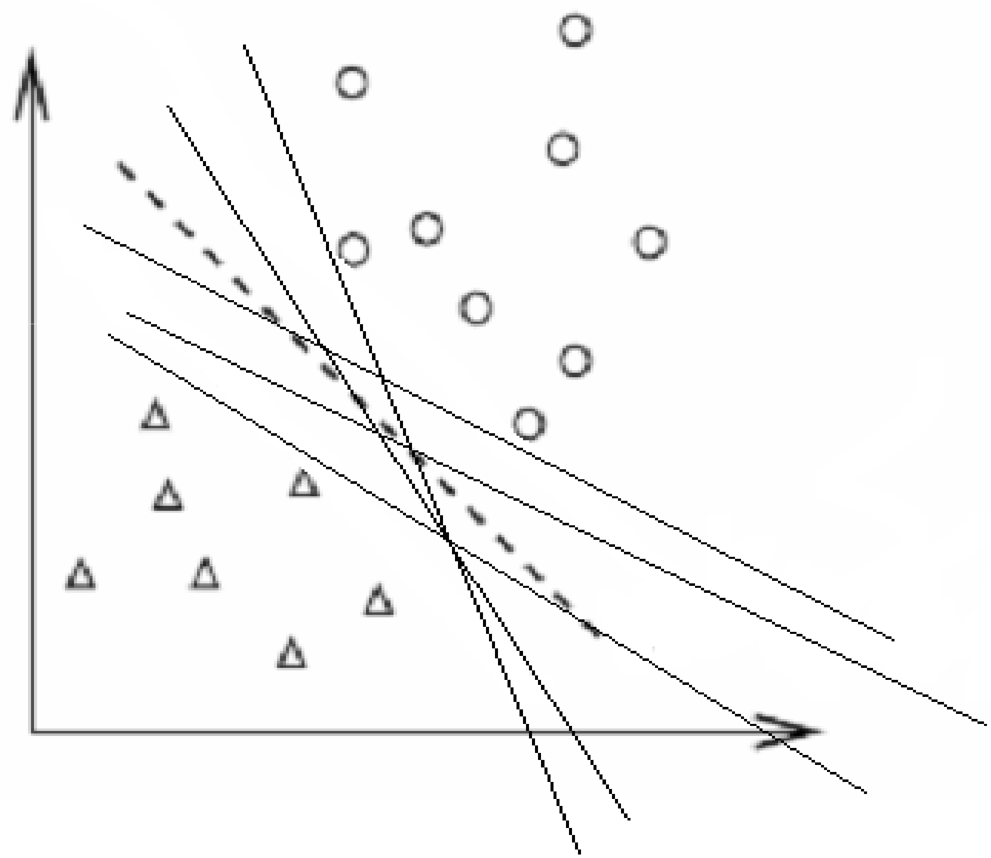
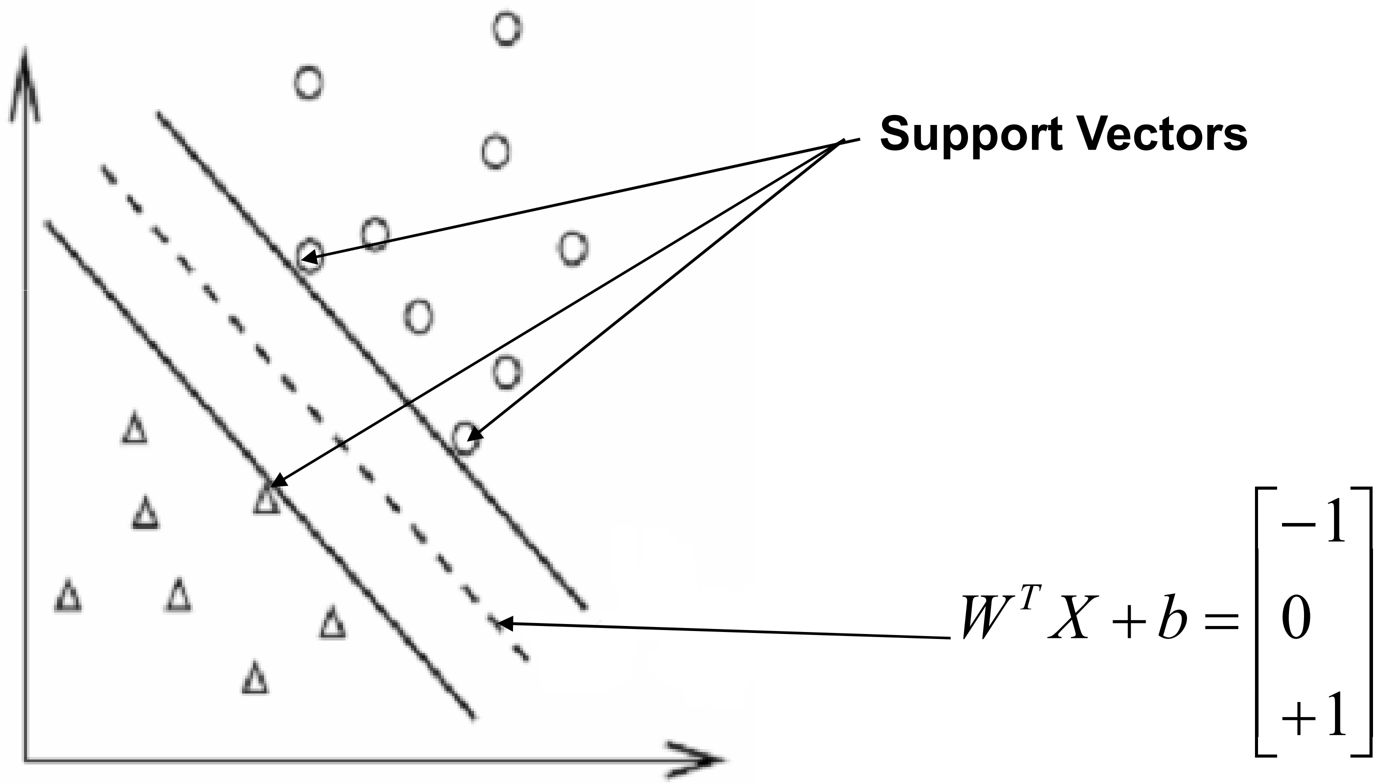
Basic Concept of SVM:



