MACHINE LEARNING

Support Vector Machine
For Classification

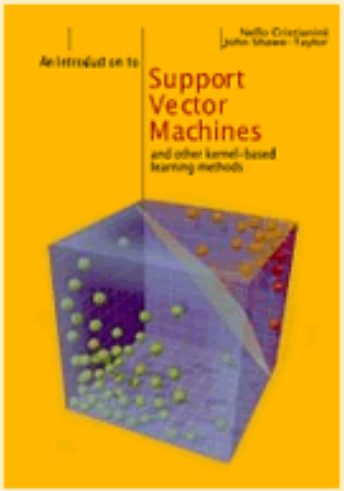
Part 1 – Linear SVM

Support Vector Machine (SVM)

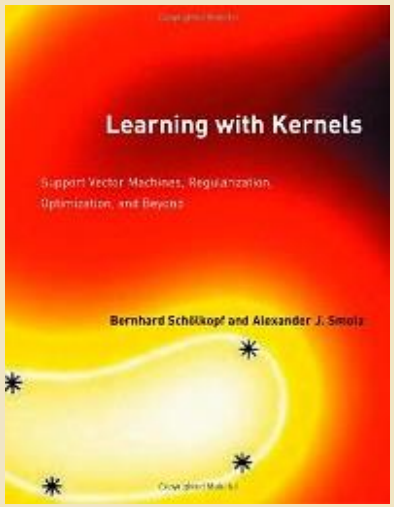
Brief history:

- SVM is traced back to the work by Vapnik and Chervonekis on statistical learning theory (Vapnik1979) and the notion of *VC dimension*.
- The current form of SVM was presented in (Boser, Guyon and Vapnik 1992) and Cortes and Vapnik (1995).

Textbooks:



A good survey of the theory behind SVM is given in *Support Vector Machines and other Kernel Based Learning methods* by Nello Cristianini and John-Shawe Taylor.



An easy introduction to SVM is given in *Learning with Kernels* by Bernhard Scholkopf and Alexander Smola.

Support Vector Machine (SVM)

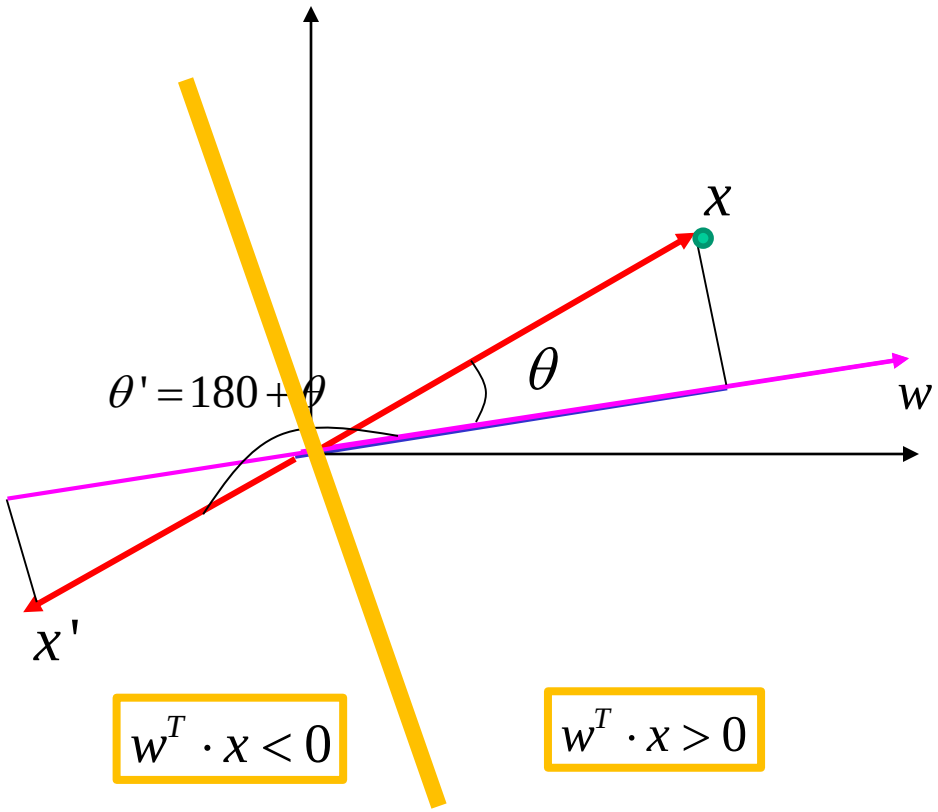
SVM was applied to numerous classification problems:

- Computer vision (face detection, object recognition, feature categorization)
- Bioinformatics (categorization of gene expression, of microarray data)
- WWW (categorization of websites)
- Production (control of quality, detection of defaults)
- Robotics (categorization of sensor readings)
- Finance (bankruptcy prediction)

The success of SVM is mainly due to:

- SVM depends on convex optimization.
- Its ease of use (lots of software available, good documentation).
- Excellent performance on variety of datasets.
- Good solvers making optimization (learning phase) very quick.
- Very fast at retrieval time – does not hinder practical applications.

Recap - Constructing a projection



Datapoint x

Projection vector w

The norm of the projection of x onto w is:

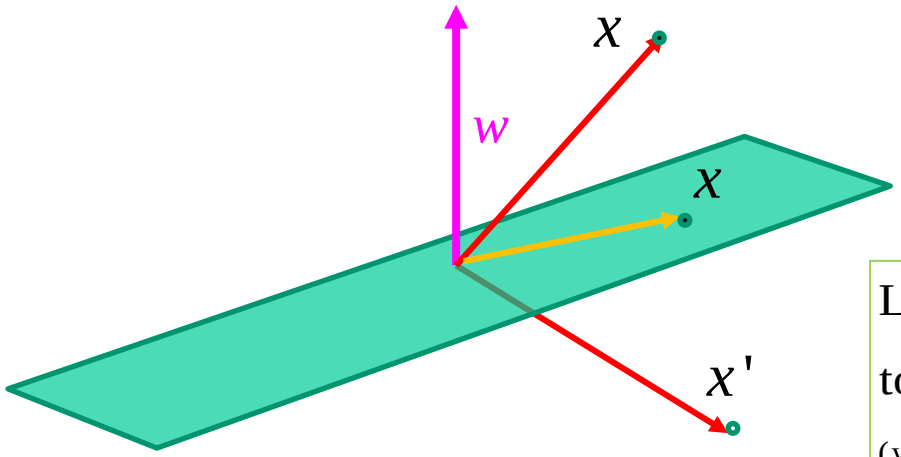
$$w^T \cdot x = \|w\| \|x\| \cos(\theta)$$

$$\cos(\theta) > 0 \quad \Rightarrow w^T \cdot x > 0$$

$$w^T \cdot x' = \|w\| \|x'\| \cos(\theta')$$

$$\cos(\theta') < 0 \quad \Rightarrow w^T \cdot x' < 0$$

Recap - Constructing a projection



Datapoint x

Normal to plane w

Looking at the sign of $\left(w^T x\right)$ allows to separate points on either side of the plane.

