5 - Support Vector Machines for supervised classification problems

G. Mastroeni and M. Passacantando

Department of Computer Science, University of Pisa

Optimization Methods and Game Theory
Master of Science in Artificial Intelligence and Data Engineering
University of Pisa – A.Y. 2023/24

Supervised pattern classification

Support Vector Machines (SVM) provide a supervised classification method for a vector of data, concerning a given problem, according to previously obtained vectors of data that have already been classified.

We are given a set of vectors of data (objects) partitioned in several classes with known labels, we want to assign to a suitable class a new object with unknown label.

Examples:

- medical diagnosis
- spam filtering
- credit card fraud detection

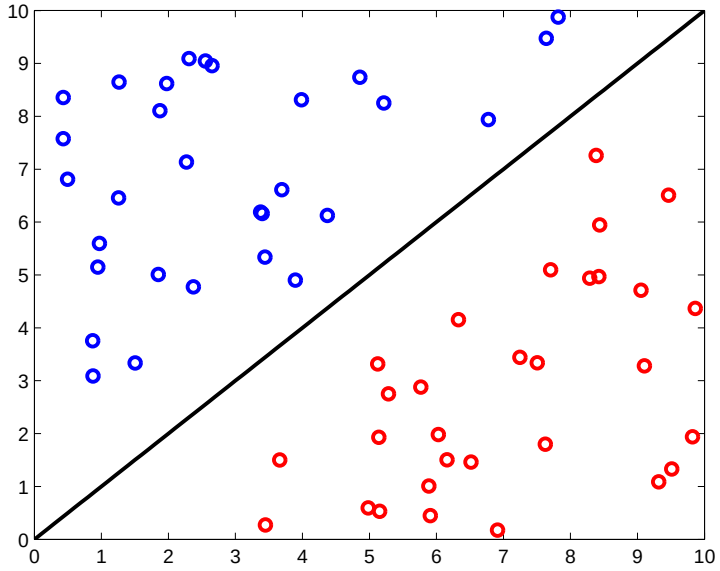
Binary classification: Linear SVM

In a binary classification, we are given 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 strictly linearly separable, i.e., there is an hyperplane $H = \{x \in \mathbb{R}^n : \ w^T x + b = 0\}$ such that

$$\begin{aligned} w^T x^i + b > 0 \qquad & \forall\ x^i \in A, \\ w^T x^j + b < 0 \qquad & \forall\ x^j \in B. \end{aligned} \tag{1}$$



We have a new test data x:

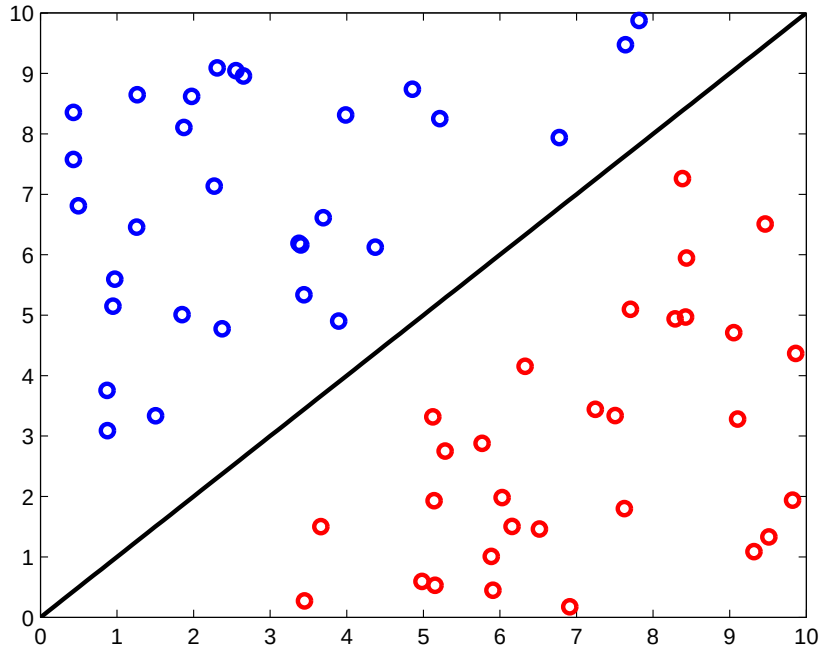
use the decision function
$f(x) = \text{sign}(w^T x + b) =$
$$\begin{cases} 1 & \text{if } w^T x + b > 0, \\ -1 & \text{if } w^T x + b < 0. \end{cases}$$

