Artificial Intelligence and Machine Learning

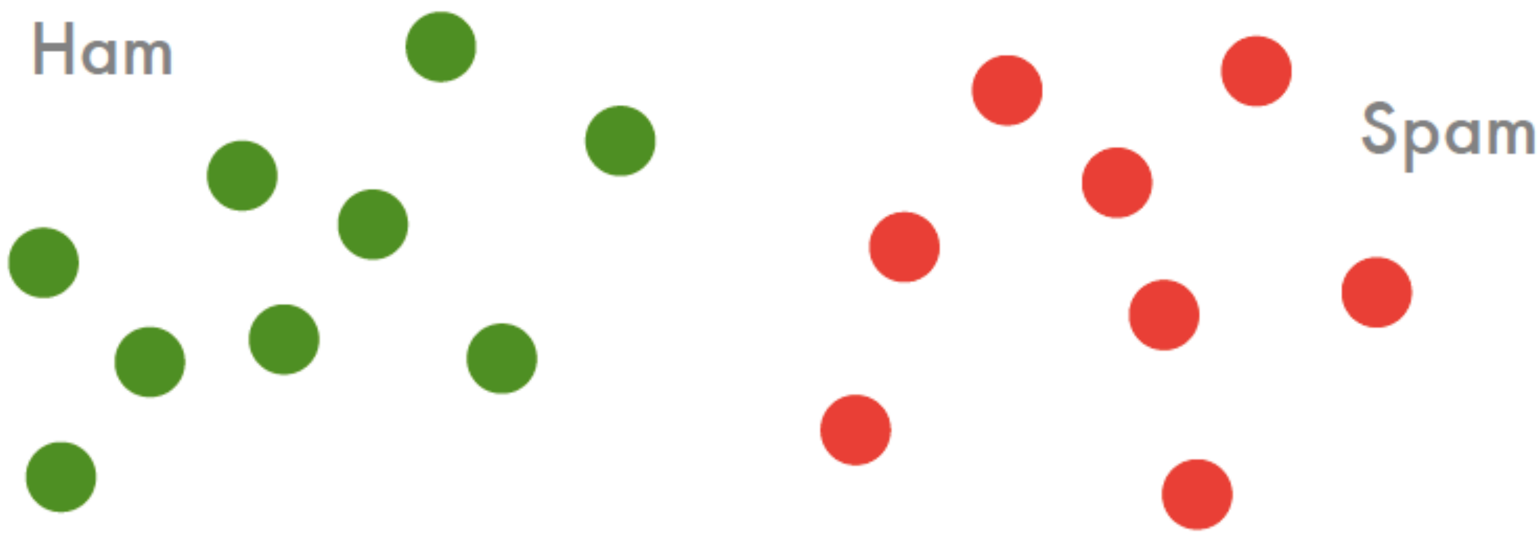
Barbara Caputo

Support Vector Machines

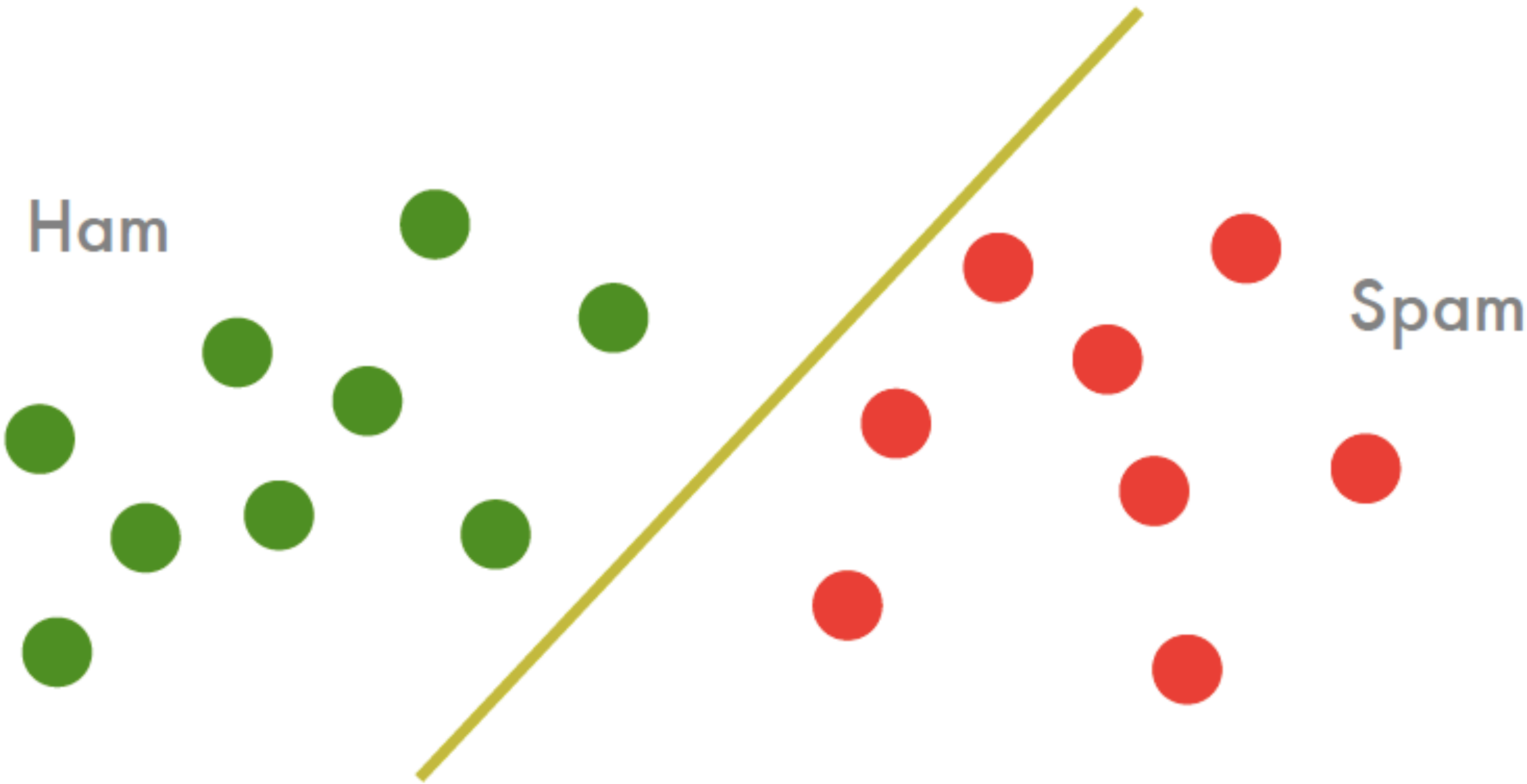
Outline

- Support Vector Classification
  Large Margin Separation, optimization problem
- Properties
  Support Vectors, kernel expansion
- Soft margin classifier
  Dual problem, robustness

