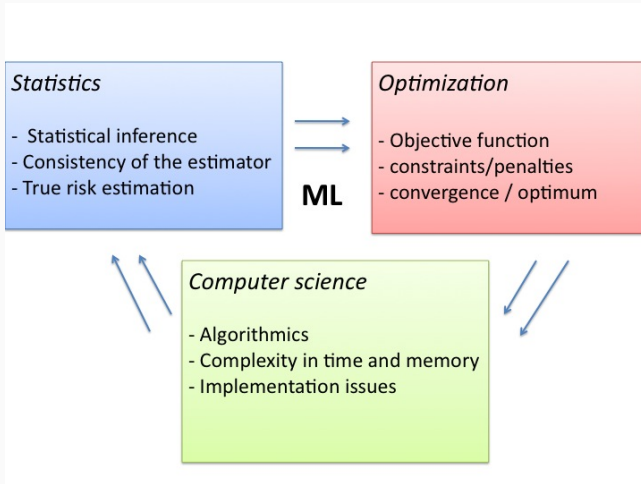
Experience, tasks and performance measure

A definition by Tom Mitchell (<http://www.cs.cmu.edu/~tom/>)

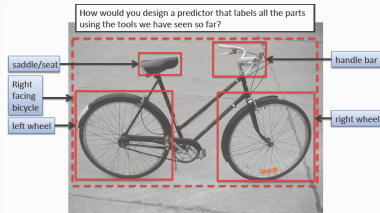
A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E.

- **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



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

- 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

Motivation

A definition of Machine Learning

Statistical learning

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

Formal setting of Supervised Learning

Empirical risk minimization

Relevance of Empirical Risk Minimization

References

- 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.

Supervised versus unsupervised learning

- **Supervised Learning (classification, regression):**
 - Goal: Learn a function f to predict a variable y from an individual x .
 y is a class label (classification), y is a real value (regression).
 - Data: Learning set (x_i, y_i)
- **Unsupervised Learning (clustering, graphical model):**
 - Goal: Discover a structure within a set 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.

- 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., and Lin, H. (2012). Learning from data: a short course.

Motivation

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

References

Lecturers

- Florence d'Alché, prof. (lecture)
- Ekhine Iruroski, associate prof. (lecture)
- Matthieu Labeau, associate prof. (lecture)
- Marta Campi, Post-doc (practical session)
- Tuan Binh Nguyen, Post-doc (practical session)
- Tamim El Ahmad, PhD student (practical session)
- Luc Brogat-Motte, PhD student (practical session)
- Jayneel Parekh, PhD student (practical session)
- Anass Aghbalou, PhD student (practical session)
- Junjie Yang, PhD student (practical session)

Evaluation of the course

- Each practical session is mandatory
- 2 practical session graded (binomes) among the 3: 3 pts each
- 1 exam: 14 pts

Planning of the course

- 1 Introduction to Statistical Machine Learning - Lecture
- 2 Practical session - Perceptron
- 3 Optimal Margin Classifier and Kernel Methods - Lecture
- 4 Practical session - SVM
- 5 Trees and ensemble methods - Lecture
- 6 Introduction to Neural Networks - Lecture
- 7 Practical session - Neural Networks
- 8 Exam

Motivation

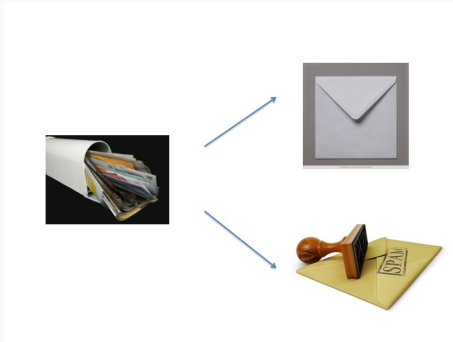
About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