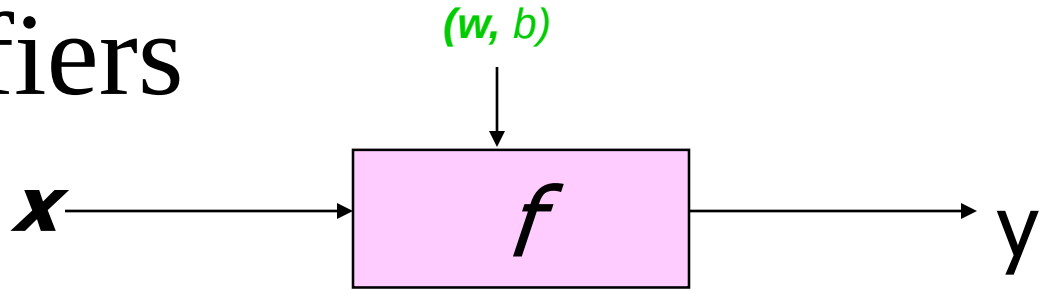
(we ignored the intercept and assumed the plane passed by the origin)

$w^T \cdot x > 0 \quad \Rightarrow$ point lies on the left handside of plane

$w^T \cdot x' < 0 \quad \Rightarrow$ point lies on the right handside of plane

$w^T \cdot x = 0$? $\quad \Rightarrow$ point on the plane

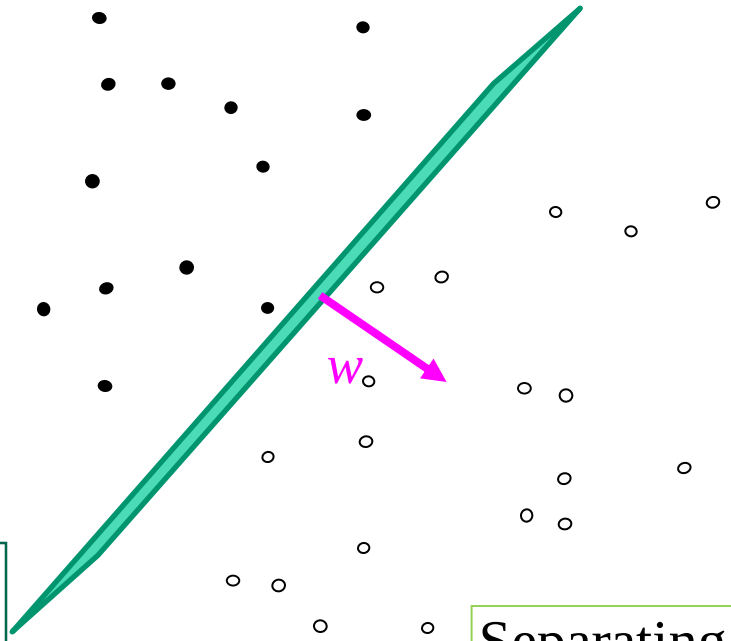
Linear Classifiers



Class label y={-1;1}

- denotes -1
- denotes +1

$$y = f(x; w, b) = \text{sgn}\left(w^T x + b\right)$$



Separating Hyperplane

Separating hyperplane is defined by:

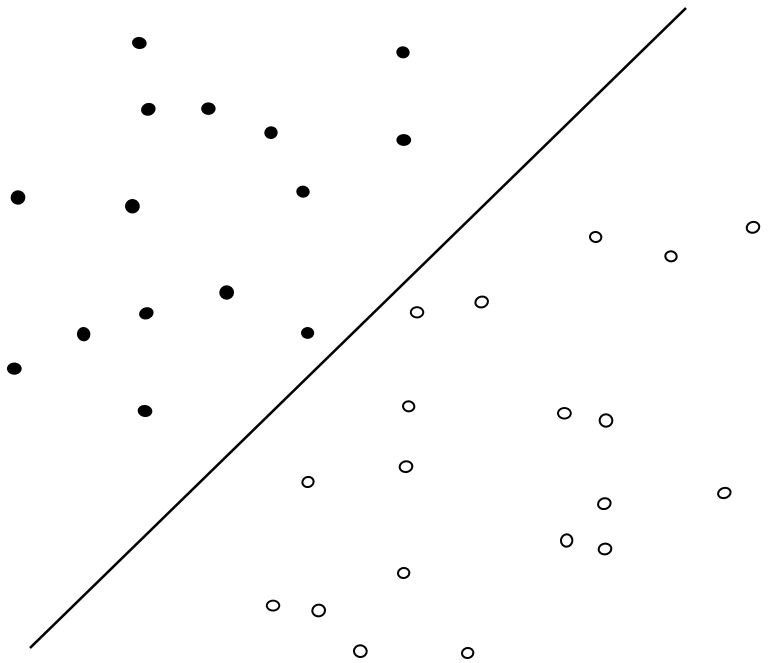
w: the normal to the plane

b: the intercept

Linear Classifiers

Class label y={-1;1}

- denotes -1
- denotes +1

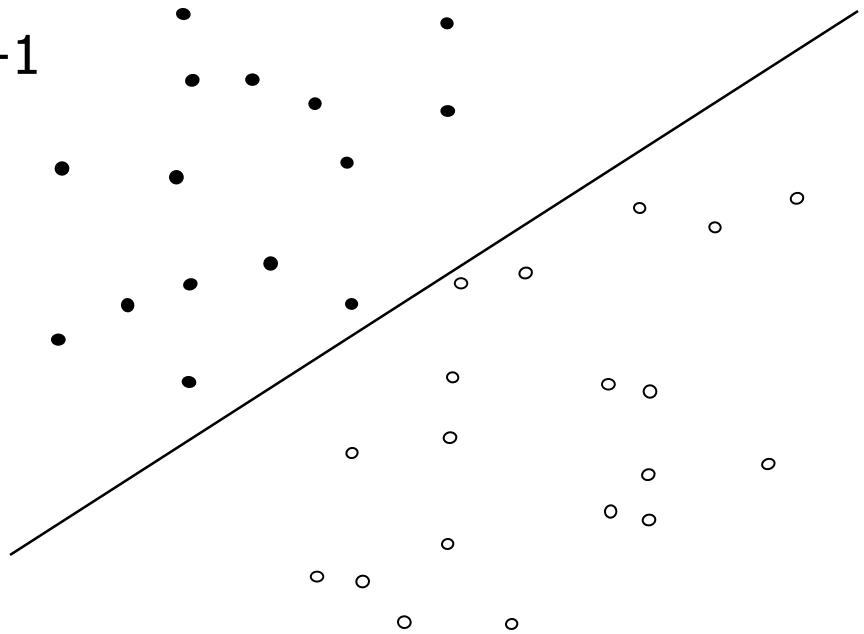


How would you classify this data?

Linear Classifiers

Class label y={-1;1}

- denotes -1
- denotes +1

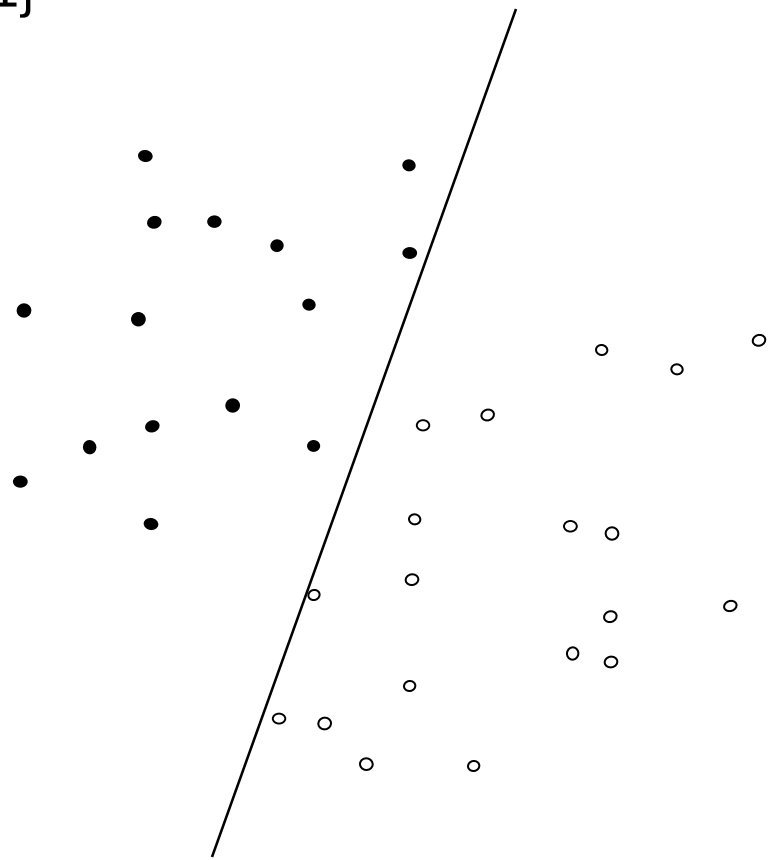


How would you classify this data?

