Data fitting problems and Support Vector Machines for classification problems

Mauro Passacantando

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

Optimization Methods
Master of Science in Embedded Computing Systems – University of Pisa
http://pages.di.unipi.it/passacantando/om/OM.html

Data fitting

m experimental data $b_1, b_2, \ldots, b_m \in \mathbb{R}$ corresponding to observations made on points $a_1, a_2, \ldots, a_m \in \mathbb{R}$.
We want to find the best approximation of experimental data with a polynomial of degree $n-1$, with $n \leq m$.
Polynomial with coefficients $x_1, \ldots, x_n$:

$$p(a) = x_1 + x_2\, a + x_3\, a^2 + \cdots + x_n\, a^{n-1}$$

The residual is $r \in \mathbb{R}^m$ s.t. $r_i = p(a_i) - b_i$.
We want to find x s.t. $\|r\|$ is minimum, i.e.

$$\begin{cases} \min \ \|A\,x - b\| \\ x \in \mathbb{R}^n \end{cases}$$

where

$$A = \begin{pmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \ldots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_m & a_m^2 & \ldots & a_m^{n-1} \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Data fitting

For any norm, the function $\|Ax - b\|$ is convex.

Euclidean norm $\|\cdot\|_2$ (least squares approximation) $\rightarrow$ quadratic programming problem:

$$\begin{cases} \min \ \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}(Ax - b)^{\mathsf{T}}(Ax - b) = \frac{1}{2}x^{\mathsf{T}}A^{\mathsf{T}}Ax - x^{\mathsf{T}}A^{\mathsf{T}}b - \frac{1}{2}b^{\mathsf{T}}b \\ x \in \mathbb{R}^n \end{cases}$$

$\text{rank}(A) = n$, thus $A^{\mathsf{T}}A$ is positive definite, the unique solution is given by the system of equations:

$$A^{\mathsf{T}}Ax = A^{\mathsf{T}}b$$

Data fitting

norm $\|\cdot\|_1 \rightarrow$ linear programming problem:

$$\begin{cases} \min \ \|Ax - b\|_1 = \sum_{i=1}^{m} |A_i x - b_i| \\ x \in \mathbb{R}^n \end{cases}$$

is equivalent to

$$\begin{cases} \min \ \sum_{i=1}^{m} y_i \\ y_i = |A_i x - b_i| \\ \quad = \max\{A_i x - b_i, b_i - A_i x\} \end{cases} \rightarrow \begin{cases} \min \ \sum_{i=1}^{m} y_i \\ y_i \geq A_i x - b_i \qquad \forall \ i = 1, \ldots, m \\ y_i \geq b_i - A_i x \qquad \forall \ i = 1, \ldots, m \end{cases}$$

Data fitting

norm $\|\cdot\|_\infty \rightarrow$ linear programming problem:

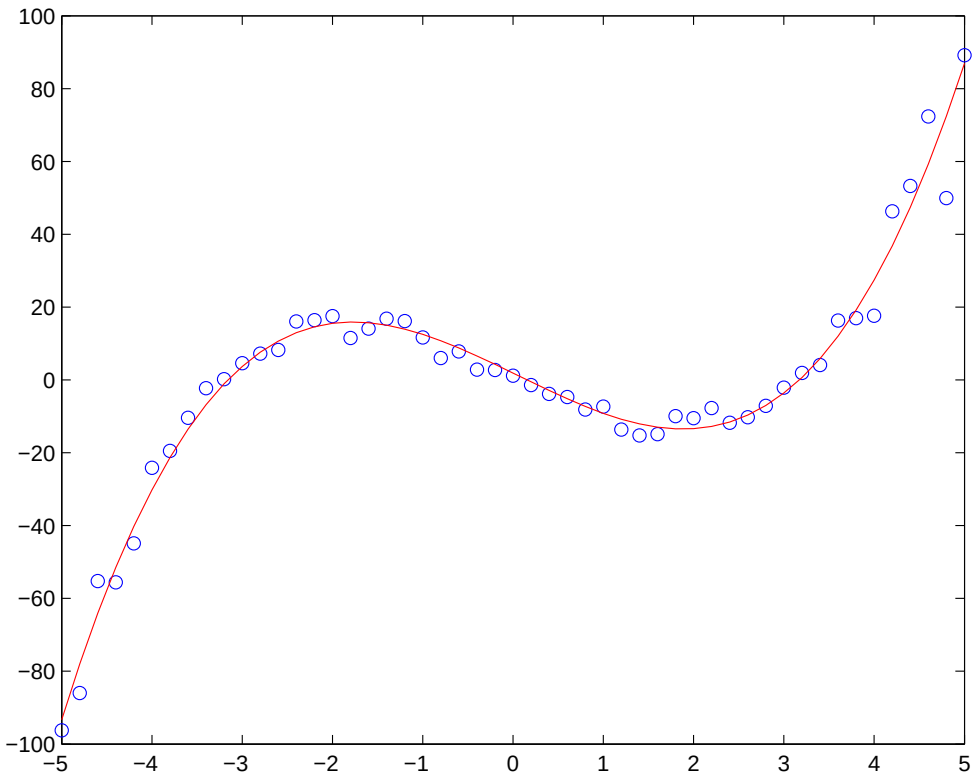
$$\begin{cases} \min \ \|Ax - b\|_\infty = \max\limits_{i=1,\dots,m} |A_i x - b_i| \\ x \in \mathbb{R}^n \end{cases}$$

is equivalent to

$$\begin{cases} \min \ y \\ y \geq A_i x - b_i \qquad \forall \ i = 1, \dots, m \\ y \geq b_i - A_i x \qquad \forall \ i = 1, \dots, m \end{cases}$$

Data fitting

Exercise. Given the experimental data in `fitting.txt`, find the best approximating polynomial of degree 3 with respect to the Euclidean norm.



Exercise. Find the best approximating polynomial of degree 3 w.r.t. $\|\cdot\|_1$.
Exercise. Find the best approximating polynomial of degree 3 w.r.t. $\|\cdot\|_\infty$.

Data fitting

Now, b depends on $n-1$ parameters, i.e., m experimental data $b_1, b_2, \ldots, b_m \in \mathbb{R}$, where b_i corresponds to an observation made on points $a_{i1}, \ldots, a_{i,n-1} \in \mathbb{R}$.
We look for the best linear approximation of experimental data.
Affine function has coefficients $x_1, \ldots, x_n$, residual is

$$r_i = \sum_{j=1}^{n-1} a_{ij}\, x_j + x_n - b_i.$$

We want to find x s.t. $\|r\|$ is minimum, i.e.

$$\begin{cases} \min \ \|A\,x - b\| \\ x \in \mathbb{R}^n \end{cases}$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1,n-1} & 1 \\ a_{21} & a_{22} & \ldots & a_{2,n-1} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{m,n-1} & 1 \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Machine learning

Extract knownledge 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
- ...

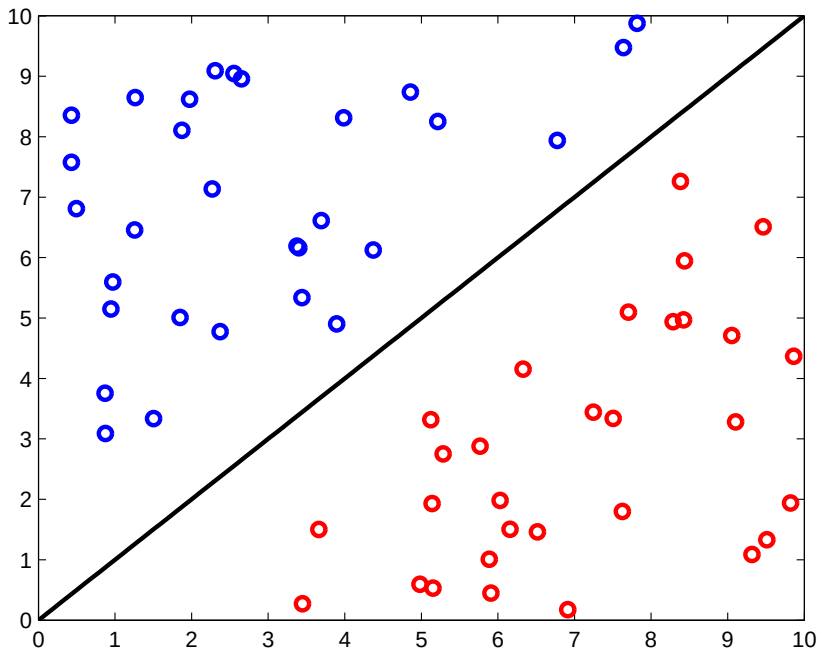
Methods:

- 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

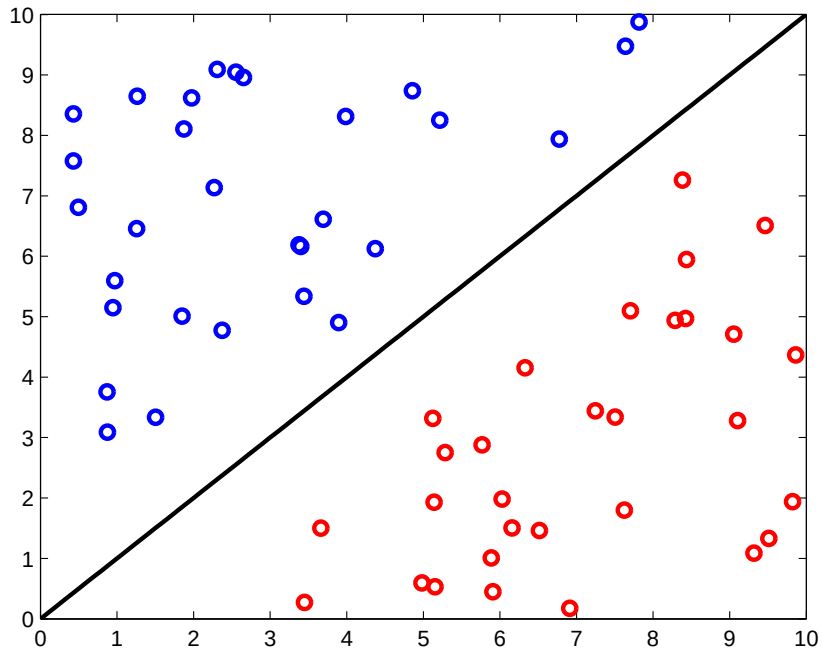
$$w^T x^i + b > 0 \qquad \forall\ x^i \in A$$
$$w^T x^j + b < 0 \qquad \forall\ x^j \in B$$

