STAT 538
Lecture 9
Support Vector Machines

mmp@stat.washington.edu

These notes supplement the reading from: C. Burges - "A tutorial on SVM for pattern recognition".

1 Linear SVM's

1.1 Notation reminder and a VC bound

The data set: inputs $x^i \in \mathbb{R}^n,\ i = 1, \dots N$, labels $y^i \in \{-1, +1\}$

Assumption: $(x, y) \sim P$, i.i.d

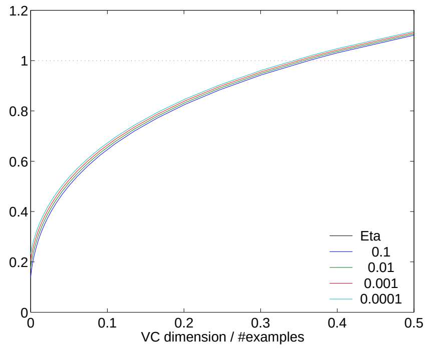
Classifier: $y = f(x, \theta)$ for new points x; $\theta =$ the parameters

The classifier family: $\mathcal{F} = \{f(., \theta)\}$

Empirical loss $\hat{L}_{01}(\theta) = \frac{1}{2N}\sum_i |y^i - \text{sgn} f(x^i, \theta)|$

Average loss $L_{01}(\theta) = \frac{1}{2}\int |y - \text{sgn} f(x, \theta)| dP(x, y)$

VC bound: $L_{01}(\theta) \leq \hat{L}_{01}(\theta) + \sqrt{\frac{h[1+\log(2N/h)]+\log(4/\delta)}{N}}$ w.p. $> 1 - \delta$, where $h = \text{VCdim}\,\mathcal{F}$ and $\delta < 1$ the confidence



1.2 Linear Maximum Margin Classifiers

The linear classifier: $f(x, w, b) \ = \ w^T x + b$

1.2.1 The margin and the classification error

Theorem Let $\mathcal{F}_{\mathcal{D}}$ be the class of hyperplanes $f(x) = w^T x$ that are R away from any data point in the training set $\mathcal{D}$. Then,

$$VCdim\, \mathcal{F}_{\mathcal{D}} \ \leq \ 1 + \min\left(N, \frac{R_{\mathcal{D}}^2}{R^2}\right) \tag{1}$$

where $R_{\mathcal{D}}$ is the radius of the smallest ball that encloses the dataset.

Theorem Let $\mathcal{F} = \{\mathrm{sgn}\,(w^T x),\ ||w|| \leq \Lambda,\ ||x|| \leq R\}$ and let $\rho > 0$ be any "margin". Then for any $f \in \mathcal{F}$, w.p $1 - \delta$ over training sets

$$R(f) \ \leq \ \nu + \sqrt{\frac{c}{N}\left(\frac{R^2\Lambda^2}{\rho^2}\ln N^2 + \ln\frac{1}{\delta}\right)} \tag{2}$$

where ν is the fraction of the training examples for which $y^i w^T x_i < \rho$ and c is a universal constant.

1.2.2 Formulating the optimization problem

Problem: min $\frac{1}{2}||w||^2$ s.t $y^i(w^T x^i + b) - 1 \geq 0$ for all i.

Optimization with Lagrange multipliers $\alpha_i \geq 0$.

minimize $L_P \;=\; \frac{1}{2}||w||^2 - \sum_i \alpha_i [y^i(w^T x^i + b) - 1]$

$w \;=\; \sum_i \alpha_i y^i x^i$

$\sum_i \alpha_i y^i \;=\; 0$

Dual optimization problem

maximize $L_D \;=\; \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y^i y_j x^{iT} x_j$ s.t $\alpha_i \geq 0$ for all i

Quadratic problem on convex domain: has unique minimum/maximum. At the optimum, $\alpha_i \;>\; 0$ for constraints that are satisfied with equality, $\alpha_i \;=\; 0$ otherwise.

