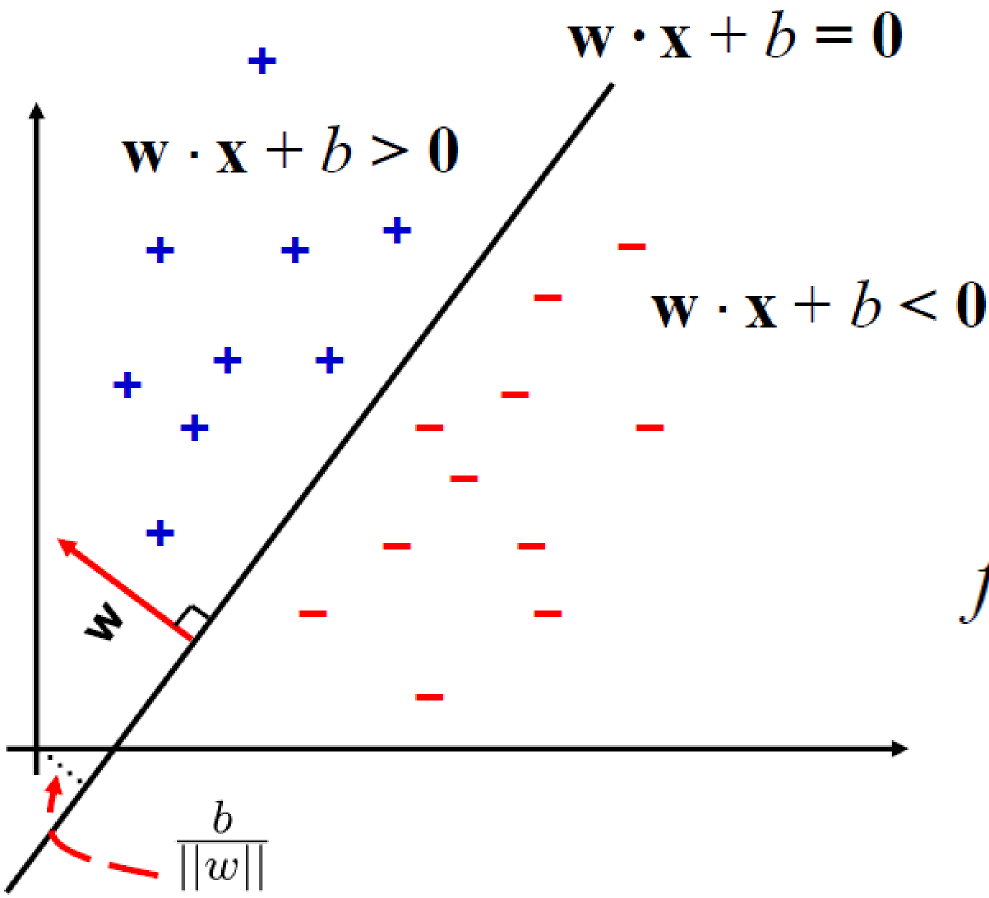
Lecture #4: Max-Margin Classification and Support Vector Machines

