Machine Learning

TSIA-SD 210 - P3

Lecture 2 - 1. A First Linear Classifier: the optimal margin hyperplane

Florence d'Alché-Buc

Contact: florence.dalche@telecom-paristech.fr,
2A Filière SD, Télécom ParisTech,Université of Paris-Saclay, France

Table of contents

Outline

Statistical learning: a methodology

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

Statistical learning for supervised classification

Two main family 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 $\hat{p}(x|Y = 1)$, $\hat{p}(x|Y = -1)$ and prior probabilities.

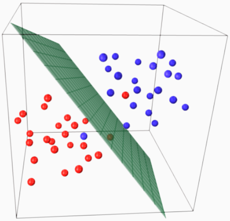
Outline

Définition

Soit $\mathbf{x} \in \mathbb{R}^p$

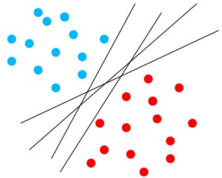
$h(\mathbf{x}) = \text{signe}(\mathbf{w}^T\mathbf{x} + b)$

L'équation : $\mathbf{w}^T\mathbf{x} + b = 0$ définit un hyperplan dans l'espace euclidien $\mathbb{R}^p$



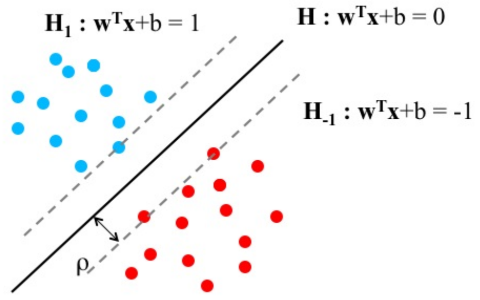
Example in 3D

Data linearly separables



What to choose ?

Margin criterion



Margin Criterion

Geometrical margin

- To separate data, let us consider a triplet of hyperplanes:
  - 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 *géométrical margin*, $\rho(\mathbf{w})$ the smallest distance between the data and Hyperplane H thus, here half of the distance between H_1 and H_{-1}
- A simple calculation gives : $\rho(\mathbf{w}) = \frac{1}{||\mathbf{w}||}$.

New objective function to optimize

How to find w and b ?

- Maximmize the margin $\rho(\mathbf{w})$ while separating the data using H_1 and H_{-1}
- Classify the blue data $(y_i = 1)$: $\mathbf{w}^T\mathbf{x}_i + b \geq 1$
- Classify the red data $(y_i = -1)$: $\mathbf{w}^T\mathbf{x}_i + b \leq -1$

Linear SVM: separable case

Optimization in the primal

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

Référence

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.

Programming under inequality constraints

Problem of the following kind:
$\min_x f(x)$
s.c. $g(x) \leq 0$

- Here: $g(x)$: linear constraints
- f is strictly convex

1. Lagrangian: $J(x, \lambda) = f(x) + \lambda g(x)$, $\lambda \geq 0$

Programming under inequality constraints

$$\underset{\mathbf{w},b}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{under constraints} \quad 1 - y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \leq 0, \ i = 1, \ldots, n.$$

Lagrangian

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

$$\forall i, \alpha_i \geq 0$$

Karush-Kuhn-Tucker conditions

In the extremum, we have:

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

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

$$\forall i, \alpha_i \geq 0$$

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

Obtaining the α_i 's : solution the dual

space

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

- Maximize $\mathcal{L}$ under the constraints $\alpha_i \geq 0$ et $\sum_i \alpha_i y_i = 0, \forall i = 1, \ldots, n$
- Call for a quadratic solver

Optimal Margin Hyperplan (linear SVM)

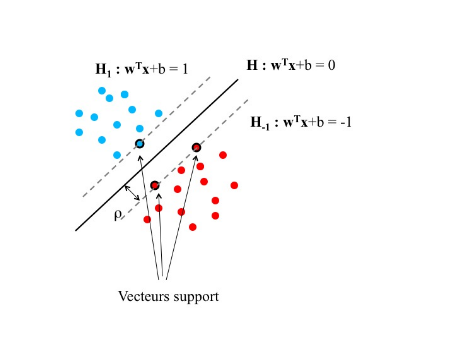
Assume the Lagrangian coefficients α_i have been found :

Linear SVM equation

$$f(\mathbf{x}) = \text{signe}(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b)$$

To classify a novel $\mathbf{x}$, this classifier makes all the support data vote with an importance weight equal to $\alpha_i \mathbf{x}_i^T \mathbf{x}$ that measures how much $\mathbf{x}$ is close to the support data.