Linear Classifiers

Class label y={-1;1}

- denotes -1
- denotes +1

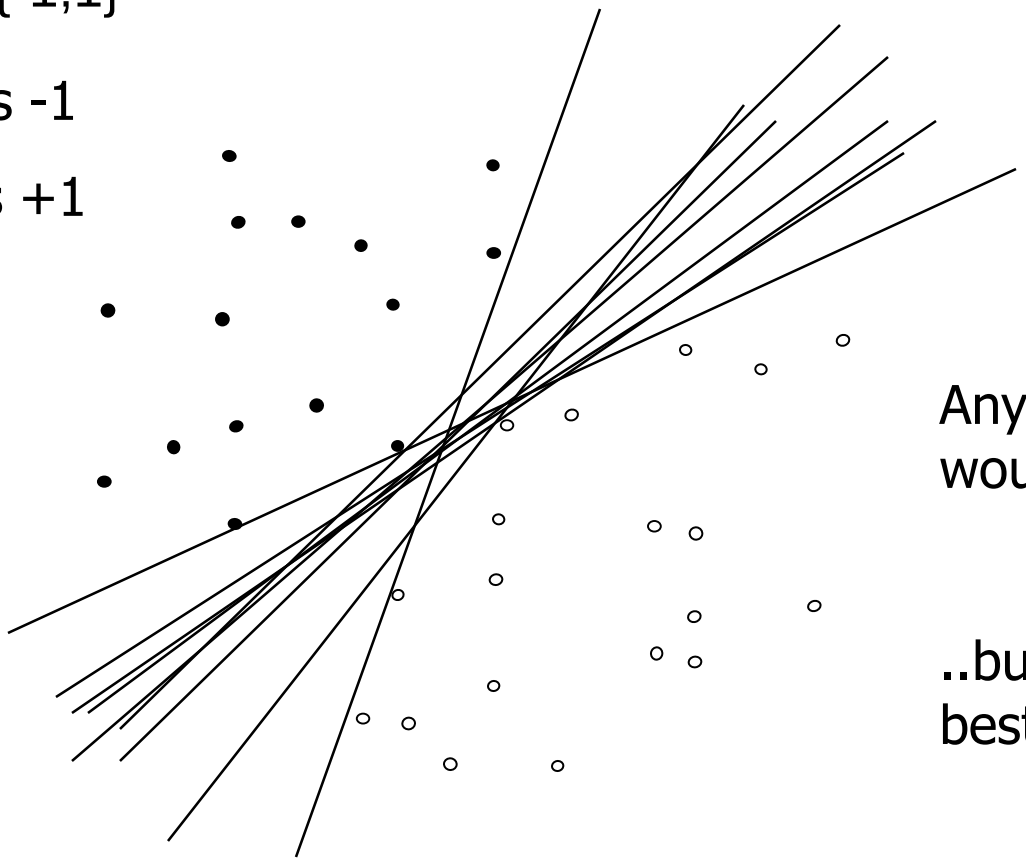


How would you classify this data?

Linear Classifiers

Class label y={-1;1}

- denotes -1
- denotes +1



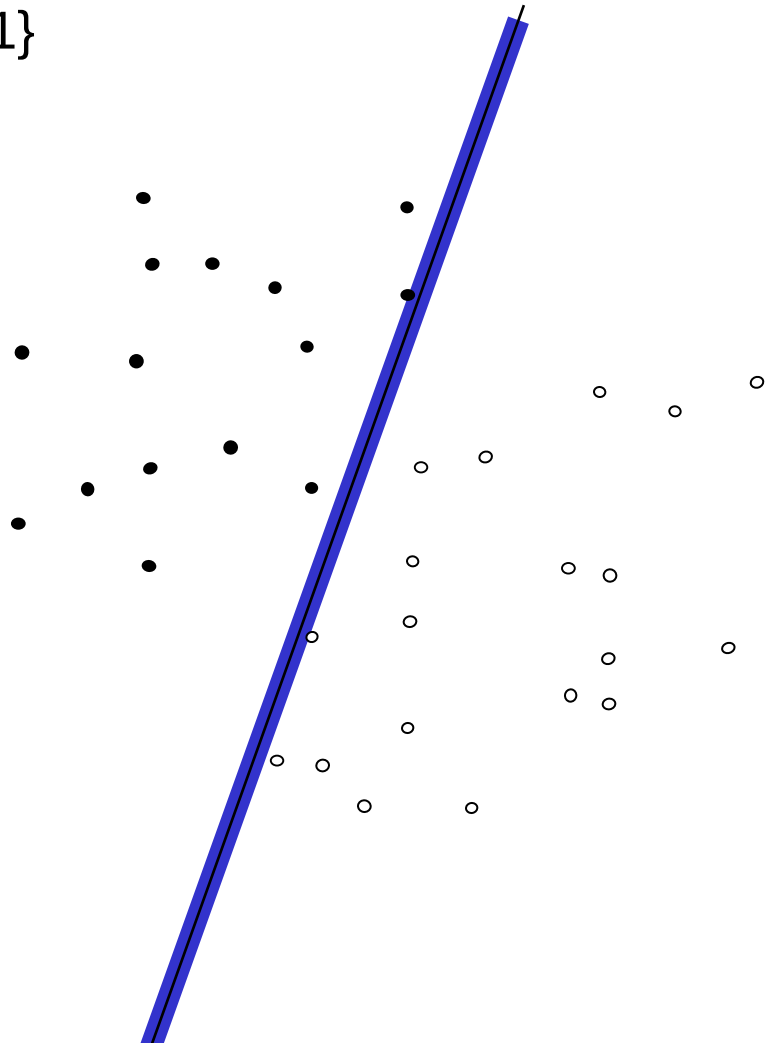
Any of these would be fine..

..but which is best?

