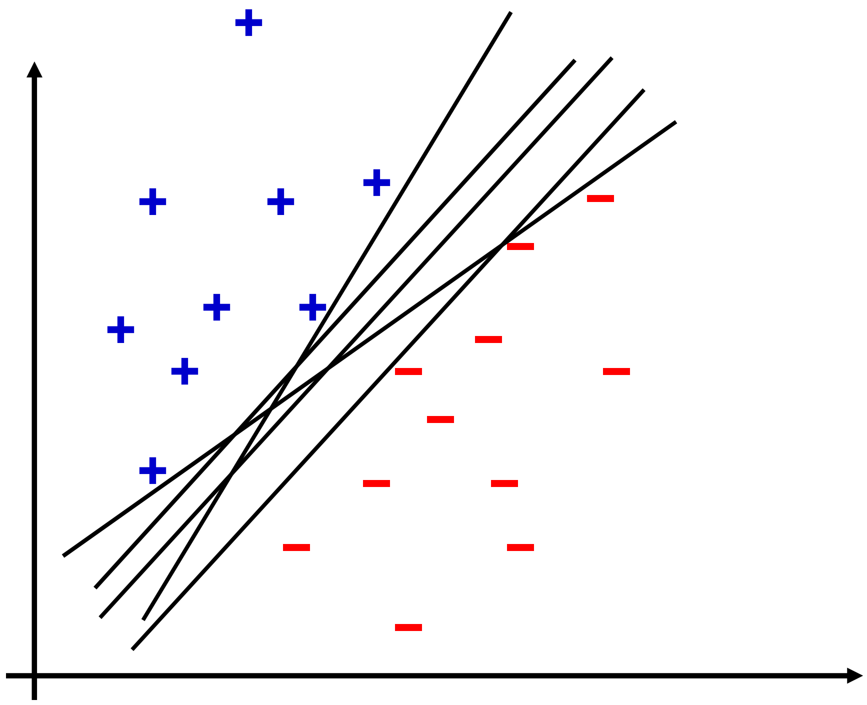
Lecture 10
Support Vector Machines

Oct - 20 - 2008

Linear Separators

- Which of the linear separators is optimal?

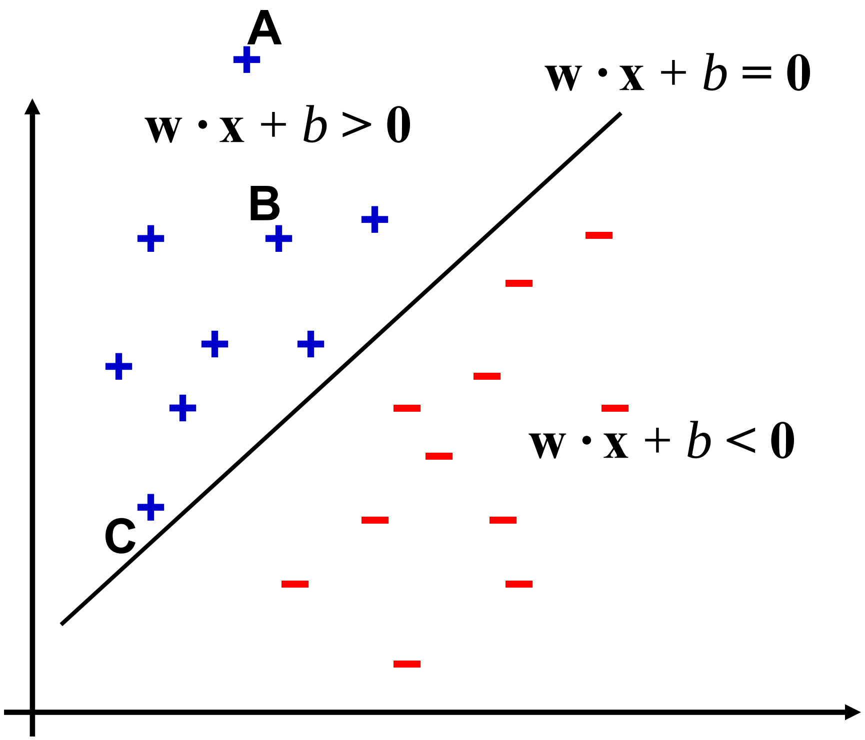


Concept of Margin

- Recall that in Perceptron, we learned that the convergence rate of the Perceptron algorithm depends on a concept called ***margin***

Intuition of Margin

- Consider points A, B, and C
- We are quite confident in our prediction for A because it is far from the decision boundary.
- In contrast, we are not so confident in our prediction for C because a slight change in the decision boundary may flip the decision.