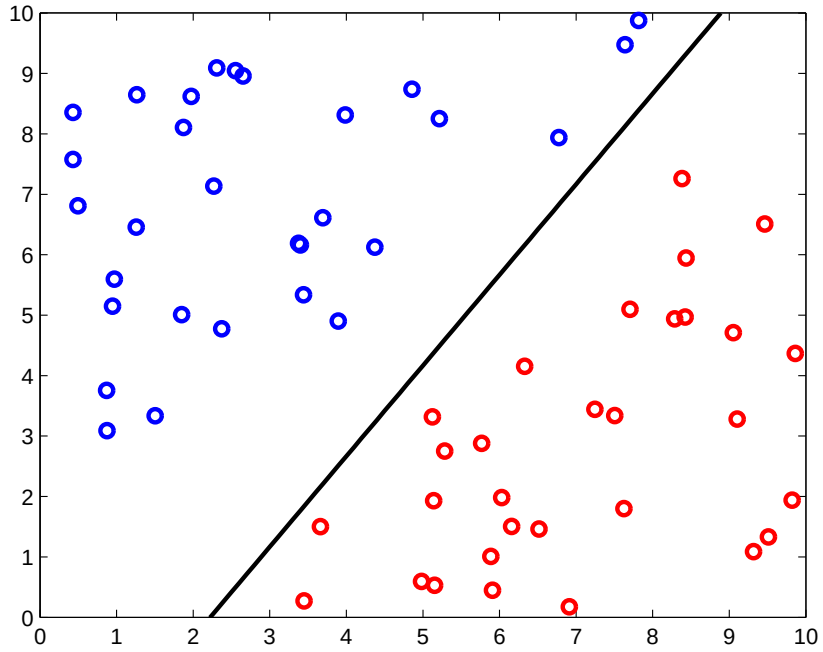
Linear SVM

- A necessary and sufficient condition for (1) to hold is

$$conv(A) \cap conv(B) = \emptyset$$

- Since there are many possible separating hyperplanes, we have to decide which hyperplane to choose.

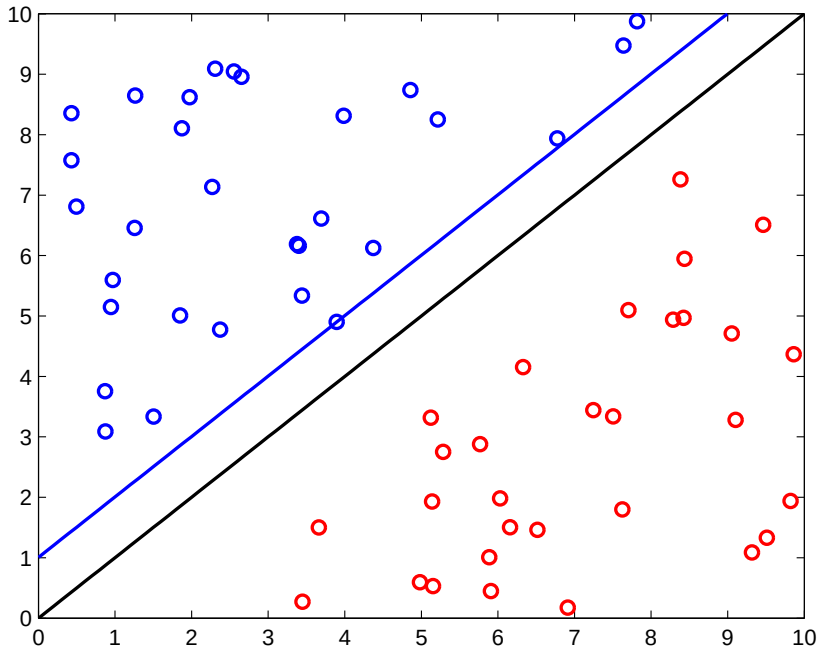


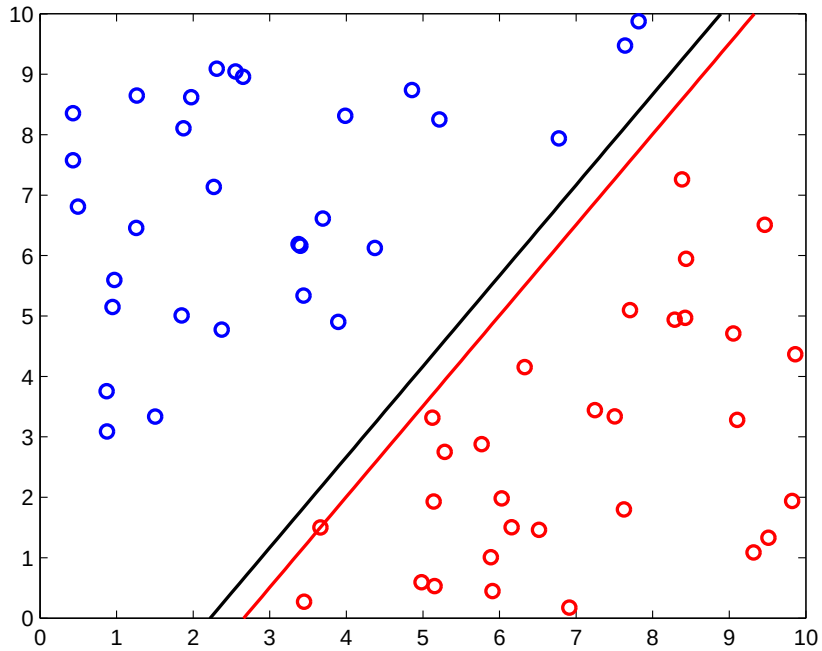


Definition

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.

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





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:

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

Proof. It is possible to show that the distance between the hyperplanes:

$$w^T x + b = 1, \qquad\qquad w^T x + b = -1$$

is $\frac{2}{\|w\|}$. In fact, consider a point $\hat{x}$ such that $w^T \hat{x} + b = 1$, then the distance between $\hat{x}$ and the other hyperplane $w^T x + b + 1 = 0$ is

$$\frac{|w^T \hat{x} + b + 1|}{\|w\|} = \frac{2}{\|w\|}.$$

Therefore, by minimizing $\|w\|$, we get two hyperplanes of maximum distance.

