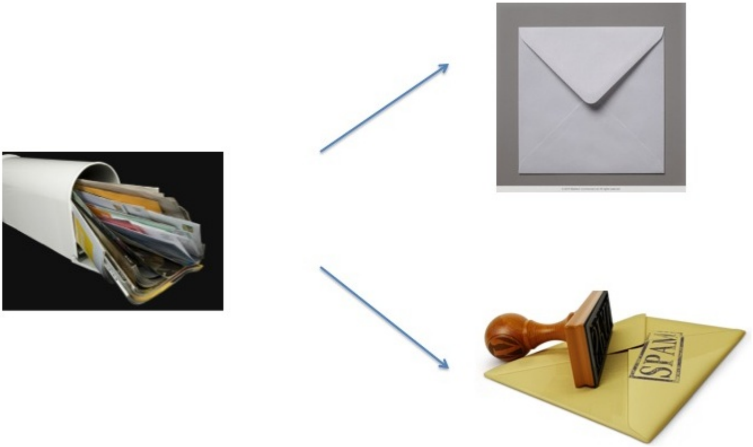
Machine Learning: from Theory to Practice
Lecture 5
Support Vector Machines

F. d'Alché-Buc and E. Le Pennec
email: florence.dalche@telecom-paristech.fr

Fall 2014

Motivation

Build a spam detector



Motivation

Supervised binary classification

Probabilistic and statistical framework

Let X be a random vector $\mathcal{X} = \mathbb{R}^p$
X describes the properties of a message (say, features)
Let Y be a binary discrete variable $\mathcal{Y} = \{-1, 1\}$
Let P be the joint probability law of (X,Y), P is supposed to be fixed but unknown
Let $S_{train} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ be a i.i.d. sample from P.

Motivation

Supervised binary classification

Probabilistic and statistical framework

- From S_{train}, determine $h \in \mathcal{H}$ that minimizes $R(h) = \mathbb{E}_P[L(X, Y, h(X))]$
- L : a local loss function that measures how the predicted class differs from the true class

Pb : the joint probability distribution$P$ is not known, $R(h)$ cannot be computed

Learning by regularizing

- Instead of minimizing $R(h)$, let us minimize the sum of two terms :
- the empirical risk : $R_n(h) = \frac{1}{n}\sum_i L(\mathbf{x}_i, y_i, h(\mathbf{x}_i))$ (the data fitting term) and a regularizing term $\Omega(h)$ that measures the complexity of f.
- We search for : $\arg\min_{h \in \mathcal{F}} R_n(h) + \lambda \Omega(h)$

Motivation

A supervised binary classification problem (ex : spam detector)

Methodology

- Define
  - a representation space of messages

Motivation

A supervised binary classification problem (ex : spam detector)

Methodology

- Define
  - a representation space of messages
  - a class of binary functions

Motivation

A supervised binary classification problem (ex : spam detector)

Methodology

- Define
  - a representation space of messages
  - a class of binary functions
  - a loss function to be minimized to obtain the best classifier in this function class.

Motivation

A supervised binary classification problem (ex : spam detector)

Methodology

- Define
  - a representation space of messages
  - a class of binary functions
  - a loss function to be minimized to obtain the best classifier in this function class.
  - a minimization algorithm for this loss (stochastic gradient, quadratic programming . . .)

Motivation

A supervised binary classification problem (ex : spam detector)

Methodology

- Define
  - a representation space of messages
  - a class of binary functions
  - a loss function to be minimized to obtain the best classifier in this function class.
  - a minimization algorithm for this loss (stochastic gradient, quadratic programming . . .)
  - a method for performance assessment in order to evalue the classifier obtained

Motivation

Encoding documents

Codage Term-Frequency-Inverse Document Frequency (TF-IDF)

- collection C of messages
- a word → a term
- Define a dictionnary D of p terms appearing in C
- a message (document) $d \longrightarrow$ a set of terms with their occurrences
- C : a collection of N documents
- $TF(t, d) = \frac{\text{nb of occurrence of t in d}}{\text{nb of terms in d}}$
- $IDF(t, C) = \log \frac{\text{N}}{\text{nb of documents of } C \text{ where t appears}}$

Motivation

Representation space of messages

TF-IDF encoding of a message d

- a vector $\mathbf{x}$ of dimension p
- $x_i = TF - IDF(t_i, d, C), i = 1, \ldots, p$
- We take : C $= S_{train}$, documents of the training sample
- $\mathcal{X} = \mathbb{R}^p$

Motivation

Supervised classification

What you already known

- Perceptron (linear classifier)
- Logistic regression (linear classifier)
- *k*-nearest neighbors
- Naive Bayes Classifier
- Decision trees
- Ensemble methods on base learners

Motivation

SVM and kernel-based methods

- Lecture 5 : Support vector machines (Oct 27)
- Lecture 6 : SVM in practise and structural risk minimization, (Nov 7)
- Lecture 7 : Kernels and Reproducing Kernel Hilbert Spaces (Nov 14)
- Lecture 8 : Kernel learning (MKL and other methods), large scale kernel-methods (Nov 20 : at Polytechnique)
- Lecture 9 : Practical session (Nov 21) : comparison of several algorithms (tree-based methods, SVM, ...)