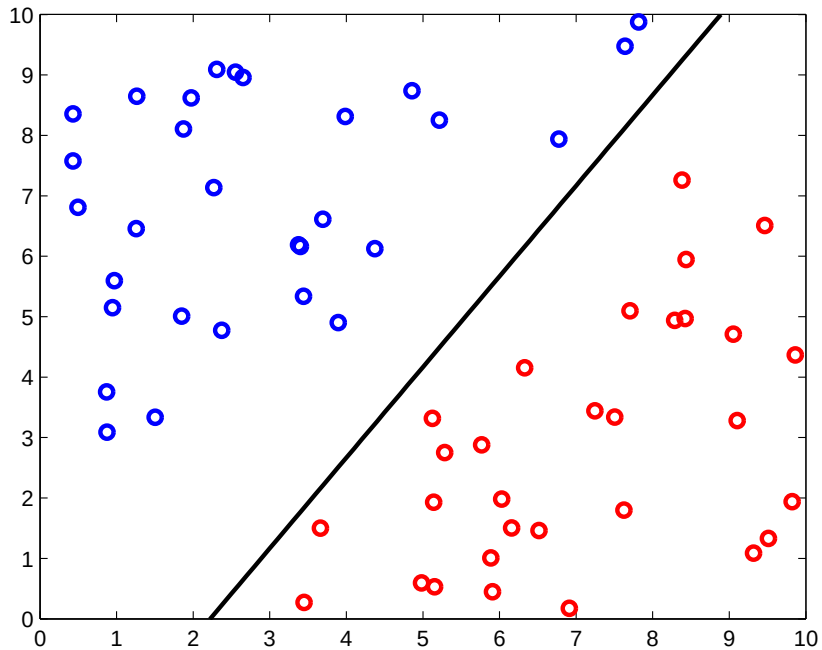


Test data x:
decision function $f(x) = \text{sign}(w^T x + b)$

Linear SVM

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



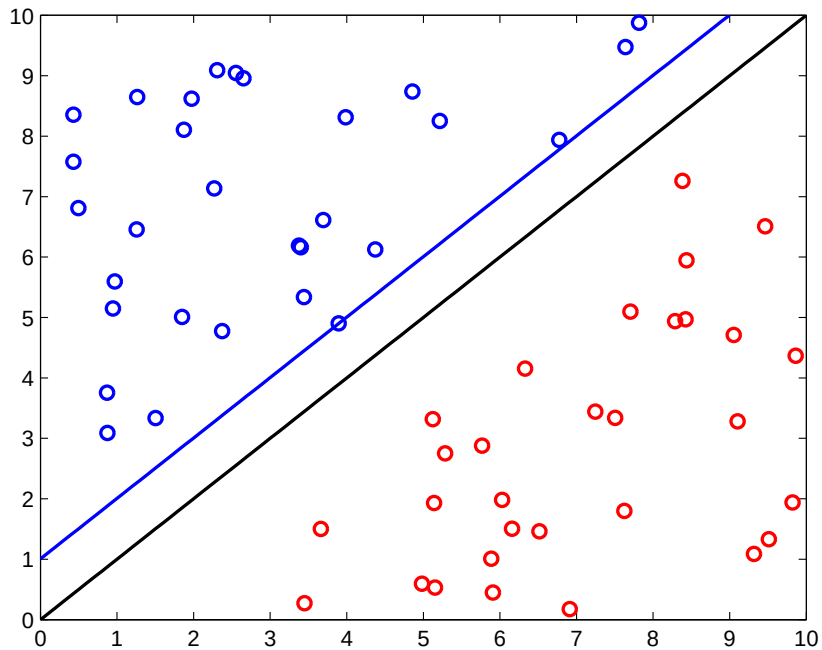


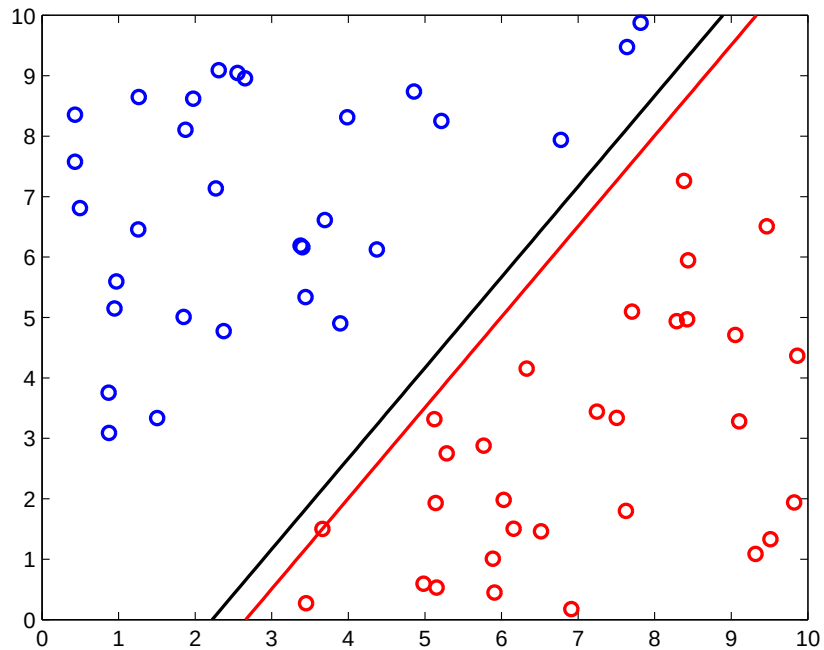
Linear SVM

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

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