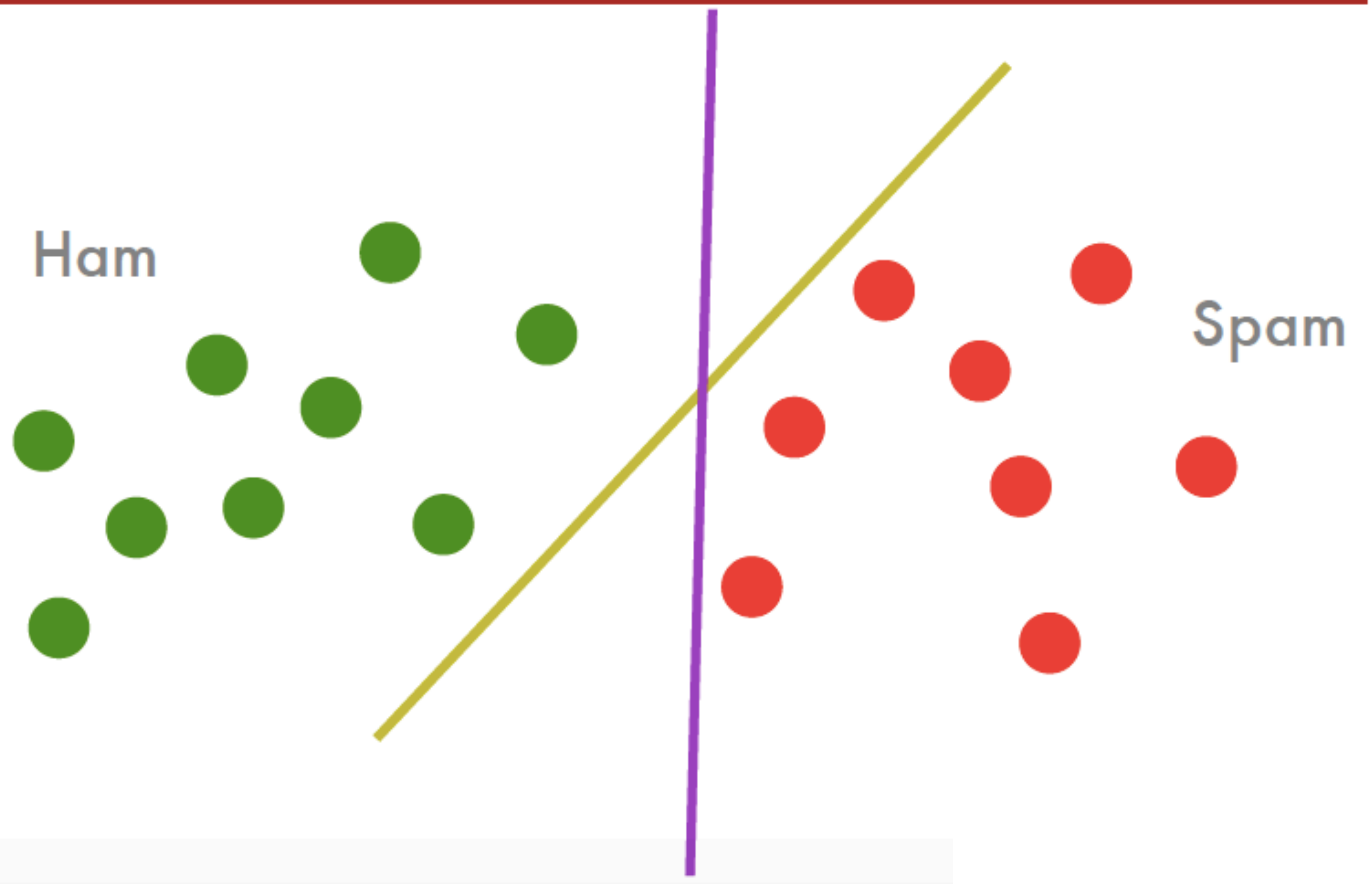
Linear Separator



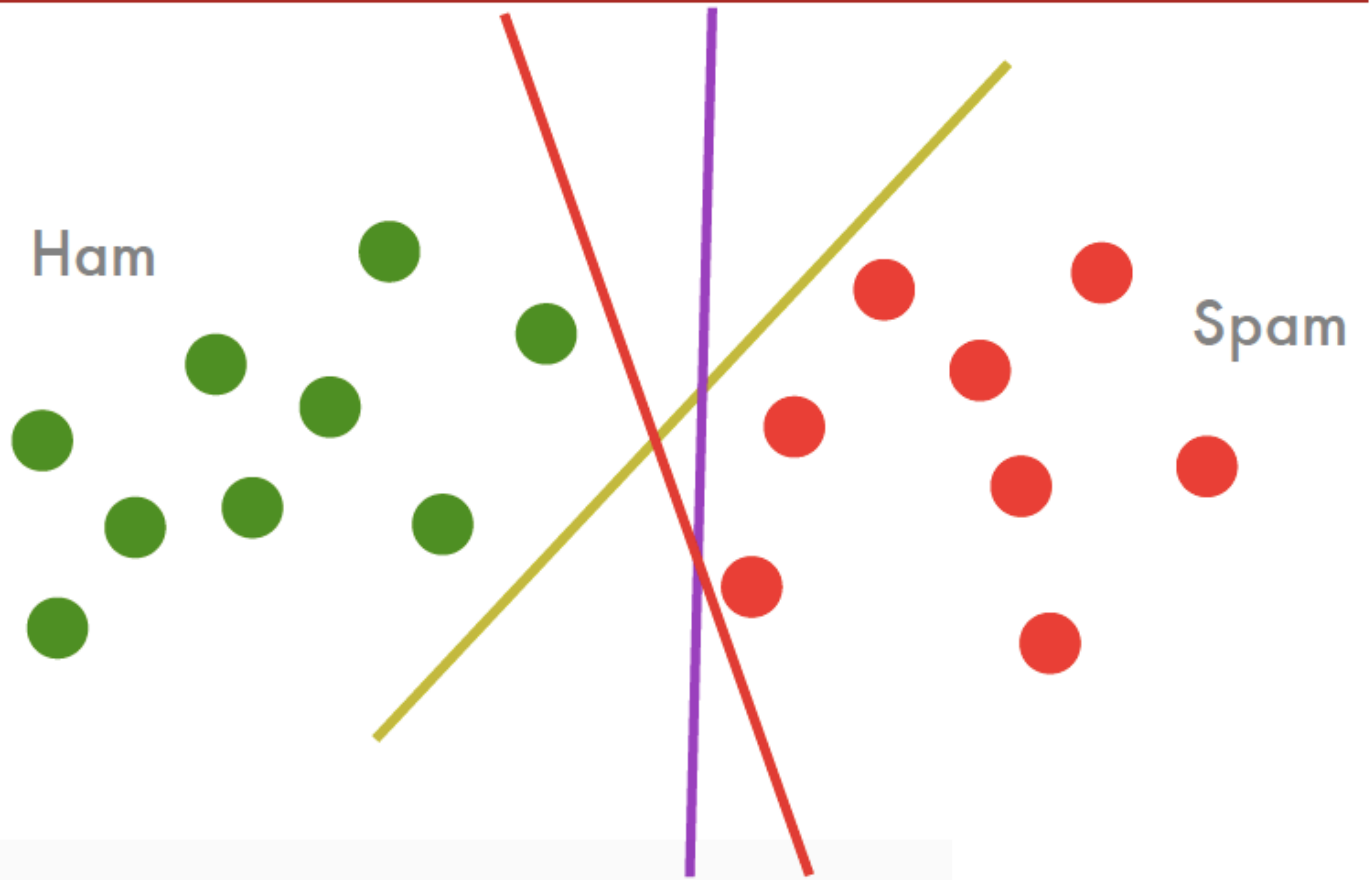
Linear Separator



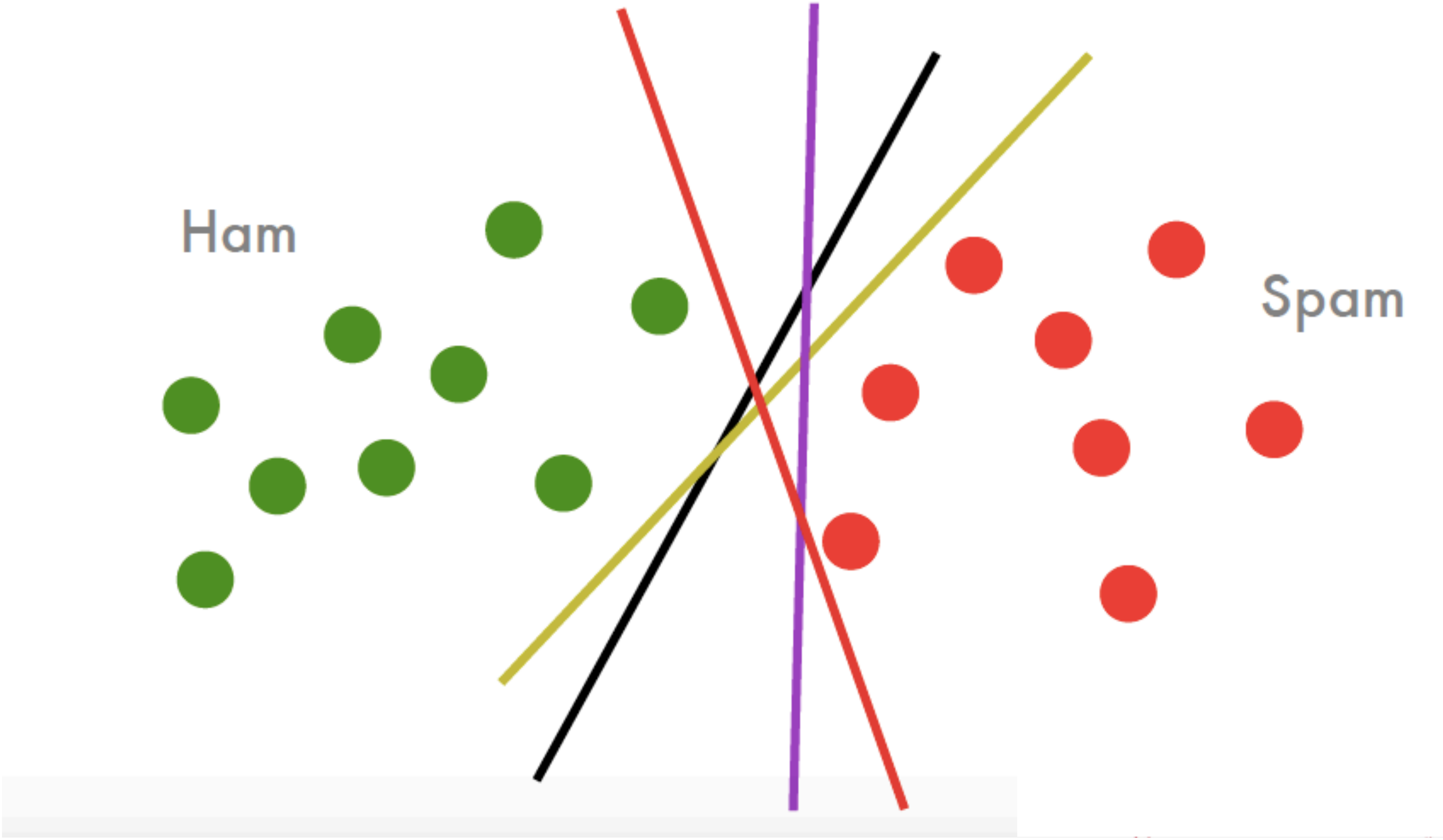
Linear Separator



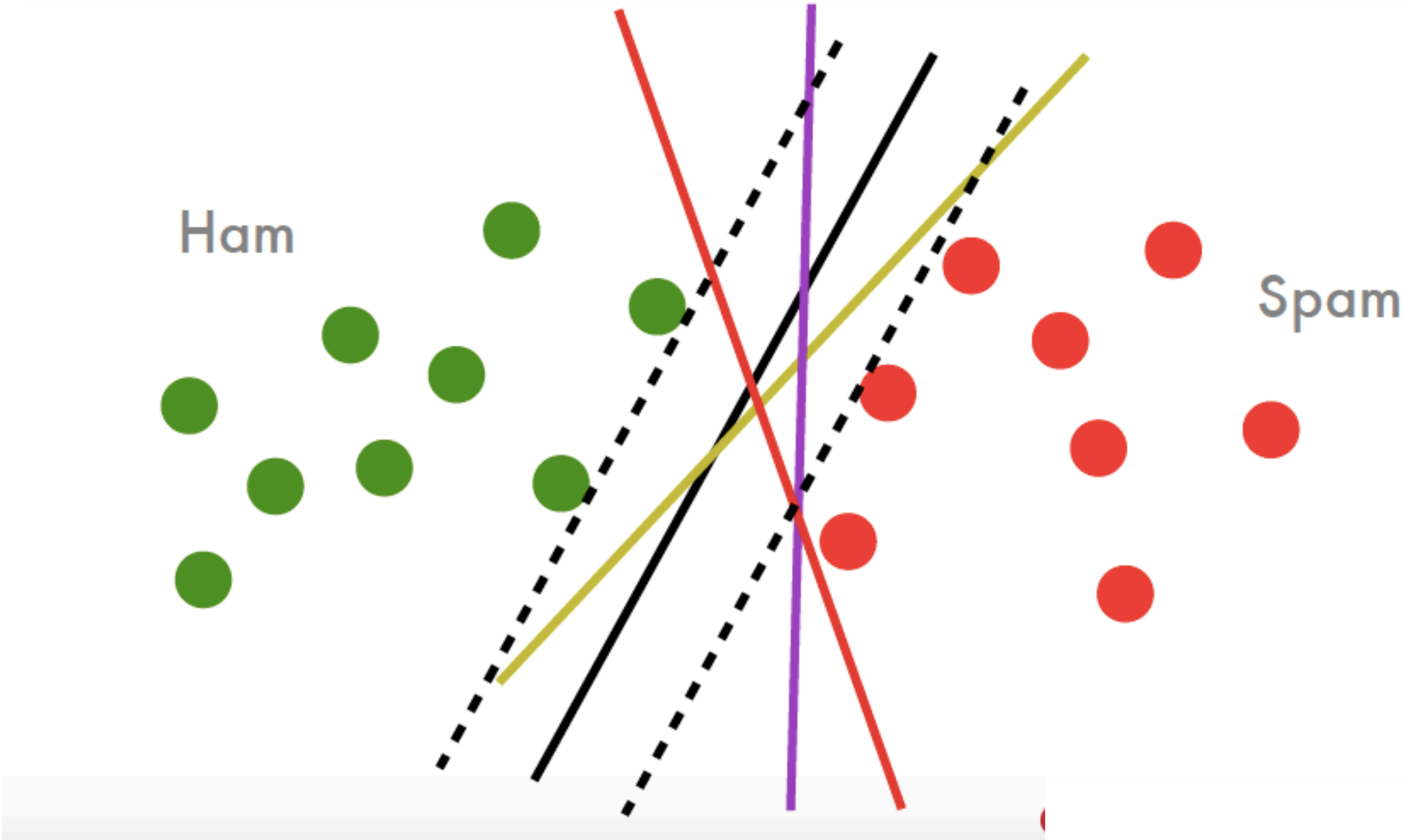
Linear Separator



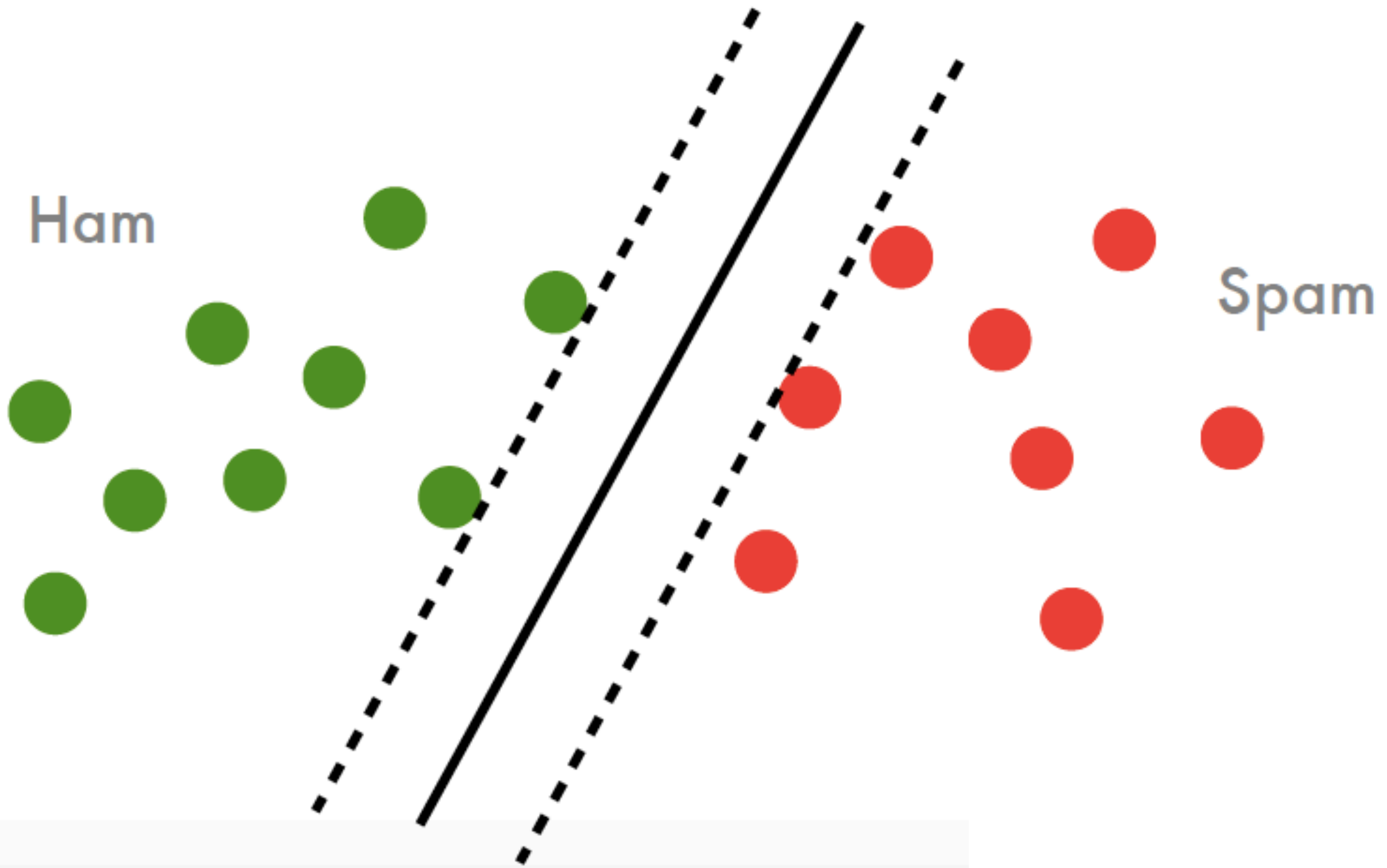
Linear Separator



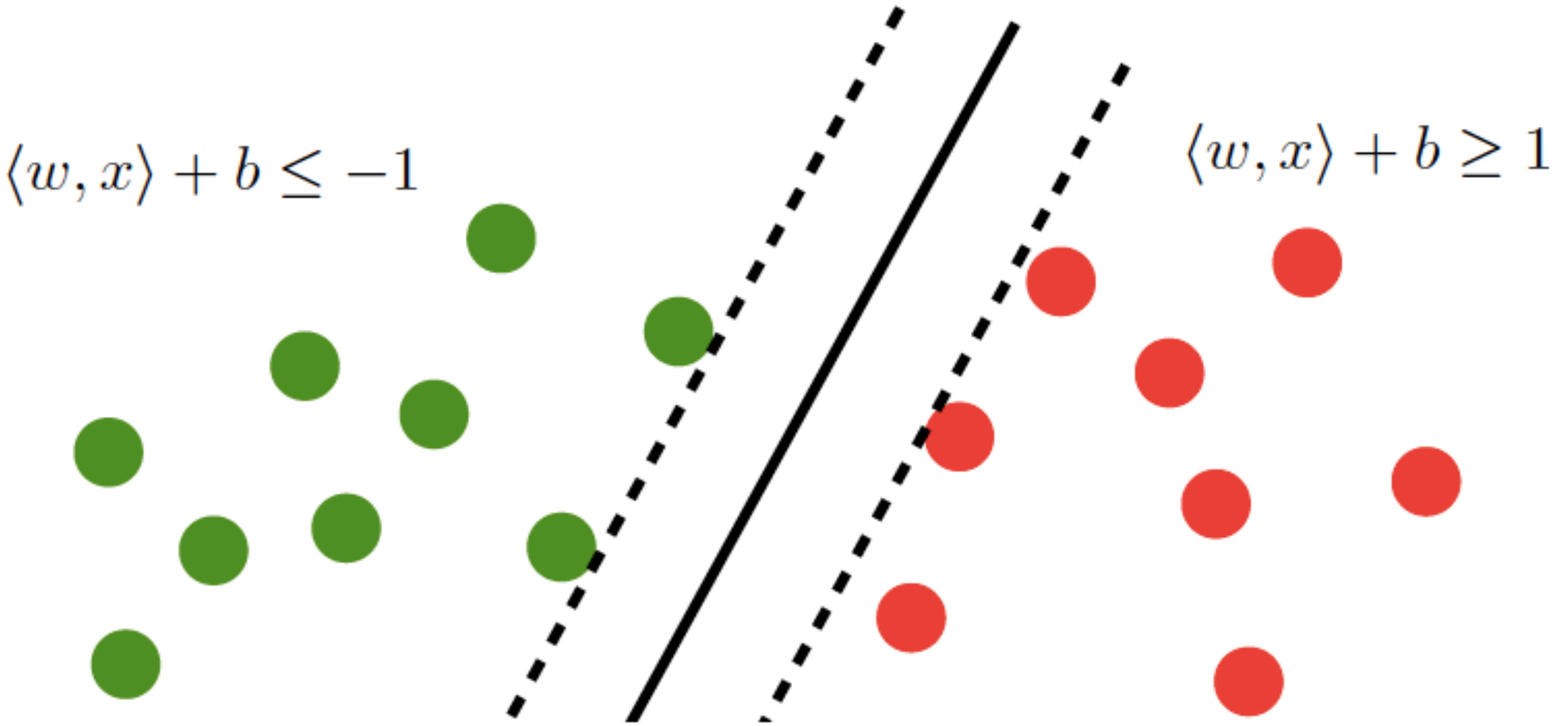
Linear Separator



Linear Separator



