S21CSE471: Statistical Methods in AI **Spring 2021**

Support Vector Machines

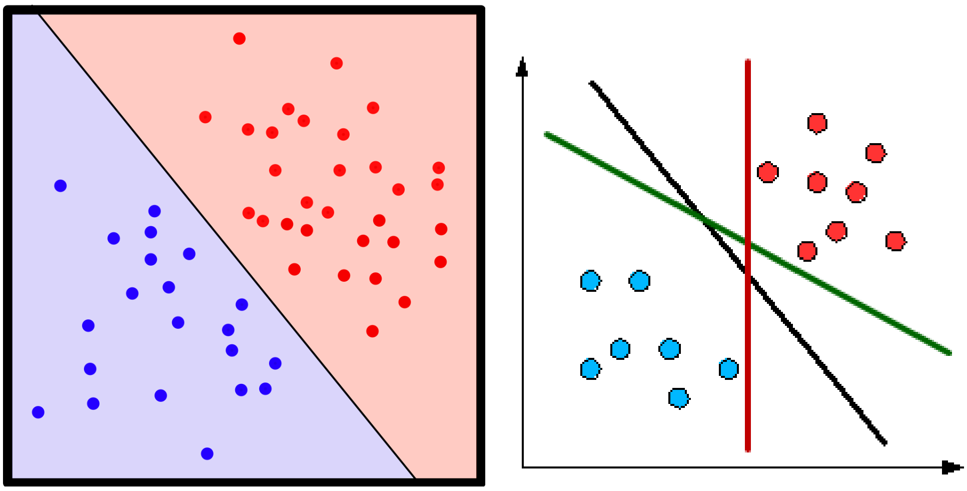
Prepared by: John, Navneet, Medha

1 What Is A Support Vector Machine?

A support vector machine algorithm can be used for both classification and regression tasks. However, it is commonly used for classification tasks. The goal of a support vector machine is to find the most suitable hyperplane in an N-dimensional space that separates clearly the two or more classes.

2 Linearly Separable Data

Multiple solutions exist for linearly separable data. In the following figures the best fit will take into account future data so the black line is the best fit as it maximizes the margin. Because our goal is not to do well on the training data but on the test or real world data so maximizing the margin between the no man's band is important.



(a) Linearly Separable Data

(b) Multiple Solutions.

Figure 1: Linearly Separable Data.

The separating hyperplane with maximum margin is likely to perform well on test data. **Support Vector** are data points that support the boundary.

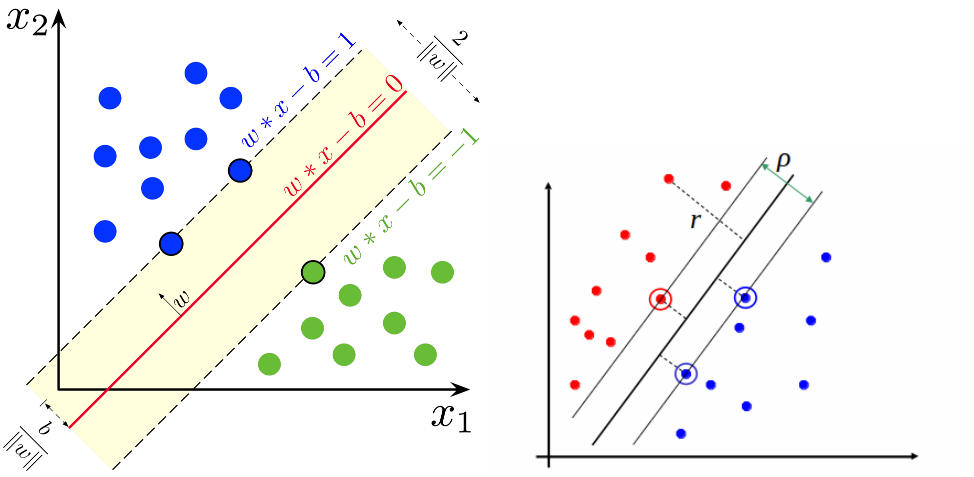
Support vectors are the data points that lie closes to the decision surface (or hyperplane) .These data points are most difficult to classify

There are 2 hyperplane for linear separable data and the distance between both hyperplane is called magin ,means Margin of the separator is the distance between support vectors.