- Which line will classify the unseen data well?
- The dotted line! Its line with Maximum Margin!

Cont…



$$W^T X + b = \begin{bmatrix} -1 \\ 0 \\ +1 \end{bmatrix}$$

Some definitions:

- Functional Margin:

  w.r.t.

1) individual examples : $\hat{\gamma}^{(i)} = y^{(i)}(W^T x^{(i)} + b)$

2)example set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots\ldots, m\}$

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}$$

- Geometric Margin:

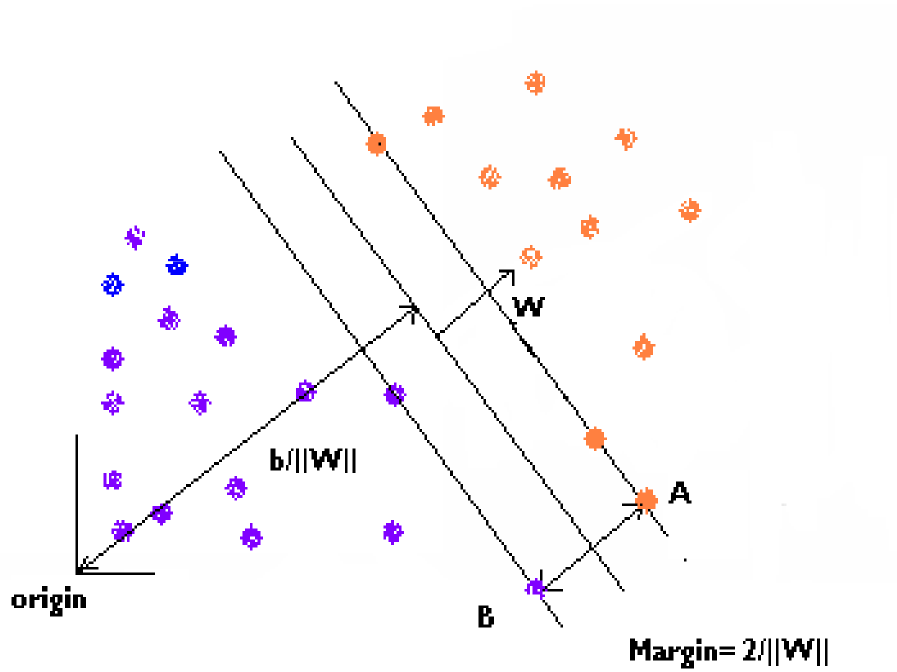
  w.r.t

  1)Individual examples: $\gamma^{(i)} = y^{(i)}\left(\left(\frac{W}{\|W\|}\right)^T x^{(i)} + \frac{b}{\|W\|}\right)$

  2) example set S,

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)}$$

Problem Formulation:



$$W^T X + b = \begin{bmatrix} -1 \\ 0 \\ +1 \end{bmatrix}$$

Cont..