Support vector: x^i such that $\alpha_i > 0$

The classifier $w \;=\; \sum_{i,\alpha_i>0} \alpha_i y^i x^i,\; b = y^i - w^T x^i$ for some support vector

1.3 Non-linearly separable problems

The **C-SVM**

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + C \sum_i \xi_i & (3) \\ \text{s.t.} \quad & y^i(w^T x^i + b) \;\geq\; 1 - \xi_i \\ & \xi_i \;\geq\; 0 \end{aligned}$$

In the above, ξ_i are the *slack variables*. Equivalent formulation:
minimize $L_P \;=\; \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i [y^i(w^T x^i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i$
s.t. $\alpha_i \geq 0,\; \xi_i \geq 0,\; \mu_i \geq 0$
Dual:

$$\begin{aligned} \text{maximize} \quad & \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y^i y_j x^{iT} x_j & (4) \\ \text{s.t.} \quad & C \;\geq\; \alpha_i \geq 0 \text{ for all } i \\ & \sum_i \alpha_i y^i \;=\; 0 \end{aligned}$$

$\Rightarrow$ two types of SV

- $\alpha_i < C$ data point x^i is "on the margin" $\Leftrightarrow$ $y^i(w^T x^i + b) = 1$ (original SV)

- $\alpha_i = C$ data point x^i cannot be classified with margin 1 (**margin error)** $\Leftrightarrow$ $y^i(w^T x^i + b) < 1$

The ν**-SVM**

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}||w||^2 - \nu\rho + \frac{1}{N}\sum_i \xi_i && (5)\\
\text{s.t.} \quad & y^i(w^T x^i + b) \;\geq\; \rho - \xi_i && (6)\\
& \xi_i \;\geq\; 0 && (7)\\
& \rho \;\geq\; 0 && (8)\\
& && (9)
\end{aligned}$$

Equivalent formulation:
minimize $L_P = \frac{1}{2}||w||^2 - \nu\rho + \frac{1}{l}\sum_i \xi_i - \sum_i \alpha_i[y^i(w^T x^i + b) - \rho + \xi_i] - \sum_i \mu_i \xi_i - \delta\rho$
s.t. $\alpha_i \geq 0,\ \delta \geq 0,\ \mu_i \geq 0$

Dual:

$$\begin{aligned}
\text{maximize} \quad & -\frac{1}{2}\sum_i \alpha_i \alpha_j y^i y_j x^{iT} x_j && (10)\\
\text{s.t.} \quad & \frac{1}{N} \;\geq\; \alpha_i \geq 0 \text{ for all } i && (11)\\
& \sum_i \alpha_i y^i \;=\; 0 && (12)\\
& \sum_i \alpha_i \;\geq\; \nu && (13)\\
& && (14)
\end{aligned}$$

Properties If $\rho > 0$ then:

- ν is an upper bound on #margin errors/N (if $\sum_i \alpha_i = \nu$)

- ν is a lower bound on #support vectors/N

- ν-SVM leads to the same w, b as C-SVM with $C = 1/\nu$

A simple error bound

$$E[L_{01}(f) \,|\, N-1] \;\leq\; E\left[\frac{\#\text{support vectors}}{N}\right] \tag{15}$$

where $E[L_{01}(f) \,|\, N]$ denotes the average loss classification error of a SVM trained on a sample of size N

2 Convex optimization and SVM

2.1 Convex optimization in a nutshell

A set $D \subseteq \mathbb{R}^n$ is **convex** iff for every two points $x^1, x^2 \in D$ the line segment defined by $x = tx^1 + (1-t)x^2,\ t \in [0,1]$ is also in D. A function $f : D \rightarrow R$ is **convex** iff, for any $x^1, x^2 \in D$ and for any $t \in [0,1]$ for which $tx^1 + (1-t)x^2 \in D$ the following inequality holds

$$f(tx^1 + (1-t)x^2) \;\leq\; tf(x^1) + (1-t)f(x^2) \tag{16}$$

If f is convex, then the set $\{\, x \,|\, f(x) \leq\; c \,\}$ is convex for any value of c. Convex functions defined on convex sets have very interesting properties which have engendered the field called **convex optimization**.

The optimization problem

$$\begin{array}{l} \min_x \; f(x) \\ \text{s.t. } f_i(x) \;\leq\; 0 \;\text{ for } i = 1, \ldots p \end{array} \tag{17}$$

is a **convex optimization problem** if all the functions f, f_i are convex. Note that in this case the **admissible domain** $D \;=\; \bigcap_i \{\, x \,|\, f_i(x) \leq\; 0 \,\}$ is a convex set.

It is known that if D has a non empty interior then the convex optimization problem has at most one optimum x^*. If D is also bounded, x^* always exists.