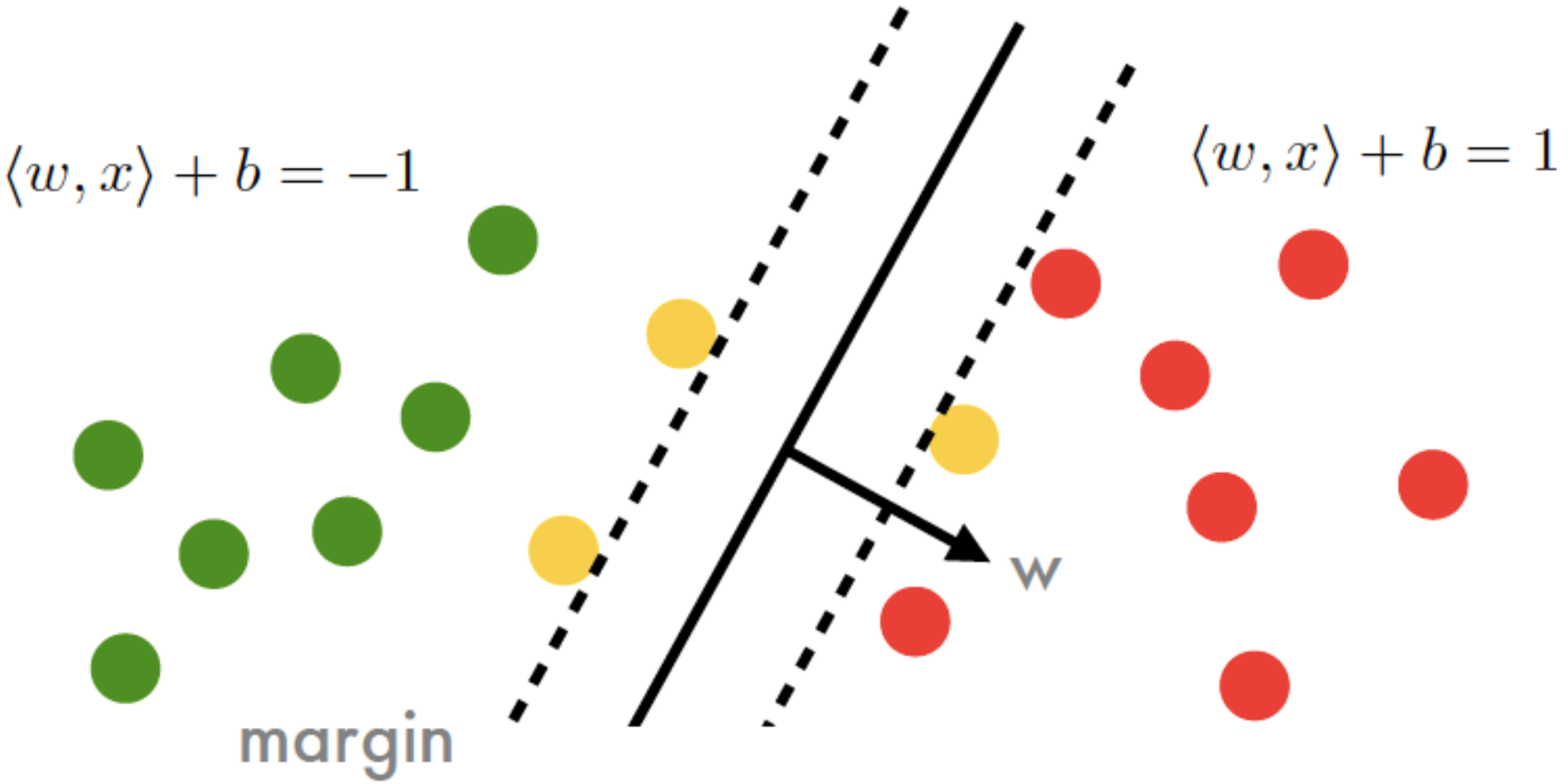
Large Margin Classifier



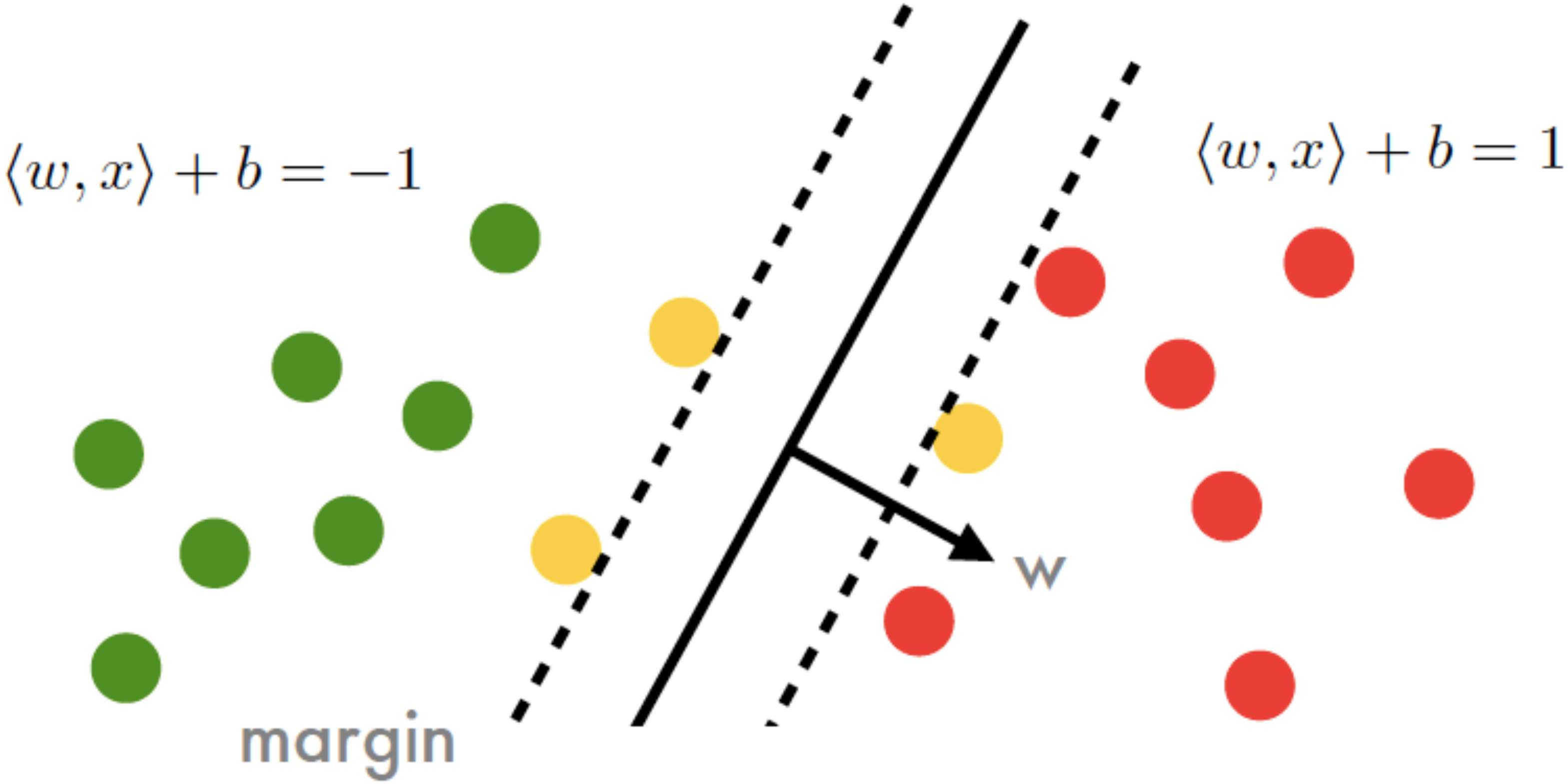
linear function

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

Large Margin Classifier

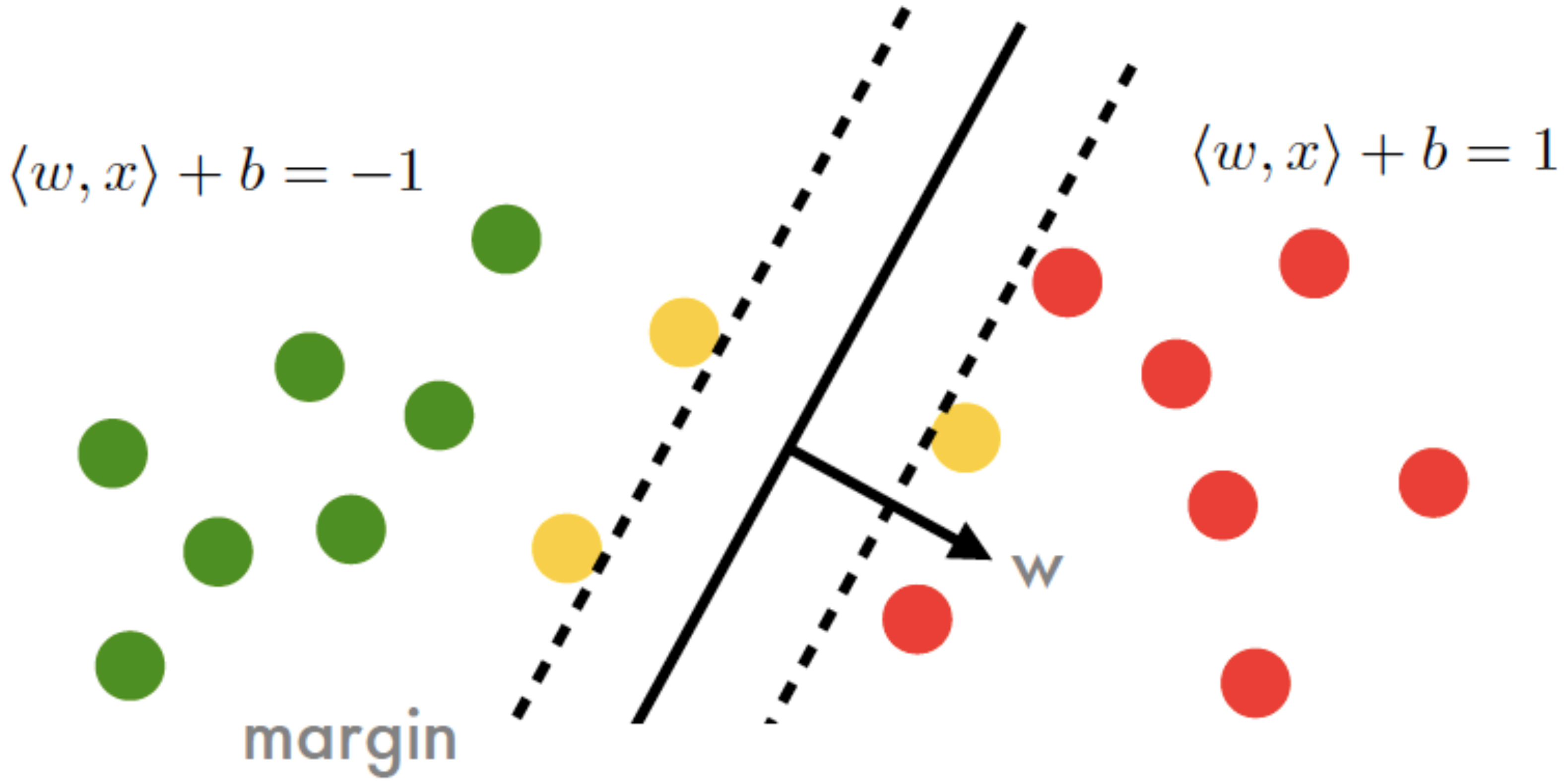


Large Margin Classifier



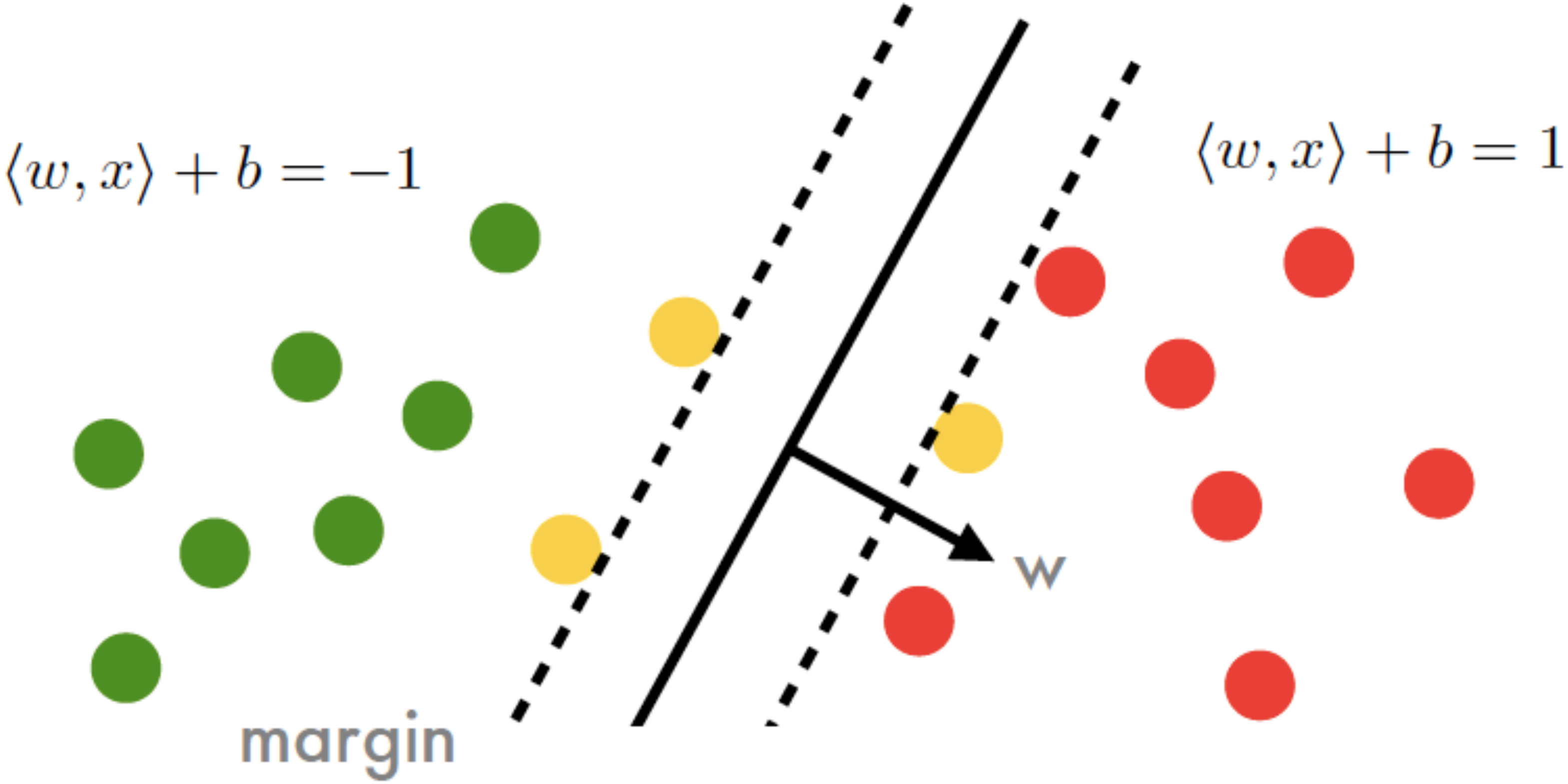
$$\frac{\langle x_+ - x_-, w \rangle}{2\,\|w\|} =$$