Moreover, we will see that problem (2) has a unique solution (w^*, b^*). □

Example 5.1 Find the separating hyperplane with maximum margin for the data set:

$$A = \begin{pmatrix} 0.4952 & 6.8088 \\ 2.6505 & 8.9590 \\ 3.4403 & 5.3366 \\ 3.4010 & 6.1624 \\ 5.2153 & 8.2529 \\ 7.6393 & 9.4764 \\ 1.5041 & 3.3370 \\ 3.9855 & 8.3138 \\ 1.8500 & 5.0079 \\ 1.2631 & 8.6463 \\ 3.8957 & 4.9014 \\ 1.9751 & 8.6199 \\ 1.2565 & 6.4558 \\ 4.3732 & 6.1261 \\ 0.4297 & 8.3551 \\ 3.6931 & 6.6134 \\ 7.8164 & 9.8767 \\ 4.8561 & 8.7376 \\ 6.7750 & 7.9386 \\ 2.3734 & 4.7740 \\ 0.8746 & 3.0892 \\ 2.3088 & 9.0919 \\ 2.5520 & 9.0469 \\ 3.3773 & 6.1886 \\ 0.8690 & 3.7550 \\ 1.8738 & 8.1053 \\ 0.9469 & 5.1476 \\ 0.9718 & 5.5951 \\ 0.4309 & 7.5763 \\ 2.2699 & 7.1371 \end{pmatrix} \qquad B = \begin{pmatrix} 7.2450 & 3.4422 \\ 7.7030 & 5.0965 \\ 5.7670 & 2.8791 \\ 3.6610 & 1.5002 \\ 9.4633 & 6.5084 \\ 9.8221 & 1.9383 \\ 8.2874 & 4.9380 \\ 5.9078 & 0.4489 \\ 4.9810 & 0.5962 \\ 5.1516 & 0.5319 \\ 8.4363 & 5.9467 \\ 8.4240 & 4.9696 \\ 7.6240 & 1.7988 \\ 3.4473 & 0.2725 \\ 9.0528 & 4.7106 \\ 9.1046 & 3.2798 \\ 6.9110 & 0.1745 \\ 5.1235 & 3.3181 \\ 7.5051 & 3.3392 \\ 6.3283 & 4.1555 \\ 6.1585 & 1.5058 \\ 8.3827 & 7.2617 \\ 5.2841 & 2.7510 \\ 5.1412 & 1.9314 \\ 6.0290 & 1.9818 \\ 5.8863 & 1.0087 \\ 9.5110 & 1.3298 \\ 9.3170 & 1.0890 \\ 6.5170 & 1.4606 \\ 9.8621 & 4.3674 \end{pmatrix}$$

Assume $A \in \mathbb{R}^{m\times n},\ B \in \mathbb{R}^{p\times n},\ w = (w_1, ..., w_n)^T$;

Problem (2) is a quadratic problem defined by:

$$\begin{cases} \min\limits_{w,b} \ \frac{1}{2}(w,b)^T C \begin{pmatrix} w \\ b \end{pmatrix} \\ D \begin{pmatrix} w \\ b \end{pmatrix} \leq d \end{cases} \tag{3}$$

where, assuming $n = 2,\ w \in \mathbb{R}^2,\ b \in \mathbb{R}$,

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad D = \begin{pmatrix} -A & -e_m \\ B & e_p \end{pmatrix} \quad d = \begin{pmatrix} -e_m \\ -e_p \end{pmatrix}$$

$$-e_m = (-1, -1, ..., -1)^T \in \mathbb{R}^m, \quad -e_p = (-1, -1, ..., -1)^T \in \mathbb{R}^p$$

The Matlab function "quadprog"

The previous problem can be solved by the Matlab function "quadprog" which solves a quadratic problem with linear constraints.

From the Matlab help

[x,fval,exitflag,output,lambda] = quadprog(H,f,A,b) attempts to solve the quadratic programming problem:
min 0.5*x'*H*x + f'*x subject to: A*x <= b x

[x,fval,exitflag,output,lambda] = quadprog(H,f,A,b,Aeq,beq) solves the problem above while additionally satisfying the equality constraints Aeq*x = beq. (Set A=[] and B=[] if no inequalities exist.)