Assuming that x^* exists, there are two possible cases: (1) The **unconstrained minimum** of f lies in D. In this case, the optimum can be found

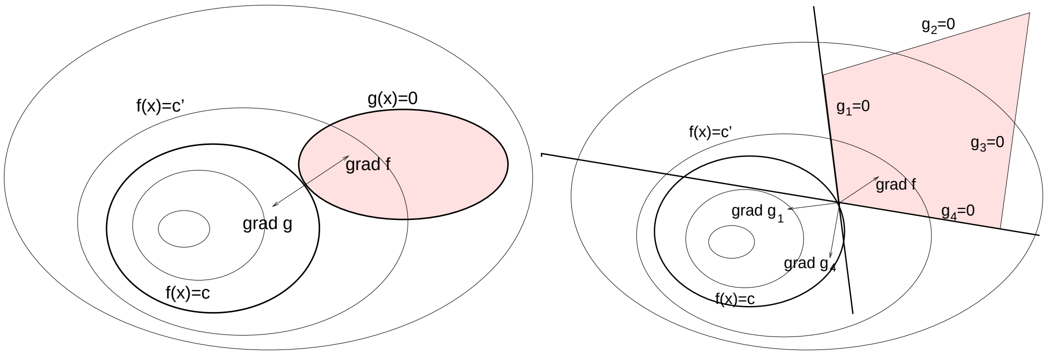


Figure 1: (a) One constraint optimization. (b) Four constraint optimization. At the optimum only constraints g_1, g_4 are active.

by solving the equations $\frac{\partial f}{\partial x} = 0$. (2) The unconstrained minimum of f lies outside D. Figure 1 depicts what happens at the optimum x^* in this case.

Assume there is only one constraint f_1. The domain D is the inside of the curve $f_1(x) = 0$. The optimum x^* is the point where a level curve $f(x) = c$ is tangent to $f_1 = 0$ from the outside. In this point, the gradients of two curves lie along the same line, pointing in opposite directions. Therefore, we can write $\frac{\partial f}{\partial x} = -\alpha \frac{\partial f_1}{\partial x}$. Equivalently, we have that at x^*, $\frac{\partial f}{\partial x} + \alpha \frac{\partial f_1}{\partial x} = 0$. Note that this is a necessary but not a sufficient condition. The above set of equations represents the **Karush-Kuhn-Tucker optimality conditions (KKT)**.

With more than one constraint, the KKT conditions are equivalent to requiring that the gradient of f lies in the subspace spanned by the gradients of the constraints.

$$\frac{\partial f}{\partial x} = -\sum_i \alpha_i \frac{\partial f_i}{\partial x} \text{ with } \alpha_i \geq 0 \text{ for all } i \tag{18}$$

Note that if a certain constraint f_i does not participate in the boundary of D at x^*, i.e if the constraint is not **active**, the coefficient α_i should be 0. Equation (18) can be rewritten as

$$\frac{\partial}{\partial x} \underbrace{[f(x) + \sum_i \alpha_i f_i(x)]}_{h(x,\alpha)} = 0 \text{ for some } \alpha_i \geq 0 \text{ for } i = 1, \ldots p \tag{19}$$

The optimum x^* has to satisfy the equation above. The new function $L(x, \alpha)$ is the **Lagrangean** of the problem and the variables α_i are called **Lagrange multipliers**. The Lagrangean is convex in x and **affine** (i.e linear + constant) in α.

The dual problem Define the function

$$g(\alpha) \;=\; \inf_x L(x, \alpha) \;\; \alpha \;=\; (\alpha_i)_i,\; \alpha_i \geq 0 \tag{20}$$

In the above, the infimum is over all the values of x for which f, f_i are defined, not just D (but everything still holds if the infimum is only taken over D). Two facts are important about g

- $g(\alpha) \leq L(x, \alpha) \leq f(x)$ for any $x \in D,\; \alpha \geq 0$, i.e g is a lower bound for f, and implicitly for the optimal value $f(x^*)$, for any value of $\alpha \geq 0$.
- $g(\alpha)$ is concave (i.e $-g(\alpha)$ is convex).