Large Margin Classifier



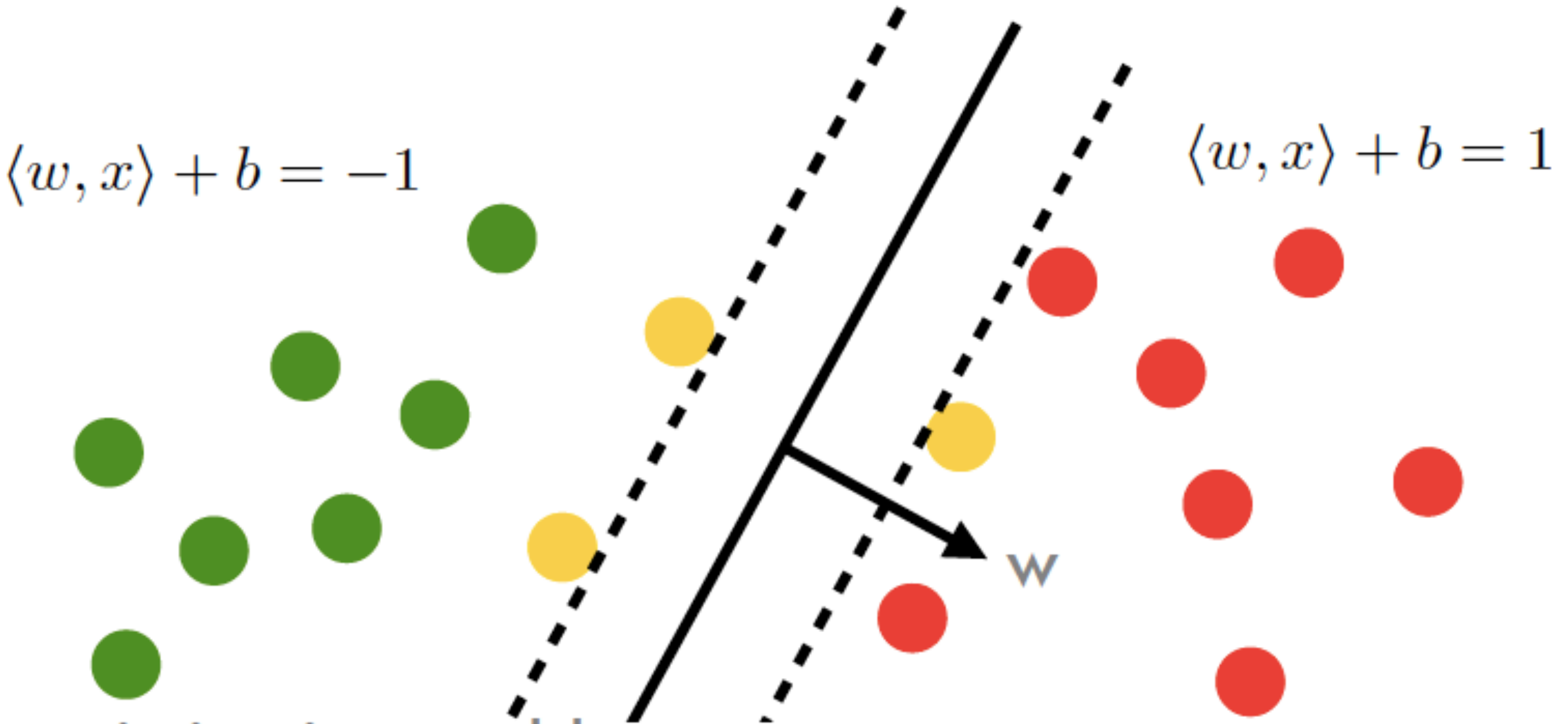
$$\frac{\langle x_+ - x_-, w \rangle}{2\,\|w\|} = \frac{1}{2\,\|w\|}\left[[\langle x_+, w \rangle + b] - [\langle x_-, w \rangle + b]\right] =$$

Large Margin Classifier



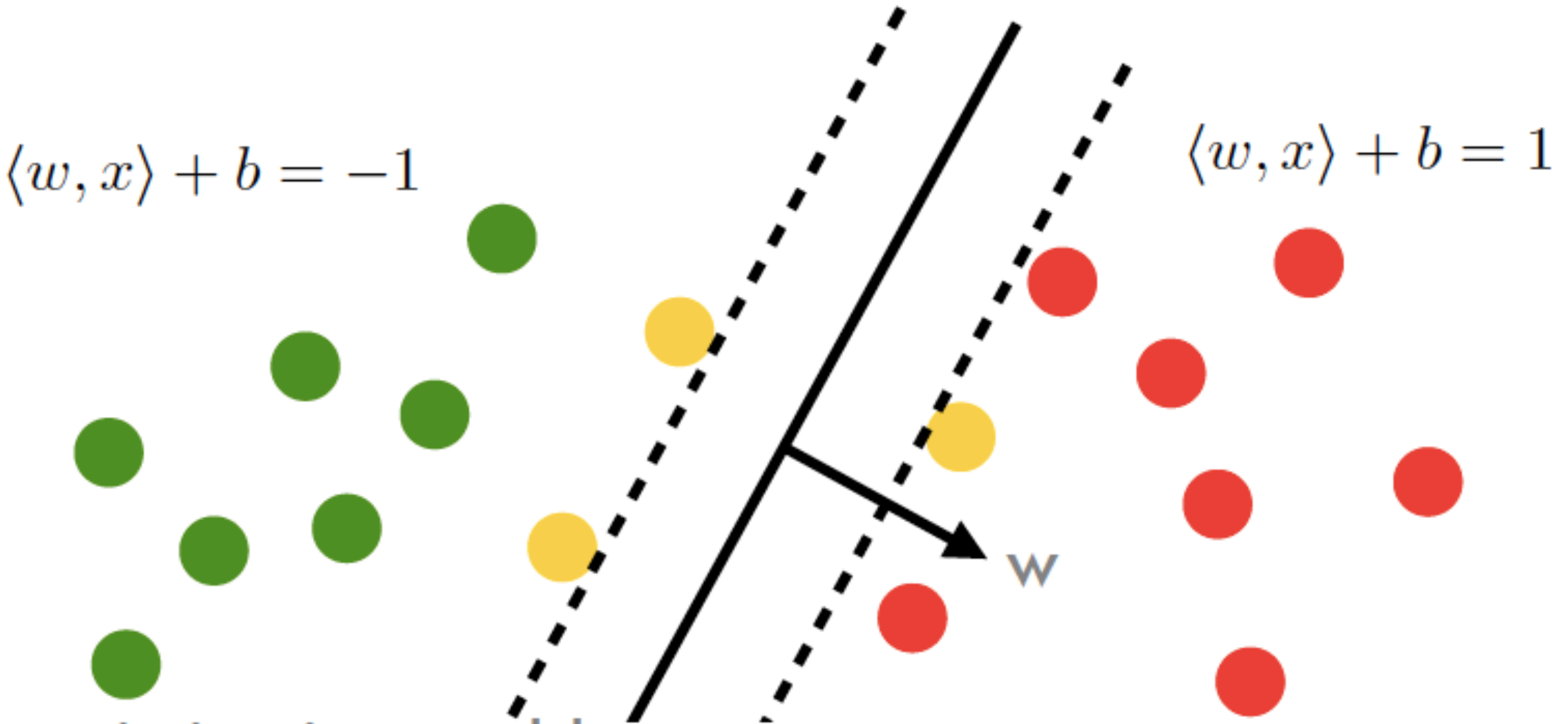
$$\frac{\langle x_+ - x_-, w\rangle}{2\,\|w\|} = \frac{1}{2\,\|w\|}\left[\left[\langle x_+, w\rangle + b\right] - \left[\langle x_-, w\rangle + b\right]\right] = \frac{1}{\|w\|}$$

Large Margin Classifier



optimization problem

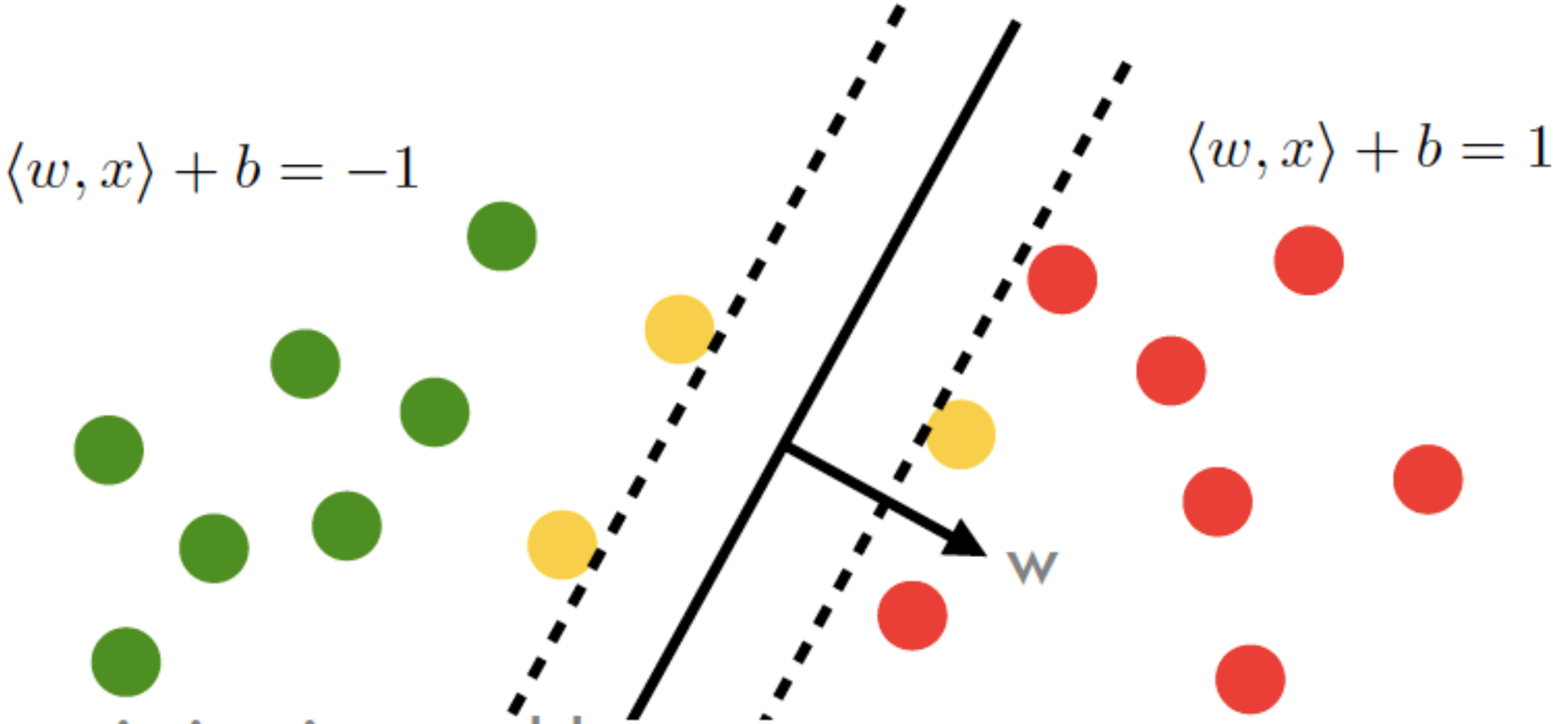
Large Margin Classifier



optimization problem

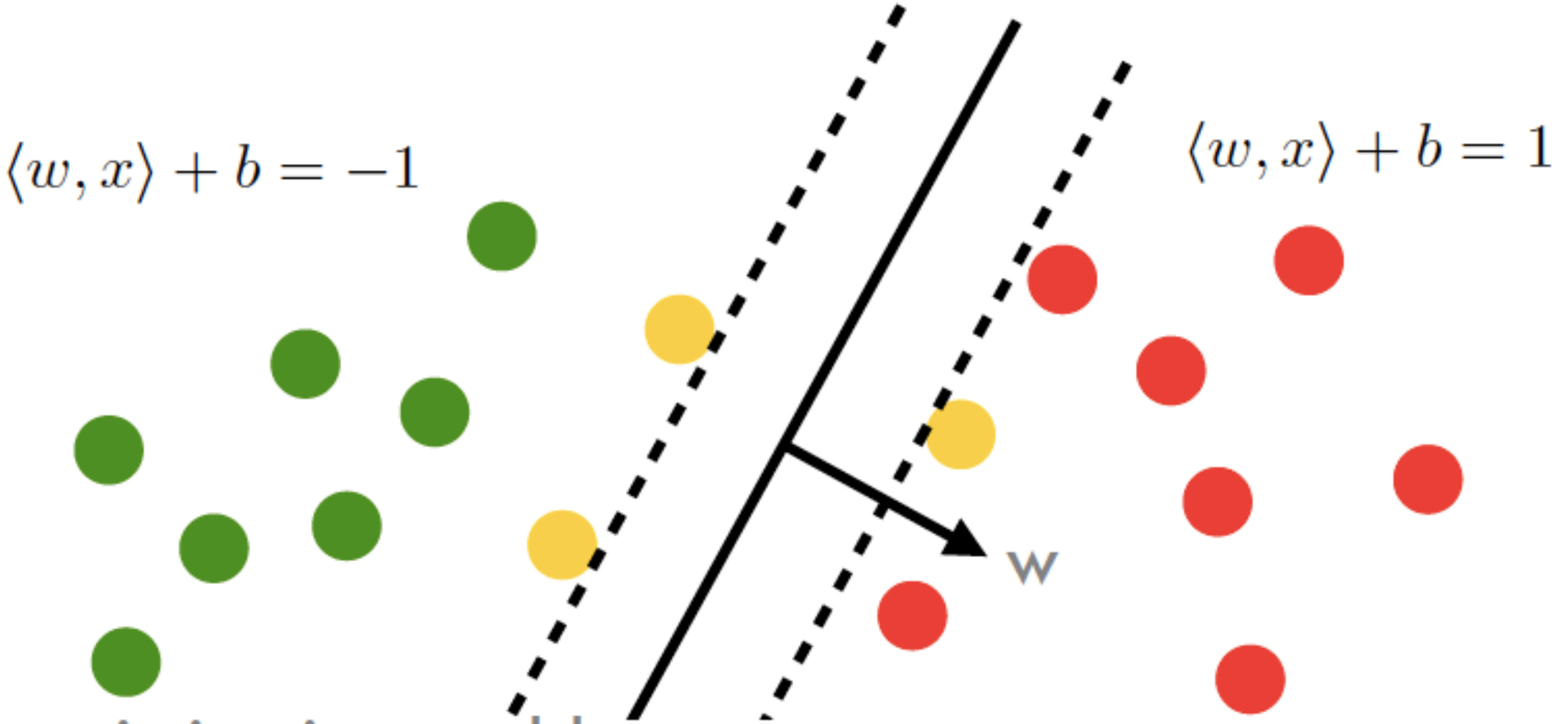
$$\underset{w,b}{\text{maximize}} \frac{1}{\|w\|} \text{ subject to } y_i \left[\langle x_i, w \rangle + b\right] \geq 1$$