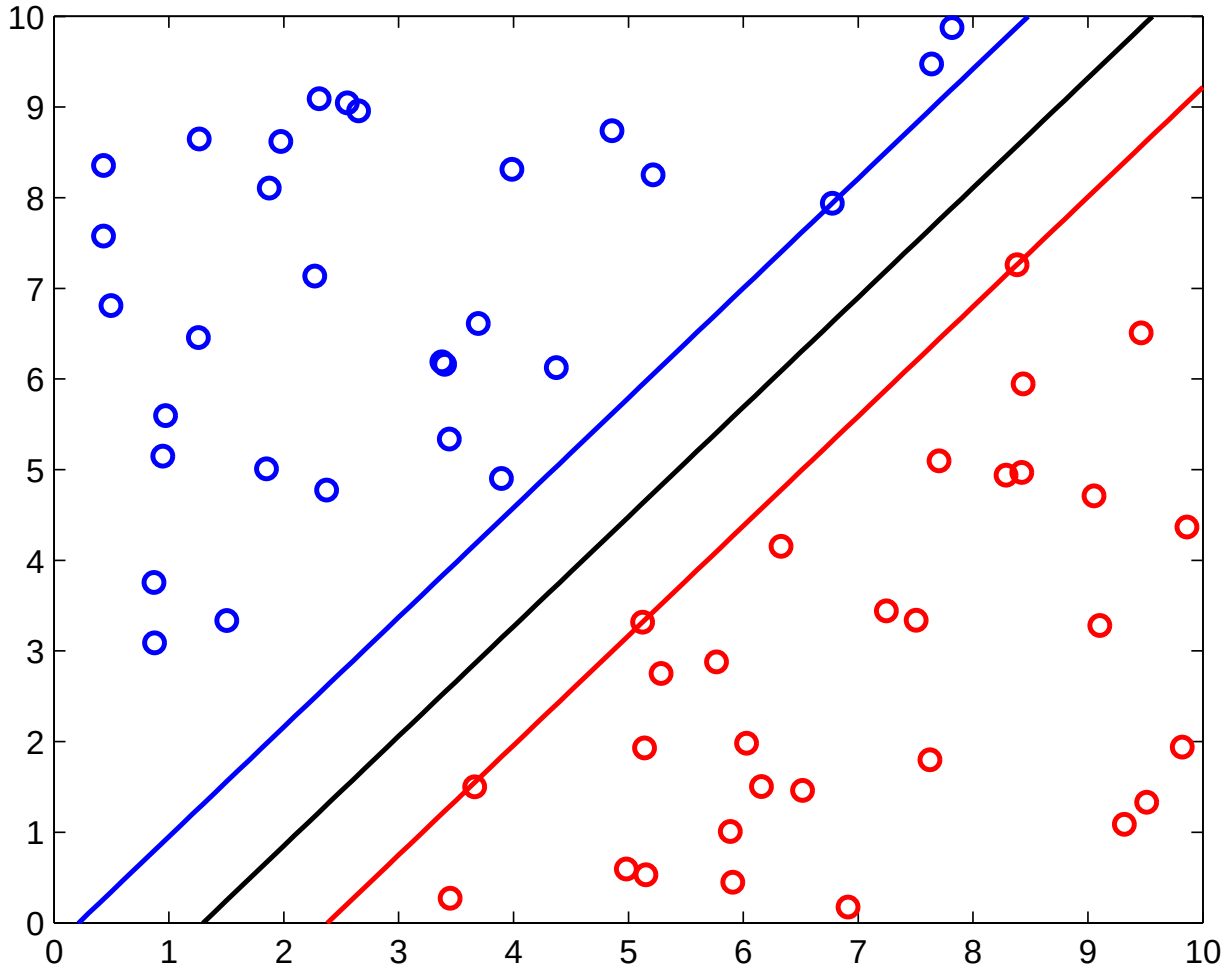
[x,fval,exitflag,output,lambda] = quadprog(H,f,A,b,Aeq,beq,LB,UB) defines a set of lower and upper bounds on the design variables, X, so that the solution is in the range LB <= X <= UB. Use empty matrices for LB and UB if no bounds exist. Set LB(i) = -Inf if X(i) is unbounded below; set UB(i) = Inf if X(i) is unbounded above.

Matlab commands: primal problem

```
A=[.....];
B=[.....];
nA = size(A,1);
nB = size(B,1);
T = [A ; B];
Q = [ 1 0 0 ; 0 1 0 ; 0 0 0 ];
D = [-A -ones(nA,1); B ones(nB,1) ] ;
d = -ones(nA+nB,1) ;
sol = quadprog(Q,zeros(3,1),D,d);
w = sol(1:2)
b = sol(3)

% Optional: plot the solution

xx = 0:0.1:10 ;
uu = (-w(1)/w(2)).*xx - b/w(2);
vv = (-w(1)/w(2)).*xx + (1-b)/w(2);
vvv = (-w(1)/w(2)).*xx + (-1-b)/w(2);

plot(A(:,1),A(:,2),'bo',B(:,1),B(:,2),'ro',xx,uu,'k-',xx,vv,'b-',xx,vvv,'r-','Linewidth', 1.5)
axis([0 10 0 10])
```

w =
-0.9229
0.7627

b =
1.1976



Equivalent formulation of problem (2)

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

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

Then the problem

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

is equivalent to

$$\text{linear SVM} \quad \begin{cases} \min\limits_{w,b} \ \dfrac{1}{2}\|w\|^2 \\ 1 - y^i(w^{\mathsf{T}} x^i + b) \leq 0 \qquad \forall\ i = 1, \ldots, \ell \end{cases} \tag{4}$$

Since problem (4) is convex, it is useful to consider the Lagrangian dual of (4).

