4 - Support Vector Machines for (supervised) classification problems

Mauro Passacantando

Department of Computer Science, University of Pisa
mauro.passacantando@unipi.it

Optimization Methods and Game Theory
Master of Science in Artificial Intelligence and Data Engineering
University of Pisa – A.Y. 2020/21

Supervised pattern classification

Given a set of objects partitioned in several classes with known labels, we want to predict the class of any new future object with unknown label.

Examples:

- handwritten digits recognition
- spam filtering
- credit card fraud detection
- marketing
- object recognition
- medical diagnosis
  (see, e.g., the following recent video (in italian)
  `https://video.repubblica.it/dossier/coronavirus-wuhan-2020/coronavirus-a-roma-si-usa-l-intelligenza-artificiale-per-abbatt`
  `356375/356940?ref=RHPPTP-BH-I251664519-C12-P6-S4.3-T1)`

Methods:

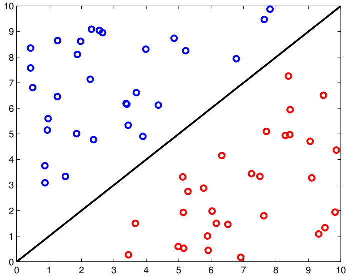
- Decision trees
- Artificial Neural Networks
- Support Vector Machines

Linear SVM

Consider binary classification.
We have two finite sets $A, B \subset \mathbb{R}^n$ with known labels (1 for points in A, -1 for points in B). $\mathbb{R}^n$ is the input space, $A \cup B$ is the training set.

Assume that A and B are linearly separable, i.e., there is an hyperplane $H = \{x \in \mathbb{R}^n : \; w^\mathsf{T} x + b = 0\}$ such that

$$\begin{aligned} w^\mathsf{T} x^i + b > 0 \qquad & \forall\, x^i \in A, \\ w^\mathsf{T} x^j + b < 0 \qquad & \forall\, x^j \in B. \end{aligned}$$



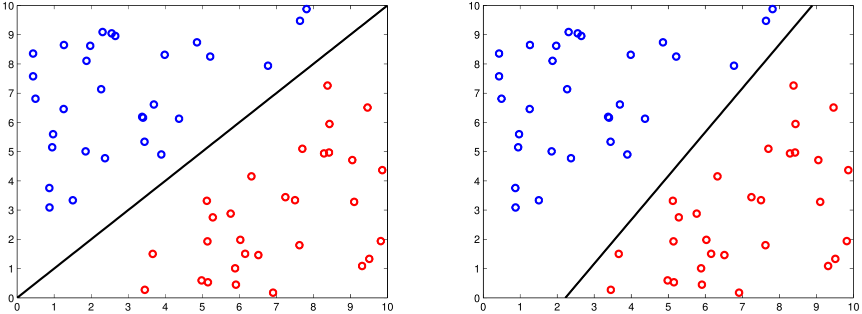
We have a new test data x:

use the decision function
$f(x) = \text{sign}(w^\mathsf{T} x + b) = \begin{cases} 1 & \text{if } w^\mathsf{T} x + b > 0, \\ -1 & \text{if } w^\mathsf{T} x + b < 0. \end{cases}$

What is a necessary and sufficient condition for A and B to be linearly separable?

Linear SVM

There are many possible separating hyperplanes. Which hyperplane do we choose?

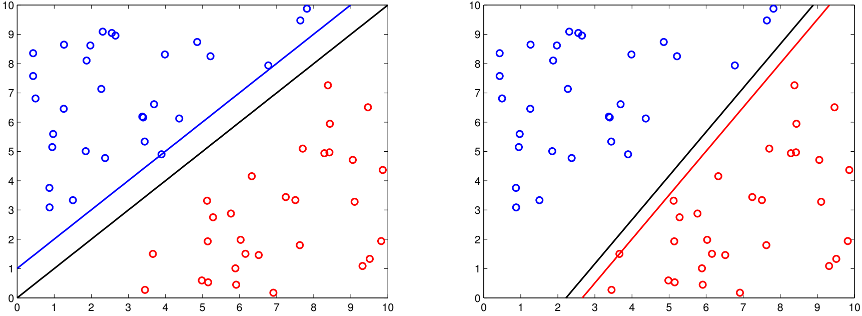


Linear SVM

Definition

If H is a separating hyperplane, then the margin of separation of H is defined as the minimum distance between H and $A \cup B$, i.e.

$$\rho(H) = \min_{x \in A \cup B} \frac{|w^\mathsf{T} x + b|}{\|w\|}.$$



Linear SVM

We look for the separating hyperplane with the maximum margin of separation.