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

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

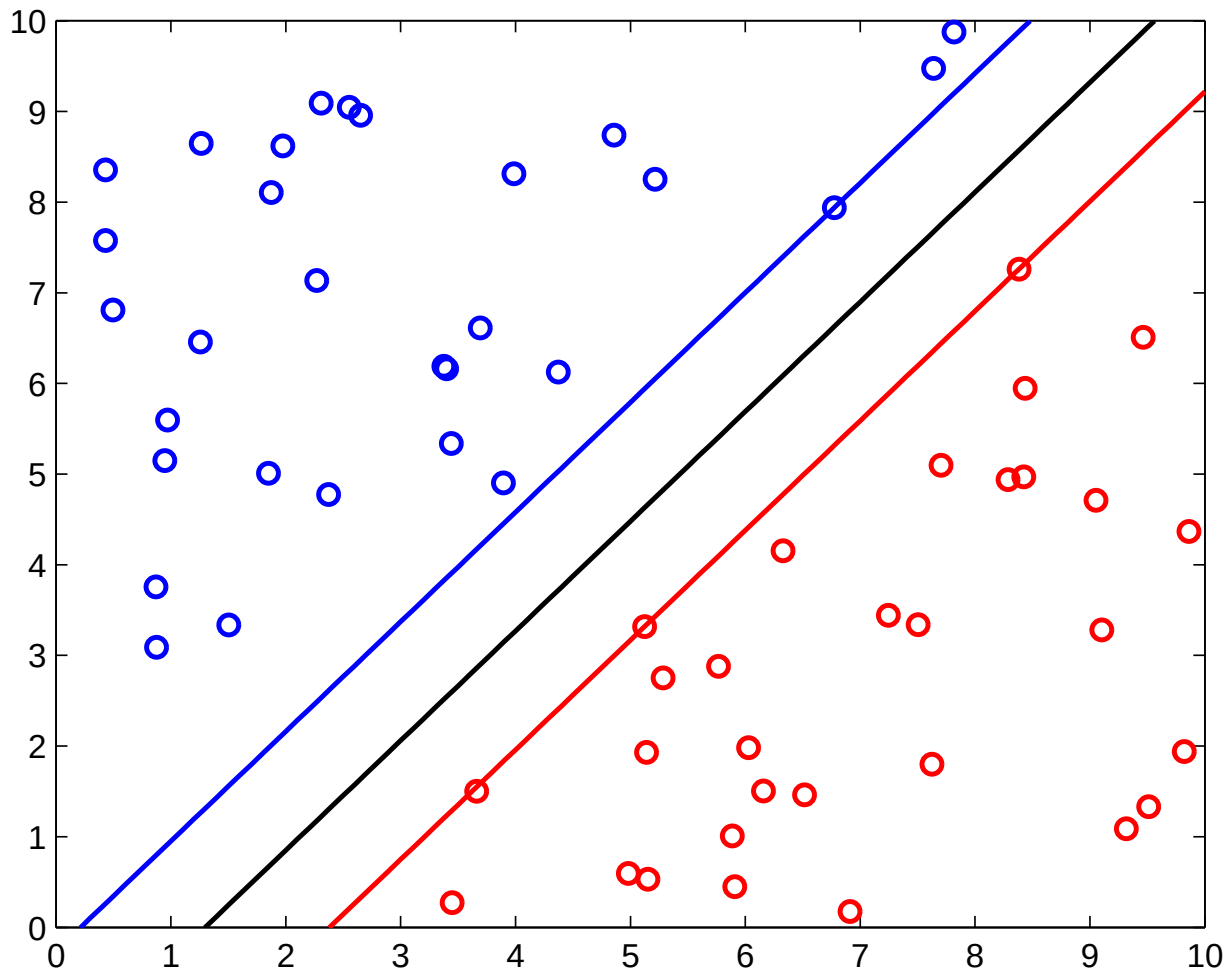
then $\tilde{H} = \{\tilde{w}^T x + \tilde{b} = 0\}$, where $\tilde{w} = 2\,w/(\alpha+\beta)$ and $\tilde{b} = (2\,b - \alpha + \beta)/(\alpha+\beta)$, is another separating hyperplane s.t.

