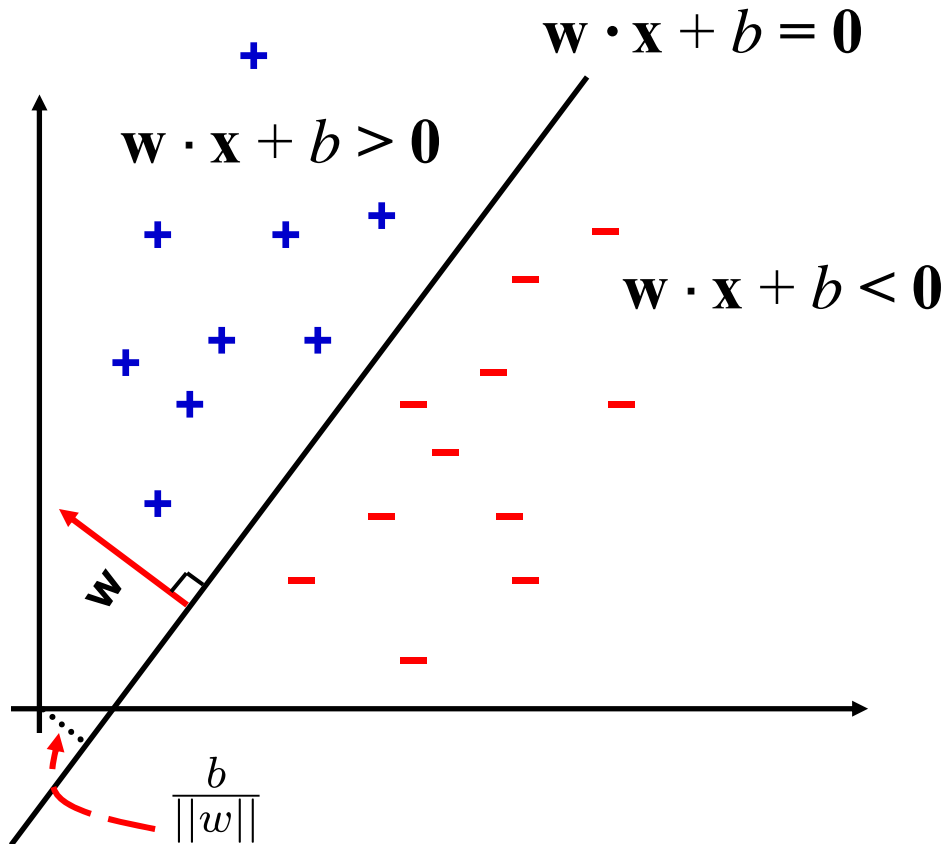


Support Vector Machines

Perceptron Revisited: Linear Separators

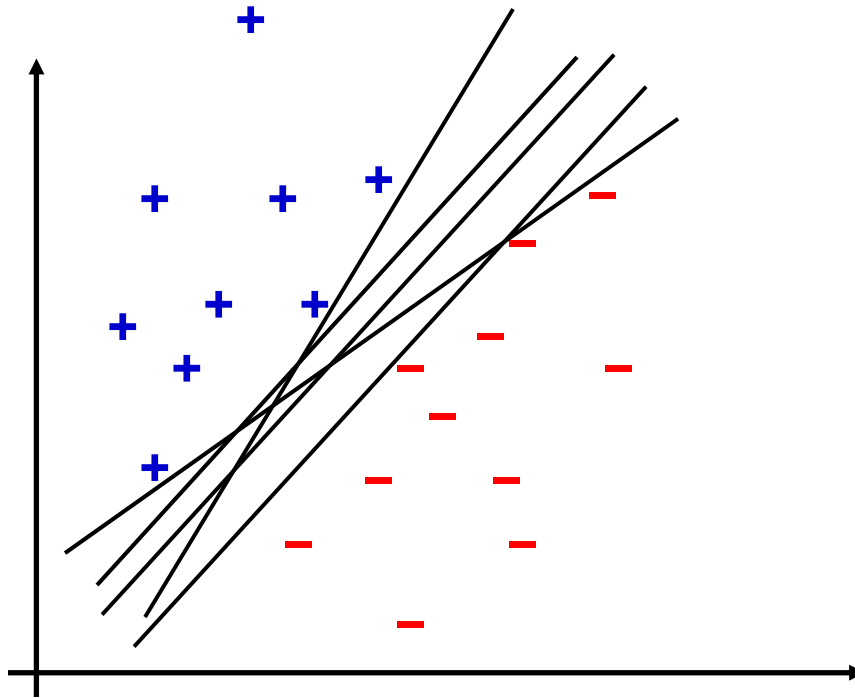
- Binary classification can be viewed as the task of separating classes in feature space:



$$f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

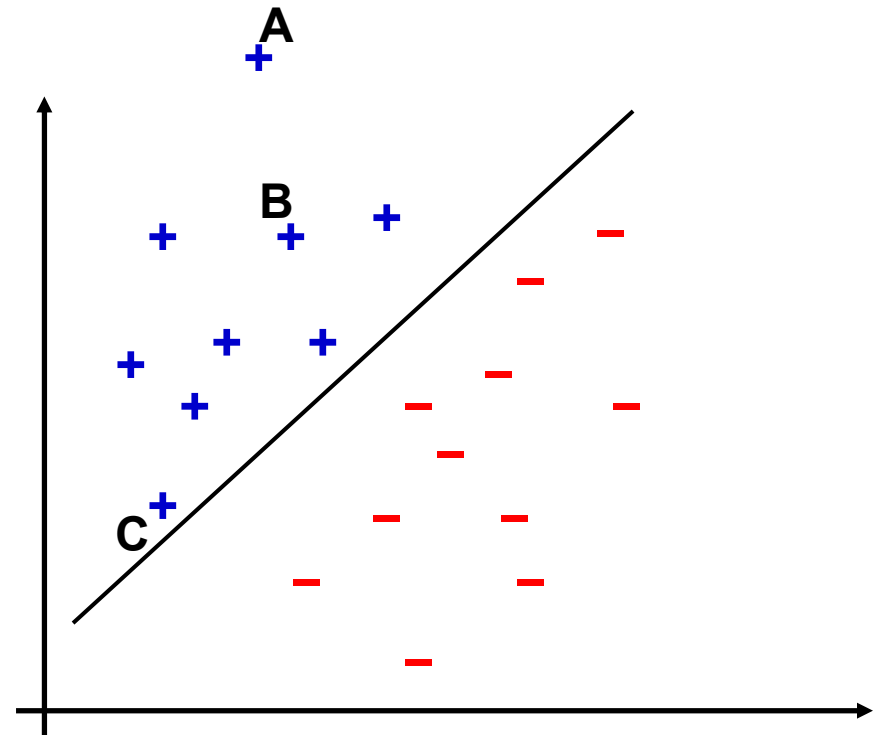
Linear Separators

- Which of the linear separators is optimal?



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 predictions correct and confident! This leads to the concept of margin.

Functional Margin

- Given a linear classifier parameterized by (\mathbf{w}, b) , we define its functional margin w.r.t training example (\mathbf{x}^i, y^i) is defined as:

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

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

- If we rescale (\mathbf{w}, b) by a factor α , functional margin gets multiplied by α
 - we can make it arbitrarily large without change anything meaningful
 - Instead, we will look at ***geometric margin***