We also can derive from (19) that if x^* exists then for an appropriate value α^* we have

$$g(\alpha^*) \;=\; L(x^*, \alpha^*) \;=\; f(x^*) + 0 \tag{21}$$

and therefore $g(\alpha^*)$ must be the unique maximum of $g(\alpha)$. The second term in L above is zero because x^* is on the boundary of D; hence for the active constraints $f_i(x^*) = 0$ and for the inactive constraints $\alpha_i^* = 0$. This surprising relationship shows that by solving the **dual problem**

$$\begin{aligned} &\max g(\alpha) \\ &\text{s.t } \alpha \;\geq\; 0 \end{aligned} \tag{22}$$

we can obtain the values α^* that plugged into (18 will allow us to find the solution x^* to our original (**primal**) problem. The constraints of the dual are simpler than the constraints of the primal. In practice, it is surprisingly often possible to compute the function $g(\alpha)$ explicitly. Below we give a simple example thereof. This is also the case of the SVM optimization problem, which will be discussed in section 2.3.

2.2 A simple optimization example

Take as an example the convex optimization problem

$$\min \frac{1}{2}x^2 \quad \text{s.t} \;\; x+1 \leq 0 \tag{23}$$

By inspection the solution is $x^* = -1$.

Let us now apply to it the convex optimization machinery. We have

$$L(x,\alpha) \;=\; \frac{1}{2}x^2 + \alpha(x+1) \tag{24}$$

defined for $x \in R$ and $\alpha \geq 0$.

$$\begin{aligned} g(\alpha) \;&=\; \inf_x \left[\frac{1}{2}x^2 + \alpha(x+1)\right] && (25)\\ &=\; \inf_x \left[\frac{1}{2}(x+\alpha)^2 - \frac{1}{2}\alpha^2 + \alpha\right] && (26)\\ &=\; -\frac{1}{2}\alpha^2 + \alpha && (27)\\ &=\; \frac{1}{2}\alpha(2-\alpha) \quad \text{attained for } x \;=\; -\alpha && (28) \end{aligned}$$

The dual problem is

$$\max \frac{1}{2}\alpha(2-\alpha) \;\; \text{s.t } \alpha \geq 0 \tag{29}$$

and its solution is $\alpha = 1$ which, using equation (28) leads to $x = -1$.

From the KKT condition

$$\frac{\partial L}{\partial x} \;=\; x + \alpha \;=\; 0 \tag{30}$$

we also obtain $x^* \;=\; -\alpha^* \;=\; -1$.

Figure 2 depicts the function L. Note that L is convex in x (a parabola) and that along the α axis the graph of L consists of lines. The areas of L that fall outside the admissible domain $x \leq -1,\; \alpha \geq 0$ are in flat (green) color. The crossection $L(x, \alpha = 0)$ represents the plot of f. The constrained minimum of f is at $x = -1$, the unconstrained one is at $x = 0$ outside the admissible

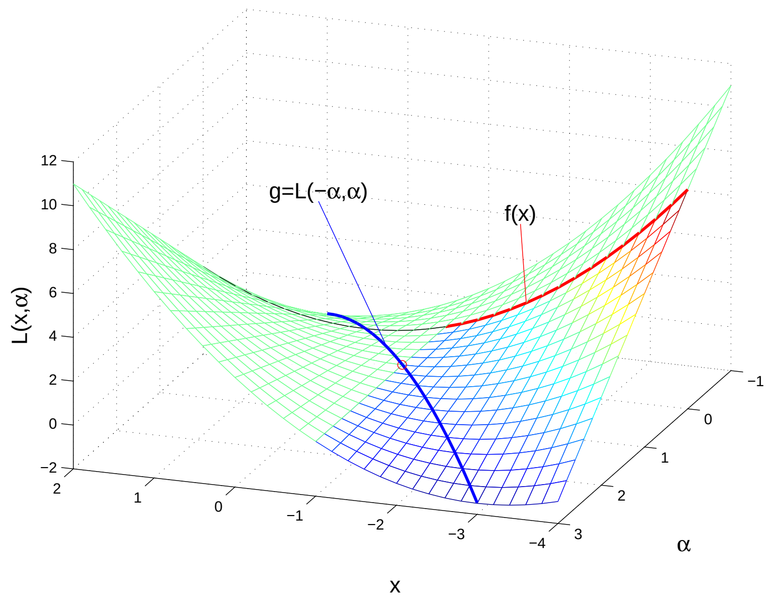


Figure 2: The surface $L(x, \alpha)$ for the problem min $\frac{1}{2}x^2$ s.t $x + 1 \leq 0$.

domain. Note that $g(\alpha) = L(-\alpha, \alpha)$ is concave, and that in the admissible domain it is always below the graph of f. The (red) dot is the optimum (x^*, α^*), which represents a **saddle point** for h. The line $L(x = -1, \alpha)$ is horizontal (because $f_1 = x + 1 = 0$) and thus $L(x^*, \alpha^*) = L(x^*,) = f(x^*)$.