Large Margin Classifier



optimization problem

Large Margin Classifier



optimization problem

$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2 \text{ subject to } y_i \left[\langle x_i, w \rangle + b\right] \geq 1$$

Dual Problem

- Primal optimization problem

Dual Problem

- **Primal optimization problem**

$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2 \text{ subject to } y_i \left[\langle x_i, w \rangle + b\right] \geq 1$$

Dual Problem

- **Primal optimization problem**

$$\underset{w,b}{\text{minimize}} \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$

- **Lagrange function**

constraint

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i\left[\langle x_i, w\rangle + b\right] - 1\right]$$

Dual Problem

- Primal optimization problem

$$\underset{w,b}{\text{minimize}} \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$

- Lagrange function

constraint

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i\left[\langle x_i, w\rangle + b\right] - 1\right]$$

Optimality in w, b is at saddle point with α

- Derivatives in w, b need to vanish

Dual Problem

- **Lagrange function**

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \left[y_i \left[\langle x_i, w \rangle + b \right] - 1 \right]$$

Dual Problem

- Lagrange function

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_i \alpha_i \left[y_i \left[\langle x_i, w \rangle + b \right] - 1 \right]$$

- Derivatives in w, b need to vanish

$$\partial_w L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0$$

Dual Problem

- Lagrange function

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i \left[\langle x_i, w\rangle + b\right] - 1\right]$$

- Derivatives in w, b need to vanish

$$\partial_w L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0$$

$$\partial_b L(w, b, a) = \sum_i \alpha_i y_i = 0$$

Dual Problem

- Lagrange function

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i \left[\langle x_i, w \rangle + b\right] - 1\right]$$

- Derivatives in w, b need to vanish

$$\partial_w L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0$$

$$\partial_b L(w, b, a) = \sum_i \alpha_i y_i = 0$$

Dual Problem

- Lagrange function

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left[y_i \left[\langle x_i, w\rangle + b\right] - 1\right]$$

- Derivatives in w, b need to vanish

$$\partial_w L(w, b, a) = w - \sum_i \alpha_i y_i x_i = 0$$

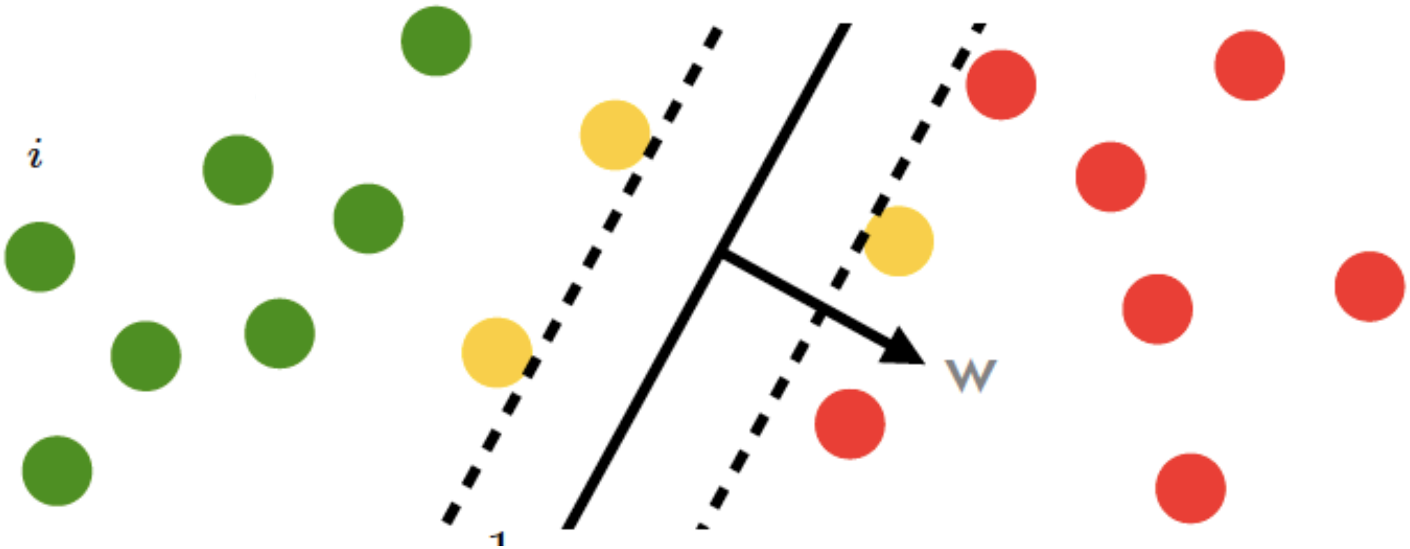
$$\partial_b L(w, b, a) = \sum_i \alpha_i y_i = 0$$

- Plugging terms back into L yields

$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i$$

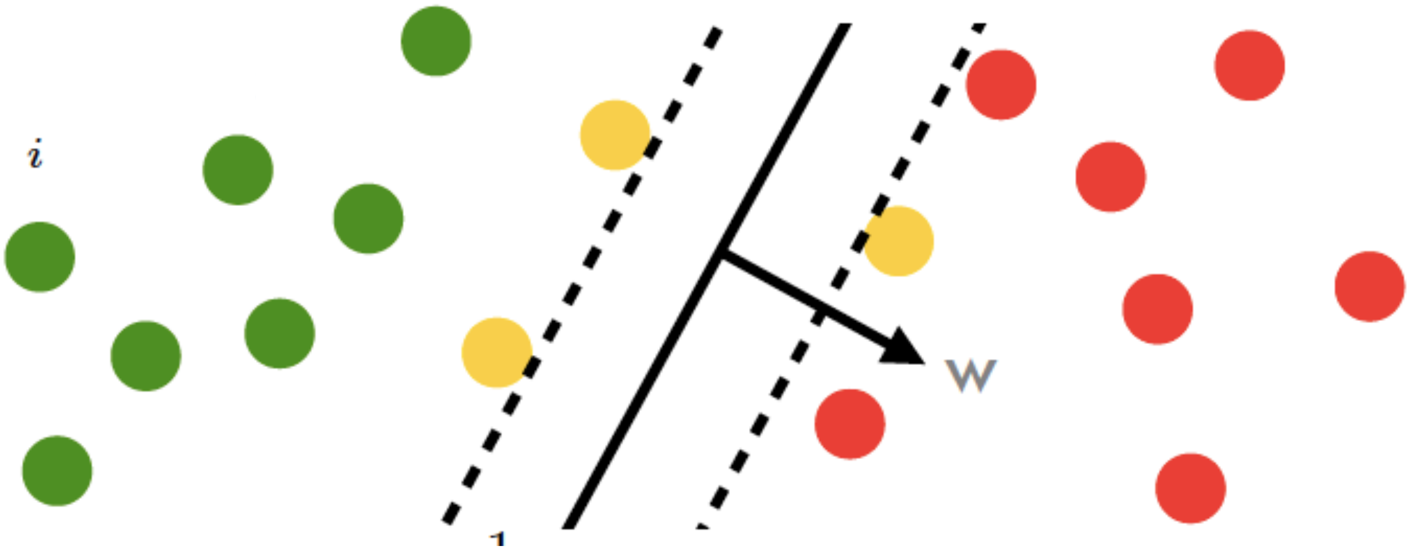
$$\text{subject to } \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

Support Vector Machines



Support Vector Machines

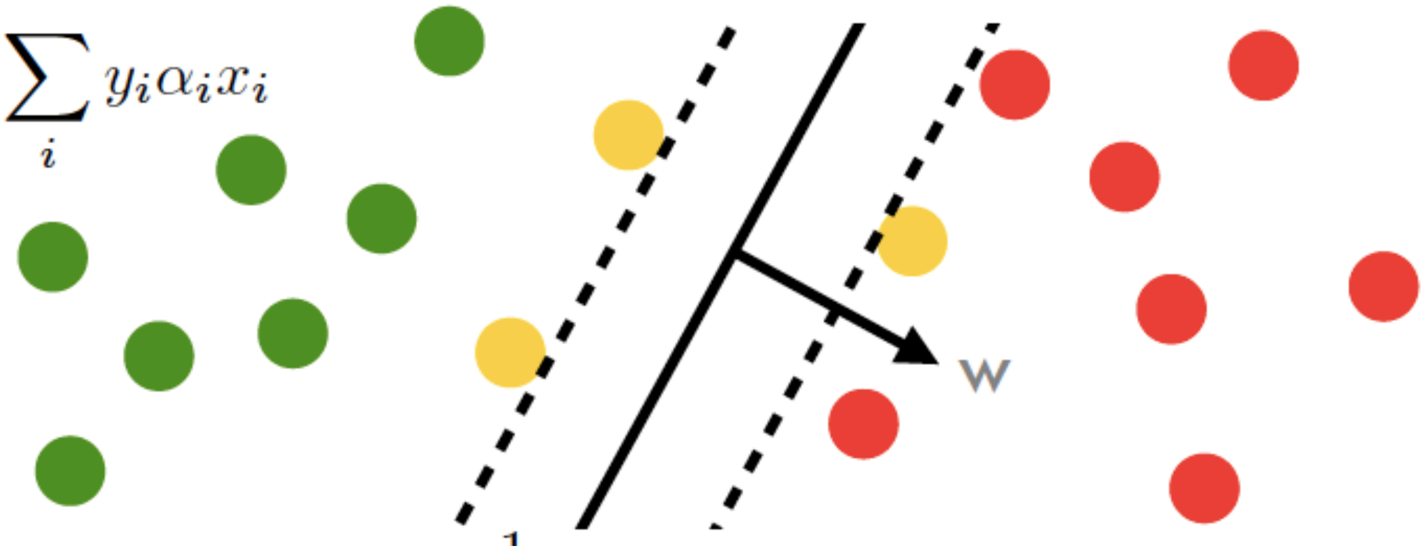
$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$



Support Vector Machines

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \ \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$

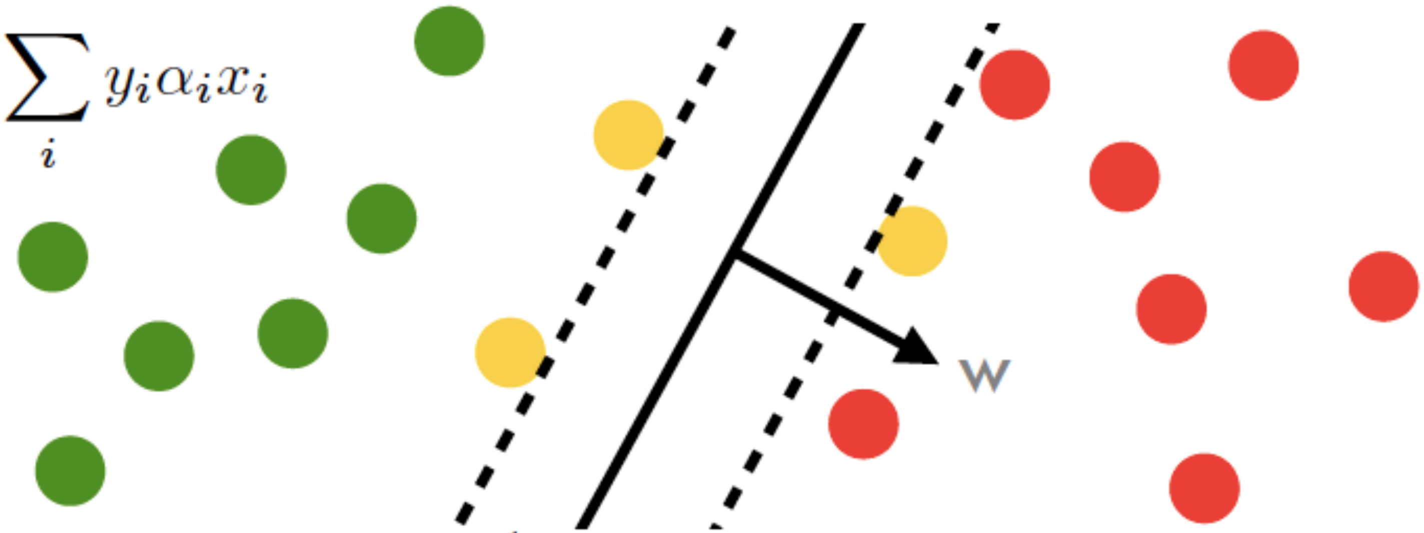
$$w = \sum_i y_i \alpha_i x_i$$



Support Vector Machines

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \ \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$