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

$$\nabla_w L(w,b,\lambda) = w - \sum_{i=1}^{\ell} \lambda_i y^i x^i = 0. \tag{5}$$

Note that, if (9) holds and $\sum_{i=1}^{\ell} \lambda_i y^i = 0$, then

$$L(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y^i w^{\mathsf{T}} x^i - b \sum_{i=1}^{\ell} \lambda_i y^i + \sum_{i=1}^{\ell} \lambda_i = \frac{1}{2} w^T w - w^T w + \sum_{i=1}^{\ell} \lambda_i$$

$$= -\frac{1}{2} w^T w + \sum_{i=1}^{\ell} \lambda_i$$

Therefore, the dual function is

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

The dual of problem (4) is

$$\begin{cases} \max\limits_{\lambda} \; -\dfrac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j (x^i)^{\mathsf{T}} x^j \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{cases} \tag{6}$$

or

$$\begin{cases} \max\limits_{\lambda} \; -\dfrac{1}{2}\lambda^{\mathsf{T}} X^{\mathsf{T}} X \lambda + e^{\mathsf{T}} \lambda \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda \geq 0 \end{cases} \tag{7}$$

where $X = (y^1 x^1, y^2 x^2, \ldots, y^\ell x^\ell)$ is a $n \times \ell$ matrix and $e^{\mathsf{T}} = (1, \ldots, 1) \in \mathbb{R}^\ell$.

Remarks

- Since $X^T X$ is always positive semidefinite then the dual problem is a convex quadratic programming problem;
- A 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, by (9), we have:

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

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

This allows us to find the separating hyperplane $(w^*)^{\mathsf{T}} x + b^* = 0$ and the decision function

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

Exercise 5.1

Determine the KKT conditions of the primal problem (4):

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

As previously seen, the Lagrangian function is:

$$L(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y^i w^\mathsf{T} x^i - b \sum_{i=1}^{\ell} \lambda_i y^i + \sum_{i=1}^{\ell} \lambda_i$$

Then, the KKT conditions are:

$$\begin{cases} \nabla_w L(w, b, \lambda) = w - \sum\limits_{i=1}^{\ell} \lambda_i y^i x^i = 0 \\ \nabla_b L(w, b, \lambda) = -\sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ \lambda_i (1 - y^i(w^\mathsf{T} x^i + b)) = 0 \qquad \forall\ i = 1, \ldots, \ell \\ \lambda_i \geq 0, \quad 1 - y^i(w^\mathsf{T} x^i + b) \leq 0 \qquad \forall\ i = 1, \ldots, \ell \end{cases}$$

Example 5.2. Find the separating hyperplane with maximum margin for the data set given in Example 5.1 by solving the dual problem (6).

We have to solve the problem:

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

Note that the generic component q_{ij} of the hessian matrix Q is given by

$$q_{ij} = y^i y^j (x^i)^{\mathsf{T}} x^j$$

Matlab commands: dual problem

```
A=[.....]; B=[.....]; nA = size(A,1); nB = size(B,1);
T = [A ; B]; y = [ones(nA,1) ; -ones(nB,1)]; l = length(y); Q = zeros(l,l);
for i = 1 : l
for j = 1 : l
Q(i,j) = y(i)*y(j)*(T(i,:))*T(j,:)' ;
end
end
la = quadprog(Q,-ones(l,1),[ ],[ ],y',0,zeros(l,1),[ ]);
wD = zeros(2,1);
for i = 1 : l
wD = wD + la(i)*y(i)*T(i,:)';
end
wD
ind = find(la > 0.001) ;
i = ind(1) ;
bD = 1/y(i) - wD'*T(i,:)'
```

Exercise

Plot the solution obtained in Example 5.2.

```
xx = 0:0.1:10 ;
uuD = (-wD(1)/wD(2)).*xx - bD/wD(2);
vvD = (-wD(1)/wD(2)).*xx + (1-bD)/wD(2);
vvvD = (-wD(1)/wD(2)).*xx + (-1-bD)/wD(2);
figure
plot(A(:,1),A(:,2),'bo',B(:,1),B(:,2),'ro',
xx,uuD,'k-',xx,vvD,'b-',xx,vvvD,'r-','Linewidth',1.5)
axis([0 10 0 10])
```

Linear SVM with soft margin

What if sets A and B are not linearly separable?

The linear system

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

has no solutions.

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

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

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. In fact,

- $x^i \in A$, with $w^\mathsf{T} x^i + b < 0$ implies

$$1 < 1 - (w^\mathsf{T} x^i + b) = 1 - y^i(w^\mathsf{T} x^i + b) \leq \xi_i$$

- $x^i \in B$, with $w^\mathsf{T} x^i + b > 0$ implies

$$1 < 1 + (w^\mathsf{T} x^i + b) = 1 - y^i(w^\mathsf{T} x^i + b) \leq \xi_i$$

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

$$\text{linear SVM with soft margin} \quad \begin{cases} \min\limits_{w,b,\xi} \ \dfrac{1}{2}\|w\|^2 + C \sum\limits_{i=1}^{\ell} \xi_i & \\ 1 - y^i(w^{\mathsf{T}} x^i + b) \le \xi_i & \forall\ i = 1, \dots, \ell \\ \xi_i \ge 0 & \forall\ i = 1, \dots, \ell \end{cases} \tag{8}$$

Exercise 5.2. Prove that the dual problem of (8) is

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

The Lagrangian function is

$$L(w,b,\xi,\lambda,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i + \sum_{i=1}^{\ell}\lambda_i\left[1 - y^i(w^{\mathsf{T}}x^i + b) - \xi_i\right] - \sum_{i=1}^{\ell}\mu_i\xi_i$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell}\lambda_i y^i w^{\mathsf{T}}x^i - b\sum_{i=1}^{\ell}\lambda_i y^i + \sum_{i=1}^{\ell}\lambda_i + \sum_{i=1}^{\ell}\xi_i(C - \lambda_i - \mu_i)$$

For a fixed (λ,μ), we aim at finding $\min_{w,b,\xi} L(w,b,\xi,\lambda,\mu)$.