Classifier Margin

Class label y={-1;1}

- denotes -1
- denotes +1

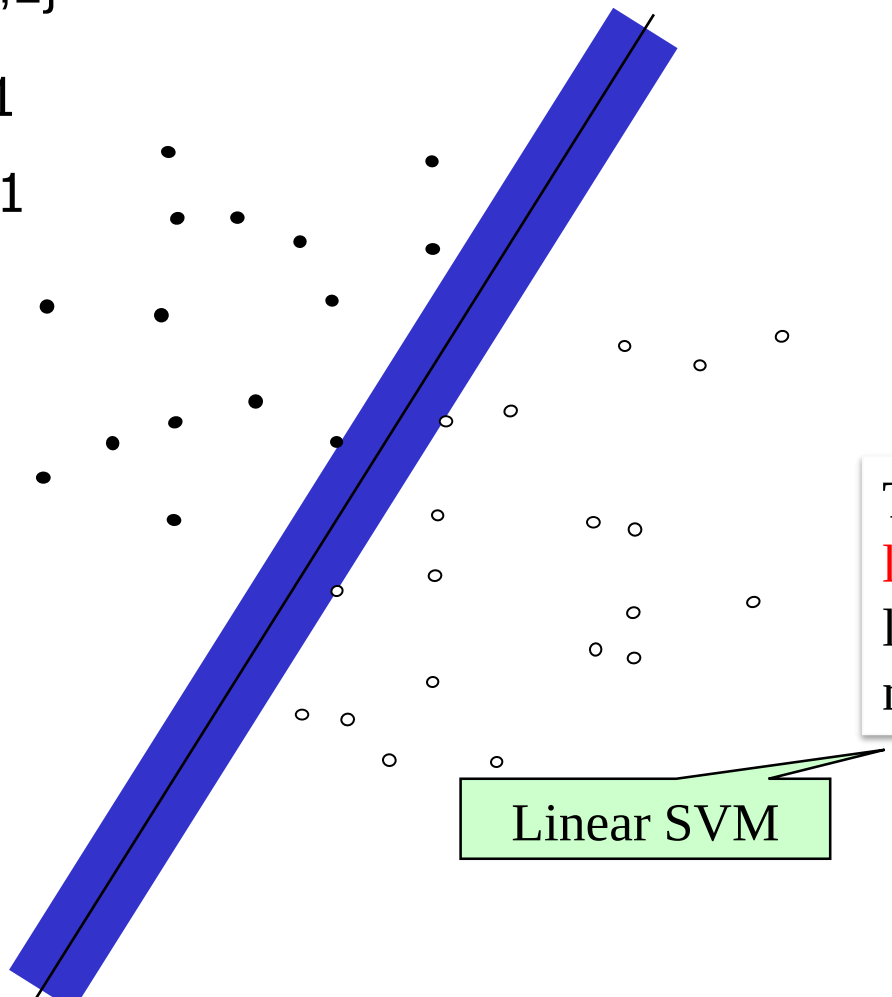


Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Classifier Margin

Class label y={-1;1}

- denotes -1
- denotes +1



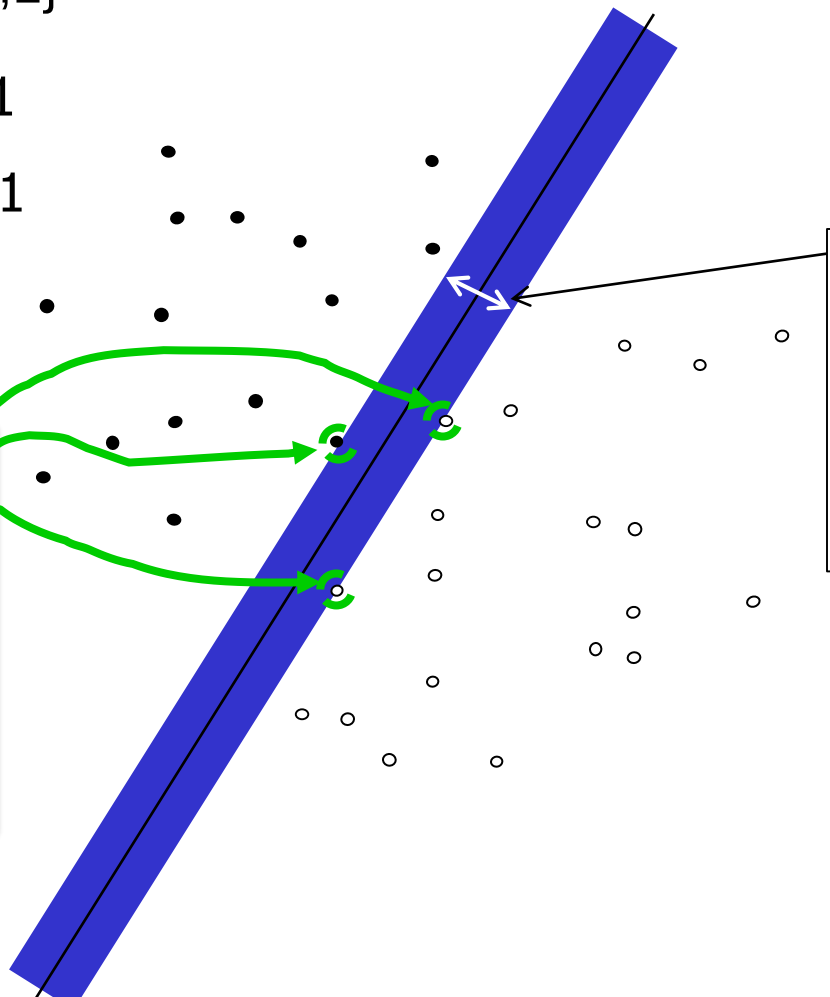
Linear SVM

The maximum margin linear classifier is the linear classifier with the maximum margin.

Classifier Margin

Class label y={-1;1}

- denotes -1
- denotes +1



Need to determine a measure of the width of the margin, so as to maximize for this measure.

Support Vectors are those datapoints that are closest to the boundary. They define the margin.

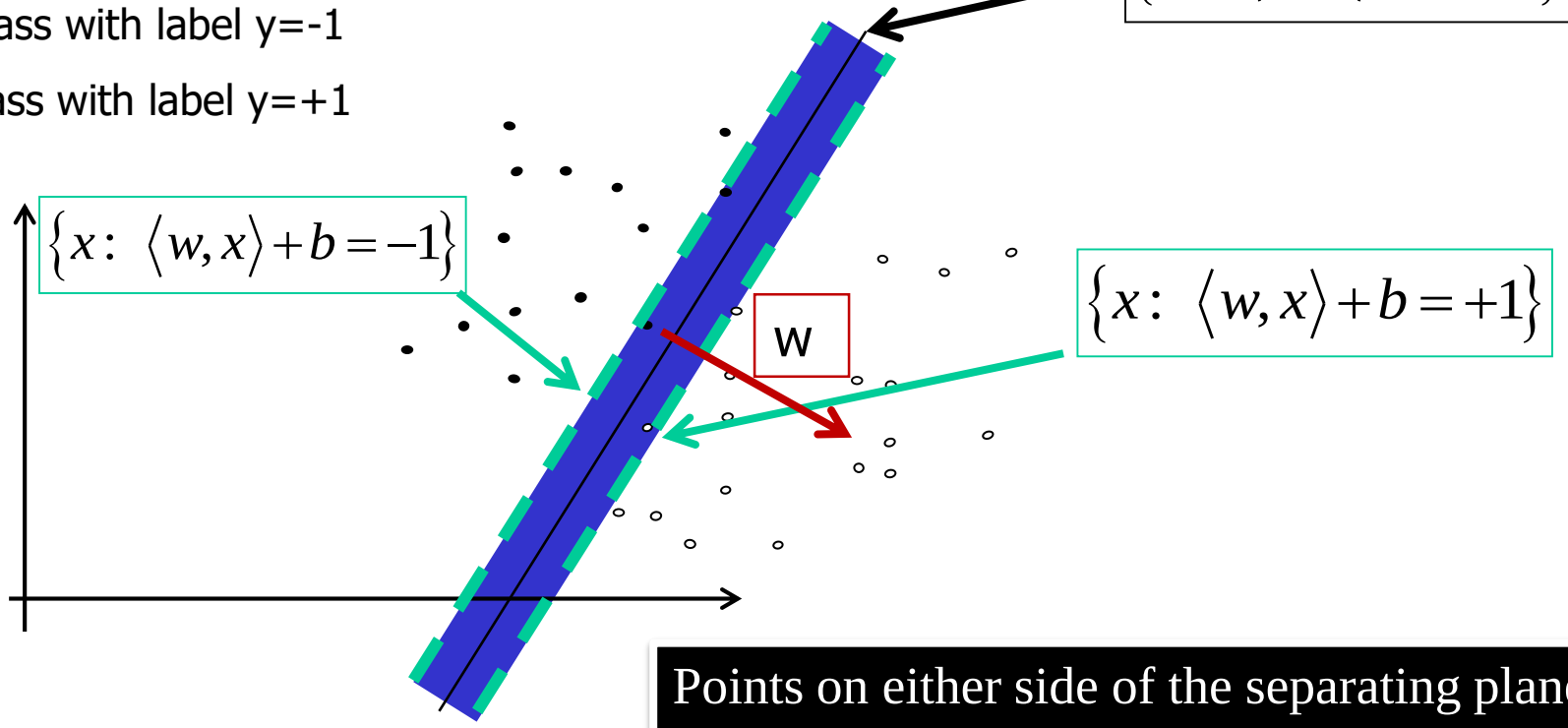
Computing the Distance to the Separating Hyperplane