$$w = \sum_i y_i \alpha_i x_i$$

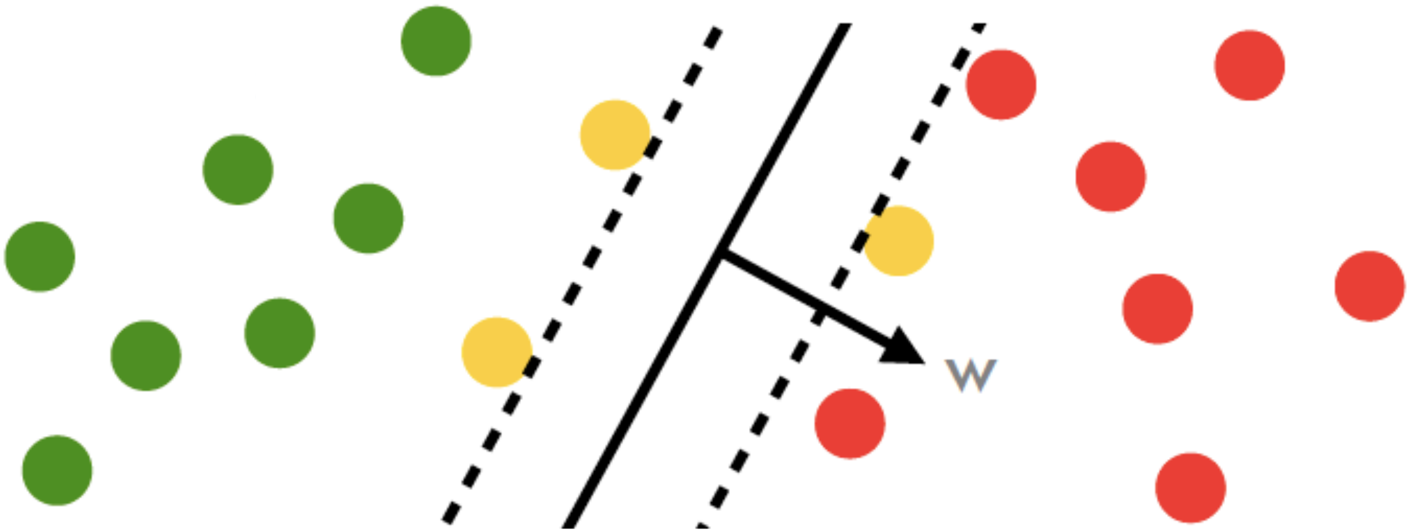


$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

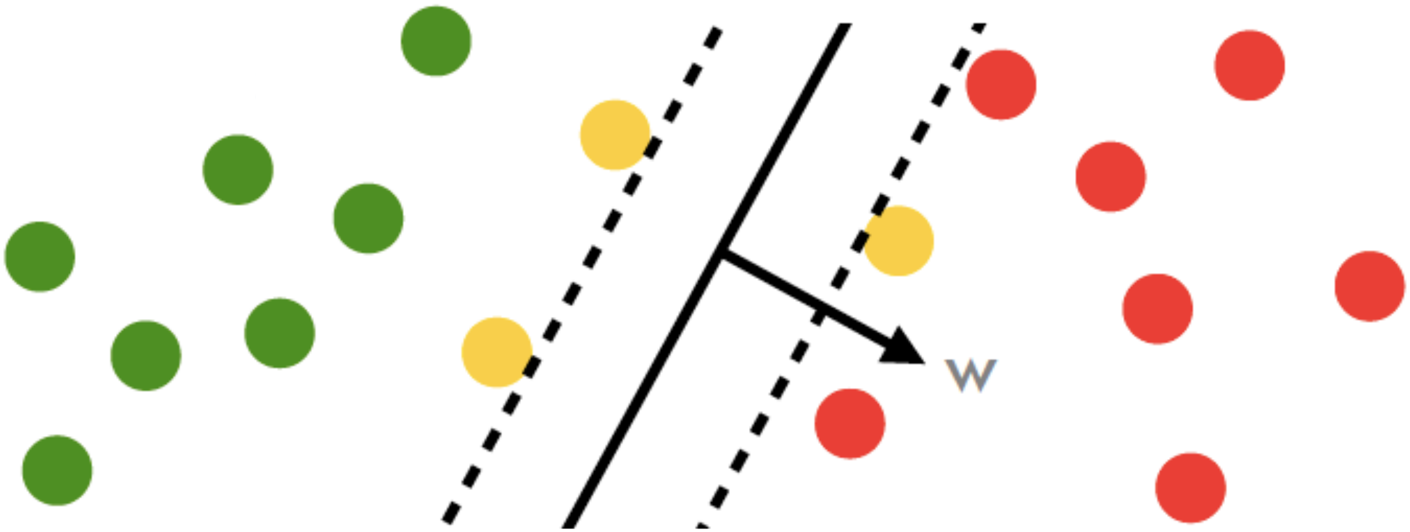
Five minutes break

Support Vector Machines



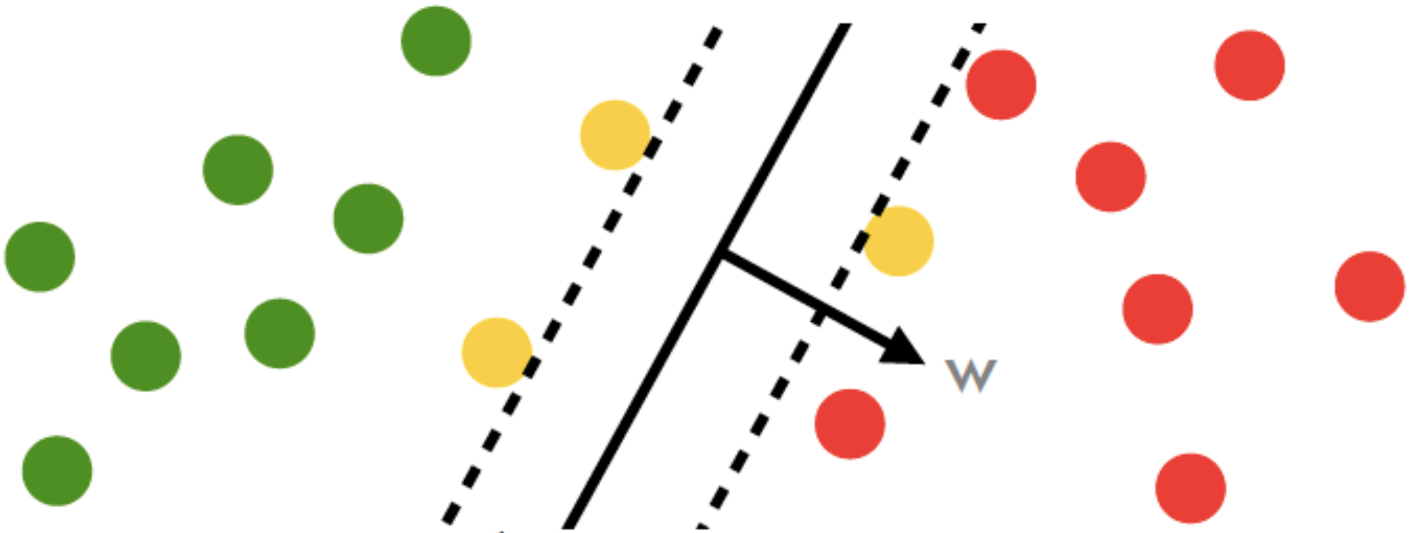
Support Vector Machines

$$\underset{w.b}{\text{minimize}} \frac{1}{2} \|w\|^2 \text{ subject to } y_i \left[\langle x_i, w \rangle + b\right] \geq 1$$



Support Vector Machines

$$\underset{w.b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$



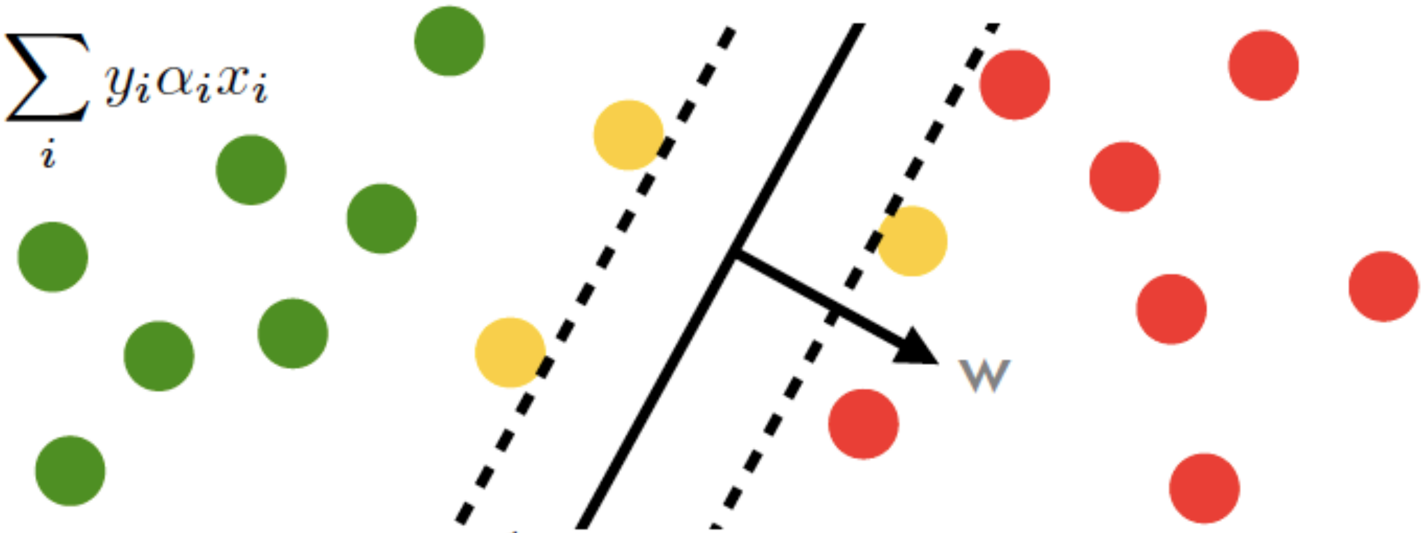
$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j\right\rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

Support Vector Machines

$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2 \text{ subject to } y_i \left[\langle x_i, w \rangle + b\right] \geq 1$$

$$w = \sum_i y_i \alpha_i x_i$$



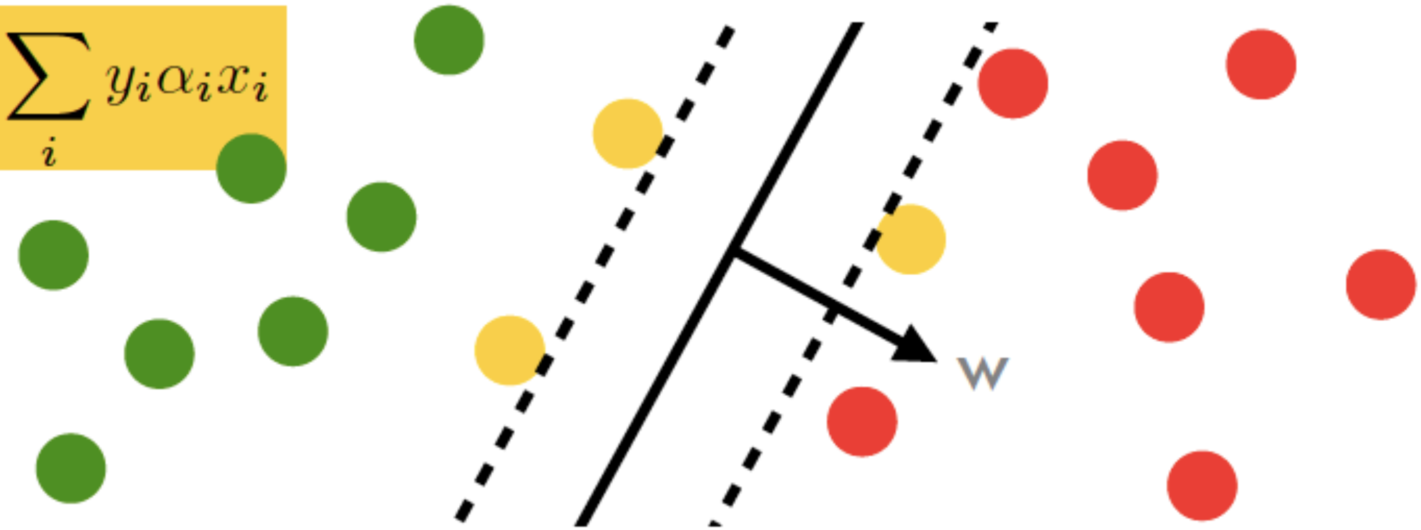
$$\underset{\alpha}{\text{maximize}} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0$$

Support Vectors

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \ \text{ subject to } y_i\left[\langle x_i, w\rangle + b\right] \geq 1$$

$$w = \sum_i y_i \alpha_i x_i$$



Karush Kuhn Tucker
Optimality condition

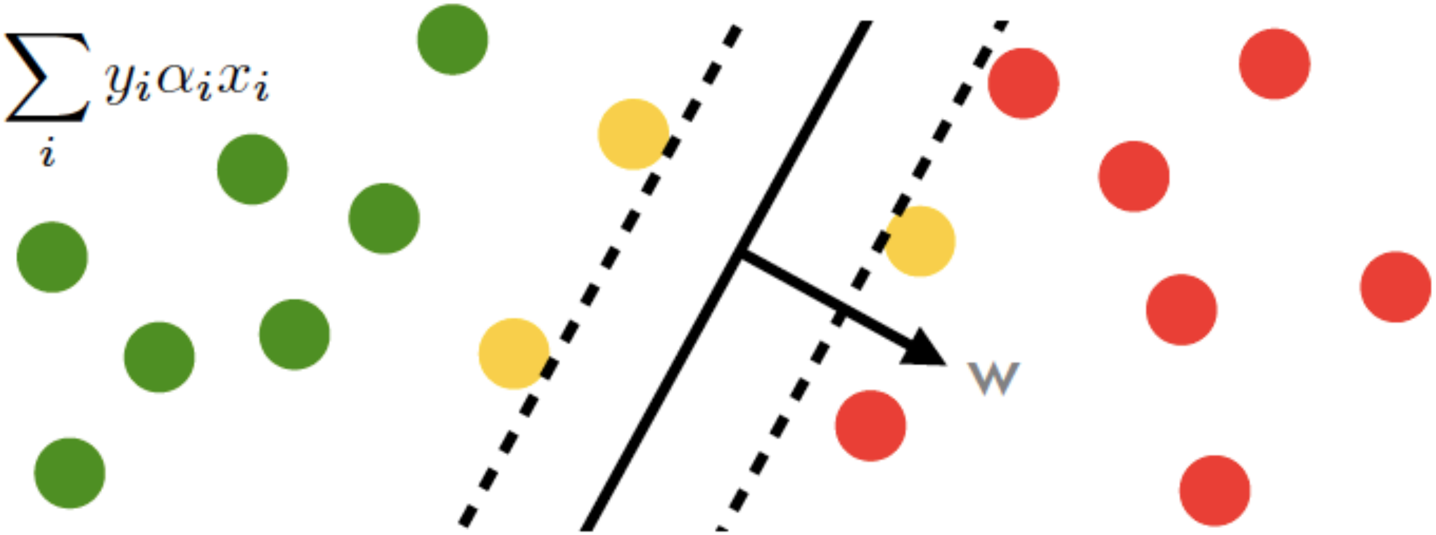
$$\alpha_i\left[y_i\left[\langle w, x_i\rangle + b\right] - 1\right] = 0$$

