

Support Vector Machines

Mauro Passacantando

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

Numerical Methods and Optimization
Master in Computer Science – University of Pisa

Machine learning

Extract knowledge from data

- ▶ classification
- ▶ regression
- ▶ clustering
- ▶ ...

Data classification

Given training data in different classes (labels **known**)

Predict test data (labels **unknown**)

Examples:

- ▶ handwritten digits recognition
- ▶ spam filtering
- ▶ credit card fraud detection
- ▶ marketing
- ▶ medical diagnosis
- ▶ image processing
- ▶ ...

Methods:

- ▶ Decision tree
- ▶ neural networks
- ▶ **support vector machines**

Linear SVM

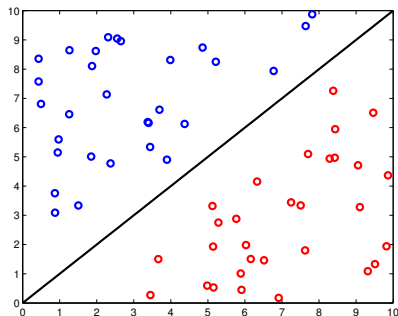
Given two sets $A, B \subset \mathbb{R}^n$ (training set).

Assume that A and B are linearly separable, i.e., there is an hyperplane

$H = \{x \in \mathbb{R}^n : w^T x + b = 0\}$ such that

$$w^T x^i + b > 0 \quad \forall x^i \in A$$

$$w^T x^j + b < 0 \quad \forall x^j \in B$$

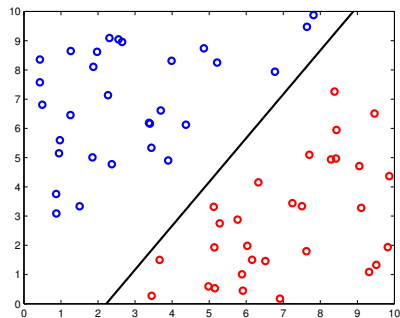
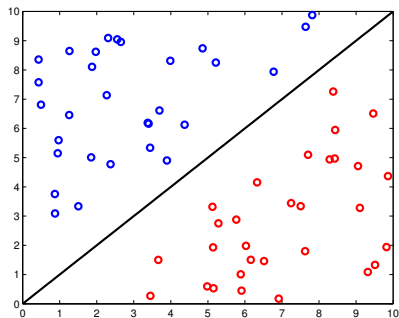


Test data x :

decision function $f(x) = \text{sign}(w^T x + b)$

Linear SVM

Many possible choices of w and b . Which hyperplane do we choose?

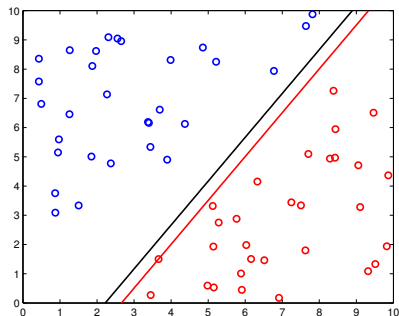
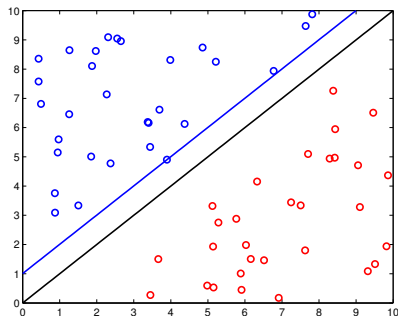


Linear SVM

Definition

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

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



Linear SVM

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

Theorem

This problem is equivalent to

$$\begin{cases} \min \|w\|^2 \\ w^T x^i + b \geq 1 & \forall x^i \in A \\ w^T x^j + b \leq -1 & \forall x^j \in B \end{cases} \quad (1)$$

Proof. If $H = \{w^T x + b = 0\}$ is a separating hyperplane, i.e.,

$$\begin{aligned} w^T x^i + b &\geq \alpha & \forall x^i \in A \\ w^T x^j + b &\leq -\beta & \forall x^j \in B \end{aligned}$$