- Distance of a point (u, v) from Ax+By+C=0, is given by |Ax+By+C|/||n||

Where ||n|| is norm of vector n(A,B)

- Distance of hyperpalne from origin = $\frac{b}{\| W \|}$
- Distance of point A from origin = $\frac{b+1}{\| W \|}$
- Distance of point B from Origin = $\frac{b-1}{\| W \|}$
- Distance between points A and B (Margin) = $\frac{2}{\| W \|}$

Cont…

We have data set $\{X^{(i)}, Y^{(i)}\}, i = 1, \ldots, m$

$$X \in R^d \quad and \quad Y \in R^1$$

separating hyperplane

$$W^T X + b = 0$$

$$s.t.$$

$$W^T X^{(i)} + b > 0 \quad if \quad Y^{(i)} = +1$$

$$W^T X^{(i)} + b < 0 \quad if \quad Y^{(i)} = -1$$

Cont…

- Suppose training data satisfy following constrains also,

$$W^T X^{(i)} + b \geq +1 \quad for \quad Y^{(i)} = +1$$

$$W^T X^{(i)} + b \leq -1 \quad for \quad Y^{(i)} = -1$$

Combining these to the one,

$$Y^{(i)}(W^T X^{(i)} + b) \geq 1 \quad for \quad \forall i$$

- Our objective is to find Hyperplane(W,b) with maximal separation between it and closest data points while satisfying the above constrains

THE PROBLEM:

$$\max_{W,b} \frac{2}{||W||}$$

such that

$$Y^{(i)}(W^T X^{(i)} + b) \geq 1 \quad for \quad \forall i$$

Also we know

$$||W|| = \sqrt{W^T W}$$

Cont..

So the Problem can be written as:

$$\min_{W,b} \quad \frac{1}{2} W^T W$$

Such that

$$Y^{(i)}(W^T X^{(i)} + b) \geq 1 \quad for \;\; \forall i$$

Notice: $W^T W = \| W \|^2$

It is just a convex quadratic optimization problem !

DUAL

- Solving dual for our problem will lead us to apply SVM for nonlinearly separable data, efficiently
- It can be shown that

$$\min \quad primal \quad = \quad \max_{\alpha \geq 0}(\min_{W,b} L(W,b,\alpha))$$

- Primal problem:

$$\min_{W,b} \quad \frac{1}{2} W^T W$$

Such that

$$Y^{(i)}(W^T X^{(i)} + b) \geq 1 \quad for \quad \forall i$$

Constructing Lagrangian

- Lagrangian for our problem:

$$L(W,b,\alpha) = \frac{1}{2} \| W \|^2 - \sum_{i=1}^{m} \alpha_i \left[Y^{(i)} (W^T X^{(i)} + b) - 1 \right]$$

Where α a Lagrange multiplier and $\alpha_i \geq 0$

- Now minimizing it w.r.t. W and b:

We set derivatives of Lagrangian w.r.t. W and b to zero

Cont…

- Setting derivative w.r.t. W to zero, it gives:

$$W - \sum_{i=1}^{m} \alpha_i Y^{(i)} X^{(i)} = 0$$

i.e.

$$W = \sum_{i=1}^{m} \alpha_i Y^{(i)} X^{(i)}$$

- Setting derivative w.r.t. b to zero, it gives:

$$\sum_{i=1}^{m} \alpha_i Y^{(i)} = 0$$

Cont…

- Plugging these results into Lagrangian gives

$$L(W,b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}Y^{(i)}Y^{(j)}\alpha_i\alpha_j(X^{(i)})^T(X^{(j)})$$

- Say it

$$D(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}Y^{(i)}Y^{(j)}\alpha_i\alpha_j(X^{(i)})^T(X^{(j)})$$

- This is result of our minimization w.r.t W and b,

So The DUAL:

- Now Dual becomes::

$$\max_{\alpha} \quad D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} Y^{(i)} Y^{(j)} \alpha_i \alpha_j \left\langle X^{(i)}, X^{(j)} \right\rangle$$

$$s.t.$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i Y^{(i)} = 0$$

- Solving this optimization problem gives us α_i
- Also Karush-Kuhn-Tucker (KKT) condition is satisfied at this solution i.e.

$$\alpha_i \left[Y^{(i)} (W^T X^{(i)} + b) - 1 \right] = 0, \quad for \quad i = 1, \ldots, m$$

Values of W and b:

- W can be found using

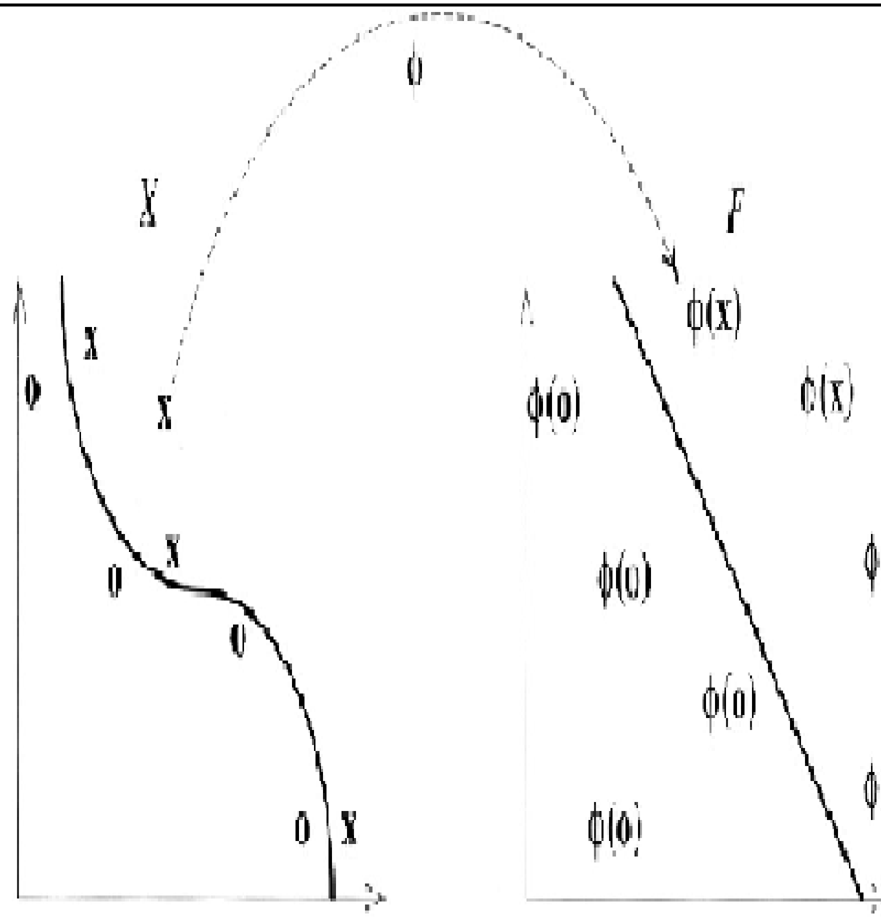
$$W = \sum_{i=1}^{m} \alpha_i Y^{(i)} X^{(i)}$$

- b can be found using:

$$b^* = -\frac{\max_{i:Y^{(i)}=-1} W^{*T} X^{(i)} + \min_{i:Y^{(i)}=1} W^{*T} X^{(i)}}{2}$$

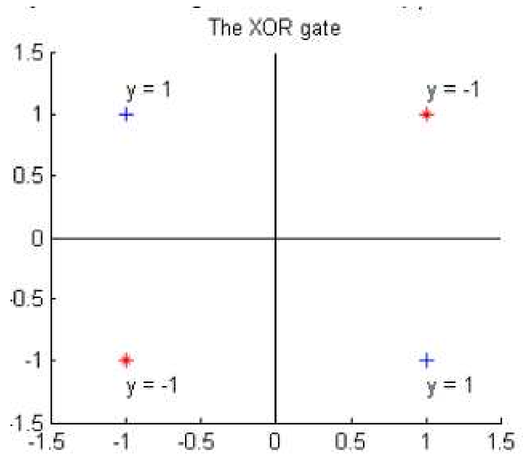
What if data is nonlinearly separable?

- The maximal margin hyperplane can classify only linearly separable data
- What if the data is linearly non-separable?
- Take your data to linearly separable (higher dimensional space) and use maximal margin hyperplane there!