Motivation

Outline

1. Motivation
2. Optimal Margin Hyperplanes (separable case)
3. Optimal Margin Hyperplanes (non separable case)
4. Kernels : a brief introduction
5. Exercise
6. References

Optimal Margin Hyperplanes (separable case)

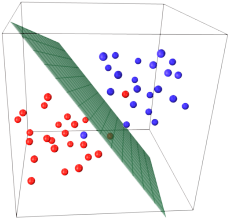
Linear classifier

Definition

$\forall x \in \mathbb{R}^p$

$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$

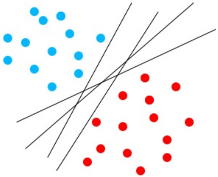
$\mathbf{w}^T\mathbf{x} + b = 0$: hyperplane in $\mathbb{R}^p$



Example in 3D

Optimal Margin Hyperplanes (separable case)

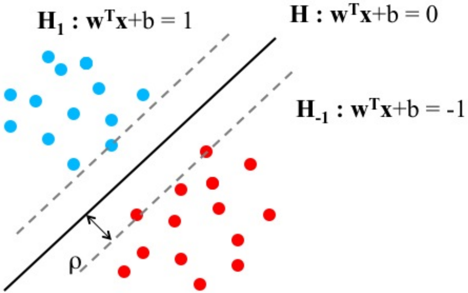
1rst case : linearly separable data



Example in 2D : what line to be chosen ?

Optimal Margin Hyperplanes (separable case)

Margin criterion


$\mathbf{H_1}$: $\mathbf{w^T x}$+b = 1
$\mathbf{H}$: $\mathbf{w^T x}$+b = 0
$\mathbf{H_{-1}}$: $\mathbf{w^T x}$+b = -1
ρ

Optimal Margin Hyperplanes (separable case)

Margin Criterion

Geometrical margin

- We consider 3 hyperplanes in order to separate the training data
  - H : $\mathbf{w}^T\mathbf{x} + b = 0$, $H_1 : \mathbf{w}^T\mathbf{x} + b = 1$, $H_{-1} : \mathbf{w}^T\mathbf{x} + b = -1$
- We call *geometrical margin*, $\rho(\mathbf{w})$ the smallest distance between the data and the hyperplane H, so hal the distance between H_1 and H_{-1}
- A simple calculation gives : $\rho(\mathbf{w}) = \frac{1}{||\mathbf{w}||}$.

Optimal Margin Hyperplanes (separable case)

New cost function to optimize

How to determine$\mathbf{w}$ and b ?

- Maximise the margin $\rho(\mathbf{w})$ while separating data usingH_1 and H_{-1}
- Classify the blue data $(y_i = 1)$: $\mathbf{w}^T\mathbf{x}_i + b \geq 1$
- Classify the red data $(y_i = -1)$: $\mathbf{w}^T\mathbf{x}_i + b \leq -1$

Optimal Margin Hyperplanes (separable case)

Optimization problem

Optimization in the primal space

$$\begin{array}{ll} \underset{\mathbf{w},b}{\text{minimize}} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{under the constraint} & y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1,\ i = 1, \ldots, n. \end{array}$$

Reference

Boser, B. E. ; Guyon, I. M. ; Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. p. 144.

Optimal Margin Hyperplanes (separable case)

Quadratic programming under inequality constraints

- We relax the constraints by introducing them into the objective function under the form of a penalty which is a linear combination of the constraints with Lagrangian coefficients all positive or non null.

Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \mathbf{alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + \mathrm{b}))$$

Optimal Margin Hyperplanes (separable case)

Karush-Kunh-Tucker Conditions

At the extremum (here minimum), we have

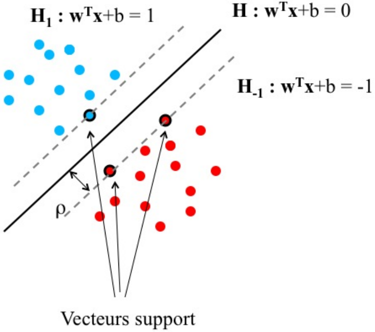
$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$

$$\nabla_{b}\mathcal{L} = - \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\forall i, \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1] = 0$$

Optimal Margin Hyperplanes (separable case)

Support Vectors



$\mathbf{H_1} : \mathbf{w^T x}+b = 1$
$\mathbf{H} : \mathbf{w^T x}+b = 0$
$\mathbf{H_{-1}} : \mathbf{w^T x}+b = -1$
ρ
Vecteurs support

Optimal Margin Hyperplanes (separable case)

Support Vectors

In the expression of $\mathbf{w}$ only some $\mathbf{x}_i$ appear in the expansion

Such vectors are called *support vectors*

These support vectors are such that $\alpha_i \neq 0$ and are on H_1 or on H_{-1} : they satisfy the last KKT condition

NB : b is calculated by applying the last KKT conditions on a given support vector.

Optimal Margin Hyperplanes (separable case)

Optimal Margin Hyperplane (linear SVM)

Once the Lagrangian coefficients are determined,

The equation of an Optimal Margin Hyperplane is :

$$h(\mathbf{x}) = \mathrm{sign}(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b)$$

Thus, to classify a new point $\mathbf{x}$, the classifier combines linearly the class labels y_i of the support vectors with weights of the form : $\alpha_i \mathbf{x}_i^T \mathbf{x}$. The term $\mathbf{x}_i^T \mathbf{x}$ measures how close are $\mathbf{x}$ and the support vectors in the sense of the inner product.

Optimal Margin Hyperplanes (separable case)

How to find the α_i's : solving the problem in the dual space

Apply all the KKT conditions

$$\mathcal{L}(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

- We need to maximize $\mathcal{L}$ under the constraints $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0, \forall i = 1, \ldots, n$
- We then call for a **quadratic solver**

Optimal Margin Hyperplanes (separable case)

Leave-one-out error (1-CV)

SVM enjoys a nice property : we can get a learning guarantee for SVMs based on the fraction of support vectors in the training set. We need for that to recall what is a *leave-one-out error*

Optimal Margin Hyperplanes (separable case)

Leave-one-out error (1-CV)

Definition

Let $h_{\mathcal{S}}$ denote the hypothesis returned by a learning algorithm $\mathcal{A}$, when trained on a fixed sample $\mathcal{S}$. Then, the leave-one-out error of $\mathcal{A}$ on a sample $\mathcal{S}$ of size n is defined by :

$$R_{LOO}(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} 1_{h_{\mathcal{S}}^{-i}(\mathbf{x}_i)\neq y_i}$$

Optimal Margin Hyperplanes (separable case)

Leave-one-out error

Lemma

The average leave-one-out error for samples of size $n \geq 2$ is an unbiased estimate of the average generalization error for samples of size (n-1) :

$$\mathbb{E}_{P^n}[R_{LOO}(\mathcal{A})] = \mathbb{E}_{P^{n-1}}[R(h_{\mathcal{S}'})],$$

where P denotes the distribution according to which points are drawn. $\mathbb{E}_{P^n}$ means that the expectation is taken over n-length samples $\mathcal{S}$ drawn from D

Proof : by linearity of expectation

Optimal Margin Hyperplanes (separable case)

Leave-one-out error for linear SVM in the case of separable data

Theorem

Let $h_{\mathcal{S}}$ be the hypothesis returned by SVM for a training sample $\mathcal{S}$, and let $N_{SV}(\mathcal{S})$ be the number of support vector that define $h_{\mathcal{S}}$. Then,

$$\mathbb{E}_{P^n}[R(h_{\mathcal{S}})] \leq \mathbb{E}_{P^{n+1}}[\frac{N_{SV}(\mathcal{S})}{n+1}]$$

Proof :

Let $\mathcal{S}$ a linearly separable sample of size $n+1$. If $\mathbf{x}$ is not a support vector then removing it does not change the solution. Therefore, $h_{\mathcal{S}-\mathbf{x}} = h_{\mathcal{S}}$ and $h_{\mathcal{S}-\mathbf{x}}$ correctly classifies $\mathbf{x}$. By contraposition, if $h_{\mathcal{S}-\mathbf{x}}$ misclassifies $\mathbf{x}$ then, $\mathbf{x}$ must be a support vector :

$$R_{LOO}(SVM) \leq \frac{N_{SV}(\mathcal{S})}{n+1}.$$

Now take the expectation of both sides and using the previous lemma gives the result.