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

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

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

is equivalent to

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

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

Linear SVM

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

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

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

Linear SVM

Dual problem is

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

or

$$\begin{cases} \max \ -\frac{1}{2}\lambda^\mathsf{T} X^\mathsf{T} X \lambda + e^\mathsf{T} \lambda \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{cases}$$

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

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:

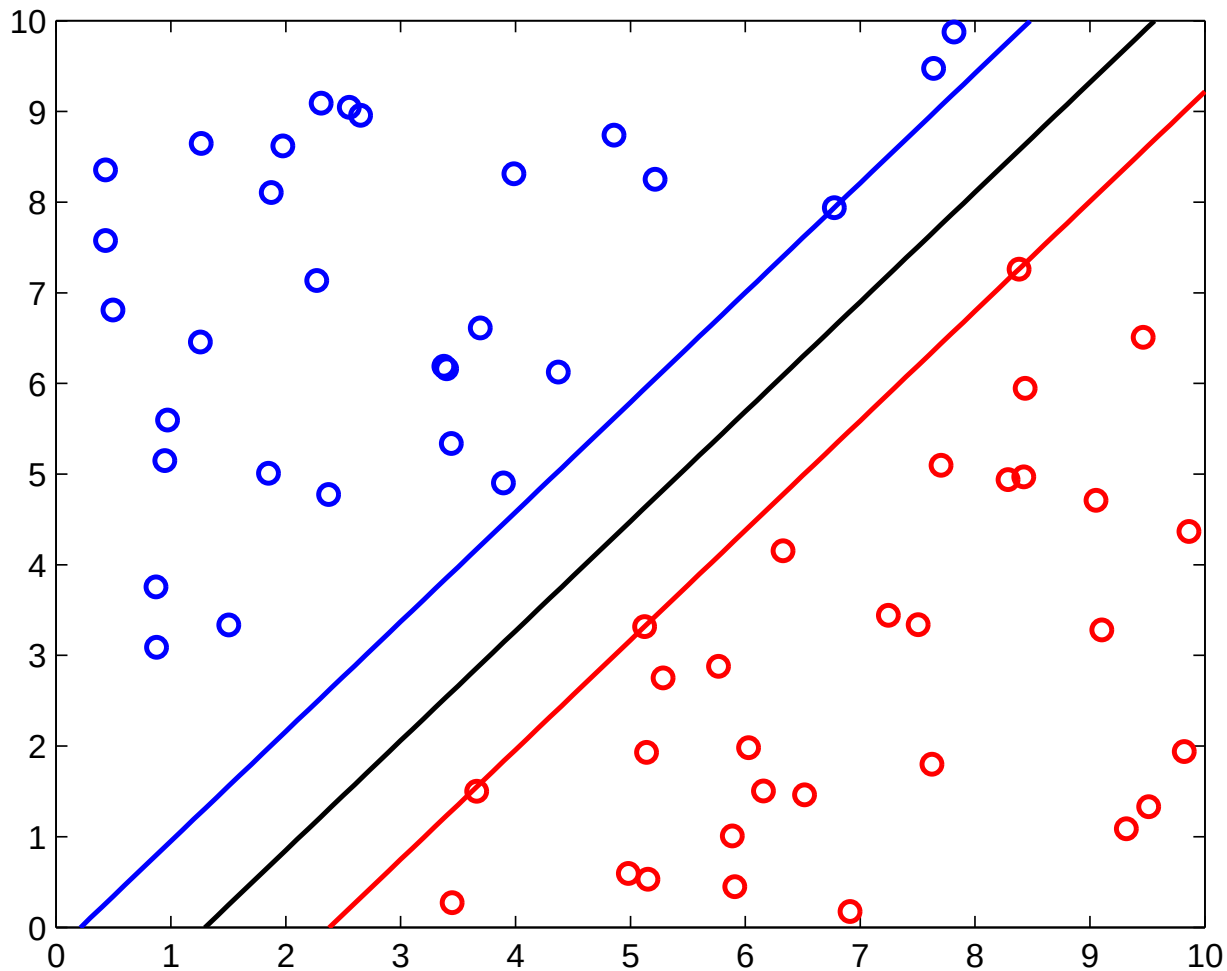
$$\lambda_i^* \left[1 - y^i((w^*)^\mathsf{T} x^i + b^*)\right] = 0.$$