Theorem

Finding the separating hyperplane with the maximum margin of separation is equivalent to solve the following convex quadratic programming problem:

$$\begin{cases} \min\limits_{w,b} \|w\|^2 \\ w^{\mathsf{T}} x^i + b \geq 1 \qquad \forall\, x^i \in A \\ w^{\mathsf{T}} x^j + b \leq -1 \qquad \forall\, x^j \in B \end{cases} \tag{1}$$

Proof. If $H = \{w^{\mathsf{T}} x + b = 0\}$ is a separating hyperplane, then there are $\alpha, \beta > 0$ s.t.

$$w^{\mathsf{T}} x^i + b \geq \alpha \qquad \forall\, x^i \in A, \qquad\qquad w^{\mathsf{T}} x^j + b \leq -\beta \qquad \forall\, x^j \in B.$$

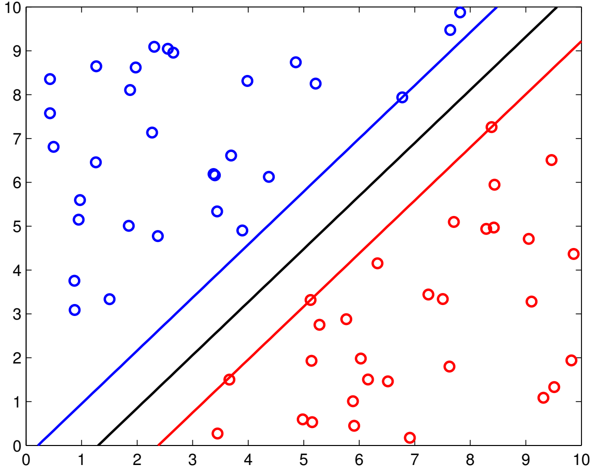
Then the hyperplane $\widetilde{H} = \{\widetilde{w}^{\mathsf{T}} x + \widetilde{b} = 0\}$, where $\widetilde{w} = 2\,w/(\alpha+\beta)$ and $\widetilde{b} = (2\,b - \alpha + \beta)/(\alpha+\beta)$, is another separating hyperplane, parallel to H, s.t.

$$\begin{aligned} &\widetilde{w}^{\mathsf{T}} x^i + \widetilde{b} \geq 1 \qquad \forall\, x^i \in A, \\ &\widetilde{w}^{\mathsf{T}} x^j + \widetilde{b} \leq -1 \qquad \forall\, x^j \in B, \\ &\rho(H) \leq \rho(\widetilde{H}) = \frac{1}{\|\widetilde{w}\|}. \end{aligned}$$

Moreover, it can be proved that problem (1) has a unique solution (w^*, b^*). □

Linear SVM

Exercise 4.1. Find the separating hyperplane with maximum margin for the data set given in the file `4-1.txt`.



Linear SVM

Let $\ell = |A \cup B|$. For any point $x^i \in A \cup B$, define a label

$$y^i = \begin{cases} 1 & \text{if } x^i \in A \\ -1 & \text{if } x^i \in B \end{cases} \qquad \forall\ i = 1, \dots, \ell.$$

Then the problem

$$\begin{cases} \min\limits_{w,b}\ \|w\|^2 \\ w^{\mathsf{T}} x^i + b \geq 1 \qquad \forall\ x^i \in A \\ w^{\mathsf{T}} x^j + b \leq -1 \qquad \forall\ x^j \in B \end{cases}$$

is equivalent to

$$\text{linear SVM} \quad \begin{cases} \min\limits_{w,b}\ \dfrac{1}{2}\|w\|^2 \\ 1 - y^i(w^{\mathsf{T}} x^i + b) \leq 0 \qquad \forall\ i = 1, \dots, \ell \end{cases} \tag{2}$$

It is useful to consider the Lagrangian dual of problem (2).