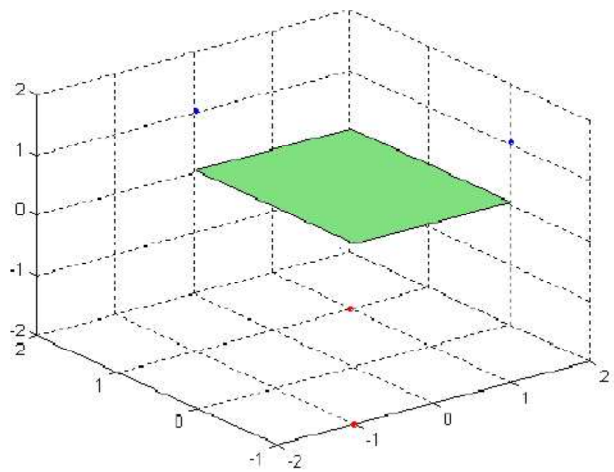


Taking it to higher dimension works! Ex. XOR

x=(x1,x2)	y
(1,1)	-1
(-1,-1)	-1
(1,-1)	1
(-1,1)	1



x=(x1,x2,x1.x2)	y
(1,1,1)	-1
(-1,-1,1)	-1
(1,-1,-1)	1
(-1,1,-1)	1



Doing it in higher dimensional space

- Let $\Phi : X \rightarrow F$ be non linear mapping from input space X (original space) to feature space (higher dimensional) F
- Then our inner (dot) product $\langle X^{(i)}, X^{(j)} \rangle$ in higher dimensional space is $\langle \phi(X^{(i)}), \phi(X^{(j)}) \rangle$
- Now, the problem becomes:

$$\max_{\alpha} \; D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} Y^{(i)} Y^{(j)} \alpha_i \alpha_j \langle \phi(X^{(i)}), \phi(X^{(j)}) \rangle$$

$$s.t.$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i Y^{(i)} = 0$$

Kernel function:

- There exist a way to compute inner product in feature space as function of original input points – Its kernel function!
- Kernel function:

$$K(x,z) = \langle \phi(x), \phi(z) \rangle$$

- We need not know ϕ to compute $K(x,z)$

An example:

$let \quad x, z \in R^n$

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

$$i.e. \quad K(x,z) = (\sum_{i=1}^{n} x_i z_i)(\sum_{j=1}^{n} x_j z_j)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j z_i z_j$$

$$= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j)$$

For n=3, feature mapping ϕ is given as :

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}$$

$$K(x,z) = \langle \phi(x), \phi(z) \rangle$$

example cont…

- Here,

$$for$$

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

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad z = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$x^T z = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$= 11$$

$$K(x,z) = (x^T z)^2 = 121$$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_2 x_1 \\ x_2 x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \end{bmatrix}$$

$$\phi(z) = \begin{bmatrix} 9 \\ 12 \\ 12 \\ 16 \end{bmatrix}$$

$$\phi(x)^T \phi(z) = \begin{bmatrix} 1 & 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} 9 \\ 12 \\ 12 \\ 16 \end{bmatrix}$$

$$= 121$$

So our SVM for the non-linearly separable data:

- Optimization problem:

$$\max_{\alpha} \; D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} Y^{(i)} Y^{(j)} \alpha_i \alpha_j K \left\langle X^{(i)}, X^{(j)} \right\rangle$$

$$s.t.$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i Y^{(i)} = 0$$

- Decision function

$$F(X) = Sign\left(\sum_{i=1}^{m} \alpha_i Y^{(i)} K(X^{(i)}, X) + b\right)$$

Some commonly used Kernel functions:

- Linear: $K(X,Y) = X^T Y$
- Polynomial of degree d: $K(X,Y) = (X^T Y + 1)^d$
- Gaussian Radial Basis Function (RBF): $K(X,Y) = e^{-\frac{\|X-Y\|^2}{2\sigma^2}}$
- Tanh kernel: $K(X,Y) = \tanh(\rho(X^T Y) - \delta)$

Implementations:

Some Ready to use available SVM implementations:

1)LIBSVM:A library for SVM by Chih-Chung Chang and chih-Jen Lin

(at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

2)SVM light : An implementation in C by Thorsten Joachims

(at: http://svmlight.joachims.org/)

3)Weka: A Data Mining Software in Java by University of Waikato

(at: http://www.cs.waikato.ac.nz/ml/weka/)

Issues:

- Selecting suitable kernel: Its most of the time trial and error
- Multiclass classification: One decision function for each class(*l*1 vs *l*-1) and then finding one with max value i.e. if X belongs to class 1, then for this and other (*l*-1) classes vales of decision functions:

$$
\begin{aligned}
&F_1(X) \geq +1 \\
&F_2(X) \leq -1 \\
&. \\
&. \\
&F_l(X) \leq -1
\end{aligned}
$$

Cont….

