Date
Page

Support Vector Machine (SVM)

$$\text{Margin} = \frac{2}{\|W\|}$$

Optimization Problem

Maximize $\frac{2}{\|W\|}$ such that $y_i(W^T x_i + b) \geq 1$
$i = 1$ to N.

(OR)

Minimize $\frac{1}{2}\|W\|^2$

A Lagrangian multiplier (α), we can combine,

$$\underset{x,\,y,\,\alpha}{\text{Minimize}}\; L(x, y, \alpha) = f(x, y) - \alpha\, g(x, y)$$

$\Downarrow$

converts constraint to unconstraint problem

Here, $f(x, y)$ is $\frac{1}{2}\|W\|^2$
$g(x, y)$ is $y_i(w^T x_i + b) \geq 1$

Now, the quadratic programming problem with linear constraints can be written as,

$$L = \frac{1}{2}\|W\|^2 - \sum_i \alpha_i \left(y_i \left(w_i x_i + b\right) - 1\right) \rightarrow ①$$

Find derivation with respect to w and b

$$\frac{\partial L}{\partial w} = \vec{w} - \sum \alpha_i y_i x_i = 0 \rightarrow (2)$$

$$\vec{w} = \sum \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} \Rightarrow - \sum \alpha_i y_i = 0 \rightarrow (3)$$

Substitute (2) and (3) in (1)

$$= \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i y_i w^T x_i - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i + w^T \left(\frac{1}{2} w - \sum_{i=1}^{n} \alpha_i y_i x_i - 0 \right)$$

$$= \sum_{i=1}^{n} \alpha_i + w^T \left(\frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i - \sum_{i=1}^{n} \alpha_i y_i x_i \right)$$

$$= \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \alpha_i^T y_i^T x_i^T \left(-\frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i \right)$$

Date
Page

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \rightarrow (4)$$

The lagrangian dual problem, instead of minimizing over w and b subject to constraints involving α's, we can maximize over α (the dual variable) subject to the relation obtained previously for w and b

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

with the constraints $\alpha_i \geq 0$ $i = 1 to n$

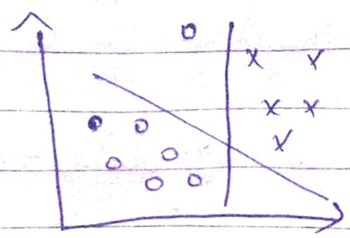
$$\sum_{i=1}^{n} \alpha_i y_i = 0 .$$

Non -Separable case and slack variable :-

→ In some cases the data points are not linearly separable becas of outliers

Result :-

Decision boundary is shifting, and resulting classifier will have small margin.



To make the algorithm work for non-linearly separable datasets as well as less sensitive to outliers, we reformulate the optimization as,

$$\min_{w,b,\alpha} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \varepsilon_{e_i}$$

subject to

$$y_i(w_i^T x_i + b) \geq 1 - \varepsilon_{e_i}, \quad i = 1 \text{ to } n$$

$$\varepsilon_{e_i} \geq 0 \quad i = 1 \text{ to } n$$

Thus the examples are now permitted to have margin less than 1. and if an example has functional margin $1-\varepsilon_{e_i}$ (with $\varepsilon_e > 0$), we would pay a cost of the objective function being increased by $C\varepsilon_{e_i}$. The parameter C contraints the ~~relatively~~ weight between the twin goals of making $\|w\|$ small and of ensuring the most examples have functional margin at least 1.

Dual form with slack variable

Non-separable problem:-

$$\min_{w,b,\alpha} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i$$

subject to $y_i(w^T x_i + b) \geq \varepsilon_{e_i}$

$$\varepsilon_{e_i} \geq 0 \quad i = 1 \text{ to } n$$

constraints transformed to,

$$g(\omega, b) = 1 - \varepsilon_{e_i} - y_i(\omega^T x_i + b) \leq 0$$

$$h(\omega, b) = -\varepsilon_{e_i} \leq 0$$

Lagrangian,

$$L(\omega, b, \varepsilon_e, \alpha, r) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \varepsilon_{e_i} - \sum_{i=1}^{n} \alpha_i (y_i(\omega^T x_i + b) + \varepsilon_{e_i} - 1) - \sum_{i=1}^{n} r_i \varepsilon_{e_i}$$

derivate This,

for ω,

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i$$

for ε_e

$$C = \alpha_i + r_i$$

$\forall i = 1$ to n

for b,

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\Rightarrow L(\omega, b, \varepsilon_e, \alpha, r) \Leftrightarrow L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n} \varepsilon_{e_i} - \sum_{i=1}^{n} \alpha_i (y_i(\omega^T x_i + b) + \varepsilon_{e_i} - 1) - \sum_{i=1}^{n} r_i \varepsilon_{e_i}$$

$$= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \varepsilon_i + \sum_{i=1}^{n} r_i \varepsilon_i$$

$$- \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1) - \sum_{i=1}^{n} \alpha_i \varepsilon_i -$$

$$\sum_{i=1}^{n} r_i \varepsilon_i$$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1)$$

$$= \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$- \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Same as previous one.

Date
Page

Non-Linear boundary

Any dataset which has a non-linear boundary would be theoretically linear separable if projected to higher dimension

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j)$$

We can write ω and other test phase equation.

$$y_{test} = \text{sign}(\omega_0 \cdot \phi(x_{test}) + b_0)$$

$$\omega_0 = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

$$\Rightarrow y_{test} = \text{sign}\left(\sum_{i=1}^{n} (\alpha_i y_i \phi(x_i) \cdot \phi(x_{test})\right) + b_0$$

Kernel trick:-

The mapping occurs as a dot product in both training as well as testing.