Decision function: $y = \mathrm{sgn}\left(\langle w, x\rangle + b\right)$

$\langle w, x\rangle = w^T x$

$\{x : \langle w, x\rangle + b = 0\}$

- Class with label y=-1
- Class with label y=+1

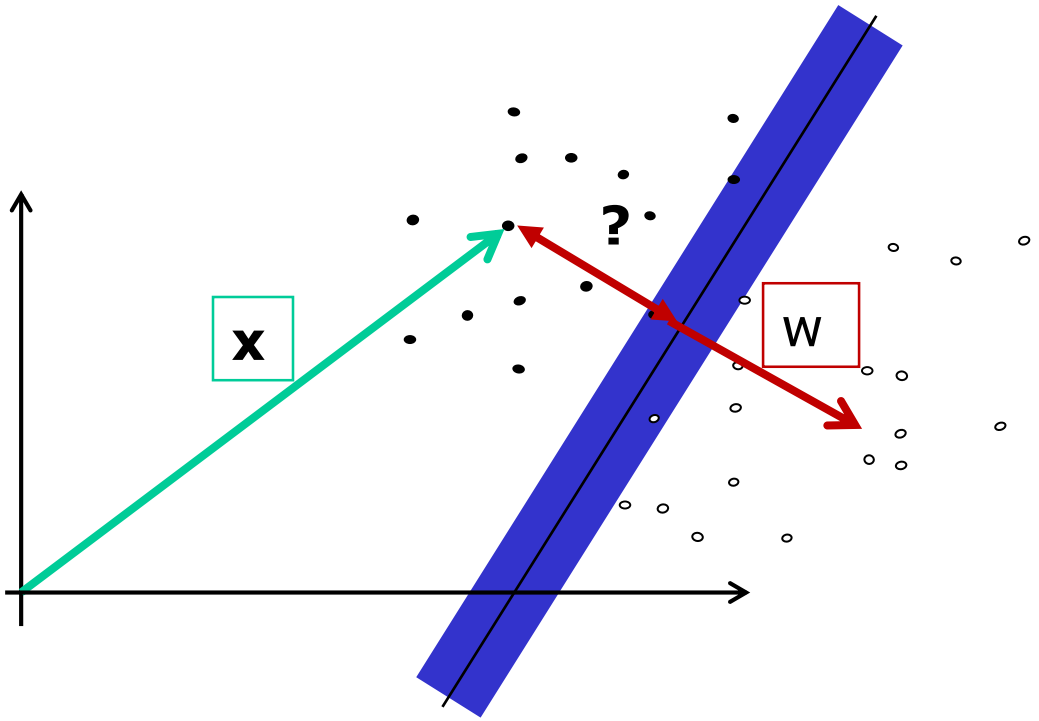


Points on either side of the separating plane have negative and positive coordinates, respectively.

Definition:

The margins on either side of the hyperplane satisfy $\langle w, x\rangle + b = \pm 1.$

Computing the Distance to the Separating Hyperplane



What is the distance from a point **x** to the separating plane $\langle \mathbf{w},\mathbf{x} \rangle + b = 0$?

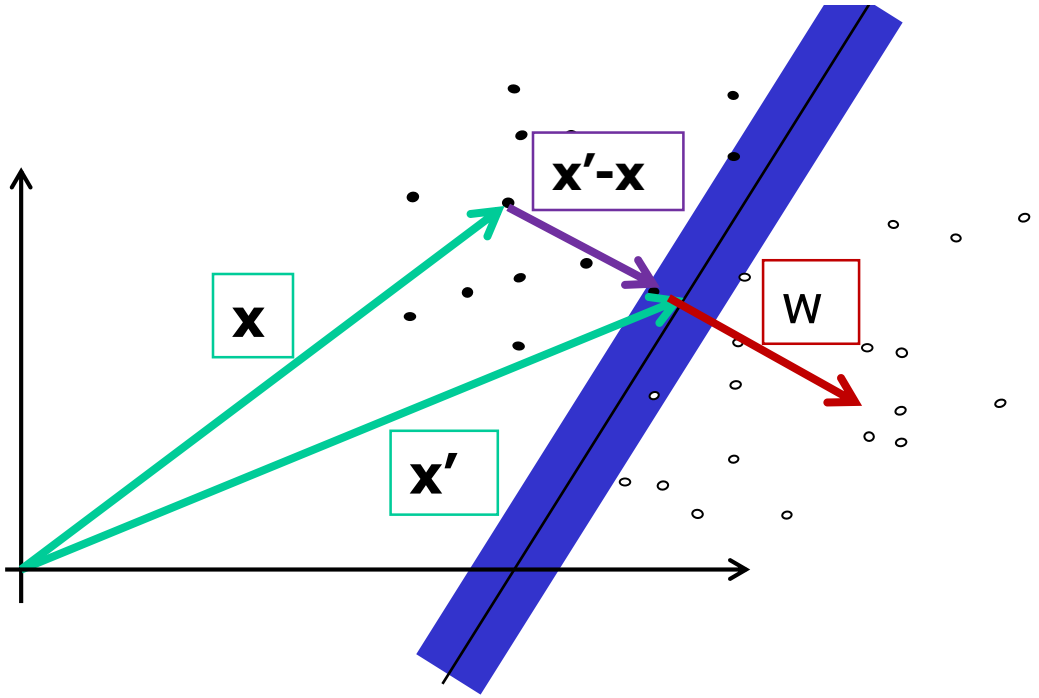
Computing the Distance to the Separating Hyperplane

Projection of $(x'-x)$ onto w is then:

$$\frac{\langle w,(x'-x)\rangle}{\|w\|}\quad\frac{w}{\|w\|}$$

unitary vector

$$\langle w, x'-x\rangle = \langle w, x'\rangle - \langle w, x\rangle$$



We know that:

$$x' \ s.t. \ \langle w, x'\rangle + b = 0$$

$$\Rightarrow \langle w, x'\rangle = -b$$

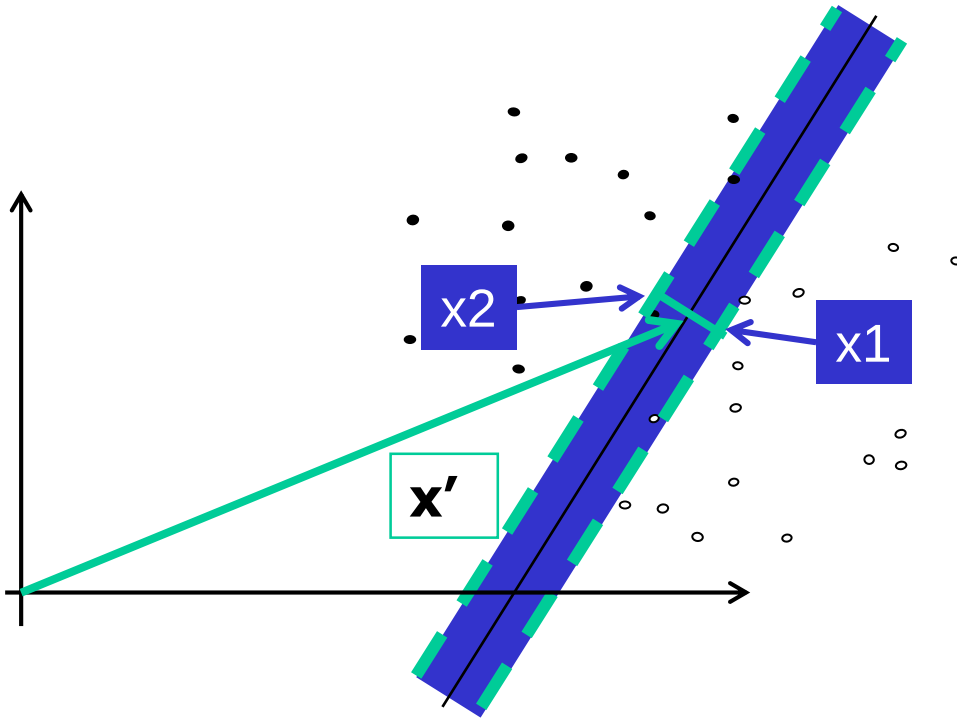
Projection of $x - x'$ onto w is then:

$$\frac{-b-\langle w, x\rangle}{\|w\|}\quad\frac{w}{\|w\|}$$

unitary vector

$$\text{Distance to plane} = \frac{\left|\langle w, x\rangle + b\right|}{\|w\|}$$

Computing the margin



Distance of each points on either side of the margin:

$$\left\|x^1 - x'\right\| = \frac{\left|\left\langle w, x^1 \right\rangle + b\right|}{\|w\|} = \frac{1}{\|w\|}$$

$$\left\|x^2 - x'\right\| = \frac{\left|\left\langle w, x^2 \right\rangle + b\right|}{\|w\|} = \frac{1}{\|w\|}$$

$$\Rightarrow \left\|x^1 - x^2\right\| = \frac{2}{\|w\|}$$

The margin between two classes is at least 2/||w||.