2.3 The SVM solution by convex optimization

The SVM optimization problem

$$\min_w \frac{1}{2}||w||^2 \quad \text{s.t. } y^i(w^T x^i + b) \geq 1 \text{ for all } i \tag{31}$$

is a convex (quadratic) optimizaton problem where

$$f(w, b) = \frac{1}{2}||w||^2 \tag{32}$$

$$g_i(w, b) = -y^i w^T x^i + 1 - y^i b \tag{33}$$

Hence,

$$h(w, b, \alpha) = \frac{1}{2}||w||^2 + \sum_i \alpha_i [1 - y^i b - y^i x^{iT} w] \tag{34}$$

Equating the partial derivatives of h w.r.t w, b with 0 we get

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y^i x^i \quad (35)$$

$$\frac{\partial L}{\partial b} = \sum_i \alpha_i y^i \quad (36)$$

or, equivalently

$$w = \sum_i \alpha_i y^i x^i \qquad 0 = \sum_i \alpha_i y^i \quad (37)$$

Hence, the normal w to the optimal separating hyperplane is a linear combination of data points. Moreover, we know that only those α_i corresponding to active constraints will be non-zero. In the case of SVM, these represent points that are classified with $yi(w^T x^i + b) = 1$. We call these points **support points** or **support vectors**. The solution of the SVM problem does not depend on all the data points, it depends only on the support vectors and therefore is **sparse**.

Computing the solution. SVM solvers use the dual problem to compute the solution. Below we derive the dual for the SVM problem. $g(\alpha)$ is computed explicitly by replacing equation (37) in (34). After a simple calculation we obtain

$$g(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y^i y_j x^{iT} x_j \alpha_i \alpha_j \quad (38)$$

or, in vector/matrix notation

$$g(\alpha) = 1^T \alpha - \frac{1}{2} \alpha^T G \alpha \quad (39)$$

where $G = [G_{ij}]_{ij} = [y^i y_j x^{iT} x_j]_{ij}$.

3 A simple SVM problem

Data: 4 vectors in the plane and their labels

$$x_1 = (-2, -2) \qquad y_1 = +1$$

$$
\begin{aligned}
x_2 &= (-1,1) & y_2 &= +1 \\
x_3 &= (1,1) & y_3 &= -1 \\
x_4 &= (2,-2) & y_4 &= -1
\end{aligned}
$$

The Gramm matrix $G = [x^{iT} x_j]_{i,j=1:l}$

$$
G = \begin{bmatrix} 8 & 0 & -4 & 0 \\ 0 & 2 & 0 & -4 \\ -4 & 0 & 2 & 0 \\ 0 & -4 & 0 & 8 \end{bmatrix}
$$

The dual function to be maximized (subject to $\alpha_i \geq 0$) is

$$
\begin{aligned}
g(\alpha) &= \sum_i \alpha_i - \frac{1}{2} \sum_i \alpha_i \alpha_j y^i y_j x^{iT} x_j \\
&= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - 4\alpha_1^2 - \alpha_2^2 - \alpha_3^2 - 4\alpha_4^2 - 4\alpha_1\alpha_3 - 4\alpha_2\alpha_4 \\
&= (2\alpha_1 + \alpha_3) - (2\alpha_1 + \alpha_3)^2 - \alpha_1 \\
&\quad +(\alpha_2 + 2\alpha_4) - (\alpha_2 + 2\alpha_4)^2 - \alpha_4
\end{aligned}
$$

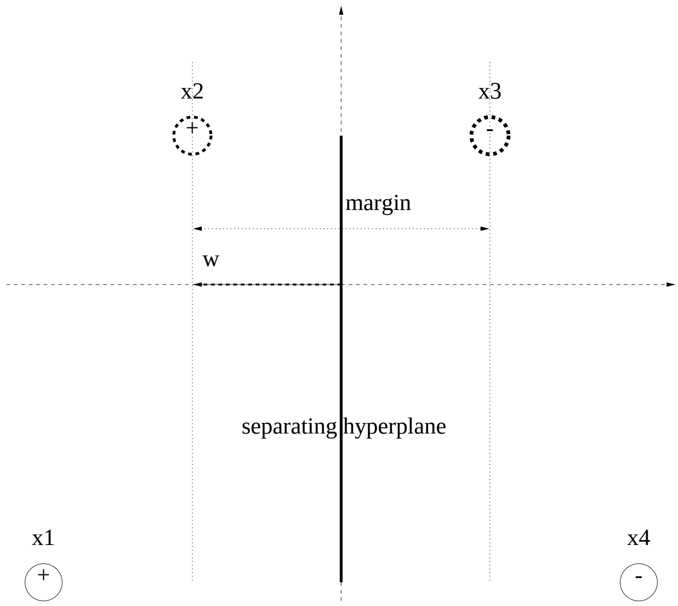
The parts depending on α_1, α_3 and α_2, α_4 can be maximized separately, and after some short calculations we obtain:

$$
\begin{aligned}
\alpha_1 &= 0 & \alpha_4 &= 0 \\
\alpha_2 &= \frac{1}{2} & \alpha_3 &= \frac{1}{2}
\end{aligned}
$$

Hence, the support vectors are x_2 and x_3. From these, we obtain

$$
\begin{aligned}
w &= \sum_i \alpha_i y^i x^i = \frac{1}{2}(x_2 - x_3) = (-1,0) \\
b &= y_2 - w^T x_2 = 0
\end{aligned}
$$

The results are depicted in the figure below:



4 Non linearly separable data: the "kernel trick"

We have seen so far how to construct a SVM classifier if the data are **linearly separable** i.e if there exist w, b such that the hyperplane $w^T x + b = 0$ leaves all the examples labeled $+1$ (called **positive examples**) on one side and all the examples labeled -1 (the **negative examples**) on its other side. If the data are not linearly separable, then no solution to the SVM optimization problem exists. Here we shall see a way of constructing SVM's that are **non linear** in the sense that they separate the positive and negative example by a (hyper)surface that is non-linear.

An old trick that allows us to use linear classifier on data that is not linearly separable is the following:

1. Map the data to a higher dimensional space $x \rightarrow z = \phi(x) \in H$, with $\dim H >> n$.
2. Construct a linear classifier $w^T z + b$ for the data in H

For example, the data $\{(x, y)\}$ below:

x		y	z		
-1	-1	1	-1	-1	1
-1	1	-1	-1	1	-1
1	-1	-1	1	-1	-1
1	1	1	1	1	1

are not linearly separable. We map them to 3 dimensions by $z = \phi(x) = [x_1\ x_2\ x_1x_2]$. Now it is easy to see that the classes can be separated by the hypeplane $z_3 = 0$ (which happens to be the maximum margin hyperplane). Hence $w = [001]$ (a vector in H) and $b = 0$ and the classification rule is $f(\phi(x)) = w^T\phi(x) + b$. If we express this rule as a function of the original x we get $f(x) = x_1x_2$ which is a quadratic classifier.

In summary, by mapping the data to H by $\phi(x)$ and then using a linear classifier, we are in fact implementing the non-linear classifier

$$f(x) \ = \ w^T\phi(x) + b \ = \ w_1\phi_1(x) + w_2\phi_2(x) + \ldots + w_m\phi_m(x) + b \qquad (40)$$

Rephrasing the non-linear classification problem in SV language we obtain:

Problem: minimize $||w||^2$ s.t $y^i(w^T\phi(x^i) + b) - 1 \geq 0$ for all i.

Note that the only difference from the linear case is that x^i is now replaced with $\phi(x^i)$. The dual Lagrangean, which is the problem that is effectively solved, is also similar to the original Lagrangean:

maximize $L_D \ = \ \sum_i \alpha_i - \frac{1}{2}\sum_i \alpha_i\alpha_j y^i y_j \phi(x^i)^T\phi(x_j)$ s.t $\alpha_i \geq 0$ for all i

How much harder has the optimization become now? Surprizingly, the optimization problem is no harder than it was before! Note that the Lagrangean has a linear term that depends only on α and a quadratic term that can be written

$$\bar{\alpha}^T G \bar{\alpha} \qquad (41)$$

where $\bar{\alpha} = [\alpha_i y^i]_{i=1:l}$ and $G = [G_{ij}]_{i,j=1}^{l}$ is the **Gram matrix**

$$G_{ij} \ = \ G_{ji} \ = \ \phi(x^i)^T\phi(x_j) \quad \text{formerly} \quad G_{ij} \ = \ G_{ji} \ = \ x^{it}x_j \qquad (42)$$

A few facts follow from this observation:

1. The ϕ vectors enter the SVM optimization problem only trough the Gram matrix, thus only as the scalar products $\phi(x^i)^T\phi(x_j)$. We denote by $K(x, x')$ the function

$$K(x, x') \;=\; K(x', x) \;=\; \phi(x)^T\phi(x') \tag{43}$$

   K is called the **kernel** function. If K can be computed efficiently, then the Gram matrix G can also be computed efficiently. This is exactly what one does in practice: we choose ϕ implicitly by choosing a kernel K. Hereby we also ensure that K can be computed efficiently.