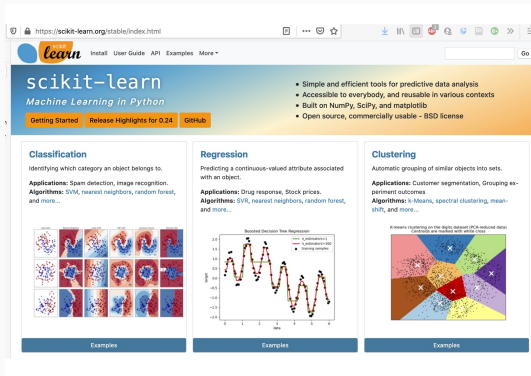
References

Goal of Supervised classification



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

Examples using Python



The screenshot shows the scikit-learn website with the URL <https://scikit-learn.org/stable/index.html>. The page features a navigation bar with links to 'Install', 'User Guide', 'API', 'Examples', and 'More'. The main header includes the 'scikit-learn' logo and the tagline 'Machine Learning in Python'. A list of features is displayed: 'Simple and efficient tools for predictive data analysis', 'Accessible to everybody, and reusable in various contexts', 'Built on NumPy, SciPy, and matplotlib', and 'Open source, commercially usable - BSD license'. Below this, three example sections are shown: 'Classification' (identifying object categories), 'Regression' (predicting continuous values), and 'Clustering' (grouping similar objects). Each section includes a brief description, applications, algorithms, and a visual example.

Classification
Identifying which category an object belongs to.
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

Regression
Predicting a continuous-valued attribute associated with an object.
Applications: Drug response, Stock prices.
Algorithms: SVM, nearest neighbors, random forest, and more...

Clustering
Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

Read more: <https://scikit-learn.org/stable/index.html>

Use a training dataset to define the classifier

Computer science/algorithmics

- Training dataset:

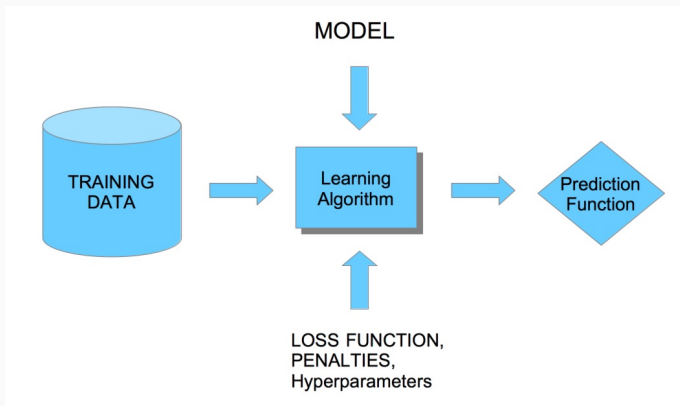
$$\mathcal{S}_n = \{(\text{document}, \text{label})\} = \{(x_i, y_i), i = 1, \dots, n\}$$

- 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]]))`

What do we need to determine a document classifier?

- Choose a way to represent a document (here an email)
- Choose a family of classification functions
- Formulate the learning problem as an optimization one (loss + constraints), take care of the unbalanced dataset (more relevant documents than spam ones)
- Define an optimization algorithm
- Evaluate the quality of the classifier learned from data

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](#) .

Building a document classifier?

- n documents available at the "training phase"
- Document $i \rightarrow$ a vector $x_i \in \mathbb{R}^p, i = 1, \dots, n$
- Label: $y_i \in \{0, 1\}$
- A linear classifier: $h(x) = s(w_0 + w_1x_1 + w_2x_2 + \dots + w_px_p)$
- with $s(z) = \frac{1}{1 + \exp(-\frac{1}{2}z)}, z \in \mathbb{R}$
- Simple example: minimization of
$$\mathcal{L}(w; x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (h(x_i) - y_i)^2$$
- Find w such that $\mathcal{L}(w; x_1, \dots, x_n)$ be minimal

Motivation

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

- Formal setting of Supervised Learning

- Empirical risk minimization

- Relevance of Empirical Risk Minimization

References

Motivation

A definition of Machine Learning

Statistical learning

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

Formal setting of Supervised Learning

Empirical risk minimization

Relevance of Empirical Risk Minimization

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 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**:= $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$: for instance, the zero-one prediction loss $\ell(y, h(x)) := 1_{y \neq h(x)}$
- **True risk** of $h \in \mathcal{H}$:= $\mathbb{E}_{(X,Y) \sim P}[\ell(h(X), Y)]$

Statistical view of the learning problem: definition

Supervised learning

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

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

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 .