Support Vectors



Training data $\mathbf{x}_i$ such that $\alpha_i \neq 0$ belong to either H_1 or H_{-1}. Only those data, called **support vectors**, are taking into account in $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

NB : b is obtained by choosing one (or all) support data such that $(\alpha_i \neq 0)$

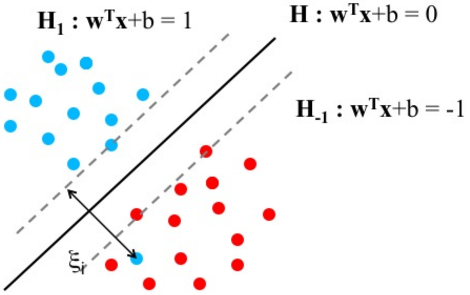
Realistic case: linear SVM in the case of nonlinearly separable data

For each training data, introduce a slack variableξ_i :

New problem in the primal space

$$
\begin{aligned}
\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{sous les contraintes} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + \mathrm{b}) \geq 1 - \xi_i \;\; i = 1, \ldots, n. \\
& \xi_i \geq 0 \;\; i = 1, \ldots, n.
\end{aligned}
$$

Realistic case: linear SVM in the case of nonlinearly separable data



Realistic case: linear SVM in the case of nonlinearly separable data

Dual problem

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

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

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

Karush-Kuhn-Tucker Conditions (KKT)

Let α^* be the solution of the ldual problem:

$$\forall i, [y_i f_{w^*,b^*}(x_i) - 1 + \xi_i^*] \leq 0 \quad (1)$$

$$\forall i, \alpha_i^* \geq 0 \quad (2)$$

$$\forall i, \alpha_i^*[y_i f_{w^*,b^*}(x_i) - 1 + \xi_i^*] = 0 \quad (3)$$

$$\forall i, \mu_i^* \geq 0 \quad (4)$$

$$\forall i, \mu_i^* \xi_i^* = 0 \quad (5)$$

$$\forall i, \alpha_i^* + \mu_i^* = C \quad (6)$$

$$\forall i, \xi_i^* \geq 0 \quad (7)$$

$$\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i \quad (8)$$

$$\sum_i \alpha_i^* y_i = 0 \quad (9)$$

$$(10)$$

Vairous cases

- if $\alpha_i^* = 0$, then $\mu_i^* = C > 0$ and thus, $\xi_i^* = 0$: x_i is well classified
- if $0 < \alpha_i^* < C$ then $\mu_i^* > 0$ and thus, $\xi_i^* = 0$: x_i is such that: $y_i f(x_i) = 1$
- if $\alpha_i^* = C$, then $\mu_i^* = 0$, $\xi_i^* = 1 - y_i f_{w^*,b^*}(x_i)$

NB : we compute b^* by using i such that $0 < \alpha_i^* < C$

Realistic case: linear SVM with soft margin

A few remarks

- some of the support are in the wrong side
- C is a hyperparameter that controls the compromise between the model complexity and the training classification error

SVM as a penalized regression problem

Optimisation dans l'espace primal

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

Avec: $(z)_+ = max(0, z)$
$f(\mathbf{x}) = \text{signe}(h(\mathbf{x}))$
Loss function: $L(\mathbf{x}, y, h(\mathbf{x})) = (1 - yh(\mathbf{x}))_+$
$yh(\mathbf{x})$ is called the classifier margin

Outline

Introduction

Linear SVM

References

References

- BOSER, Bernhard E., Isabelle M. GUYON, and Vladimir N. VAPNIK, 1992. A training algorithm for optimal margin classifiers. In: COLT 92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York, NY, USA: ACM Press, pp. 144-152.
- CORTES, Corinna, and Vladimir VAPNIK, 1995. Support-vector networks. Machine Learning, 20(3), 273297.
- Article vraiment sympa, complet (un peu de maths) : A tutorial review of RKHS methods in Machine Learning, Hofman , Schoelkopf, Smola, 2005 (https://www.researchgate.net/publication/228827159_A_Tutorial_Review_of_RKHS_Methods_in_Machine_Learning)