Linear SVM

The Lagrangian function is

$$
\begin{aligned}
L(w,b,\lambda) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{\ell} \lambda_i \left[1 - y^i(w^{\mathsf{T}} x^i + b)\right] \\
&= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y^i w^{\mathsf{T}} x^i - b \sum_{i=1}^{\ell} \lambda_i y^i + \sum_{i=1}^{\ell} \lambda_i
\end{aligned}
$$

If $\sum_{i=1}^{\ell} \lambda_i y^i \neq 0$, then $\min_{w,b} L(w,b,\lambda) = -\infty$.

If $\sum_{i=1}^{\ell} \lambda_i y^i = 0$, then L does not depend on b, L is strongly convex wrt w and $\arg\min_{w} L(w,b,\lambda)$ is given by the (unique) stationary point

$$
\nabla_w L(w,b,\lambda) = w - \sum_{i=1}^{\ell} \lambda_i y^i x^i = 0.
$$

Therefore, the dual function is

$$
\varphi(\lambda) = \begin{cases} -\infty & \text{if } \sum_{i=1}^{\ell} \lambda_i y^i \neq 0 \\ -\frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y^i y^j (x^i)^{\mathsf{T}} x^j \lambda_i \lambda_j + \sum_{i=1}^{\ell} \lambda_i & \text{if } \sum_{i=1}^{\ell} \lambda_i y^i = 0 \end{cases}
$$

Linear SVM

The dual of problem (2) is

$$\begin{cases} \max\limits_{\lambda} \ -\dfrac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j (x^i)^\mathsf{T} x^j \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{cases}$$

or

$$\begin{cases} \max\limits_{\lambda} \ -\dfrac{1}{2}\lambda^\mathsf{T} X^\mathsf{T} X \lambda + e^\mathsf{T}\lambda \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{cases} \tag{3}$$

where the $n \times \ell$ matrix $X = (y^1 x^1, y^2 x^2, \dots, y^\ell x^\ell)$ and the vector $e^\mathsf{T} = (1, \dots, 1)$.

Linear SVM

- Dual problem is a convex quadratic programming problem
- Dual constraints are simpler than primal constraints
- Dual problem has optimal solutions: each KKT multiplier λ^* associated to the primal optimum (w^*, b^*) is a dual optimum
- If $\lambda_i^* > 0$, then x^i is said support vector
- If λ^* is a dual optimum, then

$$w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i x^i.$$

- b^* is obtained using the complementarity conditions:

$$\lambda_i^* \left[1 - y^i((w^*)^\mathsf{T} x^i + b^*)\right] = 0;$$

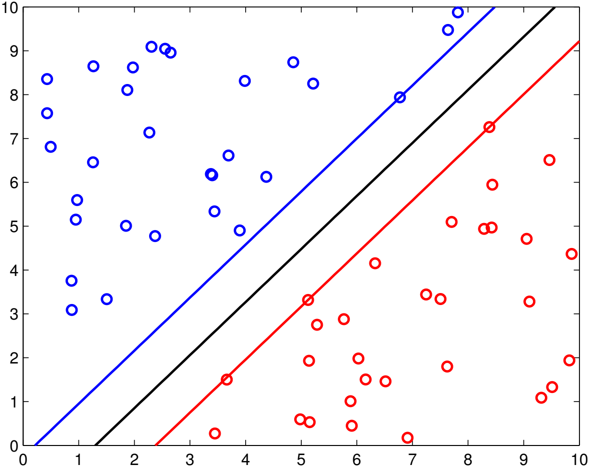
in fact, if i is such that $\lambda_i^* > 0$, then $b^* = \dfrac{1}{y^i} - (w^*)^\mathsf{T} x^i$.

- Finally, the decision function is

$$f(x) = \mathrm{sign}((w^*)^\mathsf{T} x + b^*).$$

Linear SVM

Exercise 4.2. Find the separating hyperplane with maximum margin for the data set given in the file `4-1.txt` by solving the dual problem (3).