2.1 Bounded Loss

Break through work of Vapnik and Chervonenkis was that they could put a bound on the error on test set based on training performance

$R(\alpha) \leq R_{\text{train}}(\alpha) + \sqrt{f(h)/N}$



(a) Margin reduction

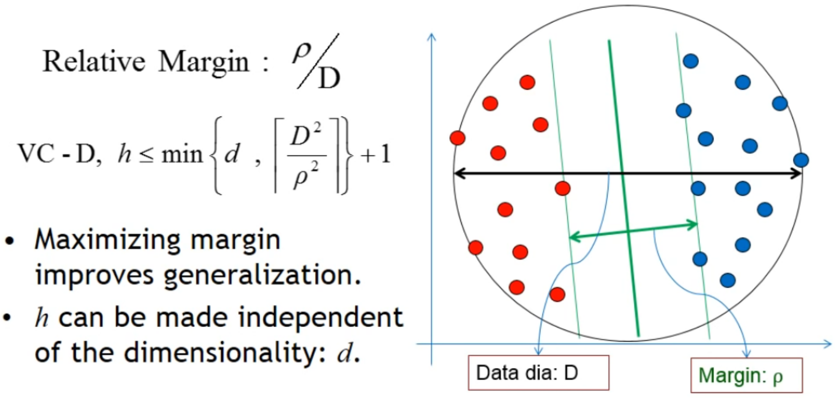
(b) Support Vectors.

Figure 2: Support Vector Machines

f(h) is a monotonically increasing function which will not take negative values. So this means that to reduce the bound on test data set we need to reduce f(h).

3 How to reduce the relative Margin ?

To reduce test error , keep trainging error low and minimise the relative margin.
Its found that h is defined as below
$h \leq min(d, \lceil D^2/\rho \rceil) + 1$
here ρ/D is relative margin



Relative Margin : ρ/D

VC - D, $h \leq \min\left\{ d \;,\; \left\lceil \frac{D^2}{\rho^2} \right\rceil \right\} + 1$

- Maximizing margin improves generalization.
- h can be made independent of the dimensionality: d.

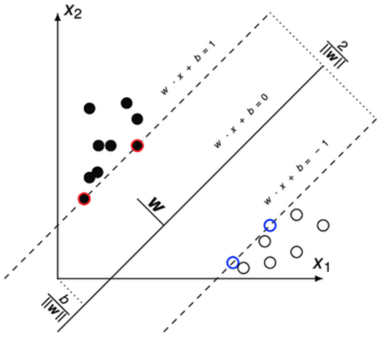
from the above information we see that if we increase the margin the relative margin increase .
This formula also help to select limited boundary with the fuction of D because margin never greater than the data diameter(D) Maximizing margin improves generalization and hence we can have a better fit on future data
Support vectors matter for computation of margin and other training examples are ignorable.

4 Formalizing the margin

Let training set (xi, yi)i=1..n, xi training data, yi-1, 1 class label , be separated by a hyperplane with margin . Then for each training example (xi, yi):

wTx$_i + b \leq -\frac{\rho}{2}\ if\ y_i = -1$
$wTx_i + b \geq \frac{\rho}{2}\ if\ y_i = 1$
$y_i(wTx_i + b)\frac{\rho}{2}$