Janardhan Rao (Jana) Doppa

School of EECS, Washington State University

Perceptron Revisited: Linear Separator

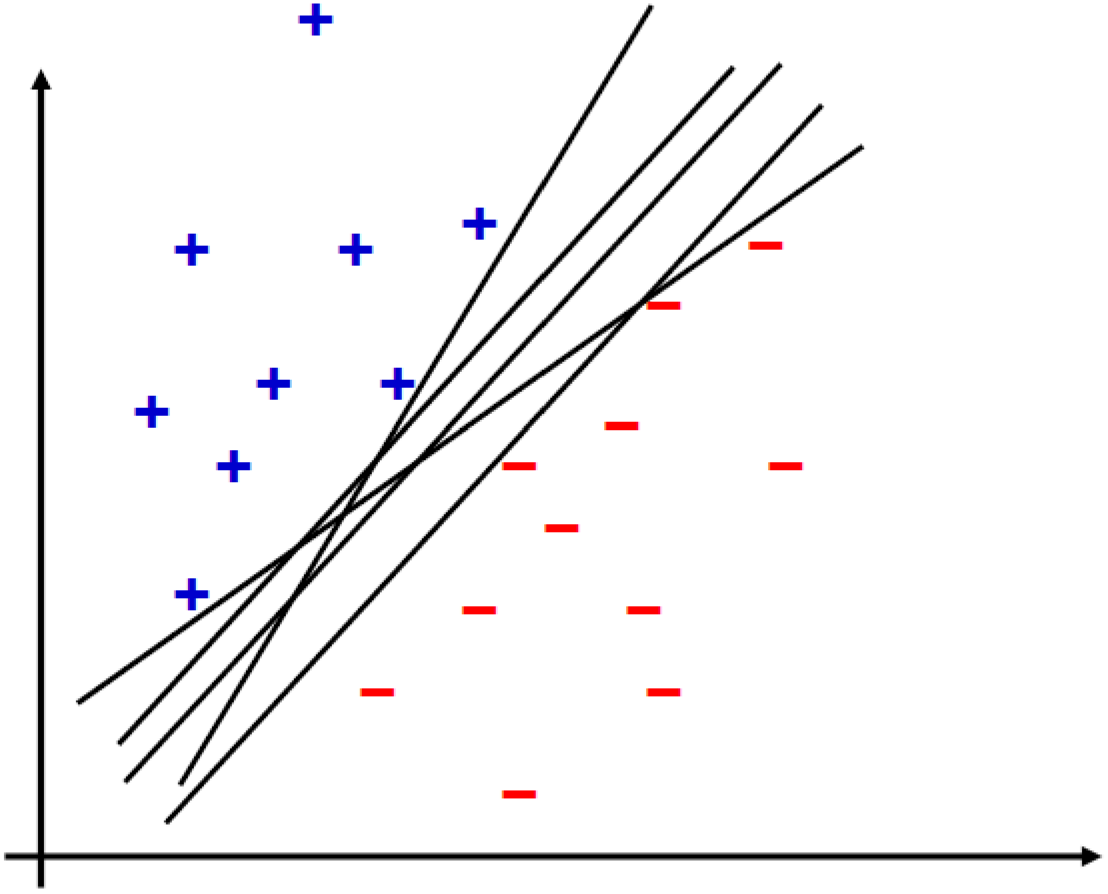
- Binary classification can be viewed as the task of separating classes in a given feature space