Since L is convex w.r.t. (w,b,ξ) then a global minimum is a stationary point of $L(\cdot,\cdot,\cdot,\lambda,\mu)$, i.e. it is a solution of the system:

$$\begin{cases} \nabla_w L(w,b,\xi,\lambda,\mu) = w - \sum_{i=1}^{\ell}\lambda_i y^i x^i = 0 \\ \nabla_b L(w,b,\xi,\lambda,\mu) = -\sum_{i=1}^{\ell}\lambda_i y^i = 0 \\ \nabla_\xi L(w,b,\xi,\lambda,\mu) = C - \lambda_i - \mu_i = 0,\ \ i = 1,..,\ell \end{cases} \tag{9}$$

Eliminating the variable w, with the same arguments used for finding the dual of (4), we have that the dual of problem (8) is

$$\begin{cases} \max\limits_{\lambda} \; -\frac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j (x^i)^\mathsf{T} x^j \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ C - \lambda_i - \mu_i = 0 \qquad i = 1, \ldots, \ell \\ \lambda \geq 0, \mu \geq 0 \end{cases} \tag{10}$$

and eliminating the variable μ, by noticing that $C - \lambda_i = \mu_i \geq 0$, we obtain the final dual formulation

$$\begin{cases} \max\limits_{\lambda} \; -\frac{1}{2}\sum\limits_{i=1}^{\ell}\sum\limits_{j=1}^{\ell} y^i y^j (x^i)^\mathsf{T} x^j \lambda_i \lambda_j + \sum\limits_{i=1}^{\ell} \lambda_i \\ \sum\limits_{i=1}^{\ell} \lambda_i y^i = 0 \\ 0 \leq \lambda_i \leq C \qquad i = 1, \ldots, \ell \end{cases} \tag{11}$$

If λ^* is optimum for (11), then

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

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

$$\begin{cases} \lambda_i^* \left[1 - y^i((w^*)^\mathsf{T} x^i + b^*) - \xi_i^*\right] = 0 \\ \mu_i^* \xi_i^* = (C - \lambda_i^*)\xi_i^* = 0 \end{cases} \tag{12}$$

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

Remark

Note that

- $0 \leq \lambda_i^* < C$ implies $\xi_i^* = 0$
- $0 < \lambda_i^* \leq C$ implies $1 - y^i((w^*)^\mathsf{T} x^i + b^*) - \xi_i^* = 0$
- $\xi_i^* > 0$ implies $\lambda_i^* = C$

The previous conditions also allow us to find the errors ξ_i^*, $i = 1, .., \ell$.

Exercise 5.3

Determine the KKT conditions of the problem :

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

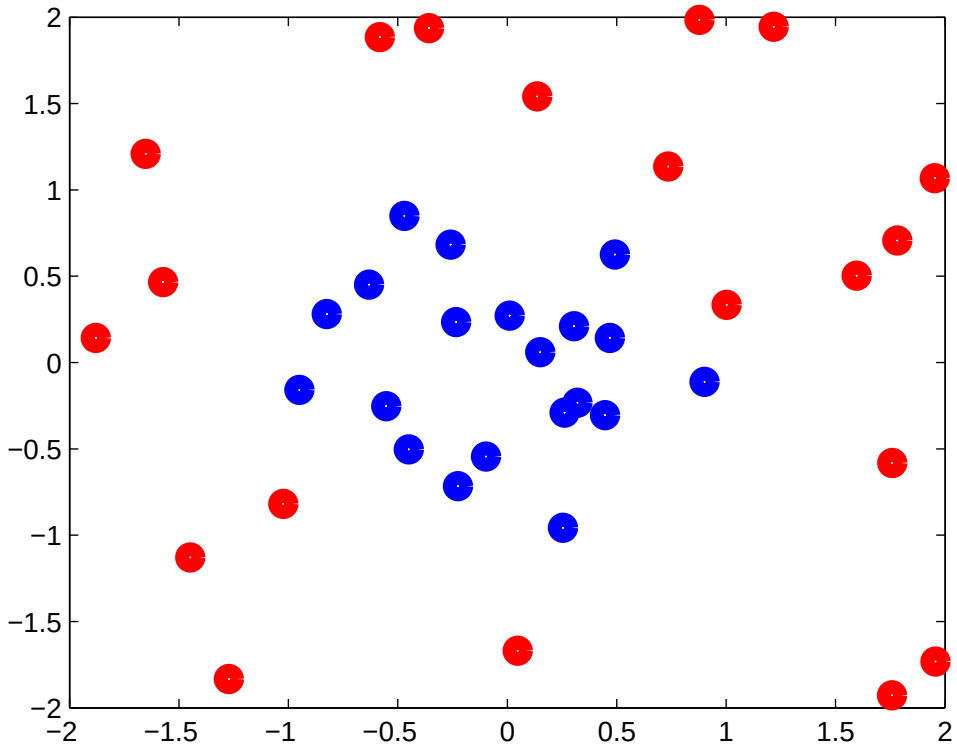
The KKT conditions are given by joining together (9), (12) plus the feasibility conditions of the given problem and the nonnegativity of the involved multipliers.