Our objective function

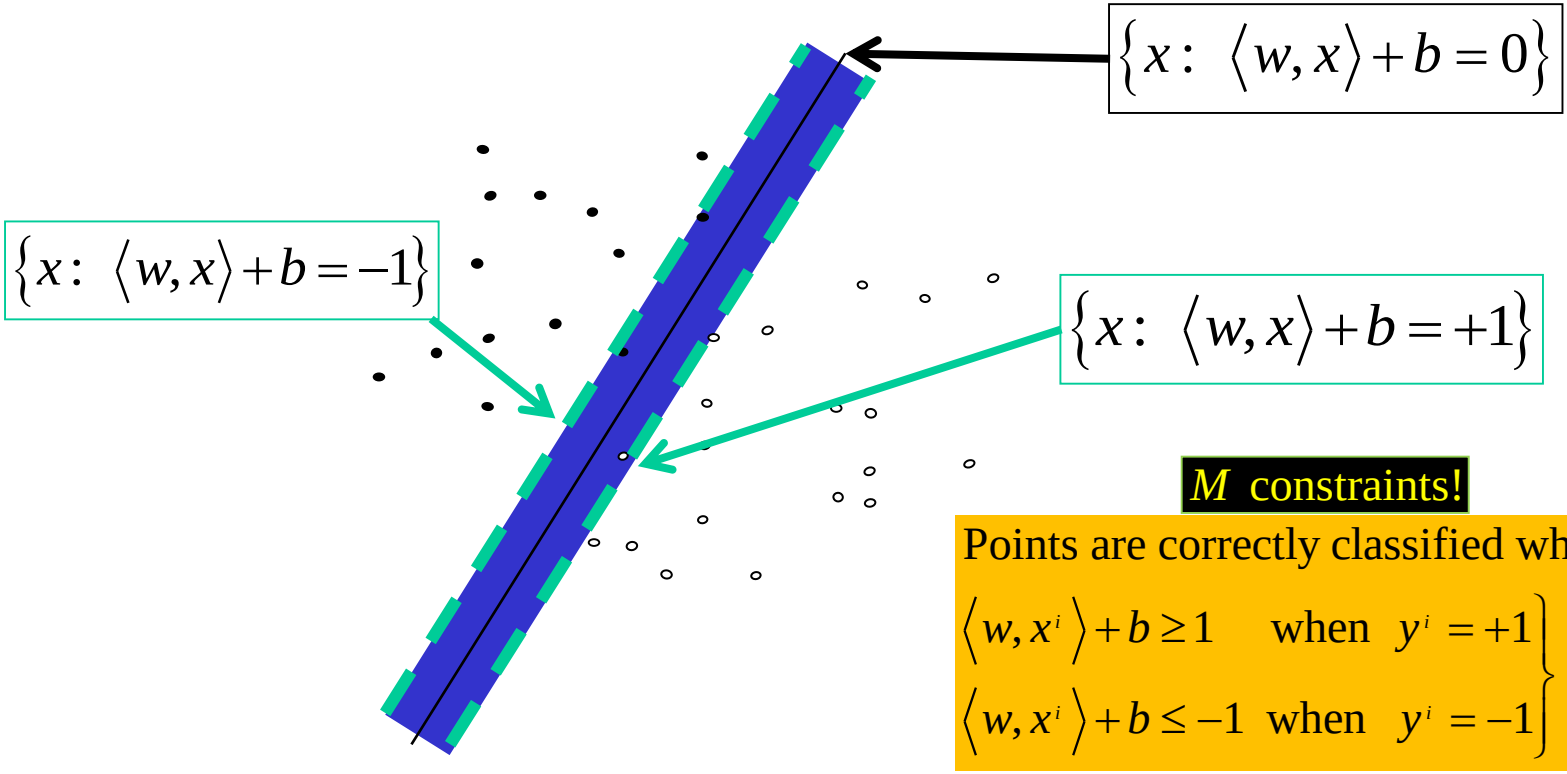
Separating condition is measured by $\frac{2}{\|w\|}$.

To maximize this condition is equivalent to minimizing $\frac{\|w\|}{2}$.

Better even is to minimize the convex form $\frac{\|w\|^2}{2}$.

How do we make sure the points sit on the correct side of the separating plane?

Determining the constraints



M constraints!

Points are correctly classified when:

$$\left.\begin{array}{ll} \langle w, x^i\rangle + b \geq 1 & \text{when } y^i = +1 \\ \langle w, x^i\rangle + b \leq -1 & \text{when } y^i = -1 \end{array}\right\} \forall i = 1, M$$

We have 3 situations:

$0 < y^i\left(\langle w, x^i\rangle + b\right) < 1 \Leftrightarrow -1 < \left(\langle w, x^i\rangle + b\right) < 1$ on the correct side but inside the margin

$1 < y^i\left(\langle w, x^i\rangle + b\right) \Leftrightarrow \left(\langle w, x^i\rangle + b\right) < -1$ for $y^i = -1$ or $\left(\langle w, x^i\rangle + b\right) > 1$ for $y^i = +1$ on the correct side *and* outside the margin

$y^i\left(\langle w, x^i\rangle + b\right) < 0 \Leftrightarrow \left(\langle w, x^i\rangle + b\right) > 0$ for $y^i = -1$ or $\left(\langle w, x^i\rangle + b\right) < 0$ for $y^i = +1$ on the wrong side!

The complete problem

Finding the Optimal Separating Hyperplane turns out to be an optimization problem of the following form:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$\left.\begin{array}{ll} \langle w, x^i \rangle + b \geq 1 & \text{when } y^i = +1 \\ \langle w, x^i \rangle + b \leq -1 & \text{when } y^i = -1 \end{array}\right\} \Rightarrow y^i \left(\langle w, x^i \rangle + b \right) \geq 1, \text{ i=1,2,....,M.}$$

- N+1 parameters (N: dimension of data)
- M constraints (M: nm of datapoints)
- It is called the *primal problem.*

Solving the constrained optimization

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

under constraints: $y^i\left(\langle w, x^i\rangle + b\right) \geq 1$, i=1,2,....,M.

This can be solved using the Lagrange method for inequality constraints:

$$L(w,b,\alpha) \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{M} \alpha_i \left(y^i\left(\langle w, x^i\rangle + b\right) - 1\right)$$

with $\alpha_i \geq 0$

We have M Lagrange multipliers a_i, i = 1, ..., M (M, # of data points), one for each of the inequality constraints.

(Minimization of convex function under linear constraints through Lagrange gives the optimal solution, see complement of information on moodle).