Given a training set, we would like to make all of our predictions correct and confident! This can be captured by the concept of margin

Functional Margin

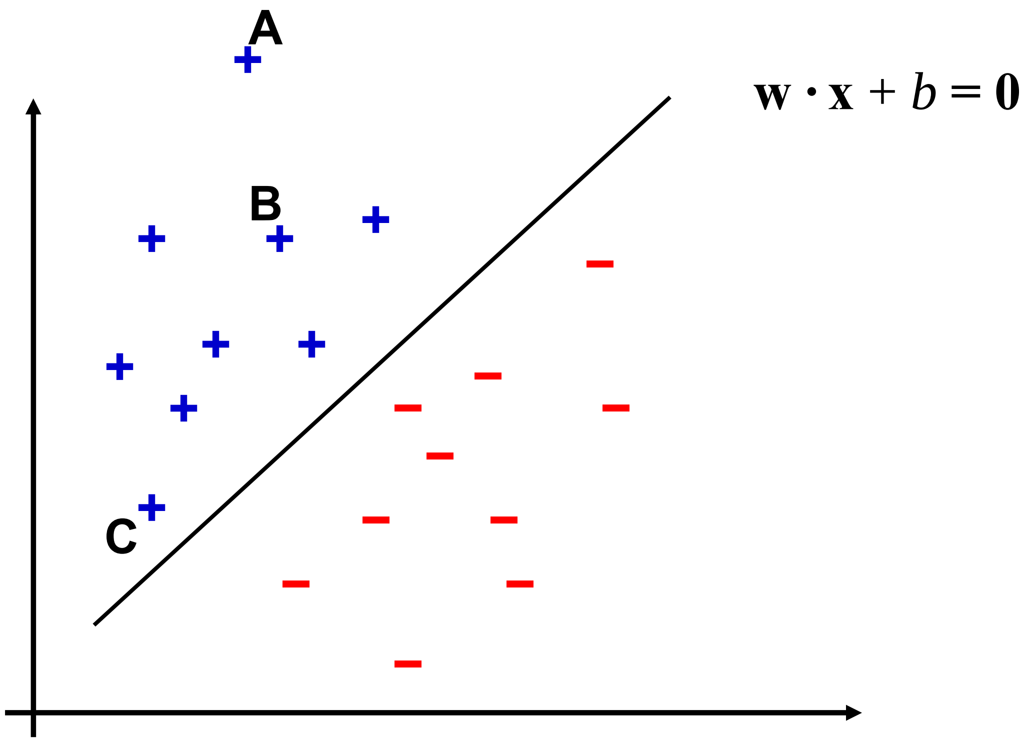
- One possible way to define margin:

$$\hat{\gamma}^i = y^i(\mathrm{w} \cdot \mathrm{x}^i + b)$$

Note that $\hat{\gamma}^i > 0$ if classified correctly

- We define this as the functional margin of the linear classifier *w.r.t training example* ($\mathbf{x}^i$, y^i)
- The large the value, the better – really?
- What if we rescale ($\mathbf{w}$, b) by a factor α, consider the linear classifier specified by ($\alpha\mathbf{w}$, αb)
  - Decision boundary remain the same
  - Yet, functional margin gets multiplied by α
  - We can change the functional margin of a linear classifier without changing anything meaningful
  - We need something more meaningful

