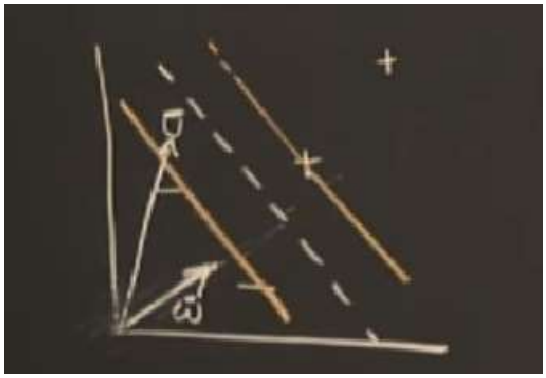
SVM

$w.u \geq C$

Si C=-b

$w.u+b \geq 0$ entonces +

Regla de Decisión.



$x_P.w+b \geq 1$

$x_N.w+b \geq 1$

y_j tal que $y_j=1$ para ejemplos positivos P y $y_j=-1$ para ejemplos negativos N. Por tanto podemos escribir:

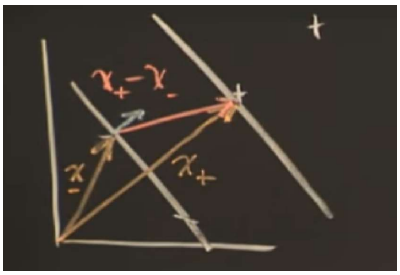
$y_j.(x_P.w+b) \geq 1$

$y_j.(x_N.w+b) \geq 1$

De forma mas compacta:

$y_j.(x.w+b)-1 \geq 0$ [1]

Ancho del canal:



Teniendo en cuenta [1], con $y_j=1$ podemos escribir:

$x_P.w+b-1 \geq 0 \; ; x_P.w=1-b$

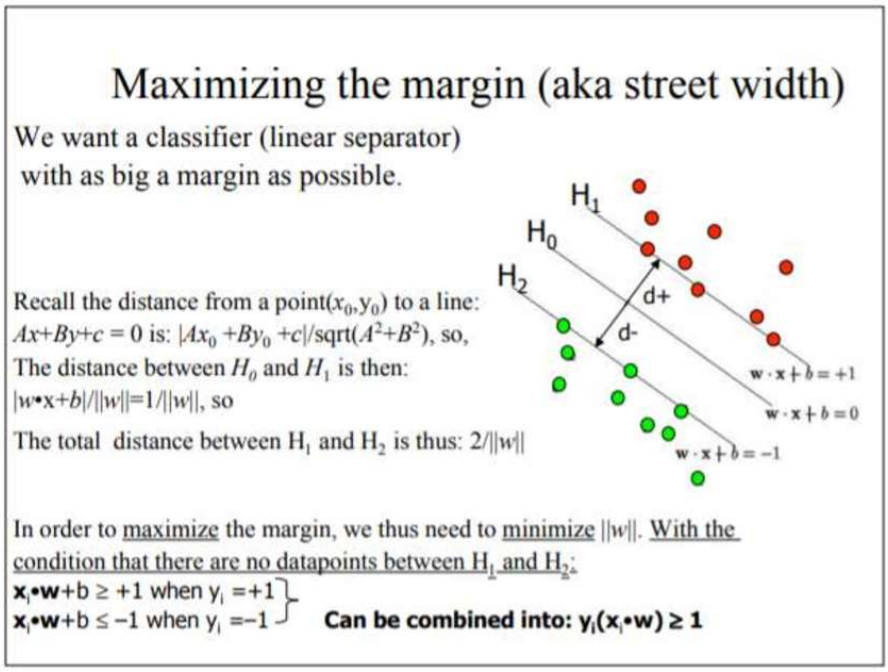
Análogamente:

$-(x_N.w+b-1) \geq 0 \; ; -x_N.w=1+b$

Por tanto, el ancho del canal es:

$$WIDTH=(x_P-x_N)\frac{w}{||w||}=\frac{1-b+1+b}{||w||}=\frac{2}{||w||} \quad [2]$$

Otra forma de calcular el ancho del canal es:



Asi que debemos maximizar $max\frac{2}{||w||}$ sujeto a la restricción [1]. Pero $max\frac{2}{||w||}$ es equivalente a $max\frac{1}{||w||}$ que a su vez es equivalente a minimizar $min||w||$ y equivalente a $min\frac{1}{2}||x||^2$

El problema de programación se puede expresar así:

$min\frac{1}{2}||x||^2$

sujeto a: $y_j.(x.w+b)-1\geq 0$

Teniendo en cuenta que la función a minimizar cumple todas las condiciones para poder usar los Multiplicadores de Lagrange, usaremos dicha herramienta para la optimización.
Un ejemplo de Multiplicadores de Lagrange:

Optimiza la función:

$$\begin{aligned} f(x,y) &= 2-x^2-2y^2 \\ g(x,y) &= x+y-1=0 \end{aligned}$$

Solución:

$$\begin{aligned} L(x,y,\lambda) &= f(x,y)-\lambda g(x,y) \\ &= 2-x^2-2y^2-\lambda(x+y-1) \end{aligned}$$

El gradiente de la función de Lagrange es 0, es decir:

$$\nabla L(x,y,\lambda)=\nabla f(x,y)-\lambda\nabla g(x,y)=0$$

Derivando L obtenemos:

$$\begin{aligned} \frac{\partial}{\partial x}L(x,y,\lambda) &= -2x-\lambda=0 \\ \frac{\partial}{\partial y}L(x,y,\lambda) &= -4y-\lambda=0 \\ \frac{\partial}{\partial \lambda}L(x,y,\lambda) &= x+y-1=0 \end{aligned}$$

Los multiplicadores de Lagrange son: $\bar{\lambda}=-\frac{4}{3}$ and $\bar{f}=\frac{4}{3}$.

La función de Lagrange para optimizar el problema de programación planteado es:

$L=\frac{1}{2}||x||^2-\sum\alpha_i[y_j.(x.w+b)-1]$ [3]

$\frac{\partial L}{\partial w}=w-\sum_i \alpha_i y_i.x_i=0$

$\frac{\partial L}{\partial b}=-\sum_i \alpha_i y_i=0$

Así que obtenemos:

$w=\sum_i \alpha_i y_i.x_i$ [4]

$$\sum_i \alpha_i y_i = 0 \quad [5]$$

Sustituyendo [4] en [3], obtenemos la siguiente expresión:

$$L = \frac{1}{2}(\sum_j \alpha_j y_j . x_{ji})(\sum_i \alpha_i y_i . x_i) - (\sum_j \alpha_j y_j . x_{ji})(\sum_i \alpha_i y_i . x_i) - \sum_i \alpha_i y_i b + \sum_i \alpha_i$$

Teniendo en cuenta la expresión [5]:

$$L = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i . x_j \quad [6]$$

En la expresion [6] podemos observar que el proceso de obtención de los coeficientes de Lagrange depende del producto de los datos $x_i . x_j$

Problema.

Determina la recta que divide al conjunto de puntos A=[0,0] que pertenece a la clase Y=1 y B=[4,4] que pertenece a la clase Y=-1 usando SVM. Calcula K (usamos el kernel K(u,v)=u.v, dot product), restricciones SVM, los parametros de Lagrange α , w, las ecuaciones del hiperplano y la de los planos positivos y negativos, el ancho del margen.
¿A que clase pertenece el punto [5,6]? ¿Y el punto [1,-4]?
Solución:
Calculamos los valores del kernel:

$$K(\vec{A}, A) = A.A = 0;\ K(A, B) = A.B = 0;\ K(B, A) = B.A = 0;\ K(B, B) = A.B = 32$$

La ecuaciones que usaremos son:

1. $\sum_{i=1}^{n} \alpha_i Y_i = 0$
2. $\sum_{i=1}^{n} \alpha_i Y_i K(X_i, X) + b = +1$
3. $\sum_{i=1}^{n} \alpha_i Y_i K(X_i, X) + b = -1$

La ecuacion 2 la cumple los puntos cuya clase es Y=1, mientras que la tercera ecuacion la cumple los puntos de la clase Y=-1.
Para nuestro caso concreto, se puede escribir:

1. $y_A \alpha_A + y_B \alpha_B = 0$
2. $y_A * K(A, A) * \alpha_A + y_B * K(B, A) * \alpha_B + b = +1$
3. $y_A * K(A, B) * \alpha_A + y_B * K(B, B) * \alpha_B + b = -1$

Sustituyendo por sus valores, tenemos el sistema de ecuaciones:

$$\alpha_A + \alpha_B = 0$$
$$b = 1$$
$$-32\alpha_B + b = -1$$

Cuya soluciona es $b = 1; \alpha_A = \alpha_B = \frac{1}{16}$

La ecuación [4] nos permite obtener w:

$$w = y_A \alpha_A x_A + y_B \alpha_B X_B$$

Sustituyendo:

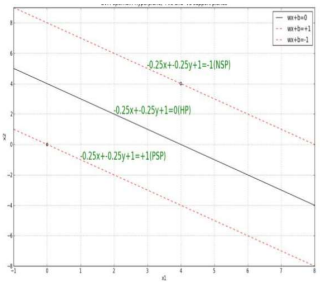
$$w = (+1)(\frac{1}{16})[\frac{0}{0}] + (-1)(\frac{1}{16})[\frac{4}{4}] = [\frac{-0.25}{-0.25}]$$

La ecuación del hiperplano es $w_1 x + w_2 y + b = 0$, sustituyendo:

$-0.25x - 0.25y + 1 = 0$; (Hiperplano)

$-0.25x - 0.25y + 1 = 1$; (Hiperplano constituido por los vectores soporte con Y=1)

$-0.25x - 0.25y + 1 = -1$; (Hiperplano constituido por los vectores soporte con Y=-1)



$$Ancho = \frac{1}{\|w\|} = 4\sqrt{2}$$

¿A que clase pertenece el punto [5,6]? ¿Y el punto [1,-4]?