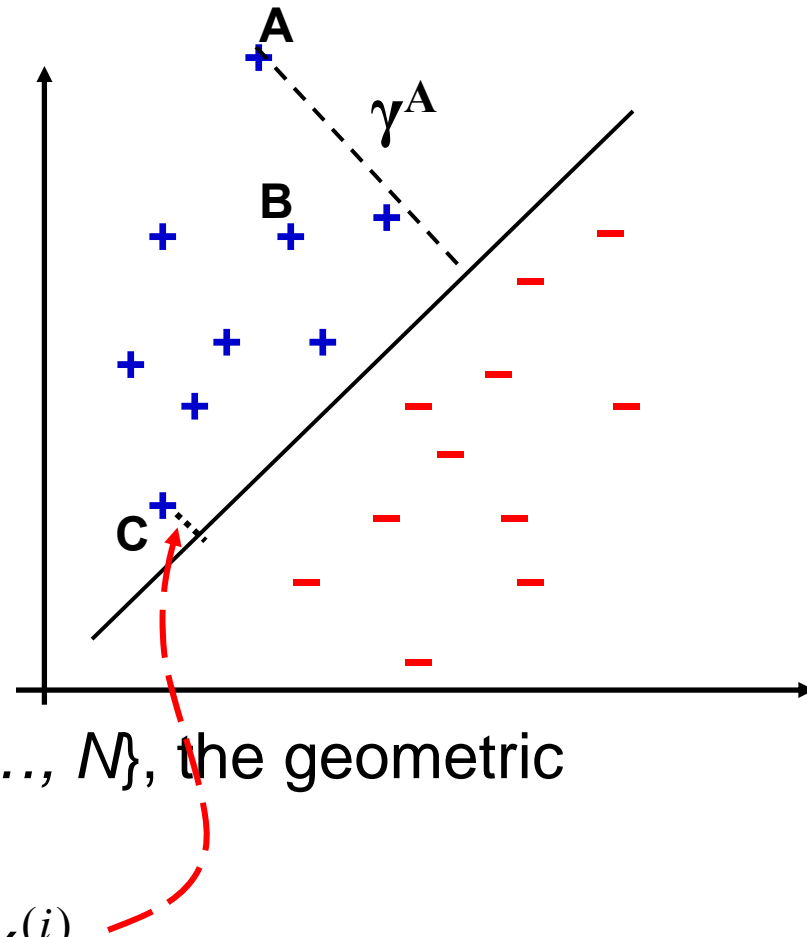
Geometric Margin

- The geometric margin of (\mathbf{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 $S = \{(\mathbf{x}^i, y^i) : i=1, \dots, N\}$, the geometric margin of the classifier w.r.t. S is

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



Points closest to the boundary are called Support vectors – we will see that these are the points that really matters

Maximum Margin Classifier

- Given a linearly separable training set $S = \{(\mathbf{x}^{(i)}, y^{(i)}) : i = 1, \dots, N\}$, we would like to find a linear classifier with maximum margin.
- This can be represented as an optimization problem.

$$\max_{\mathbf{w}, b, \gamma}$$

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

Nasty optimization problem! Let's make it look nicer!

- Let $\gamma' = \gamma \cdot \|\mathbf{w}\|$, this is equivalent to

$$\max_{\mathbf{w}, b, \gamma'} \frac{\gamma'}{\|\mathbf{w}\|}$$

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

Maximum Margin Classifier

- Note that rescaling \mathbf{w} and b by $(1/\gamma')$ will not change the classifier, we can thus further reformulate the optimization problem

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



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

Maximizing the geometric margin is equivalent to minimizing the magnitude of \mathbf{w} subject to maintaining a functional margin of at least 1

Solving the Optimization Problem

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

subject to : $y^i (\mathbf{w} \cdot \mathbf{x}^i + b) \geq 1, \quad i = 1, \dots, N$

- 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.
 - One could solve for w using any of these methods
- We will see that it is useful to first formulate an equivalent dual optimization problem and solve it instead
 - This requires a bit of machinery

Aside: Constrained Optimization

- To solve the following optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad g_i(\mathbf{x}) \leq 0 \quad \text{for } i = 1, \dots, m$$

- Consider the following function known as the Lagrangian

$$\mathcal{L}(x, \alpha) = f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x})$$

- Under certain conditions it can be shown that for a solution x' to the above problem we have

$$f(x') = \underbrace{\min_x \max_{\alpha} \mathcal{L}(x, \alpha)}_{\text{Primal form}} = \underbrace{\max_{\alpha} \min_x \mathcal{L}(x, \alpha)}_{\text{Dual form}}$$

Primal form

Dual form

subject to $\alpha_i \geq 0$

Back to the Original Problem

$$\text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{subject to: } 1 - y^i (\mathbf{w} \cdot \mathbf{x}^i + b) \leq 0, \quad i = 1, \dots, N$$

- The Lagrangian is

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + \sum_{i=1}^N \alpha_i \{1 - y^i (\mathbf{w} \cdot \mathbf{x}^i + b)\}, \text{ subject to } \alpha_i \geq 0$$

- We want to solve $\max_{\boldsymbol{\alpha}} \min_{w, b} \mathcal{L}(w, b, \boldsymbol{\alpha}) \quad s.t. \quad \alpha_i \geq 0$

- Setting the gradient of \mathcal{L} w.r.t. \mathbf{w} and b to zero, we have

$$\mathbf{w} - \sum_{i=1}^N \alpha_i y^i \mathbf{x}^i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^N \alpha_i y^i \mathbf{x}^i$$

$$\sum_{i=1}^N \alpha_i y^i = 0$$

The Dual Problem

- If we substitute $\mathbf{w} = \sum_{i=1}^N \alpha_i y^i \mathbf{x}^i$ to \mathcal{L} , we have