- Sensitive to noise: Mislabeled data can badly affect the performance
- Good performance for the applications like-

  1)computational biology and medical applications (protein, cancer classification problems)

  2)Image classification

  3)hand-written character recognition

  And many others…..
- Use SVM :High dimensional, linearly separable data (strength), for nonlinearly depends on choice of kernel

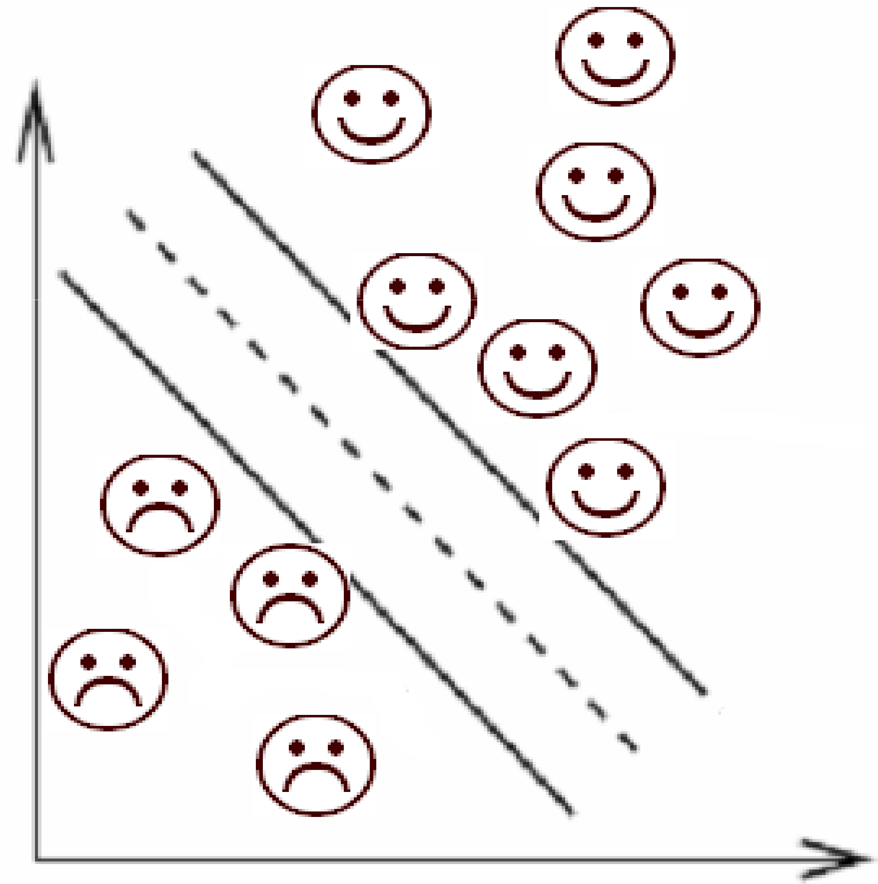
Conclusion:

Support Vector Machines provides very simple method for linear classification. But performance, in case of nonlinearly separable data, largely depends on the choice of kernel!

References:

- Nello Cristianini and John Shawe-Taylor (2000)??
  An Introduction to Support Vector Machines and Other Kernel-based Learning Methods
  Cambridge University Press
- Christopher J.C. Burges (1998)??
  A tutorial on Support Vector Machines for pattern recognition
  Usama Fayyad, editor, *Data Mining and Knowledge Discovery,* 2, 121-167.
  Kluwer Academic Publishers, Boston.
- Andrew Ng (2007)
  CSS229 Lecture Notes
  Stanford Engineering Everywhere, Stanford University .
- *Support Vector Machines* <http://www.svms.org > (Accessed 10.11.2008)
- *Wikipedia*
- *Kernel-Machines.org*<http://www.kernel-machines.org >(Accessed 10.11.2008)

Thank You!



prakash@cdacmumbai.in ;
pbpimpale@gmail.com