Exercise 5.4. Find the separating hyperplane with soft margin for the following data set by solving the dual problem (11) with $C = 10$.
Compute the vector ξ of the errors.

$$A = \begin{pmatrix} 2.6505 & 8.9590 \\ 3.4403 & 5.3366 \\ 3.4010 & 6.1624 \\ 5.2153 & 8.2529 \\ 7.6393 & 9.4764 \\ 1.5041 & 3.3370 \\ 3.9855 & 8.3138 \\ 1.8500 & 5.0079 \\ 1.2631 & 8.6463 \\ 3.8957 & 4.9014 \\ 1.9751 & 8.6199 \\ 1.2565 & 6.4558 \\ 4.3732 & 6.1261 \\ 0.4297 & 8.3551 \\ 3.6931 & 6.6134 \\ 7.8164 & 9.8767 \\ 4.8561 & 8.7376 \\ 6.7750 & 7.9386 \\ 2.3734 & 4.7740 \\ 0.8746 & 3.0892 \\ 2.3088 & 9.0919 \\ 2.5520 & 9.0469 \\ 3.3773 & 6.1886 \\ 0.8690 & 3.7550 \\ 1.8738 & 8.1053 \\ 0.9469 & 5.1476 \\ 0.4309 & 7.5763 \\ 2.2699 & 7.1371 \end{pmatrix} \qquad B = \begin{pmatrix} 7.7030 & 5.0965 \\ 5.7670 & 2.8791 \\ 3.6610 & 1.5002 \\ 9.4633 & 6.5084 \\ 9.8221 & 1.9383 \\ 8.2874 & 4.9380 \\ 5.9078 & 0.4489 \\ 4.9810 & 0.5962 \\ 5.1516 & 0.5319 \\ 8.4363 & 5.9467 \\ 8.4240 & 4.9696 \\ 7.6240 & 1.7988 \\ 3.4473 & 0.2725 \\ 9.0528 & 4.7106 \\ 9.1046 & 3.2798 \\ 6.9110 & 0.1745 \\ 5.1235 & 3.3181 \\ 7.5051 & 3.3392 \\ 6.3283 & 4.1555 \\ 6.1585 & 1.5058 \\ 8.3827 & 7.2617 \\ 5.2841 & 2.7510 \\ 5.1412 & 1.9314 \\ 5.8863 & 1.0087 \\ 9.5110 & 1.3298 \\ 6.5170 & 1.4606 \\ 9.8621 & 4.3674 \\ 6.0000 & 8.0000 \end{pmatrix}$$

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.

Primal problem:

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

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

Dual problem:

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

number of variables = number of training data

- Let λ^* be a solution of the dual problem.
- Then $w^* = \sum_{i=1}^{\ell} \lambda_i^* y^i \phi(x^i)$.
- By any λ_i^* s.t. $0 < \lambda_i^* < C$ we can find b^*, by the complementarity relations that now become:

$$\begin{cases} \lambda_i^* \left[1 - y^i((w^*)^\mathsf{T}\phi(x^i) + b^*) - \xi_i^*\right] = 0 \\ \mu_i^* \xi_i^* = (C - \lambda_i^*)\xi_i^* = 0 \end{cases} \tag{13}$$

Then,

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

The decision function is given by:

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

Note that $f(x)$ depends on

- λ^* (that can be found knowing $\phi(x^i)^\mathsf{T}\phi(x^j)$)
- $\phi(x^i)^\mathsf{T}\phi(x)$
- b^* (that can be found knowing $\phi(x^i)^\mathsf{T}\phi(x^j)$)

Definition (Kernel functions)

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 the features space $\mathcal{H}$.

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

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 defined as follows

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

is positive semidefinite.

The dual problem depends on the kernel k:

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

Nonlinear SVM

The method:

- 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

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

The separating surface $f(x) = 0$ is

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

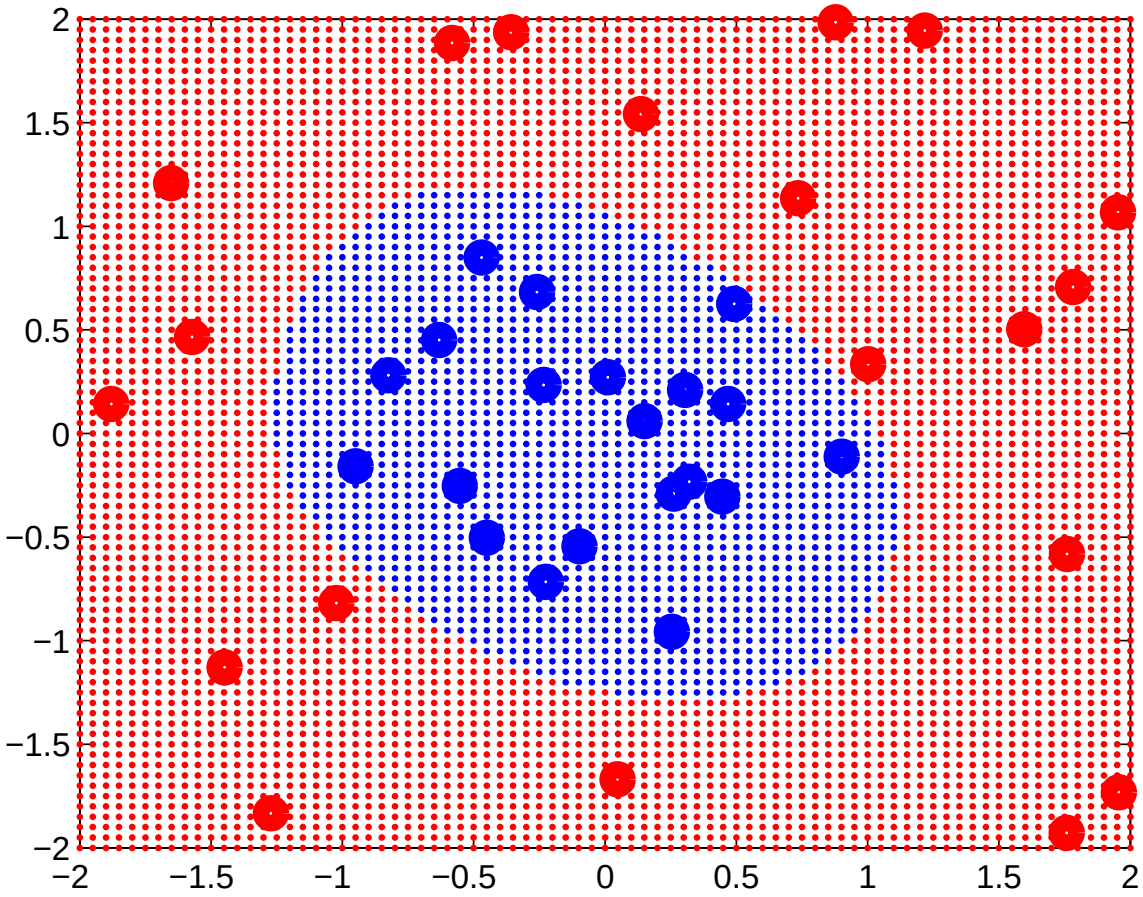
Exercise 5.5. Find the optimal separating surface for the following data set using a Gaussian kernel with parameters $C = 1$ and $\gamma = 1$.

$$A = \begin{pmatrix} 0.0113 & 0.2713 \\ 0.9018 & -0.1121 \\ 0.2624 & -0.2899 \\ 0.3049 & 0.2100 \\ -0.2255 & -0.7156 \\ -0.9497 & -0.1578 \\ -0.6318 & 0.4516 \\ -0.2593 & 0.6831 \\ 0.4685 & 0.1421 \\ -0.4694 & 0.8492 \\ -0.5525 & -0.2529 \\ -0.8250 & 0.2802 \\ 0.4463 & -0.3051 \\ 0.3212 & -0.2323 \\ 0.2547 & -0.9567 \\ 0.4917 & 0.6262 \\ -0.2334 & 0.2346 \\ 0.1510 & 0.0601 \\ -0.4499 & -0.5027 \\ -0.0967 & -0.5446 \end{pmatrix} \qquad B = \begin{pmatrix} 1.2178 & 1.9444 \\ -1.8800 & 0.1427 \\ -1.6517 & 1.2084 \\ 1.9566 & -1.7322 \\ 1.7576 & -1.9273 \\ 0.7354 & 1.1349 \\ 0.1366 & 1.5414 \\ 1.5960 & 0.5038 \\ -1.4485 & -1.1288 \\ -1.2714 & -1.8327 \\ -1.5722 & 0.4658 \\ 1.7586 & -0.5822 \\ -0.3575 & 1.9374 \\ 1.7823 & 0.7066 \\ 1.9532 & 1.0673 \\ -1.0233 & -0.8180 \\ 1.0021 & 0.3341 \\ 0.0473 & -1.6696 \\ 0.8783 & 1.9846 \\ -0.5819 & 1.8850 \end{pmatrix}$$

Matlab commands

```
A=[.....]; B=[.....]; nA = size(A,1); nB = size(B,1);
T = [A ; B]; y = [ones(nA,1) ; -ones(nB,1)]; l = length(y); C=1;

gamma = 1 ;                                   % Gaussian kernel
K = zeros(l,l);
for i = 1 : l
for j = 1 : l
K(i,j) = exp(-gamma*norm(T(i,:)-T(j,:))^ 2);
end
end

Q = zeros(l,l);                               % define the problem
for i = 1 : l
for j = 1 : l
Q(i,j) = y(i)*y(j)*K(i,j) ;
end
end

la = quadprog(Q,-ones(l,1),[ ],[ ],y',0,zeros(l,1),C*ones(l,1));    % solve the problem

ind = find((la > 0.001) & (la < C-0.001));                          % compute b
i = ind(1);
```

```
b = 1/y(i);
for j = 1 : l
b = b - la(j)*y(j)*K(i,j);
end

                                % plot the surface f(x)=0

for xx = -2 : 0.01 : 2
for yy = -2 : 0.01 : 2
s = 0;
        for i = 1 : l
                        s = s + la(i)*y(i)*exp(-gamma*norm(T(i,:)-[xx yy])^ 2);
        end
s = s + b;
        if (abs(s)< 0.01)
                plot(xx,yy,'g.');
                hold on
        end
end
end

plot(A(:,1),A(:,2),'bo',B(:,1),B(:,2),'ro','Linewidth',5)
```


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2