For every support vector xs the above inequality is an equality. After rescaling w and b by$\rho/2$ in the equality, we obtain that distance between each xs and the hyperplane is
$r = \frac{(y_s(wTx_s+b))}{||w||} = \frac{1}{||w||}$
Then the margin can be expressed through (rescaled) w and b as :
$\rho = 2r = \frac{2}{||w||}$


x_2
$\mathbf{w}\cdot\mathbf{x}+b=1$
$\frac{2}{||\mathbf{w}||}$
$\mathbf{w}\cdot\mathbf{x}+b=0$
$\mathbf{w}\cdot\mathbf{x}+b=-1$
$\mathbf{w}$
x_1
$\frac{b}{||\mathbf{w}||}$

$$\underset{\mathbf{w}}{\text{Minimize}}\ \frac{1}{2}||\mathbf{w}||^2$$
$$\text{subject to}\ y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1$$

$$\mathbf{Y}\cdot(\mathbf{X}\mathbf{w}+b) \geq \mathbf{1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \cdot \left(\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} + b \right) \geq \mathbf{1}^n$$

After that , we can formulate the problem:
Find w and b such that $\rho = \frac{2}{||w||}$ is maximized and for all (xi, yi), i=1..n :
$y_i(wTx_i + b) \geq 1$

$\phi(x) = ||w|| = wTw$

so we can rewrite above equation as : $\rho = \frac{2}{||w||}\ maximized$
$means\ that\ we\ minimise\ ||w||$

minimise : $\phi(x) = ||w|| = wTw$
$subject\ to : y_i(wTx_i + b) \geq 1$

NOTE: Want to look for solution point p where
$\nabla f(p) = \nabla \lambda g(p)$
$g(x) = 0$

Combining these two as the Langrangian L requiring derivative of L be zero:
$L(x, a) = f(x) - ag(x)$
$\nabla(x, a) = 0$ Partial derivatives wrt x recover the parallel normal constraint
Partial derivatives wrt recover the g(x,y)=0
$L(x, a) = f(x) + a_i \sum_i \lambda g_i(x)$

Now convert the Our SVM equation in Langrangian form because coputation become easier.
$f(x) : \frac{1}{2}||w||^2$
$g(x) : y_i(w \cdot x_i + b)–1 = 0$

so Lagrangian is:

$minJ(w,b,a) = \frac{1}{2}w^Tw - \sum a_i[y_i(w \cdot x_i + b) – 1]wrt\ w\ ,\ b$

expand the last to get the following

$min\ J(w,b,a) = \frac{1}{2}w^Tw - \sum a_iy_i(w \cdot x_i + b) + \sum_i a_i$

$s.t.\ \forall_i, a_i \geq 0$

From the property that the derivatives at min = 0

$(a)\ \frac{\delta J}{\delta w} = 0$

$(b)\ \frac{\delta J}{\delta b} = 0$

$from\ this\ sloutions\ we\ get$

(1) $w^* = \sum_{i=1} \alpha_i y_i x_i$

(2) $\sum_i \alpha_i y_i = 0$

(3) $\alpha_i[d_i(w^Tx_i + b_0) - 1]$

By substituting for w and b back in the original equation we can get rid of the dependence on w and b. i.e. $Q(\alpha) = \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{i=1}\sum j = 1\alpha_i\alpha_j y_i y_j x_i^T x_j$

$subject\ to\ :\ \alpha_i \geq \forall_i\ and\ \sum_{i=1} \alpha_i d_i = 0$

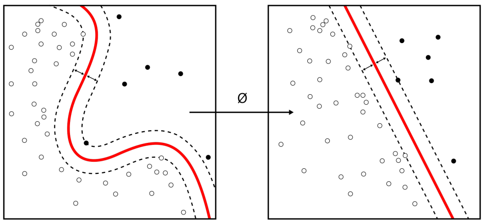
After solving .....

$b_0^* = 1 - w_0^T \cdot x_s +$

5 Non-Linear Classification Using SVMs

Non-linear classification can be done using kernel trick. In this algorithm, in place of the dot products, a nonlinear kernel function is used. Using this variation, a transformed feature space is used to fit the hyperplane.

Using a tranformed feature space results in a higher value of generalization error, however this error diminishes on providing a larger number of samples.



(a) Kernel Machine

References

[1] Images, Link: https://en.wikipedia.org/wiki/Support-vector_machine

[2] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72