- Decision function is

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

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

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

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

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

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 $C > 0$:

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

Linear SVM with soft margin

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

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

If λ^* is dual optimum, then

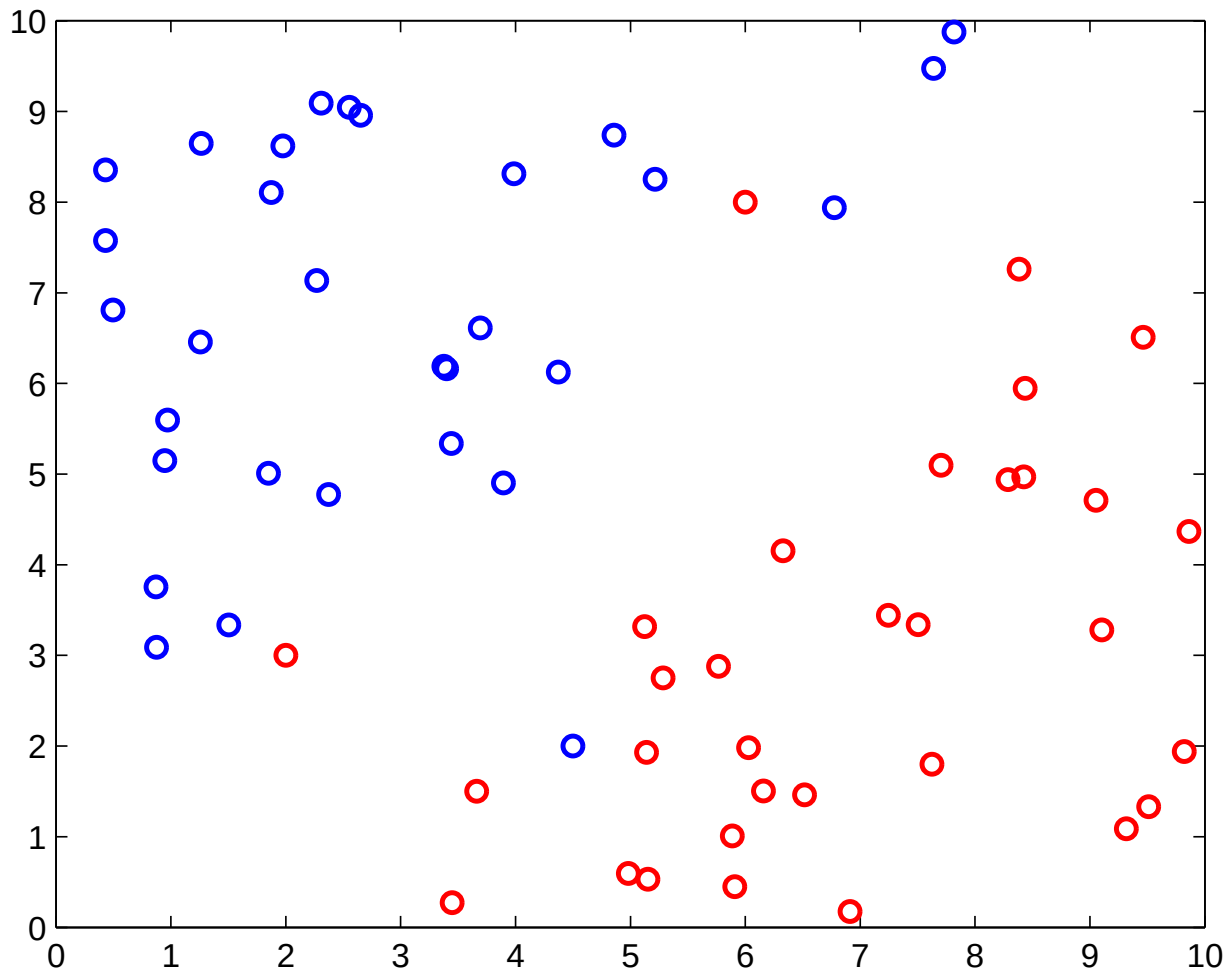
$$w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i x^i.$$