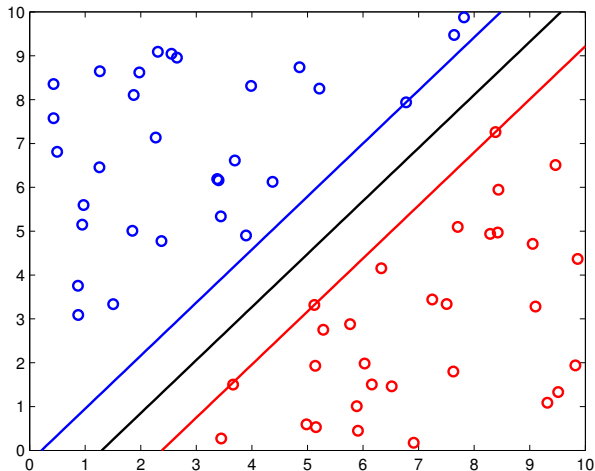
then $\tilde{H} = \{\tilde{w}^T x + \tilde{b} = 0\}$, where $\tilde{w} = 2w/(\alpha + \beta)$ and $\tilde{b} = (2b - \alpha + \beta)/(\alpha + \beta)$, is another separating hyperplane s.t.

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

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

Linear SVM

Exercise. Find the separating hyperplane with maximum margin (data in svm1.txt).



Linear SVM

Define labels

$$y^i = \begin{cases} 1 & \text{if } x^i \in A \\ -1 & \text{if } x^i \in B \end{cases}$$

Then the problem

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

is equivalent to

$$\begin{cases} \min \frac{1}{2} \|w\|^2 \\ 1 - y^i (w^T x^i + b) \leq 0 & \forall i = 1, \dots, \ell \end{cases} \quad (2)$$

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

Linear SVM

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

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

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

Dual function

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

Linear SVM

Dual problem is

$$\left\{ \begin{array}{l} \max -\frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y^i y^j (x^i)^T x^j \lambda_i \lambda_j + \sum_{i=1}^{\ell} \lambda_i \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{array} \right.$$

or

$$\left\{ \begin{array}{l} \max -\frac{1}{2} \lambda^T X^T X \lambda + e^T \lambda \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{array} \right.$$

where $X = (y^1 x^1, y^2 x^2, \dots, y^{\ell} x^{\ell})$ and $e^T = (1, \dots, 1)$.

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

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

and b^* is obtained using one of the complementarity conditions:

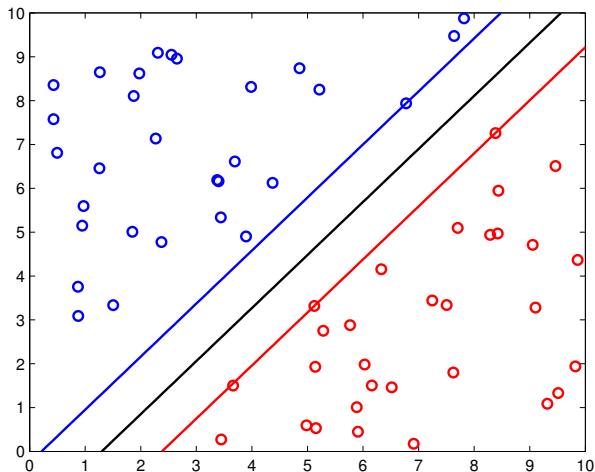
$$\lambda_i^* [1 - y^i ((w^*)^T x^i + b^*)] = 0.$$

- ▶ Decision function is

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

Linear SVM

Exercise. Find the separating hyperplane with maximum margin by solving the dual problem (data in svm1.txt).



Linear SVM with soft margin

If sets A and B are not linearly separable?