Solving the constrained optimization

The solution of this problem is found when maximizing over α and minimizing over *w* and *b*:

$$\max_{\alpha \geq 0} \left(\min_{w,b} L(w,b,\alpha) \right)$$

where

$$L(w,b,\alpha) \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{M} \alpha_i \left(y^i \left(\langle w, x^i \rangle + b \right) - 1 \right)$$

Solving the constrained optimization

The solution of this problem is found when maximizing over α and minimizing over *w* and *b*:

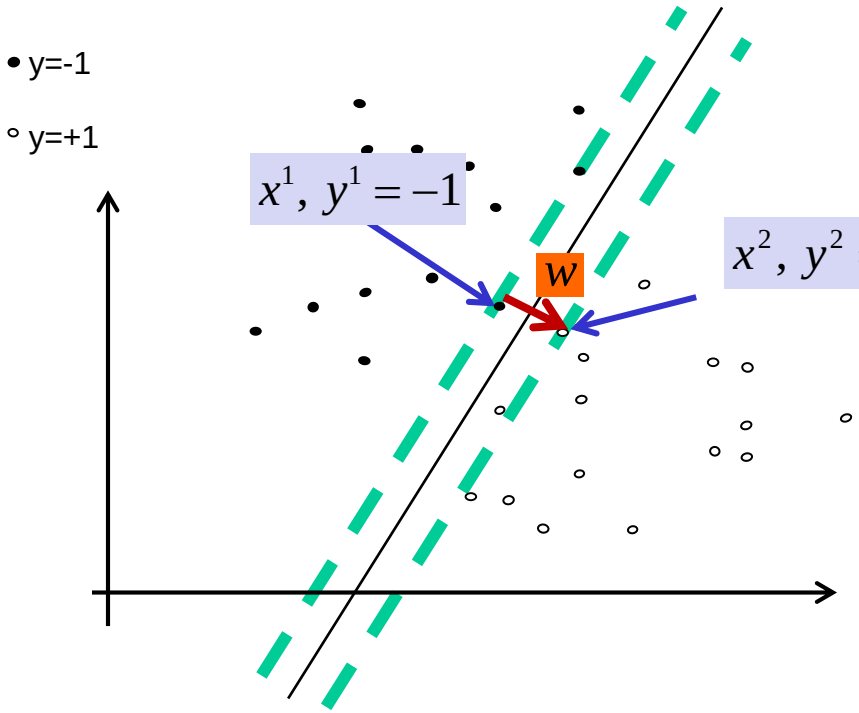
$$\max_{\alpha \geq 0} \left(\min_{w,b} L(w,b,\alpha) \right)$$

Requesting that the gradient of *L* vanishes with *w.*

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{M} \alpha_i y^i x^i$$

(Take the partial derivatives on each coordinate of w)

Interpreting the solution



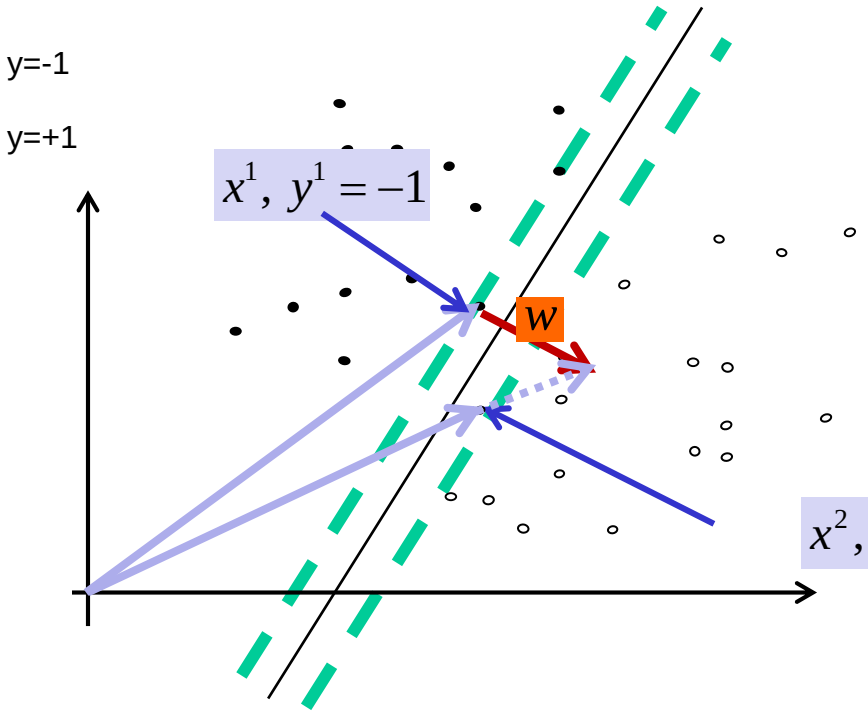
If $\alpha_1 = \alpha_2 = 1$

The vector defining the hyperplane is determined by the training points.

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{M} \alpha_i y^i x^i$$

Note that while w is unique (minimization of convex function), the alpha-s are not unique.

Interpreting the solution



Then $\alpha_1 < \alpha_2$

The vector defining the hyperplane is determined by the training points.

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{M} \alpha_i y^i x^i$$

Note that while w is unique (minimization of convex function), the alpha-s are not unique.

Interpreting the solution

Requesting that the gradient of *L* vanishes with *b.*

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{M} \alpha_i y^i = 0$$

Requires at minimum one datapoint in each class.

The dual optimization

Taking the definition of w and plugging it to the Lagrangian

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

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

Dual optimization problem

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

$$\text{subject to: } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{M} \alpha_i y_i = 0$$

This is usually solved through the:
Sequential Minimal Optimization algorithm (SMO)

The Karush-Kuhn-Tucker (KKT) Conditions

The KKT conditions ensure that our primal and dual optimization problems have the same optimal solutions

Complete optimization problem:

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{M} \alpha_i y^i x^i \qquad \text{(Dual feasibility)}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{M} \alpha_i y^i = 0 \qquad \text{(Dual feasibility)}$$

$$\frac{\partial L(w,b,\alpha)}{\partial \alpha} \leq 0 \quad \Leftrightarrow \quad y^i \left(\langle w, x^i \rangle + b \right) \geq 1 \qquad \text{(Primal feasibility)}$$

Karush-Kuhn-Tucker conditions:

$$\alpha_i \left(y^i \left(\langle w, x^i \rangle + b \right) - 1 \right) = 0 \quad \forall \; i = 1,..M \qquad \text{(Complementarity conditions)}$$

$$\alpha_i \geq 0, \qquad \forall \; i = 1,..M$$

Interpreting the conditions

All the pairs of data points $\left(x^i, y^i\right)$ for which $\alpha_i > 0$ are the support vectors.

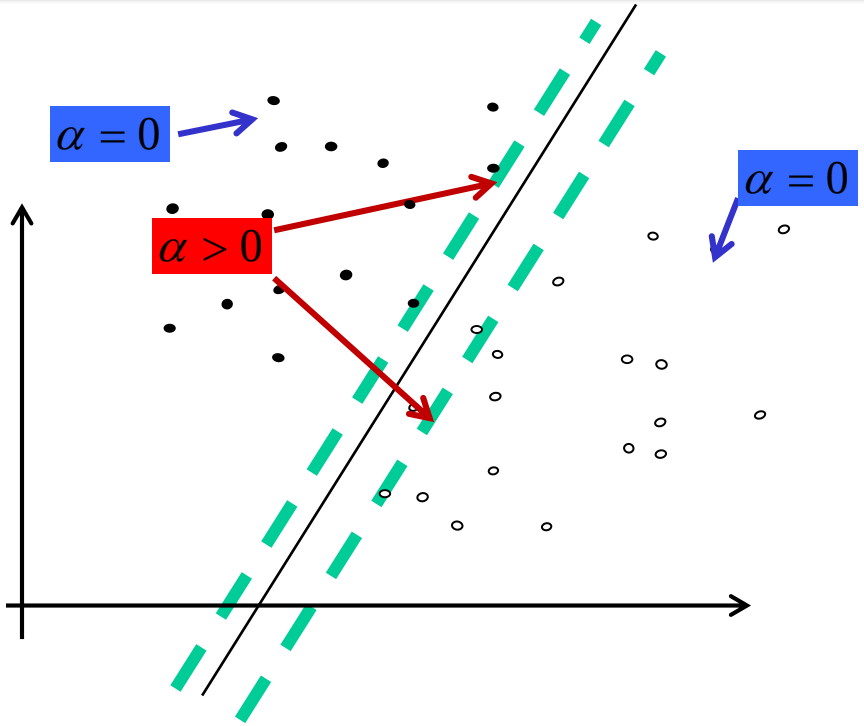
All the pairs of data points $\left(x^i, y^i\right)$ for which $\alpha_i = 0$ are "ignored".