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

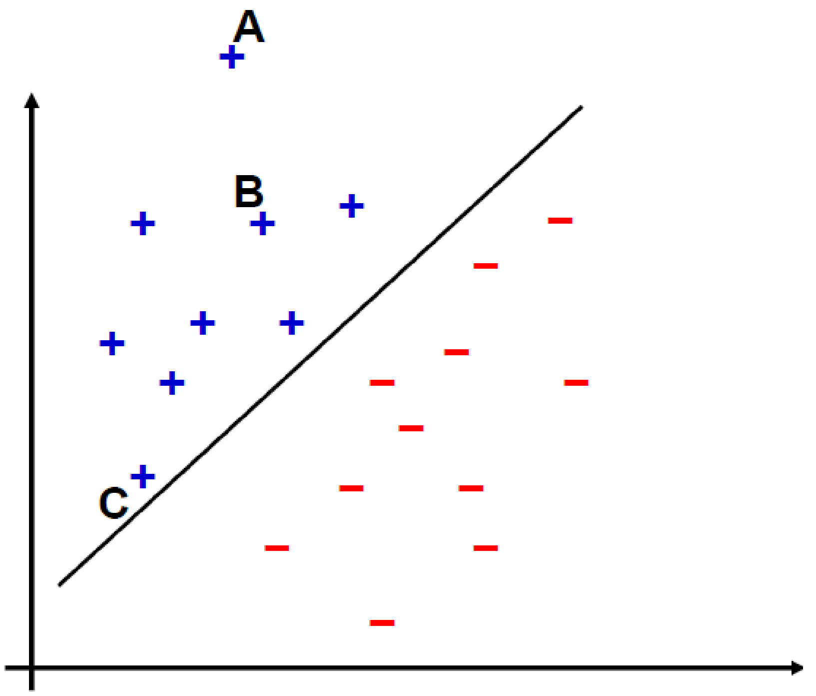
Linear Separators

- 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

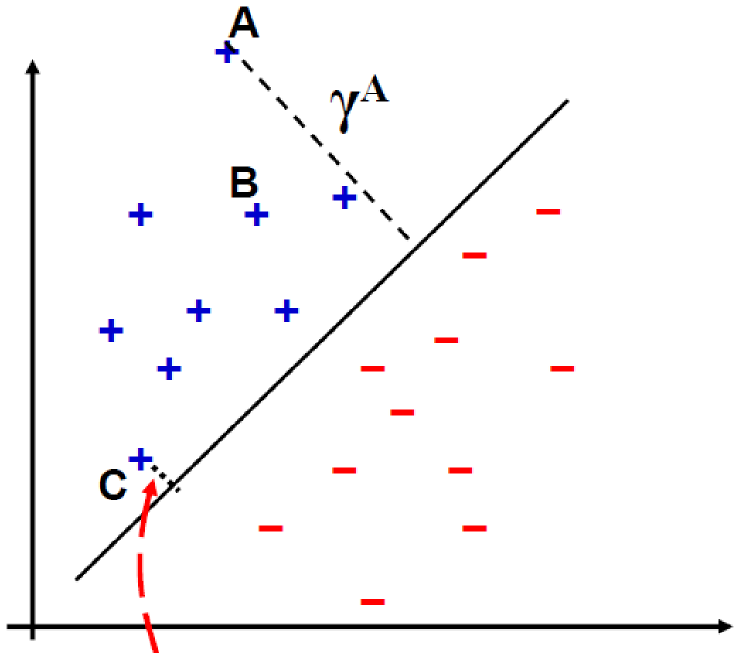


Given a training set, we would like to make all predictions correct and confident! This leads to the concept of margin.

Geometric Margin

- The geometric margin of $(\boldsymbol{w}, b)$ w.r.t. example (x_i, y_i) is the distance from x_i to the decision surface, which can be computed as:

$$\gamma_i = \frac{y_i(w \cdot x_i + b)}{\|w\|}$$



- Given a training set $S = (x_i, y_i): i = 1,2, . . N$ the geometric margin of the classifier w.r.t. S is

$$\gamma = min_{i=1,2,..N} \quad \gamma_i$$

Maximum Margin Classifier

- Given a linearly separable training set $(x_i, y_i): i = 1,2, .. N$, we would like to find a linear classifier with maximum margin

- This can be represented as an optimization problem

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

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

Maximum Margin Classifier

- Let $\gamma' = \gamma \|w\|$, we can rewrite the optimization problem as follows:

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

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

$$\Downarrow$$

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

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

Maximum Margin Classifier

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

$$\boxed{\begin{array}{l} \max\limits_{\mathbf{w},b} \dfrac{\gamma'}{\|\mathbf{w}\|} \\ \text{subject to}: \ y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq \gamma', \quad i = 1,\cdots,N \end{array}}$$

$$\Downarrow$$

$$\boxed{\begin{array}{l} \max\limits_{\mathbf{w},b} \dfrac{1}{\|\mathbf{w}\|} \ \text{(or equivalently } \min\limits_{\mathbf{w},b}\|\mathbf{w}\|^2\text{)} \\ \text{subject to}: \ y^i(\mathbf{w}\cdot\mathbf{x}^i + b) \geq 1, \quad i = 1,\cdots,N \end{array}}$$

Maximum Margin Classifier

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

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

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

$$\Downarrow$$

$$\max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|} \quad \text{(or equivalently } \min_{\mathbf{w},b} \|\mathbf{w}\|^2 \text{)}$$

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

Maximum Margin Classifier: Formulation

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

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

- This results in a **quadratic optimization (QP) 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

Characteristics of Solution

- Weights w can be represented as a linear combination of the training examples

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

Characteristics of Solution

- Many of the $\alpha's$ are zero
  - Weights w is a linear combination of small number of data points
- x_i with non-zero α_i are called support vectors (SVs)
  - The decision boundary is determined only by the SVs
  - Let $t_j (j = 1,2, \ldots, s)$ be the indices of the s support vectors. We can write

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

Characteristics of Solution

- For classifying a new input example z, compute

$$\mathbf{w}^T\mathbf{z} + b = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} (\mathbf{x}_{t_j}^T \mathbf{z}) + b$$

  - Classify z as positive if the sum is positive, and negative otherwise
  - Note: w need not be formed explicitly, rather we can classify z by taking inner products with the support vectors (useful when we generalize the notion of inner product later)