Optimal Margin Hyperplanes (separable case)

Outline

1. Motivation
2. Optimal Margin Hyperplanes (separable case)
3. Optimal Margin Hyperplanes (non separable case)
4. Kernels : a brief introduction
5. Exercise
6. References

Optimal Margin Hyperplanes (non separable case)

More realistic : Linear SVM when data are not separable

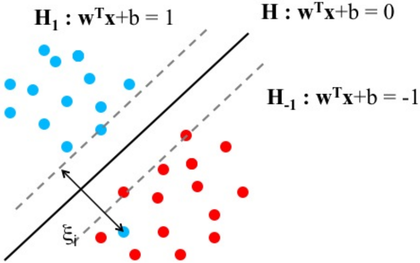
Let us introduce a slack variable ξ_i for each training data :

Solving the problem in the primal space

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\xi} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i \\ \text{under the constraints} & y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1 - \xi_i \; i = 1, \ldots, n. \\ & \xi_i \geq 0 \; i = 1, \ldots, n. \end{array}$$

Optimal Margin Hyperplanes (non separable case)

Soft Margin


$\mathbf{H_1} : \mathbf{w^T x}+b = 1$
$\mathbf{H} : \mathbf{w^T x}+b = 0$
$\mathbf{H_{-1}} : \mathbf{w^T x}+b = -1$
ξ_i

Optimal Margin Hyperplanes (non separable case)

Optimization problem

Solving the pb in the dual

$$\begin{aligned} \max_{\alpha} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{under the constraints} \quad & 0 \leq \alpha_i \leq C \; i = 1, \ldots, n. \\ & \sum_i \alpha_i y_i \; i = 1, \ldots, n. \end{aligned}$$

Optimal Margin Hyperplanes (non separable case)

Remarks

- some vector support can be on the wrong side of H_0 : we talk about "soft margin"
- C is an hyperparameter that controls the compromise between the model complexity and the number of training errors

Optimal Margin Hyperplanes (non separable case)

SVM as a regularized approach

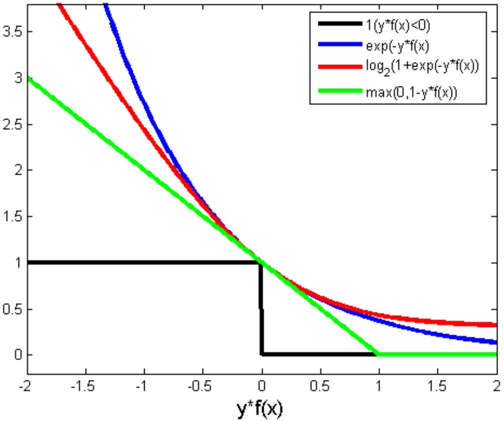
In the primal

$$\min_{\mathbf{w},b} \quad \sum_{i=1}^{n}(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))_+ + \lambda\frac{1}{2}\|\mathbf{w}\|^2$$

with : $(z)_+ = max(0,z)$
$h(\mathbf{x}) = \text{sign}(f(\mathbf{x}))$
Loss : $L(\mathbf{x},y,f(\mathbf{x})) = (1 - yf(\mathbf{x}))_+$
$yf(\mathbf{x})$ is the classifier margin

Optimal Margin Hyperplanes (non separable case)

Different losses



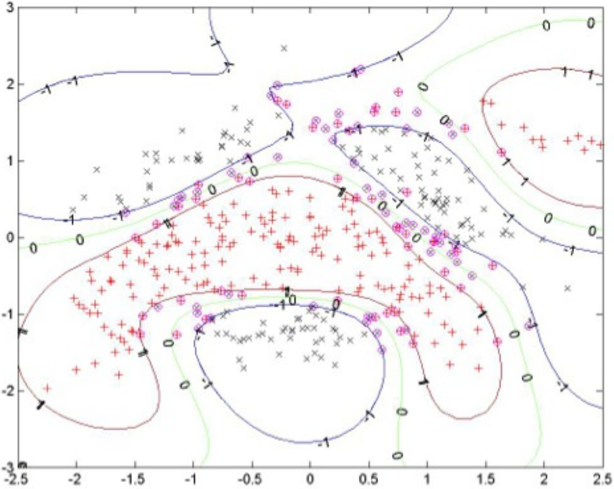
Optimal Margin Hyperplanes (non separable case)

Outline

1. Motivation
2. Optimal Margin Hyperplanes (separable case)
3. Optimal Margin Hyperplanes (non separable case)
4. Kernels : a brief introduction
5. Exercise
6. References