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

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

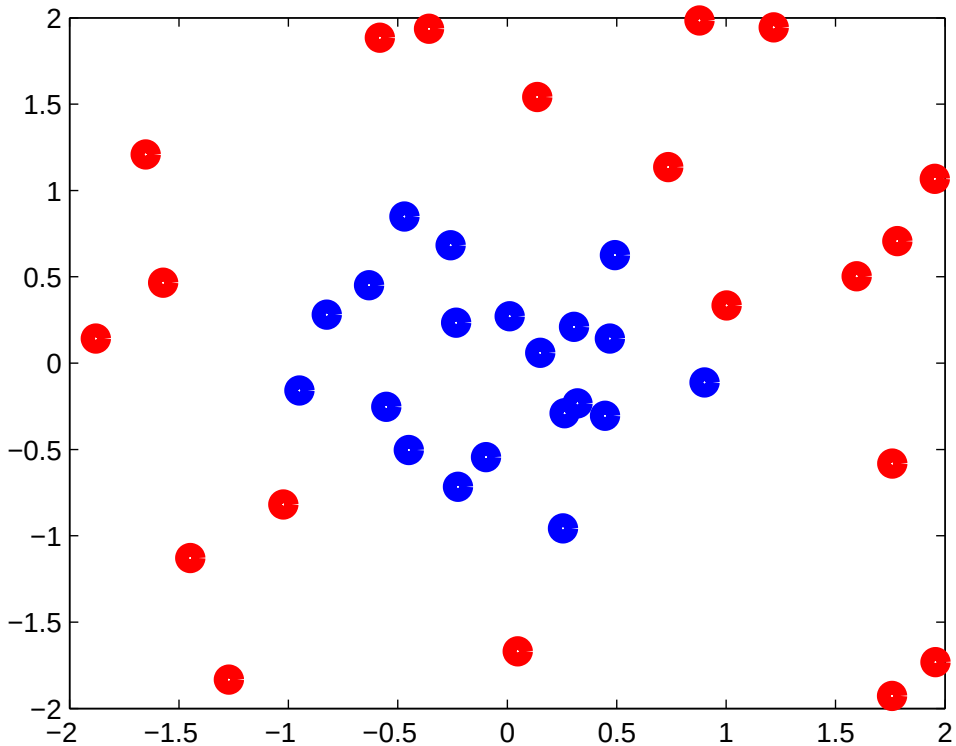
Linear SVM with soft margin

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

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

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

Nonlinear SVM

Primal problem:

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

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

Dual problem:

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

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

- $\lambda^* \rightarrow$ know $\phi(x^i)^\mathsf{T} \phi(x^j)$
- $\phi(x^i)^\mathsf{T} \phi(x)$
- $b^* \rightarrow$ 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 a kernel if there is a function $\phi : \mathbb{R}^n \to \mathcal{H}$ 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 $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a kernel and $x^1, \ldots, x^\ell \in \mathbb{R}^n$, then the matrix

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

is positive semidefinite.

Dual problem is convex:

$$\begin{cases} \max \; -\frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y^i y^j k(x^i, x^j) \lambda_i \lambda_j + \sum_{i=1}^{\ell} \lambda_i \\ \sum_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \leq \lambda_i \leq C \qquad i = 1, \ldots, \ell \end{cases}$$

Nonlinear SVM

In practice:

- choose a kernel k
- solve the dual $\rightarrow \lambda^*$
- find b^*

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

- 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).