The linear system

$$1 - y^i(w^T x^i + b) \leq 0 \quad i = 1, \dots, \ell$$

has no solutions. We introduce slack variables $\xi_i \geq 0$ and consider the system:

$$\begin{aligned} 1 - y^i(w^T x^i + b) &\leq \xi_i & i = 1, \dots, \ell \\ \xi_i &\geq 0 & i = 1, \dots, \ell \end{aligned}$$

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 $C \sum_{i=1}^{\ell} \xi_i$, with $C > 0$:

$$\begin{cases} \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \\ 1 - y^i(w^T x^i + b) \leq \xi_i & \forall i = 1, \dots, \ell \\ \xi_i \geq 0 & \forall i = 1, \dots, \ell \end{cases} \quad (3)$$

Linear SVM with soft margin

Exercise. Prove that the dual problem of (3) is

$$\left\{ \begin{array}{l} \max -\frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y^i y^j (x^i)^T x^j \lambda_i \lambda_j + \sum_{i=1}^{\ell} \lambda_i \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \leq \lambda_i \leq C \quad i = 1, \dots, \ell \end{array} \right.$$

If λ^* is dual optimum, then

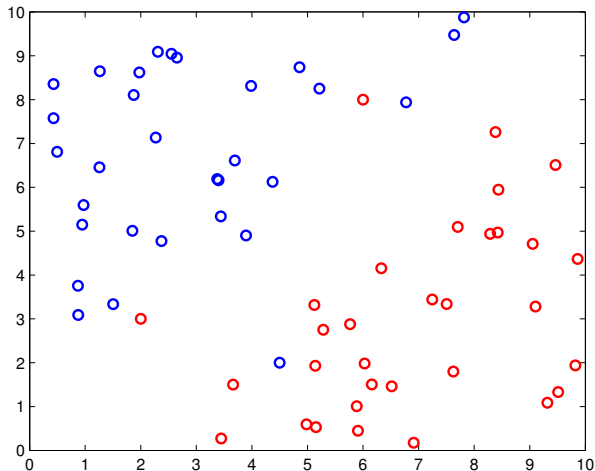
$$w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i x^i.$$

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

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

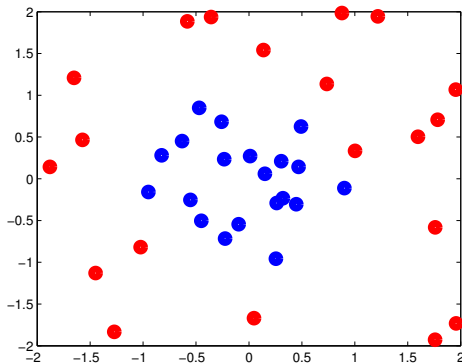
Linear SVM with soft margin

Exercise. Find the optimal hyperplane by solving the dual problem with $C = 10$ (data in svm2.txt).



Nonlinear SVM

Consider sets A and B which are not linearly separable.



Are they linearly separable in other spaces?

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

\mathcal{H} is the **features space**

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

Nonlinear SVM

Primal problem:

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

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

Dual problem:

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

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^* :

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

Decision function

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

depends on

- ▶ $\lambda^* \rightarrow$ know $\phi(x^i)^T \phi(x^j)$
- ▶ $\phi(x^i)^T \phi(x)$
- ▶ $b^* \rightarrow$ know $\phi(x^i)^T \phi(x^j)$

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

Nonlinear SVM

We use **kernel functions**

Definition

A function $k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a kernel if there is a function $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ s.t.

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

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

Examples:

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

Nonlinear SVM

Theorem

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

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

is positive semidefinite.

Dual problem is convex:

$$\left\{ \begin{array}{l} \max -\frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y^i y^j k(x^i, x^j) \lambda_i \lambda_j + \sum_{i=1}^{\ell} \lambda_i \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \leq \lambda_i \leq C \quad i = 1, \dots, \ell \end{array} \right.$$

Nonlinear SVM

In practice:

- ▶ choose a kernel k
- ▶ solve the dual \rightarrow find λ^*
- ▶ find b^* :

$$b^* = \frac{1}{y^i} - \sum_{j=1}^{\ell} \lambda_j^* y^j k(x^i, x^j), \quad \text{for some } i \text{ s.t. } 0 < \lambda_i^* < C$$

- ▶ 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. Find the separating surface using a Gaussian kernel with $C = 1$, $\gamma = 1$ (data in svm3.txt).