Statistical view of the learning problem: a first discussion

Is this problem well-posed ? **No !**

- We do not know how to calculate $R(h) = \mathbb{E}_{(X,Y) \sim P}[\ell(h(X), Y)]$ as we do not have access to the probability distribution P
- **Solution:** We therefore will look for a proxy of this true risk to define a problem amenable to optimization in practice

Binary Supervised Classification

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

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

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

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 $P(Y = 1) + P(Y = -1) = 1$

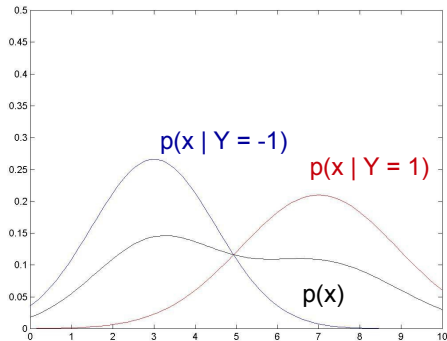
What is the best Classifier for the zero-one loss ?

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

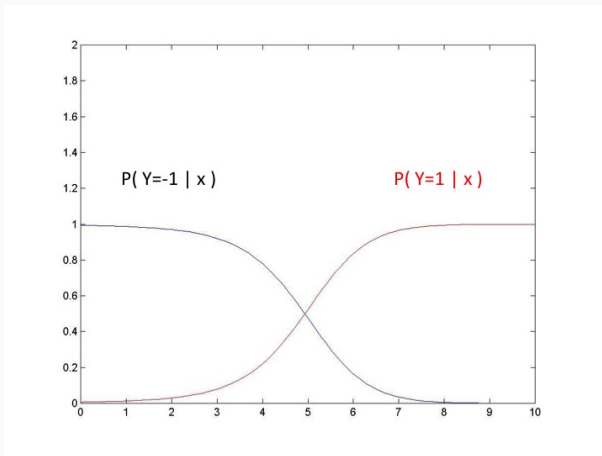
$$h_{\text{bayes}}(x) = 1_{\eta(x) \geq 1/2}$$

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

A 1D example with Gaussian probability distribution

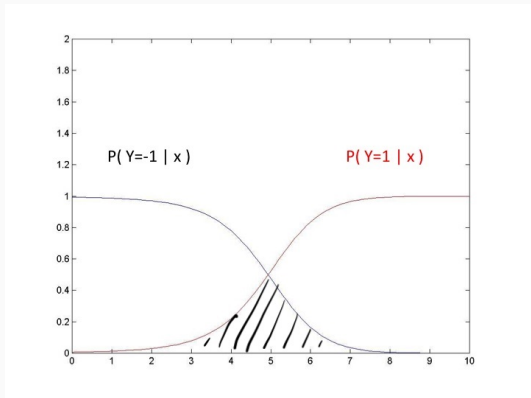


Bayesian classifier: Gaussian probability distribution



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

Bayesian classifier: Gaussian probability distribution



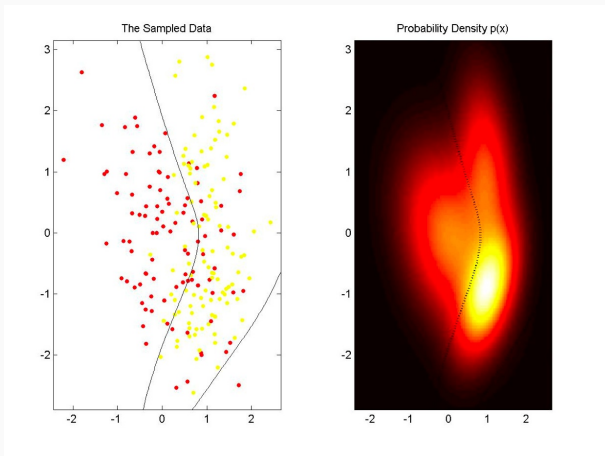
Exercise: what is the true risk of the Bayes Classifier ?