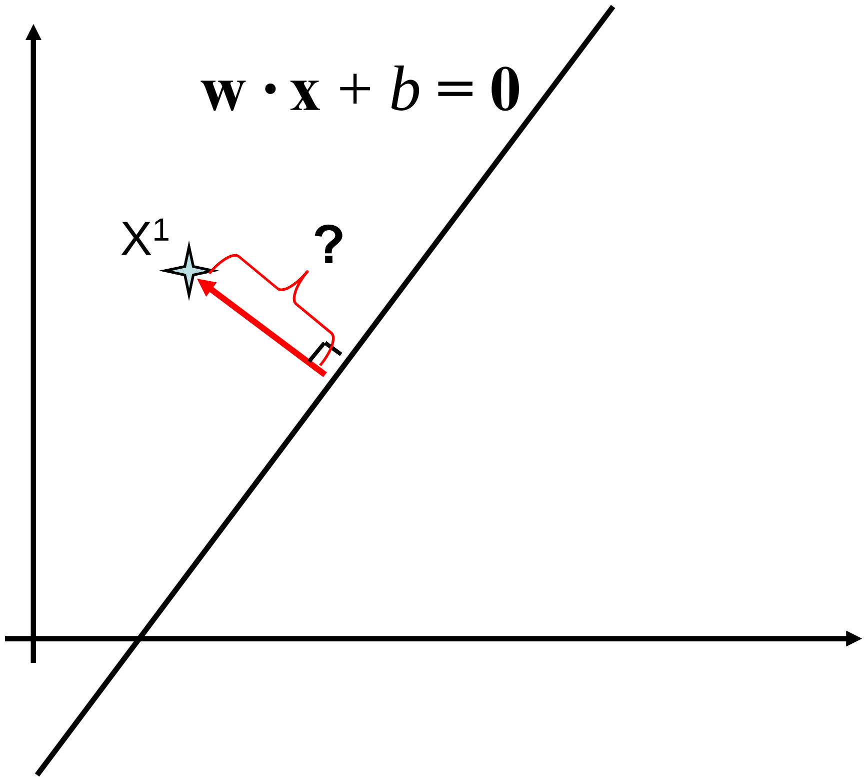
What we really want



We want the distances between the examples and the decision boundary to be large – this quantity is what we call geometric margin

But how do we compute the geometric margin of a data point w.r.t a particular line (parameterized by w and b)?

Some basic facts about lines

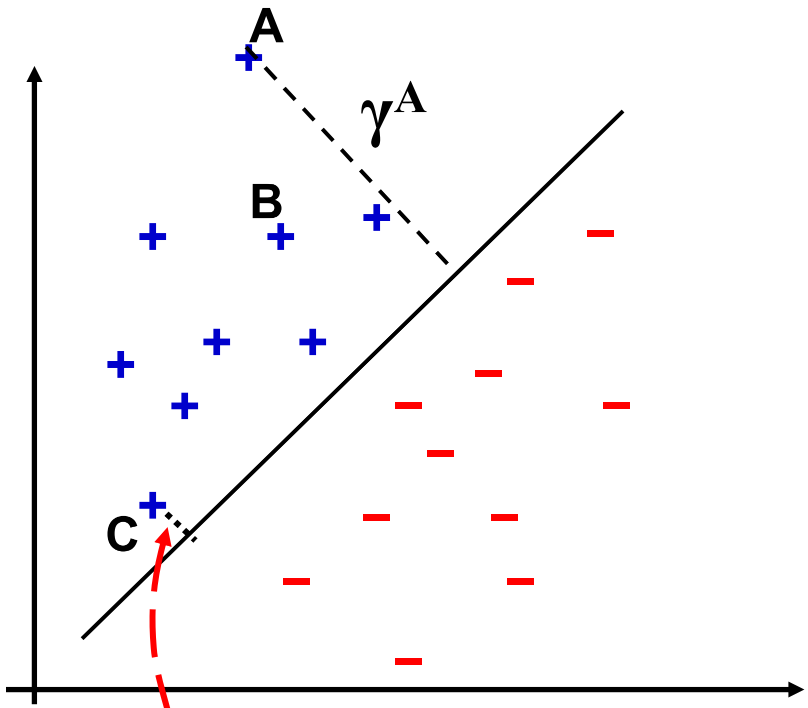


$$\boxed{\frac{w \cdot x^1 + b}{\|w\|}}$$

Geometric Margin

- The geometric margin of (**w**, *b*) w.r.t. $\mathbf{x}^{(i)}$ is the distance from $\mathbf{x}^{(i)}$ to the decision surface
- This distance can be computed as