$signo(f([5,6]^T)) = -0.25(5) - 0.25*(6) = signo(-1.75) = -1$; (clase -)

$signo(f([1,4]^T)) = -0.25(1) - 0.25*(4) = signo(+1.75) = +1$ (clase +)

Sea el punto A=[2,0] que pertenece a la clase Y=1 y B=[0,0], C=[1,1] que pertenece a la clase Y=-1. Calcula K (usamos el kernel K(u,v)=u.v, dot product), restricciones SVM, los parámetros de Lagrange α , w, las ecuaciones del hiperplano y la de los planos positivos y negativos, el ancho del margen.

¿A que clase pertenece el punto [5,6]? ¿Y el punto [1,-4]?

Example3: We are given the **positively labeled data** points at: [2, 2], [2, -2], [-2, -2], [-2, 2] in R^2. We are given the **negatively labeled data points** at: [1, 1], [1, -1], [-1, -1], [-1, 1] in R^2 and we are asked to solve for equation for the decision boundary.

Our goal, again, is to discover a separating hyperplane that accurately discriminates the two classes. Of course, it is obvious that no such hyperplane exists in the input space (that is, in the space in which the original input data live). Therefore, we must use a nonlinear SVM (that is, one whose mapping function Ø is a nonlinear mapping from input space into some feature space).
Define:

$$\emptyset\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4 - x_2 + |x_1 - x_2|\\4 - x_1 + |x_1 - x_2|\end{pmatrix} & \textit{if } \sqrt{x_1^2 + x_2^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & \textit{otherwise}\end{cases}$$

We can see how Ø transforms our data before the dot products are performed. Therefore, we can rewrite the data in feature space as [2, 2], [10, 6], [6, 6], [6, 10] for the positive examples and [1, 1], [1, -1], [-1, -1], [-1, 1] for the negative examples. Now we can once again easily identify the support vectors.

Let $X_1 = [2, 2]$, $X_2 = [10, 6]$, $X_3 = [6, 6]$, $X_4 = [6, 10]$ belong to class $Y = +1$ and
$Y_1 = [1, 1]$, $Y_2 = [1, -1]$, $Y_3 = [-1, -1]$, $Y_4 = [-1, 1]$ belong to class $Y = -1$. In general $X_i = [x, y]$, $Y_i = [x_0, y_0]$
Step1: Determine the support vector points from minimum distance between each point of class $Y = +1$ and each point of class $Y = -1$, by the formula $\sqrt{(x - x_0)^2 + (y - y_0)^2}$.

Distance between points X_1 & $Y_1 = \sqrt{(2-1)^2 + (2-1)^2} = \sqrt{2}$, similarly
Distance between points X_1 & $Y_2 = \sqrt{10}$, Distance between points X_1 & $Y_3 = \sqrt{18}$
Distance between points X_1 & $Y_4 = \sqrt{10}$, Distance between points X_2 & $Y_1 = \sqrt{106}$
Distance between points X_2 & $Y_2 = \sqrt{130}$, Distance between points X_2 & $Y_3 = \sqrt{170}$
Distance between points X_2 & $Y_4 = \sqrt{146}$, Distance between points X_3 & $Y_1 = \sqrt{50}$
Distance between points X_3 & $Y_2 = \sqrt{74}$, Distance between points X_3 & $Y_3 = \sqrt{98}$
Distance between points X_3 & $Y_4 = \sqrt{74}$, Distance between points X_4 & $Y_1 = \sqrt{106}$
Distance between points X_4 & $Y_2 = \sqrt{146}$, Distance between points X_4 & $Y_3 = \sqrt{170}$
Distance between points X_4 & $Y_4 = \sqrt{130}$
Distance between X_1 and Y_1 is minimum and therefore X_1, Y_1 are the support vectors. $X_1 = [2, 2]$ is positive class support vector and $Y_1 = [1, 1]$ is negative class support vector. Once the support vectors are got rest of the procedure is same as in Example1.

Problema

To understand the first two questions, let's consider $x, y \in \mathbb{R}^2, x = (x_1, x_2), y = (y_1, y_2)$ and examine the polynomial kernel of degree 2:

$$K(x, y) = (x^T y)^2$$

Which can be rewritten as:

$$K(x, y) = (x_1 y_1 + x_2 y_2)^2 = x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2$$

We know that the kernel function is $K(x, y) = \Phi(x)^T \Phi(y)$, therefore we try to find a feature map Φ that will be equivalent to the above. Let

$$\Phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$$

From this, we can see that $\Phi(x)^T \Phi(y) = x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2$, which is the kernel function!

Notice that by using Φ we mapped the input vectors from $\mathbb{R}^2$ to $\mathbb{R}^3$, therefore when we compute $K(x, y)$, this mapping will be implicitly performed.