First take-home message

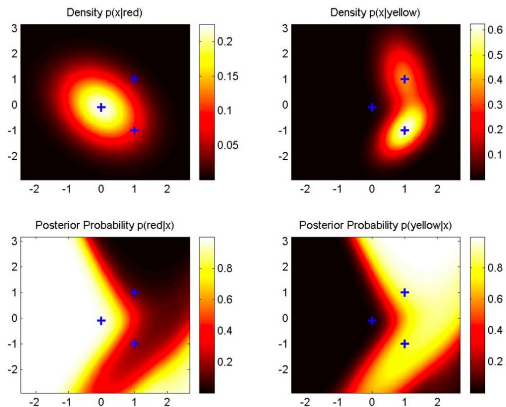
- The target function in supervised classification is the Bayes classifier for the 0 – 1 loss
- The target function in regression is $h(x) = \mathbb{E}[Y|x]$ for the square loss
- Now we call h_{target} the true target function

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

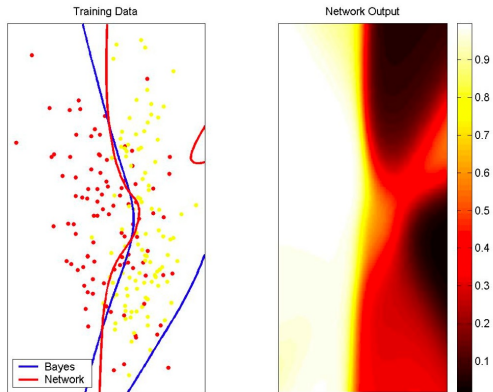
Example in 2D



Example in 2D



Using training set



Motivation

A definition of Machine Learning

Statistical learning

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

Formal setting of Supervised Learning

Empirical risk minimization

Relevance of Empirical Risk Minimization

References

A new 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.
- $\mathcal{S}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$.
- Statistical learning can be defined by:
 - Define a learning algorithm $\mathcal{A} : \mathcal{S}_n \rightarrow \mathcal{A}(\mathcal{S}_n) \in \mathcal{H}$ such that $\forall P, \mathcal{S}_n$ drawn from P , $R(\mathcal{A}(\mathcal{S}_n))$ converges towards $R(h_{target})$ in probability

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.
- $\mathcal{S}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$.
- 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 $\min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]$

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

Motivation

A definition of Machine Learning

Statistical learning

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of Supervised Learning

Formal setting of Supervised Learning

Empirical risk minimization

Relevance of Empirical Risk Minimization

References

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

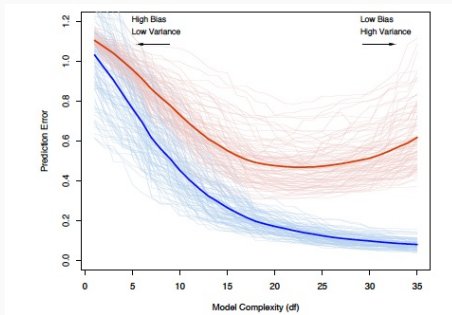
$$R(h_n) - R_{\text{Bayes}} = R(h_n) - R_{\mathcal{H}} + R_{\mathcal{H}} - R_{\text{Bayes}}$$

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}
- $R_{\mathcal{H}} - R_{\text{Bayes}}$: 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 dilemma

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

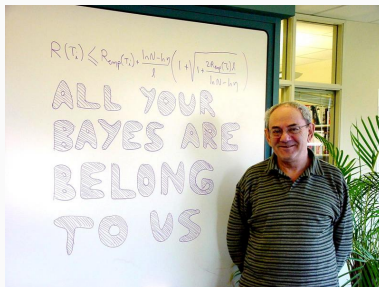


Figure 1: Vladimir Vapnik in front of a white board, claiming for statistical learning against Bayesian inference (frequentist against bayesian stat.)

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 $R(h)$?

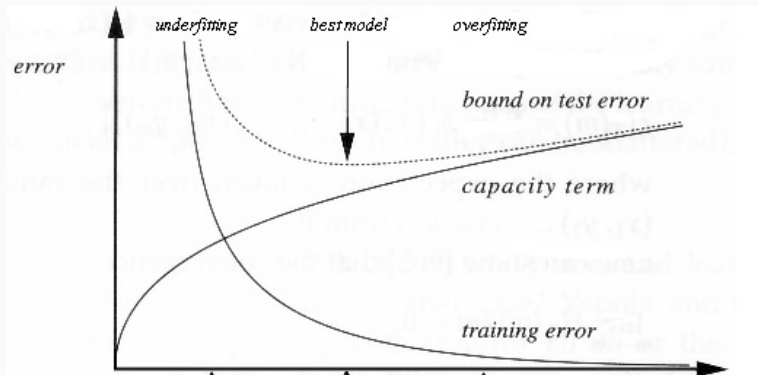
Read more: [▶ Link towards a small tutorial with proof](#)