$$\gamma^i = \frac{y^i(\mathbf{w} \cdot \mathbf{x}^i + b)}{\|\mathbf{w}\|}$$



- Given training set $S=\{(\boldsymbol{x}^i, y^i): i=1,\ldots, N\}$, the geometric margin of the classifier w.r.t. S is

$$\gamma = \min_{i=1\cdots N} \gamma^{(i)}$$

Note that the points closest to the boundary are called the ***support vectors*** – in fact these are the only points that really matters, other examples are ignorable

What we have done so far

- We have established that we want to find a linear decision boundary whose margin is the largest
- We know how to measure the margin of a linear decision boundary
- Now what?
- We have a new learning objective
  - Given a ***linearly separable*** (will be relaxed later) training set $S=\{(\boldsymbol{x}^i, y^i): i=1,\ldots, N\}$, we would like to find a linear classifier ($\mathbf{w}$, b) with maximum margin.

Maximum Margin Classifier

- This can be represented as a constrained optimization problem.

$$\max_{\mathbf{w},b} \gamma$$
$$\text{subject to: } \; y^{(i)} \frac{(\mathbf{w}\cdot\mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|} \ge \gamma, \quad i = 1, \cdots, N$$

- This optimization problem is in a nasty form, so we need to do some rewriting
- Let $\gamma' = \gamma \cdot ||w||$, we can rewrite this as

$$\max_{\mathbf{w},b} \frac{\gamma'}{\|\mathbf{w}\|}$$
$$\text{subject to: } \; y^{i}(\mathbf{w}\cdot\mathbf{x}^{i} + b) \ge \gamma', \quad i = 1, \cdots, N$$

Maximum Margin Classifier

- Note that we can arbitrarily rescale **w** and *b* to make the functional margin γ' large or small
- So we can rescale them such that γ'=1

$$\max_{\mathbf{w},b} \frac{\gamma'}{\|\mathbf{w}\|}$$
$$\text{subject to}: \ y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq \gamma', \quad i = 1,\cdots,N$$

$$\Downarrow$$

$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|} \ \text{(or equivalently } \min_{\mathbf{w},b} \|\mathbf{w}\|^2\text{)}$$
$$\text{subject to}: \ y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq 1, \quad i = 1,\cdots,N$$

Maximizing the geometric margin is equivalent to minimizing the magnitude of **w** subject to maintaining a functional margin of at least 1

Solving the Optimization Problem

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to}: \ y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq 1, \quad i = 1,\cdots,N$$

- This results in a **quadratic optimization problem** with linear inequality constraints.
- This is a well-known class of mathematical programming problems for which several (non-trivial) algorithms exist.
  - In practice, we can just regard the QP solver as a "black-box" without bothering how it works
- You will be spared of the excruciating details and jump to