Since we don't know the mapping, we can find a function $k(x,y)$ which is equivalent to the dot product of the mapping.

We can avoid explicit mapping to the higher dimension.

Let us consider an example of quadratic kernel to understand better,

$$\phi(x) = \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1\, x_1 \\ x_1\, x_2 \\ x_2\, x_1 \\ x_2\, x_2 \end{bmatrix} \qquad R^2 \rightarrow R^4$$

$$k(x,y) = \phi(x) \cdot \phi(y)$$

$$= \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2 x_1 \\ x_2^2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_1 y_2 \\ y_2 y_1 \\ y_2^2 \end{bmatrix}$$

$$x_1^2 y_1^2 + 2 x_1 y_1 x_2 y_2 + x_2^2 y_2^2 = (x_1 y_1 + x_2 y_2)^2$$

$$\Downarrow \qquad = (x \cdot y)^2$$

Reverse

$$[x_1^2 \;\; \sqrt{2} x_1 x_2 \;\; x_2^2]^T \, [y_1^2 \;\; \sqrt{2} y_1 y_2 \;\; y_2^2]$$

$$\Rightarrow \phi(x)^T \cdot \phi(y)$$

n dimensional mapping | kernel,

$$k(x,y) = (x \cdot y)^n$$

$$k(x,y) = \phi(x)^T \cdot \phi(y)$$

~~Let~~

$$k = \phi(x)^T \cdot \phi(x)$$

$$= \begin{bmatrix} \phi(x_1)^T \cdot \phi(x_1) & \phi(x_1)^T \cdot \phi(x_2) & \phi(x_1)^T \cdot \phi(x_3) \\ \phi(x_2)^T \cdot \phi(x_1) & \phi(x_2)^T \cdot \phi(x_2) & \phi(x_2)^T \cdot \phi(x_3) \\ \phi(x_3)^T \cdot \phi(x_1) & \phi(x_3)^T \cdot \phi(x_2) & \phi(x_3)^T \cdot \phi(x_3) \end{bmatrix}$$

Example:-

$$x = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 6 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

d-degree polynomial kernel

$$k(x_i, x_j) = (1 + x_i^T x_j)^d$$

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

- Let $H = k * (y_i^T * y_j)$

$$\begin{bmatrix} -1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \begin{bmatrix} -1 & -1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 \end{bmatrix}$$

$$k = \begin{bmatrix} (1\times1+1)^2 & (1\times2+1)^2 & (1\times5+1)^2 & (1\times6+1)^2 \\ (2\times1+1)^2 & (2\times2+1)^2 & (2\times5+1)^2 & (2\times6+1)^2 \\ (5\times1+1)^2 & (5\times2+1)^2 & (5\times5+1)^2 & (5\times6+1)^2 \\ (6\times1+1)^2 & (6\times2+1)^2 & (6\times5+1)^2 & (6\times6+1)^2 \end{bmatrix}$$

$$k = \begin{bmatrix} 4 & 9 & 36 & 49 \\ 9 & 25 & 121 & 169 \\ 36 & 121 & 676 & 961 \\ 49 & 169 & 961 & 1369 \end{bmatrix}$$

$$H = \begin{bmatrix} 4 & 9 & -36 & 49 \\ 9 & 25 & -121 & 169 \\ -36 & -121 & 679 & -961 \\ 49 & 169 & -961 & 1369 \end{bmatrix}$$

$$\min_{\alpha} L(\alpha) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^{n} \alpha$$

$$= \frac{1}{2} H \alpha^T \alpha - \sum_{i=1}^{n} \alpha$$

$$= \frac{1}{2} [\alpha_1 \; \alpha_2 \; \alpha_3 \; \alpha_4] \begin{bmatrix} 4 & 9 & -36 & 49 \\ 9 & 25 & -121 & 169 \\ -36 & -121 & 679 & -961 \\ 49 & 169 & -961 & 1369 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} -$$

$$[\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4]$$

Date
Page

Slove $L(\alpha)$ using quadratic prgming process to solve Lagrangian variables

$\alpha_1 = 0 \quad \alpha_2 = 2.5 \quad \alpha_3 = 7.333 \quad \alpha_4 = 4.83$

$$y = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \, \phi(x_i) \cdot \phi(x_{test}) + b_0\right)$$

$$= \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \, (x_i^T x + 1)^2 + b_0\right)$$

$$= \text{sign}\big((-1) \cdot 0 \cdot (x+1)^2 +$$

$$(-1)(2.5) \cdot (2x+1)^2 +$$

$$(1)(7.33)(5x+1)^2 +$$

$$(-1)(4.833)(6x+1)^2 + b_0\big)$$

$$= \text{sign}(0.667x^2 + 5.33x + b_0)$$

Find bias b_0 by considering one of the support vector $x=2$ and $y=-1$

$$(-0.667x^2 + 5.33x + b_0) = -1$$

$$-0.667 \times 4 + 5.33 \times 2 + b_0 = -1$$

$$8 + b_0 = -1$$

$$b_0 = -9$$

$$\therefore y = -0.667x^2 + 5.33x - 9$$

(Q1)

$$y = \text{sign}(-0.667x^2 + 5.33x - 9)$$