Theorem:

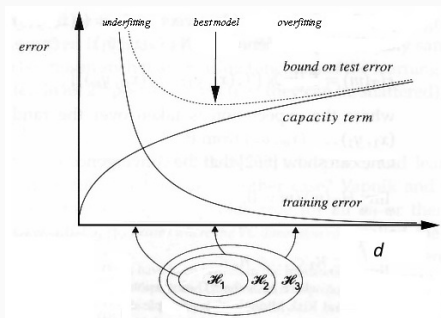
Let \mathcal{H} be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d_{VC} . Then, for any $\delta > 0$, the following holds for all $h \in \mathcal{H}$ with probability greater than $1 - \delta$

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

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.

Definition: **Shattering**

\mathcal{H} is said to shatter a set of data points (x_1, x_2, \dots, 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.

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} :

$$d_{VC}(\mathcal{H}) = \max\{m : \exists(x_1, \dots, x_m) \in \mathcal{X}^m \text{ that are shattered by } \mathcal{H}\}$$

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) ?

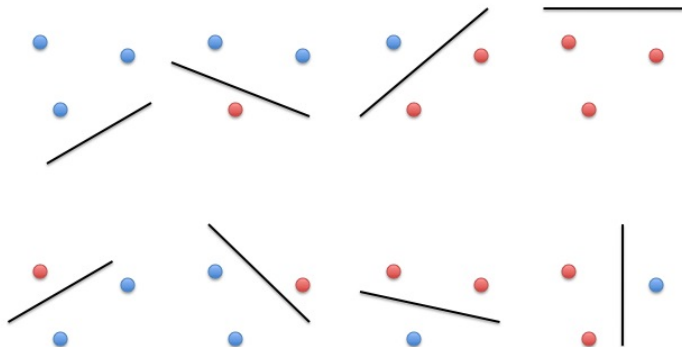
Obviously $d_{VC}(\mathcal{H}_2) \geq 2$

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) ?

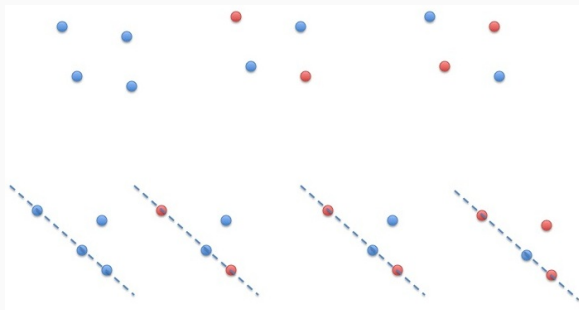
Let us consider the following triplet of points



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.



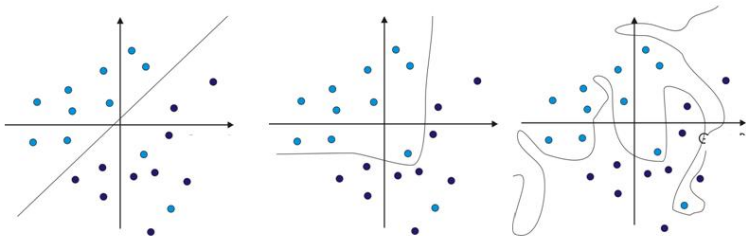
We can show that it is not possible for \mathcal{H}_2 to shatter 4 points.

Then $d_{VC}(\mathcal{H}_2) = 3$.

More generally, one can prove :

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

In practice, how to avoid overfitting



Optimization problem in practice: regularization

Pb1

$$\text{Min}_h R_n(h) \text{ s.t. } \Omega(h) \leq C$$

Pb2

$$\text{Min}_h \Omega(h) \text{ s.t. } R_n(h) \leq C$$

Pb3

$$\text{Min}_h R_n(h) + \lambda \Omega(h)$$

- $\Omega(h)$: measures the complexity of a single function h

A practical methodology of machine learning

- Four 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
 - **Model selection:** among many candidate models of various complexities, chose the best
 - **Performance Evaluation:** provide a performance estimate

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.

Motivation

About this course

A practical and computational introduction to Supervised Learning

A probabilistic and statistical view of 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., and Lin, H. (2012). Learning from data: a short course.