2. Once G is obtained, the SVM optimization is independent of the dimension of x and of the dimension of $z = \phi(x)$. The complexity of the SVM optimization depends only on l the number of examples. This means that we can choose a very high dimensional ϕ without any penalty on the optimization cost.

3. Classifying a new point x. As we know, the SVM classification rule is

$$f(x) \;=\; w^T\phi(x) + b \;=\; \sum_{i=1}^{l} \alpha_i y^i \phi(x^i)^T\phi(x) \;=\; \sum_{i=1}^{l} \alpha_i y^i K(x^i, x) \tag{44}$$

   Hence, the classification rule is expressed in terms of the support vectors and the kernel only. No operations other than scalar product are performed in the high dimensional space H.

The above describes the celebrated **kernel trick** of the SVM literature.

5 Kernels

The previous section shows why SVMs are often called **kernel machines**. If we choose a kernel, we have all the benefits of a mapping in high dimensions, without ever carrying on any operations in that high dimensional space. The most usual kernel functions are

$K(x, x') \;=\; (1 + x^T x')^p$	the polynomial kernel of degree p
$K(x, x') \;=\; e^{-\frac{\|\|x-x'\|\|^2}{\sigma^2}}$	the Gaussian or **radial basis function** (RBF) kernel it's ϕ is ∞-dimensional
$K(x, x') \;=\; \tanh(\sigma x^T x' - \beta)$	the "neural network" kernel

How do we verify that a symmetric function K is a valid kernel, i.e that there is a mapping ϕ for which K is the scalar product? This is ensured by the **Mercer condition** which is a positivity condition

$$\int K(x,x')g(x)g(x')dxdx' \geq 0 \quad \text{for all } g \text{ such that } ||g(x)||_{L_2} < \infty \tag{45}$$

6 Extensions to other problems

6.1 Multi-class SVM

For a problem with K possible classes, we construct K separating hyperplanes $w_r^T x + b_r = 0$.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\sum_{r=1}^{K}||w_r||^2 + \frac{C}{l}\sum_{i,r}\xi_{i,r} & (46) \\ \text{s.t.} \quad & w_{y^i}^T x^i + b_{y^i} \geq w_r^T x^i + b_r + 1 - \xi_{i,r} \quad \text{for all i} = 1:\text{l}, \text{r} \neq \text{y}^\text{i} & (47) \\ & \xi_{i,r} \geq 0 & (48) \end{aligned}$$

6.2 One class SVM

This SVM finds the “support regions” of the data, by separating the data from the origin by a hyperplane. It’s mostly used with the Gaussian kernel, that projects the data on the unit sphere. The formulation below is identical to the ν-SVM where all points have label 1.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 - \nu\rho + \frac{1}{N}\sum_i \xi_i & (49) \\ \text{s.t.} \quad & w^T x^i + b \geq \rho - \xi_i & (50) \\ & \xi_i \geq 0 & (51) \\ & \rho \geq 0 & (52) \end{aligned}$$

6.3 SV Regression

The idea is to construct a "tolerance interval" of $\pm\epsilon$ around the regressor f and to penalize data points for being outside this tolerance margin. In words, we try to construct the smoothest function that goes within ϵ of the data points.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + C\sum_i(\xi_i^+ + \xi_i^-) && (53) \\ \text{s.t.} \quad & \epsilon + \xi_i^+ \;\geq\; w^T x^i + b - y^i \;\geq\; -\epsilon - \xi_i^- && (54) \\ & \xi_i^\pm \;\geq\; 0 && (55) \\ & \rho \;\geq\; 0 && (56) \end{aligned}$$

The above problem is a linear regression, but with the kernel trick we obtain a kernel regressor of the form $f(x) = \sum_i(\alpha_i^- - \alpha_i^+)K(x^i, x) + b$