$$\alpha_i = 0$$

$$\alpha_i > 0 \Longrightarrow y_i\left[\langle w, x_i\rangle + b\right] = 1$$

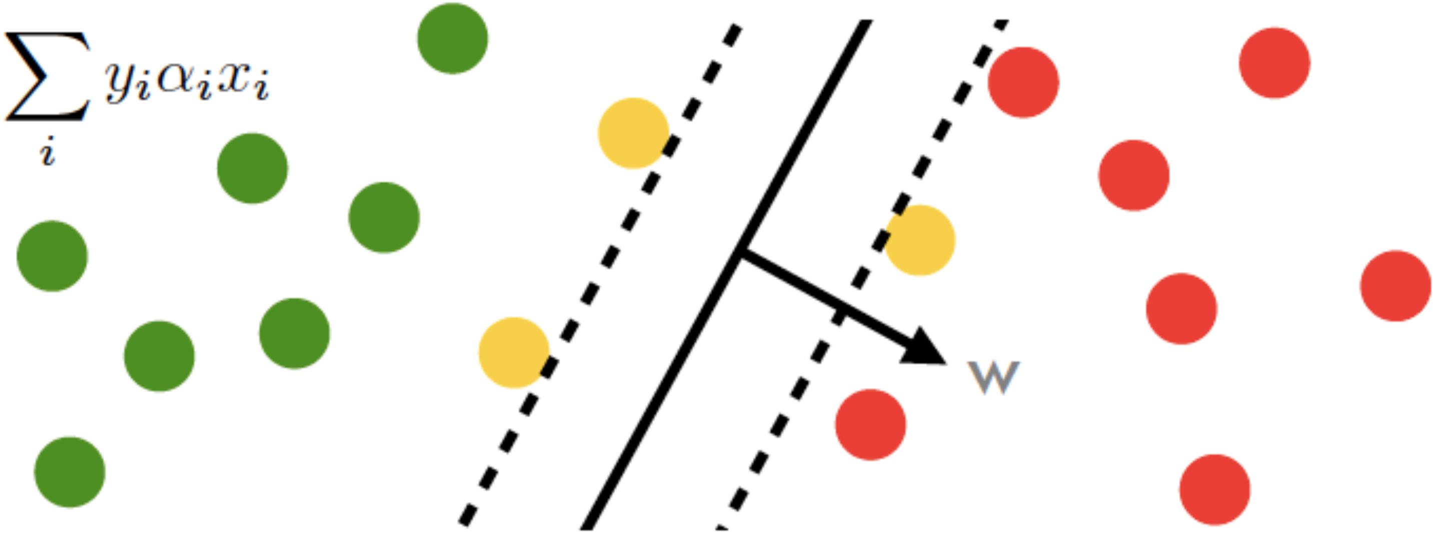
Properties

$$w = \sum_i y_i \alpha_i x_i$$



Properties

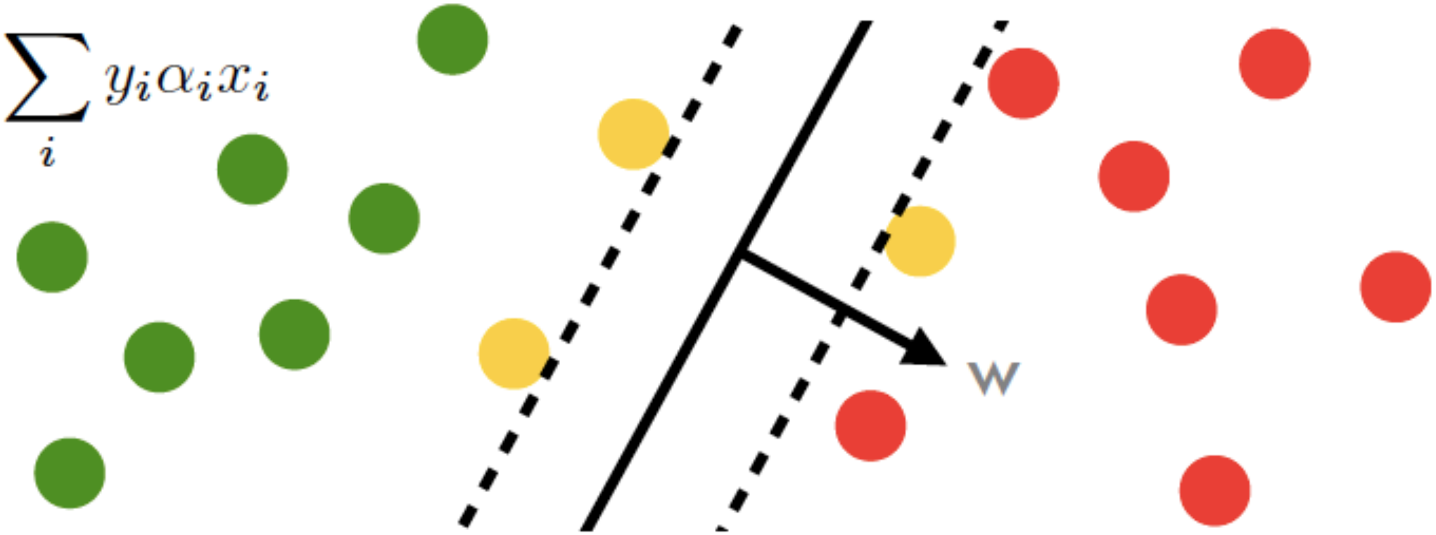
$$w = \sum_i y_i \alpha_i x_i$$



- Weight vector w as weighted linear combination of instances

Properties

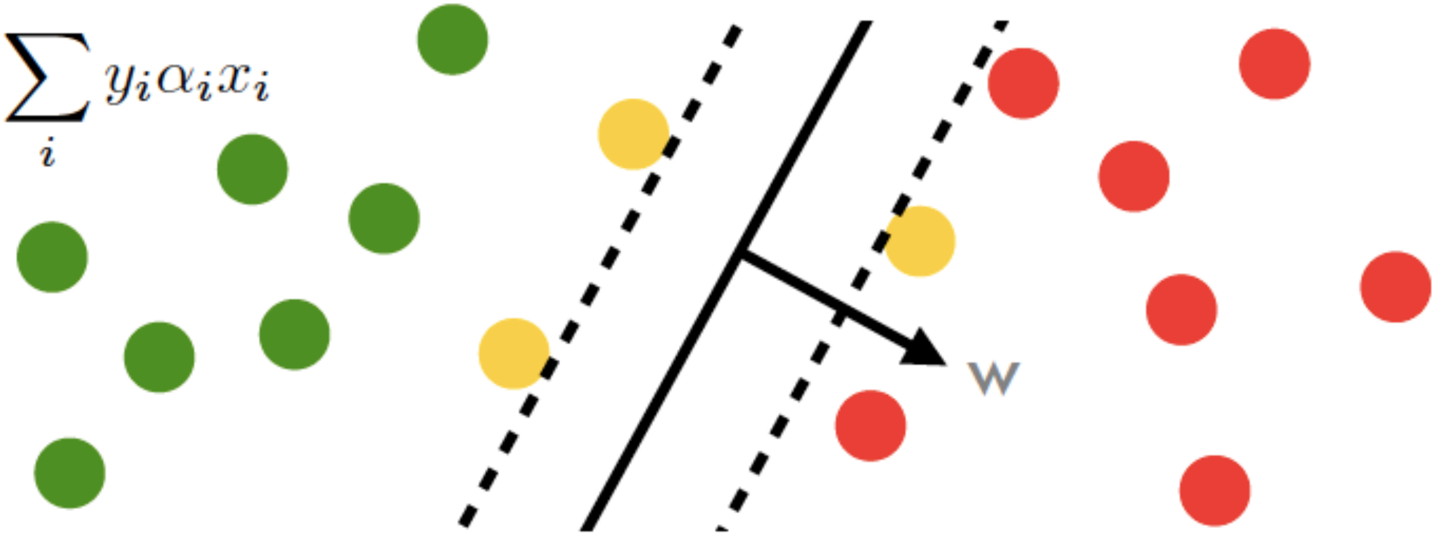
$$w = \sum_i y_i \alpha_i x_i$$



- Weight vector w as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)

Properties

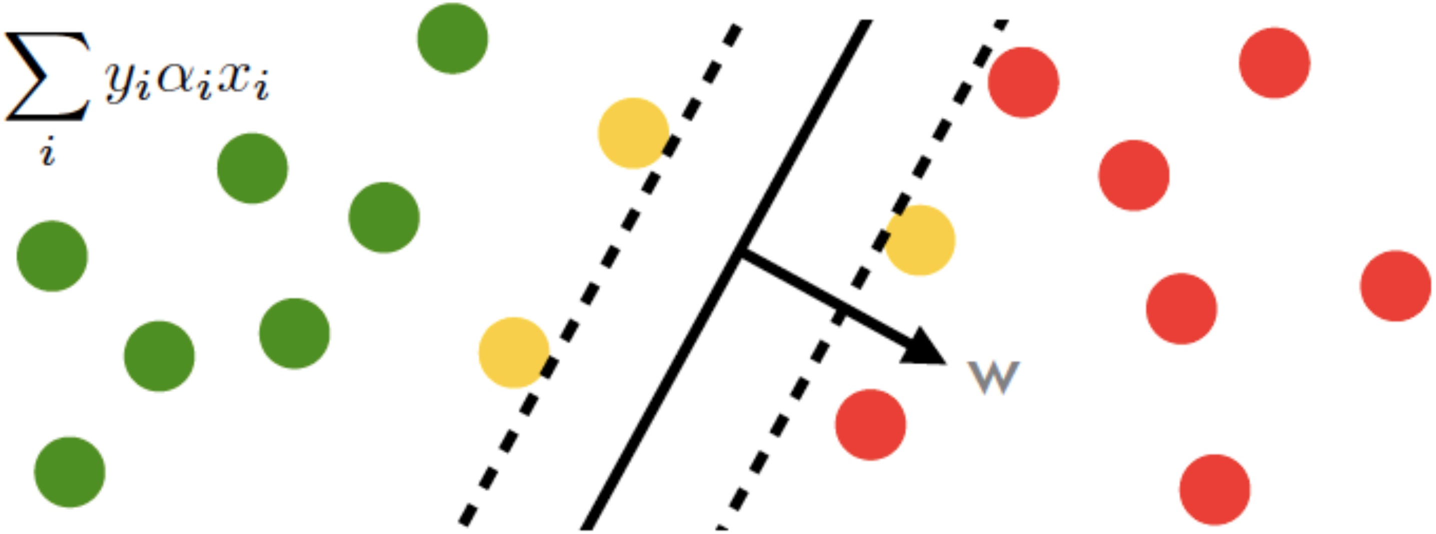
$$w = \sum_i y_i \alpha_i x_i$$



- Weight vector w as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)
- Only inner products matter

Properties

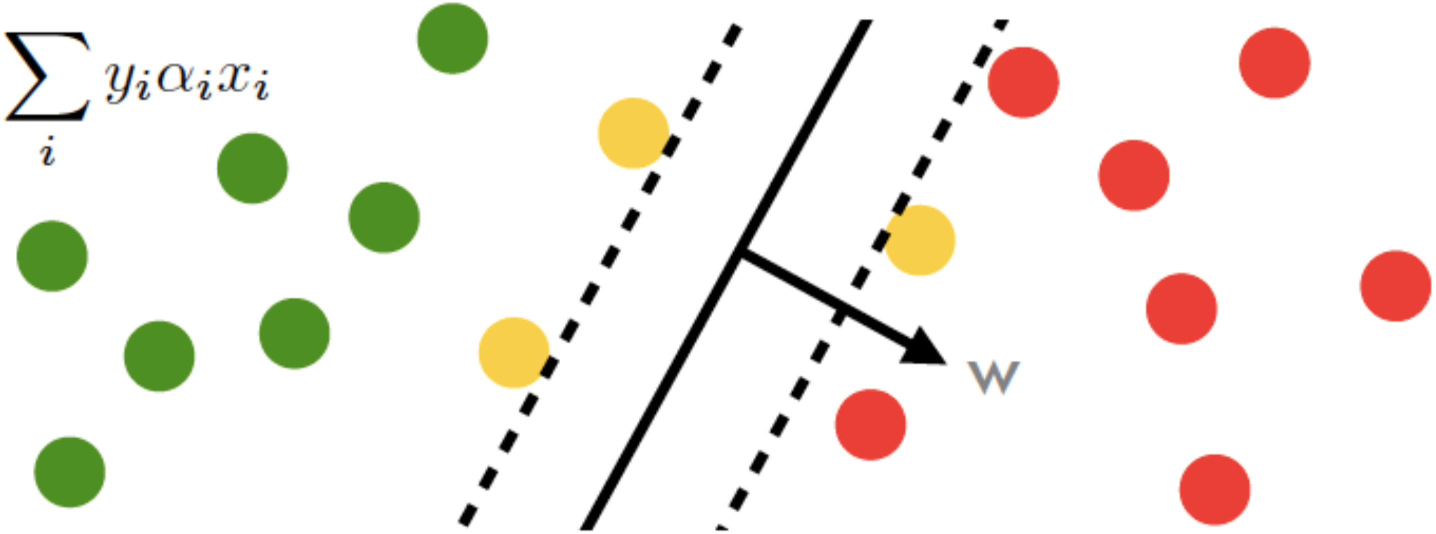
$$w = \sum_i y_i \alpha_i x_i$$



- Weight vector w as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)
- Only inner products matter
  - Quadratic program

Properties

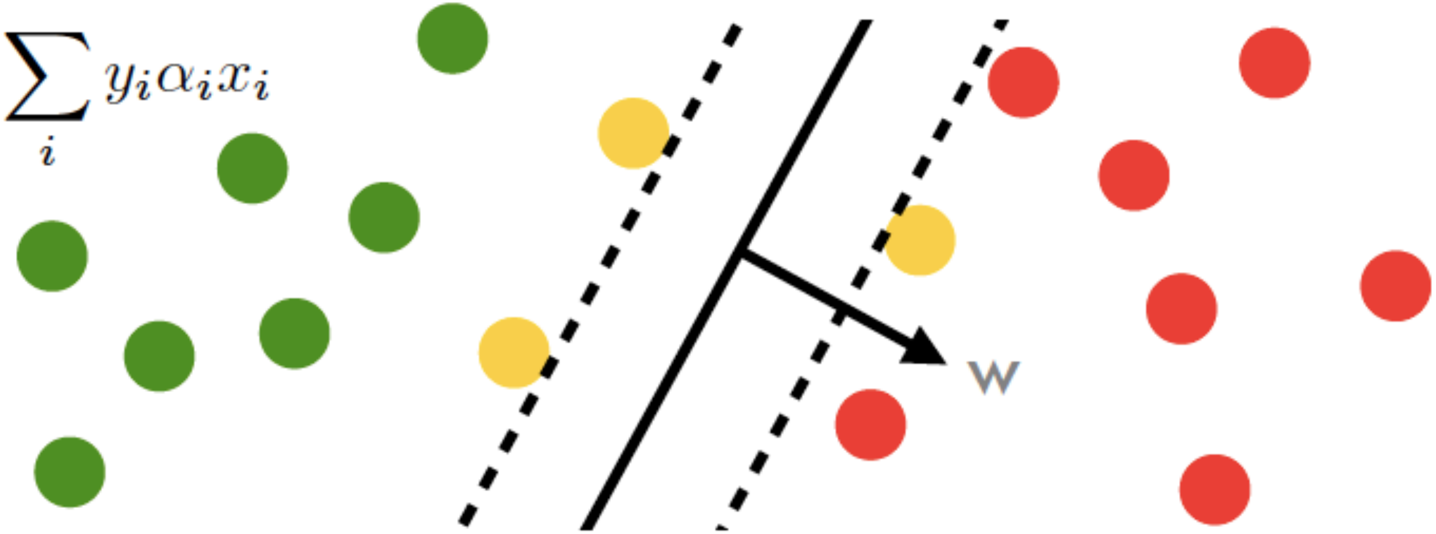
$$w = \sum_i y_i \alpha_i x_i$$



- Weight vector w as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)
- Only inner products matter
  - Quadratic program
  - We can replace the inner product by a kernel