Linear SVM with soft margin

What if sets A and B are not linearly separable?

The linear system

$$1 - y^i(w^\mathsf{T} x^i + b) \le 0 \qquad i = 1, \dots, \ell$$

has no solutions.

We introduce slack variables $\xi_i \ge 0$ and consider the (relaxed) system:

$$\begin{array}{ll} 1 - y^i(w^\mathsf{T} x^i + b) \le \xi_i & i = 1, \dots, \ell \\ \xi_i \ge 0 & i = 1, \dots, \ell \end{array}$$

If x^i is misclassified, then $\xi_i > 1$, thus $\sum_{i=1}^{\ell} \xi_i$ is an upper bound of the number of misclassified points.

We add to the objective function the term $C \sum_{i=1}^{\ell} \xi_i$, where $C > 0$ is a parameter:

linear SVM
with sotf margin

$$\begin{cases} \min\limits_{w,b,\xi} \ \dfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \\ 1 - y^i(w^\mathsf{T} x^i + b) \le \xi_i & \forall\ i = 1, \dots, \ell \\ \xi_i \ge 0 & \forall\ i = 1, \dots, \ell \end{cases} \tag{4}$$

Linear SVM with soft margin

Exercise 4.3. Prove that the dual problem of (4) is

$$\begin{cases} \max\limits_{\lambda} \ -\dfrac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j (x^i)^\mathsf{T} x^j \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \le \lambda_i \le C \qquad i = 1, \dots, \ell \end{cases} \tag{5}$$

If λ^* is optimum for (5), then

$$w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i x^i.$$

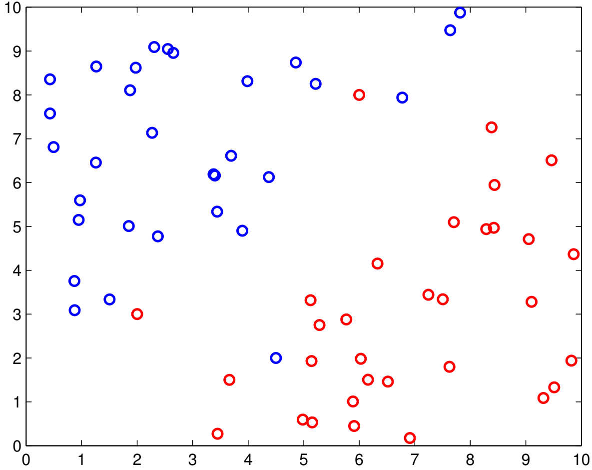
Find b^* choosing i s.t. $0 < \lambda_i^* < C$ and using the complementarity conditions:

$$\begin{cases} \lambda_i^* \left[1 - y^i((w^*)^\mathsf{T} x^i + b^*) - \xi_i^* \right] = 0 \\ (C - \lambda_i^*)\xi_i^* = 0 \end{cases}$$

Thus $b^* = \dfrac{1}{y^i} - (w^*)^\mathsf{T} x^i$.

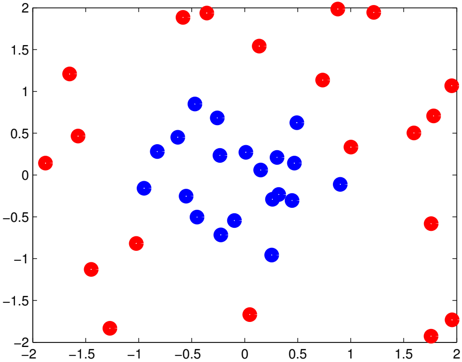
Linear SVM with soft margin

Exercise 4.4. Find the separating hyperplane for the data set given in the file `4-4.txt` by solving the dual problem (5) with $C = 10$.
What is the value of λ_i corresponding to the misclassified points?



Nonlinear SVM

Consider now two sets A and B which are not linearly separable.



Are they linearly separable in other spaces?

Use a map $\phi : \mathbb{R}^n \to \mathcal{H}$, where $\mathcal{H}$ is an higher dimensional (maybe infinite) space.