Karush-Kuhn-Tucker conditions:

$$\alpha_i \left(y^i \left(\left\langle w, x^i \right\rangle + b \right) - 1 \right) = 0 \quad \forall\ i = 1,..M \qquad \text{(Complementarity conditions)}$$

$$\alpha_i \geq 0, \qquad \forall\ i = 1,..M$$

Interpreting the conditions



Data points $\left(x^i, y^i\right)$ for which $\alpha_i > 0$ are the support vectors.

They participate in defining the hyperplane: $w = \sum_{i=1}^{M} \alpha_i y^i x^i$

Data points $\left(x^i, y^i\right)$ for which $\alpha_i = 0$ are "ignored".

The decision function in SVM

The decision function is then expressed in terms of the support vectors:

$$f(x) = sgn\left(\langle w, x\rangle + b\right)$$

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^{M} \alpha_i y^i x^i$$

$$= sgn\left(\sum_{i=1}^{M} \alpha_i y^i \langle x, x^i \rangle + b\right)$$

Use $\left(y^i\left(\left\langle \sum_{i=1}^{M} \alpha_i y^i x^i, x^i \right\rangle + b\right) - 1\right) = 0$

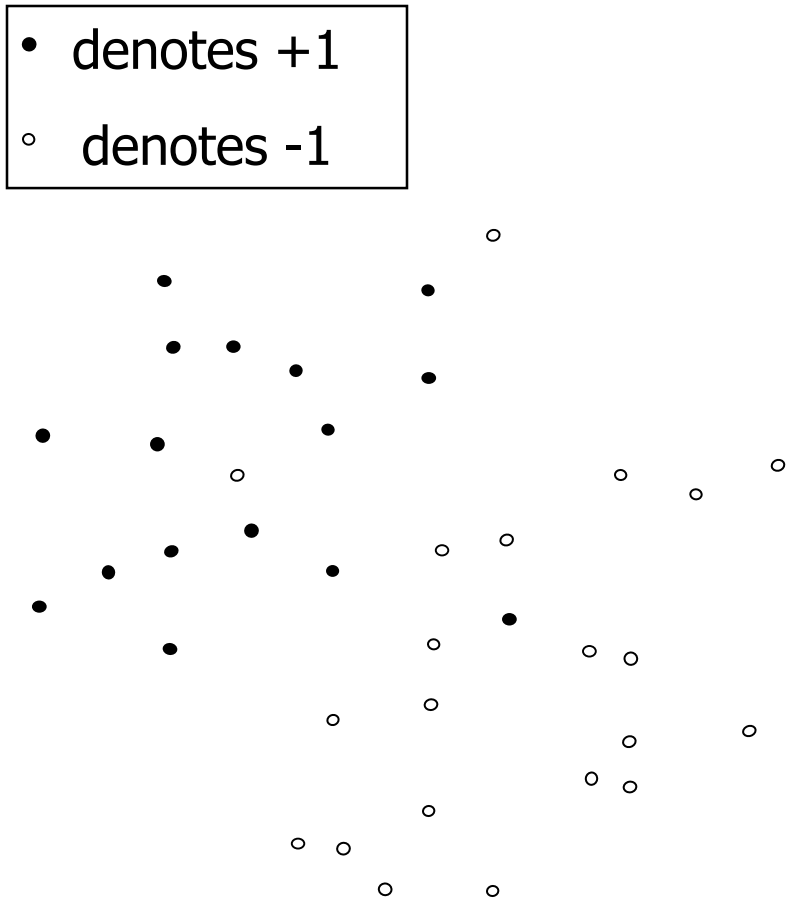
to compute b.

Non-Separable Data Sets

What should we do?

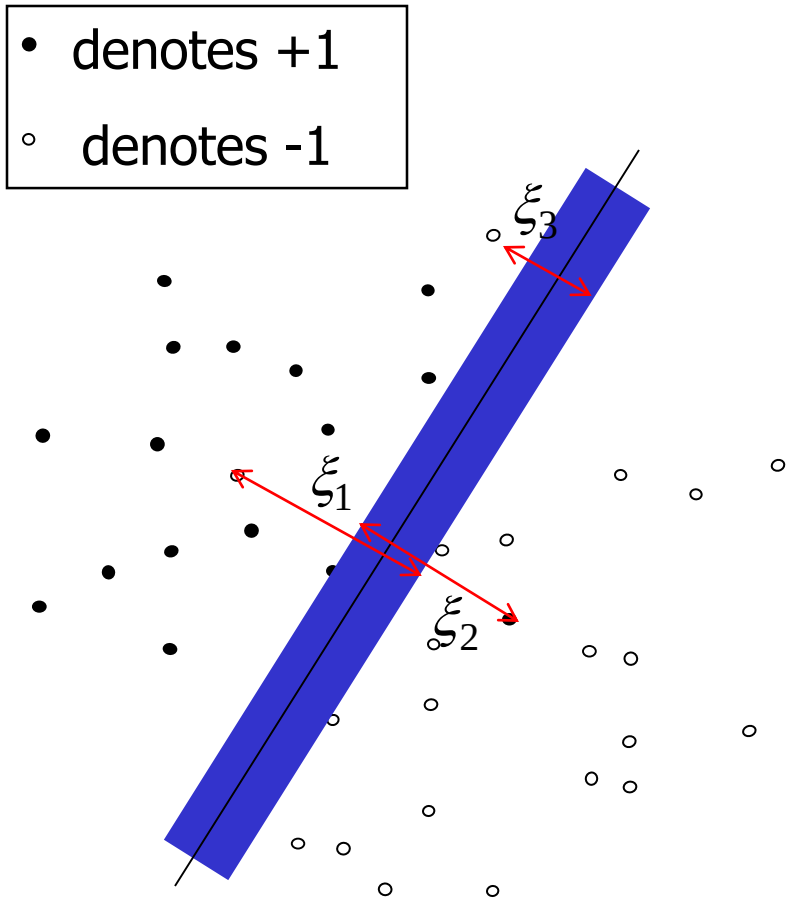
Idea :

Introduce some slack on the constraints



Support Vector Machine for non-separable datasets

Introduce slack variables: $\xi_i \geq 0$



Relax the constraints:

$y_i(< w, x > + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$

Support Vector Machine for non-separable datasets

Tradeoff

Maximize Margin

Minimize total slack

Optimization problem:

$$\min_{w,\xi}\left(\frac{1}{2}\|w\|^2+\frac{C}{M}\sum_{j=1}^{M}\xi_j\right)$$

u.c: $y_i(<w,x>+b)\geq 1-\xi_i$ and $\xi_i\geq 0$

Three cases for ξ:

$\xi_m = 0 \rightarrow$ correct classification and outside the margin

$0 < \xi_m < 1 \rightarrow$ correct classification inside margin

$\xi_m \geq 1 \rightarrow$ missclassification

Support Vector Machine for non-separable datasets

The Dual Form is given by:

$$\max_{\alpha} L_D(\alpha) \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \left\langle x^i, x^j \right\rangle$$

Subject to these constraints:

$$0 \le \alpha_j \le \frac{C}{M} \quad \forall j = 1,\ldots,M \qquad \sum_{j=1}^{M} a_j y_j = 0$$

Datapoints with $\alpha_j > 0$ will be the support vectors.

The hyperplane has the same solution:

$$w = \sum_{j=1}^{M} \alpha_j y_j x^j$$