Kernels : a brief introduction

Non linear SVM

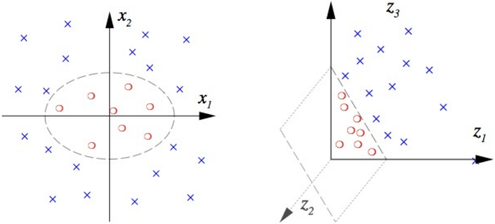


Kernels : a brief introduction

Support Vector Machine and "feature map"

- Transform a problem of nonlinear saparation into a linear one
- How ? In using φ a feature map that maps data into a high-dimensional space where the separation will be easier

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\,x_1 x_2, x_2^2)$$



Kernels : a brief introduction

Kernel trick for the polynomial case

We notice that $\varphi(\mathbf{x}_1)^T\varphi(\mathbf{x}')$ can be calculated without working directly in $\mathbb{R}^3$
We have : $\varphi(\mathbf{x})^T\varphi(\mathbf{x}') = (\mathbf{x}^T\mathbf{x}')^2$
In this case , there is a function, let us call it $k : \mathbb{R}^2\mathbb{R}^2 \rightarrow \mathbb{R}$ defined by :

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}')^2,$$

that satisfies :

$$k(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x})^T\varphi(\mathbf{x}')$$

Definition

Let $\mathcal{X}$ be a non-empty set. Let k :$\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function. k is a positive definite kernel if and only if for any finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ de $\mathcal{X}$ and the column vector $\mathbf{c}$ of $\mathbb{R}^m$, $\mathbf{c}^T K \mathbf{c} = \sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$

Mercer Theorem (a first version)

Theorem

Let k :$\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a symmetric function. k is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{F}$ and a feature map : $\varphi : \mathcal{X} \rightarrow \mathcal{F}$ such that : $k(x, x') =< \varphi(x), \varphi(x') >$

Example : $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma ||\mathbf{x} - \mathbf{x}'||^2)$

Kernels : a brief introduction

Kernel methods

In any (learning) algorithm that only requires inner products, $\mathbf{x}_i^T\mathbf{x}_j$ can be replaced by $k(\mathbf{x}_i, \mathbf{x}_j)$, and the algorithm can be used to work in a high-dimensional space where the separation frontiers are simpler.

Solving the pb in the dual

$$\max_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$$

under the constraints $\quad 0 \leq \alpha_i \leq C \; i = 1, \ldots, n.$

$$\sum_i \alpha_i y_i \; i = 1, \ldots, n.$$

Kernels : a brief introduction

Kernel-based SVM

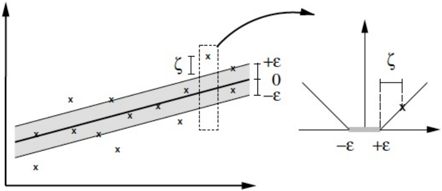
How to do nonlinear separation frontiers with SVM ?

Replace any $\mathbf{x}_i$ by $\varphi(\mathbf{x}_i)$ in the SVM algorithm
When predicting, h uses as well $\varphi(\mathbf{x})^T\varphi(\mathbf{x}_i)$:
$h(\mathbf{x}) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b)$

Exercise - Lecture 5 -

Extend the same ideas to regression

As an exercise, you can address a linear regression problem where the loss is the ε-insensitive loss depicted in the figure below :



Exercise - Lecture 5 -

To do for Nov 7

(1) With all the training data in the ε-tube (no ξ variables), the problem states as :
how to find $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ such that

$$\begin{array}{ll} \min\limits_{\mathbf{w},b} & \frac{1}{2}||\mathbf{w}||^2 \\ \text{under the constraints} & y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \varepsilon \; i = 1, \ldots, n. \\ & \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \varepsilon \; i = 1, \ldots, n. \end{array}$$

(2) However as in the classification case, you may want relax this problem using slack variables (be careful that you will need two kinds of slack variables to deal with each kind of the constraints. Write the Lagrangian and solve the problem in the dual space in both cases.

References

Bibliography

Video-lectures :

- MIT course (Patrick Winston)
  https://www.youtube.com/watch?v=_PwhiWxHK8o
- Stanford (Andrew Ng)
  https://www.youtube.com/watch?v=s8B4A5ubw6c
- Caltech (Yaser Avu-Mostapha)
  https://www.youtube.com/watch?v=XUj5JbQihlU
- Books :
  - Foundations of Machine Learning, Mohri, Rostamizadeh, Talwalkar, MIT Press, 2012
  - Pattern recognition, Chris Bishop, MIT Press.