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \frac{1}{2} \mathbf{w} \cdot \mathbf{w} - \sum_{i=1}^N \alpha_i \{y^i (\mathbf{w} \cdot \mathbf{x}^i + b) - 1\} \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle - \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle - b \sum_{i=1}^N \alpha_i y^i + \sum_{i=1}^N \alpha_i \\ &= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle \end{aligned}$$

- Note that $\sum_{i=1}^N \alpha_i y^i = 0$
- This is a function of α_i only

The Dual Problem

- The new objective function is in terms of α_i only
- It is known as the dual problem: if we know all α_i , we know \mathbf{w}
- The original problem is known as the primal problem
- The objective function of the dual problem needs to be maximized!
- The dual problem is therefore:

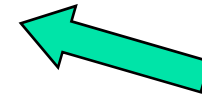
$$\max L(\boldsymbol{\alpha}) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle$$

subject to $\alpha_i \geq 0, i = 1, \dots, n,$



Properties of α_i when we introduce the Lagrange multipliers

$$\sum_{i=1}^N \alpha_i y^i = 0$$



The result when we differentiate the original Lagrangian w.r.t. \mathbf{b}

The Dual Problem

$$\max L(\boldsymbol{\alpha}) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle$$

$$\text{subject to } \alpha_i \geq 0, i = 1, \dots, n, \quad \sum_{i=1}^N \alpha_i y^i = 0$$

- This is also quadratic programming (QP) problem
 - A global maximum of α_i can always be found

- \mathbf{w} can be recovered by $\mathbf{w} = \sum_{i=1}^N \alpha_i y^i \mathbf{x}^i$

- b can also be recovered as well (wait for a bit)

Characteristics of the Solution

- Many of the α_i are zero
 - \mathbf{w} is a linear combination of only a small number of data points
- In fact, optimization theory requires that the solution to satisfy the following KKT conditions:

$$\alpha_i \geq 0, i = 1, \dots, n,$$

$$y^i \left(\sum_{j=1}^N \alpha_j y^j \langle \mathbf{x}^j \cdot \mathbf{x}^i \rangle + b \right) \geq 1$$

Functional margin ≥ 1

$$\alpha_i \left\{ y^i \left(\sum_{j=1}^N \alpha_j y^j \langle \mathbf{x}^j \cdot \mathbf{x}^i \rangle + b \right) - 1 \right\} = 0$$

α_i is nonzero only when
functional margin = 1

- \mathbf{x}_i with non-zero α_i are called support vectors (SV)
 - The decision boundary is determined only by the SV
 - Let t_j ($j=1, \dots, s$) be the indices of the s support vectors. We can write

$$\mathbf{w} = \sum_{j=1}^s \alpha_{t_j} y^{t_j} \mathbf{x}^{t_j}$$



Solve for b

- Note that we know that for support vectors the functional margin = 1
- We can use this information to solve for b
- We can use any support vector to achieve this

$$y^i \left(\sum_{j=1}^s \alpha_{t_j} y^{t_j} \langle \mathbf{x}^{t_j} \cdot \mathbf{x}^i \rangle + b \right) = 1$$

- A numerically more stable solution is to use all support vectors (details in the book)

Classifying new examples

- For classifying with a new input \mathbf{z}

- Compute $\mathbf{w}^T \mathbf{x} + b = \sum_{j=1}^s \alpha_{t_j} y^{t_j} \langle \mathbf{x}^{t_j} \cdot \mathbf{x} \rangle + b$ and classify \mathbf{z} as positive if the sum is positive, and negative otherwise

- Note: \mathbf{w} need not be formed explicitly, rather we can classify \mathbf{z} by taking a weighted sum of the inner products with the support vectors

(useful when we generalize from inner product to kernel functions later)



The Quadratic Programming Problem

- Many approaches have been proposed
 - Loqo, cplex, etc. (see <http://www.numerical.rl.ac.uk/qp/qp.html>)
- Most are “interior-point” methods
 - Start with an initial solution that can violate the constraints
 - Improve this solution by optimizing the objective function and/or reducing the amount of constraint violation
- For SVM, sequential minimal optimization (SMO) seems to be the most popular
 - A QP with two variables is trivial to solve
 - Each iteration of SMO picks a pair of (α_i, α_j) and solve the QP with these two variables; repeat until convergence
- In practice, we can just regard the QP solver as a “black-box” without bothering how it works

A Geometrical Interpretation