$\mathcal{H}$ is called the features space

We try to linearly separate the images $\phi(x^i)$, $i = 1, \ldots, \ell$ in the feature space.

Nonlinear SVM

Primal problem:

$$\begin{cases} \min\limits_{w,b,\xi} \ \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell} \xi_i \\ 1 - y^i(w^{\mathsf{T}}\phi(x^i) + b) \leq \xi_i & \forall\ i = 1,\dots,\ell \\ \xi_i \geq 0 & \forall\ i = 1,\dots,\ell \end{cases}$$

w is a vector in a high dimensional space (maybe infinite variables)

Dual problem:

$$\begin{cases} \max\limits_{\lambda} \ -\dfrac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j \phi(x^i)^{\mathsf{T}}\phi(x^j)\lambda_i\lambda_j + \sum\limits_{i=1}^{\ell}\lambda_i \\ \sum\limits_{i=1}^{\ell}\lambda_i y^i = 0 \\ 0 \leq \lambda_i \leq C & \forall\ i = 1,\dots,\ell \end{cases}$$

number of variables = number of training data

Nonlinear SVM

- Solve dual problem λ^*
- Compute $w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i \phi(x^i)$
- Use any λ_i^* s.t. $0 < \lambda_i^* < C$ for finding b^*:

$$y^i \left[\sum_{j=1}^{\ell} \lambda_j^* y^j \phi(x^j)^\mathsf{T} \phi(x^i) + b^* \right] - 1 = 0$$

Decision function

$$f(x) = \text{sign}((w^*)^\mathsf{T} \phi(x) + b^*) = \text{sign} \left(\sum_{i=1}^{\ell} \lambda_i^* y^i \phi(x^i)^\mathsf{T} \phi(x) + b^* \right)$$

depends on

- $\lambda^* \to$ know $\phi(x^i)^\mathsf{T} \phi(x^j)$
- $\phi(x^i)^\mathsf{T} \phi(x)$
- $b^* \to$ know $\phi(x^i)^\mathsf{T} \phi(x^j)$

No need to explicitly know $\phi(x)$, but only $\phi(x)^\mathsf{T} \phi(y)$

Nonlinear SVM

We use kernel functions.

Definition

A function $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called kernel if there exists a map $\phi : \mathbb{R}^n \to \mathcal{H}$ such that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is a scalar product in $\mathcal{H}$.

Examples:

- $k(x, y) = x^{\mathsf{T}} y$
- $k(x, y) = (x^{\mathsf{T}} y + 1)^p$, with $p \geq 1$ (polynomial)
- $k(x, y) = e^{-\gamma \|x-y\|^2}$ (Gaussian)
- $k(x, y) = \tanh(\beta x^{\mathsf{T}} y + \gamma)$, with suitable β and γ

Nonlinear SVM

Theorem

If $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a kernel and $x^1, \dots, x^\ell \in \mathbb{R}^n$, then the matrix K defined as follows

$$K_{ij} = k(x^i, x^j)$$

is positive semidefinite.

The dual problem depends on the kernel k:

$$\begin{cases} \max\limits_{\lambda} \ -\dfrac{1}{2} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{\ell} y^i y^j \color{red}{k(x^i, x^j)} \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \le \lambda_i \le C \qquad i = 1, \dots, \ell \end{cases}$$

Nonlinear SVM

In practice:

- choose a kernel k
- find an optimal solution λ^* of the dual
- choose i s.t. $0 < \lambda_i^* < C$ and find b^*:

$$b^* = \frac{1}{y^i} - \sum_{j=1}^{\ell} \lambda_j^* y^j k(x^i, x^j)$$

- Decision function

$$f(x) = \text{sign}\left(\sum_{i=1}^{\ell} \lambda_i^* y^i k(x^i, x) + b^* \right)$$

Separating surface $f(x) = 0$ is

- linear in the features space
- nonlinear in the input space

Nonlinear SVM

Exercise 4.5. Find the optimal separating surface for the data set given in the file `4-5.txt` using a Gaussian kernel with parameters $C = 1$ and $\gamma = 1$.