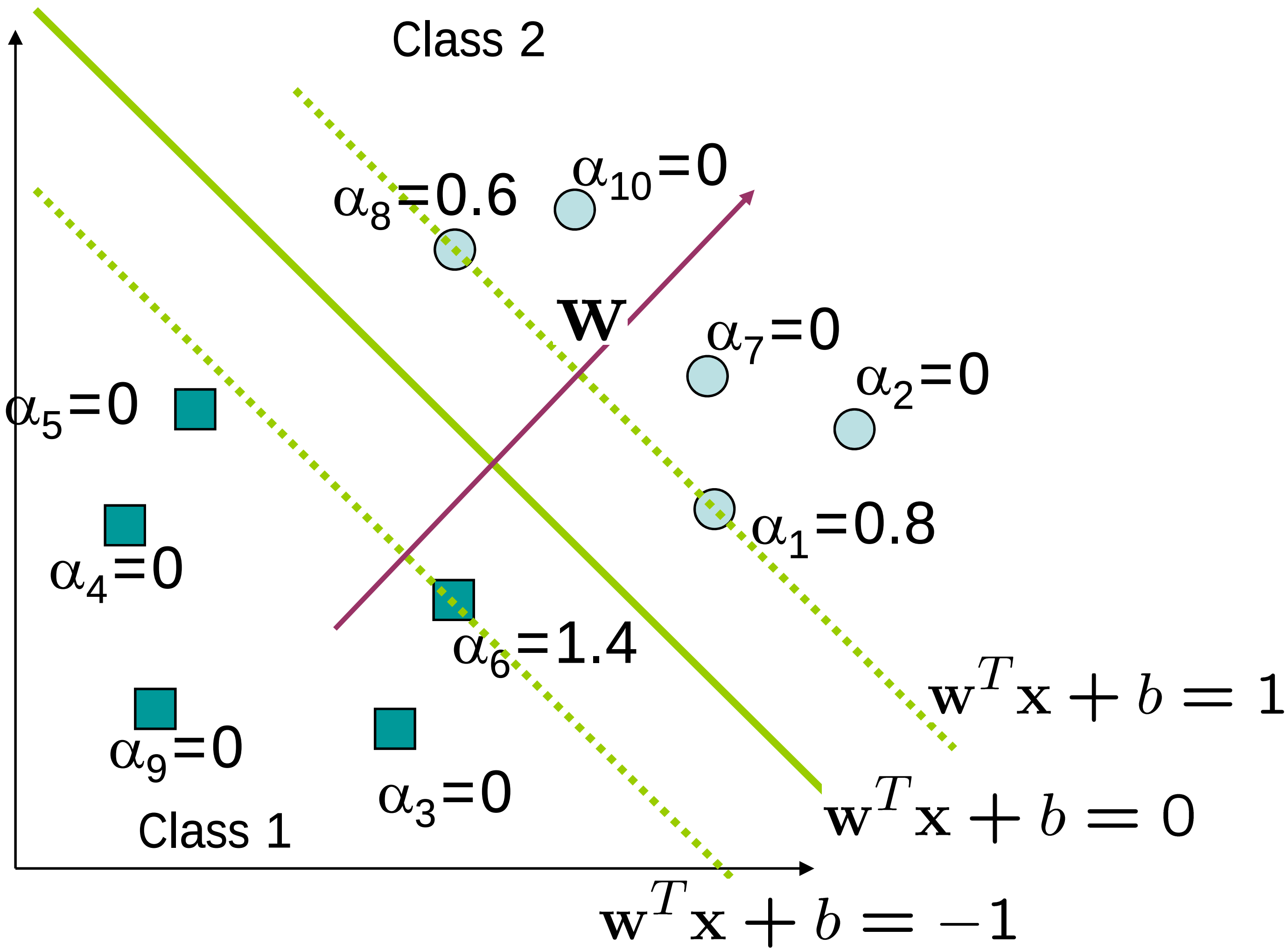
The solution

- We can not give you a close form solution that you can directly plug in the numbers and compute for an arbitrary data sets
- But, the solution can always be written in the following form

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y^i x^i, \text{s.t.} \sum_{i=1}^{N} \alpha_i y^i = 0$$

- This is the form of **w**, b can be calculated accordingly using some additional steps
- The weight vector is a linear combination of all the training examples
- Importantly, many of the α_i's are zeros
- These points that have non-zero α_i's are the **support vectors**

A Geometrical Interpretation



A few important notes regarding the geometric interpretation

- $\mathbf{w}^T\mathbf{x} + b = 0$ gives the decision boundary
- $\mathbf{w}^T\mathbf{x} + b = 1$ positive support vectors lie on this line
- $\mathbf{w}^T\mathbf{x} + b = -1$ negative support vectors lie on this line
- We can think of a decision boundary now as a tube of certain width, no points can be inside the tube
  - Learning involves adjusting the location and orientation of the tube to find the largest fitting tube for the given training set