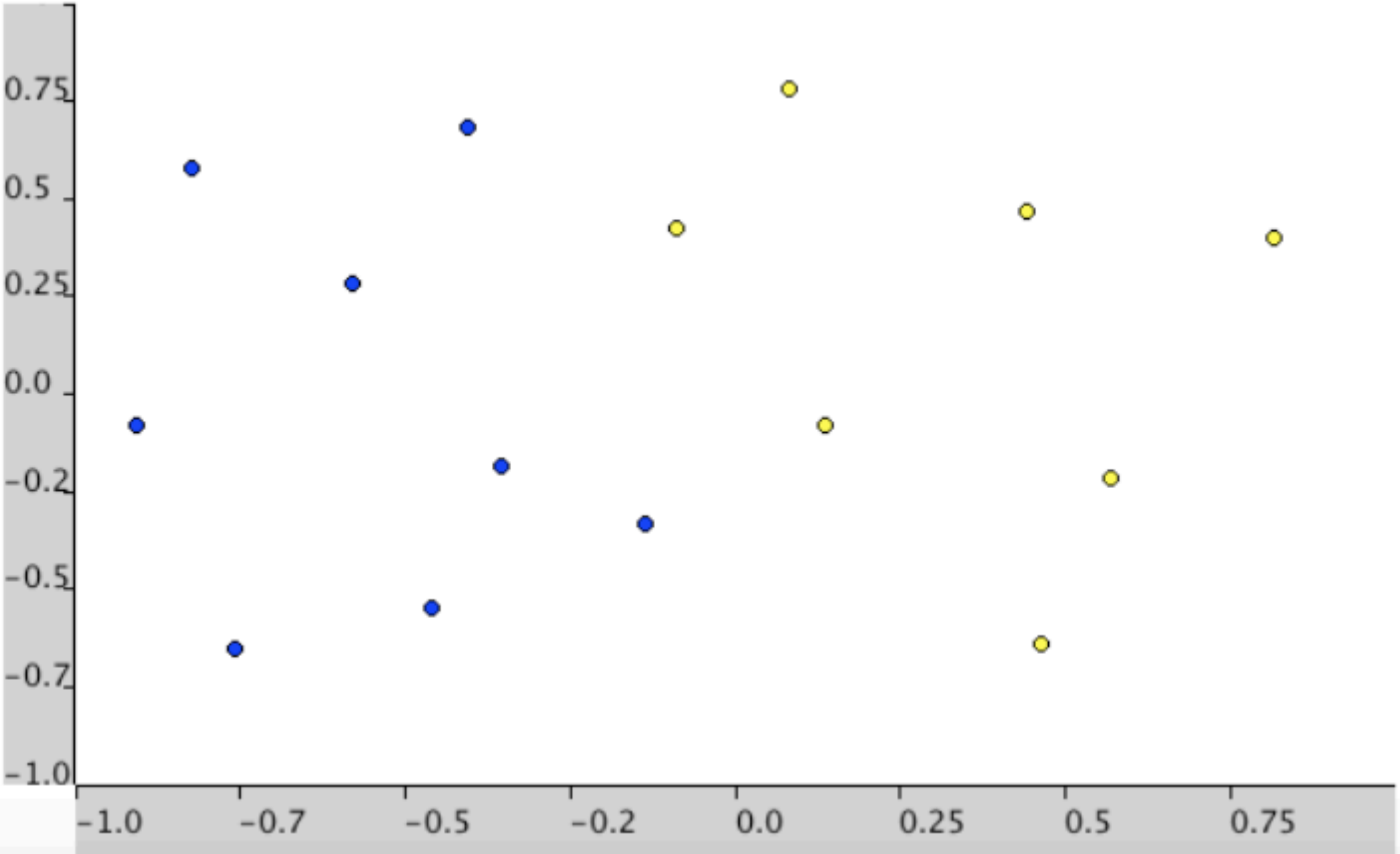
Properties

$$w = \sum_i y_i \alpha_i x_i$$



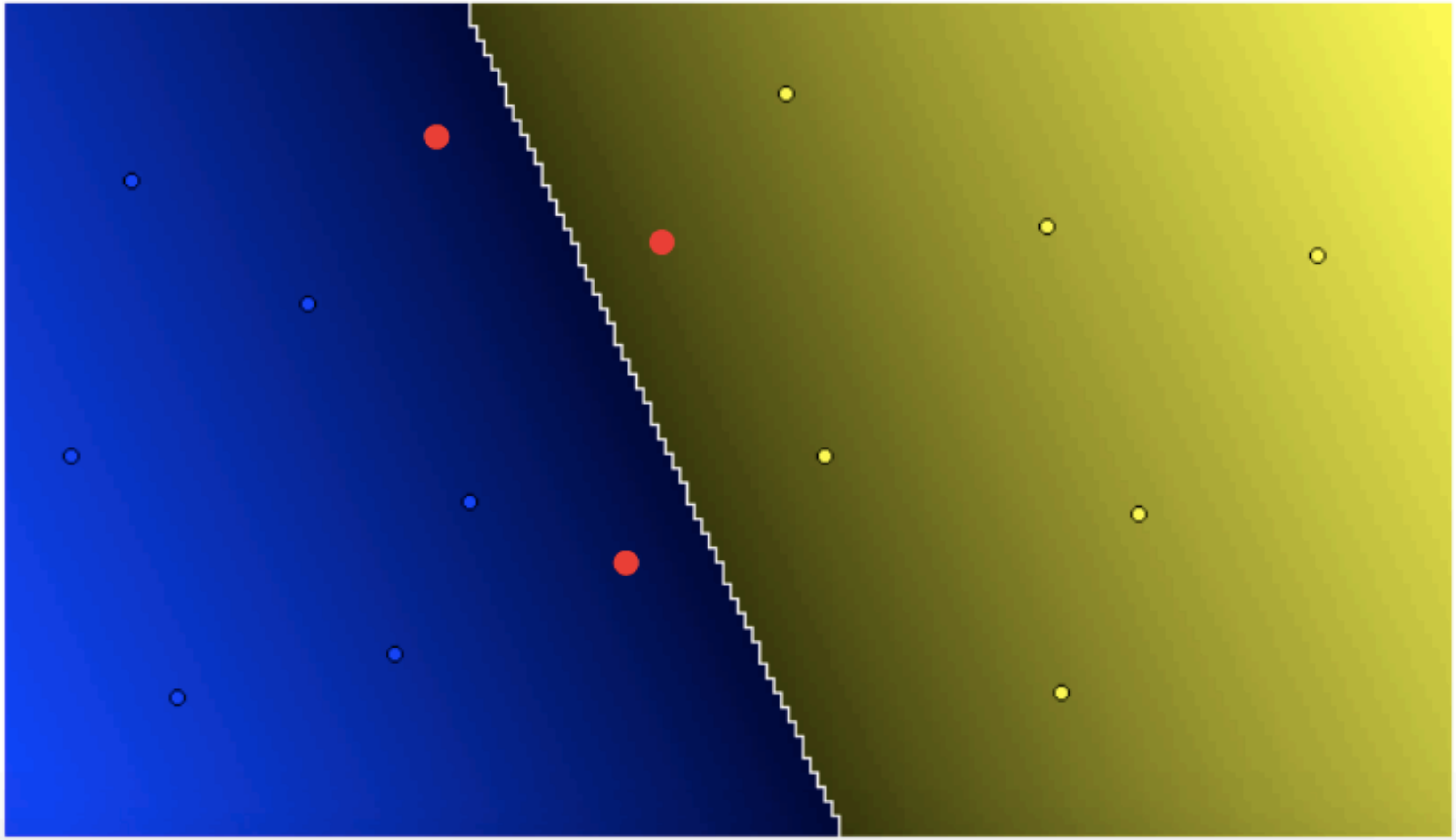
- Weight vector w as weighted linear combination of instances
- Only points on margin matter (ignore the rest and get same solution)
- Only inner products matter
  - Quadratic program
  - We can replace the inner product by a kernel
- Keeps instances away from the margin

Example

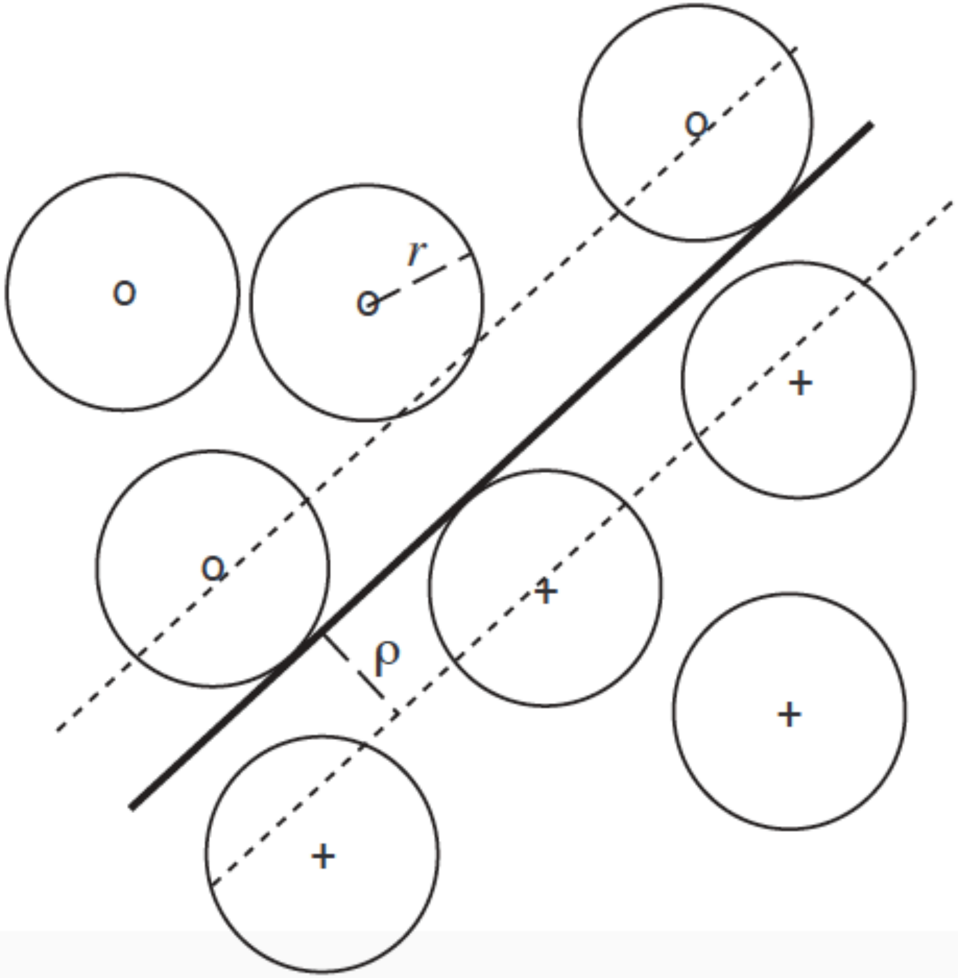


Example

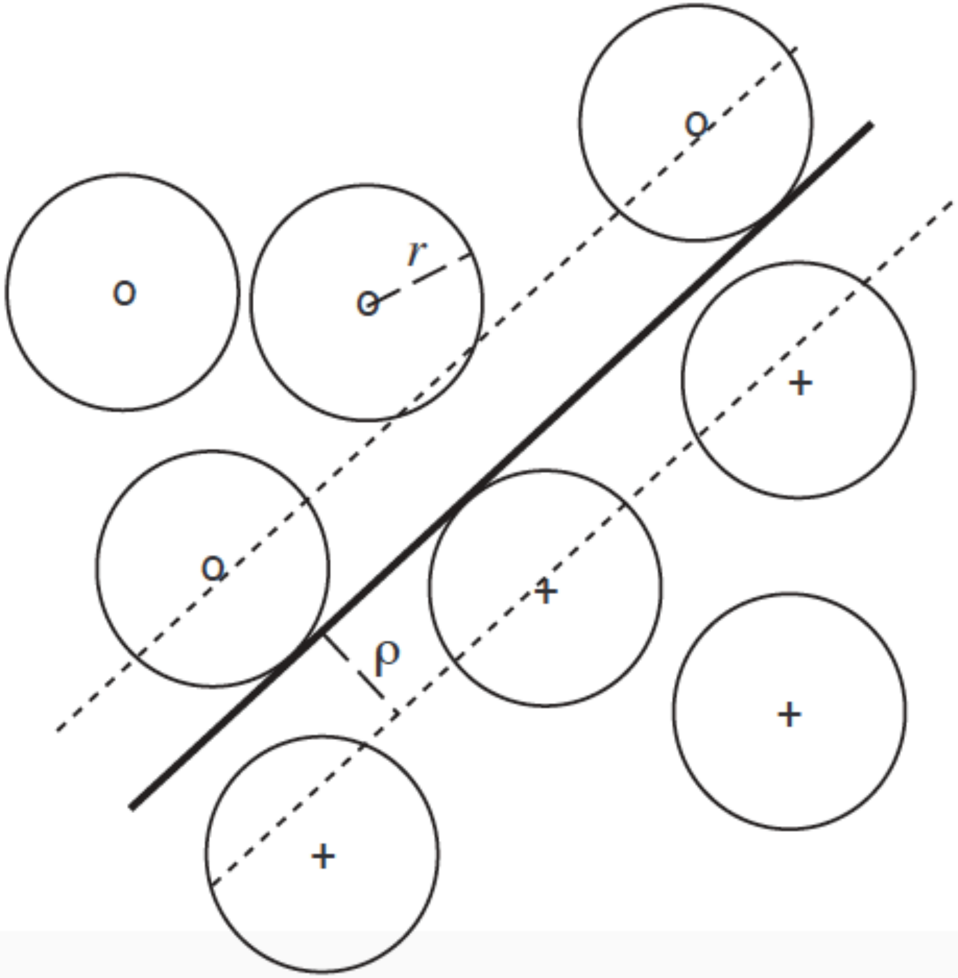
Number of Support Vectors: **3** (-ve: 2, +ve: 1) Total number of points: 15



Why large margins?

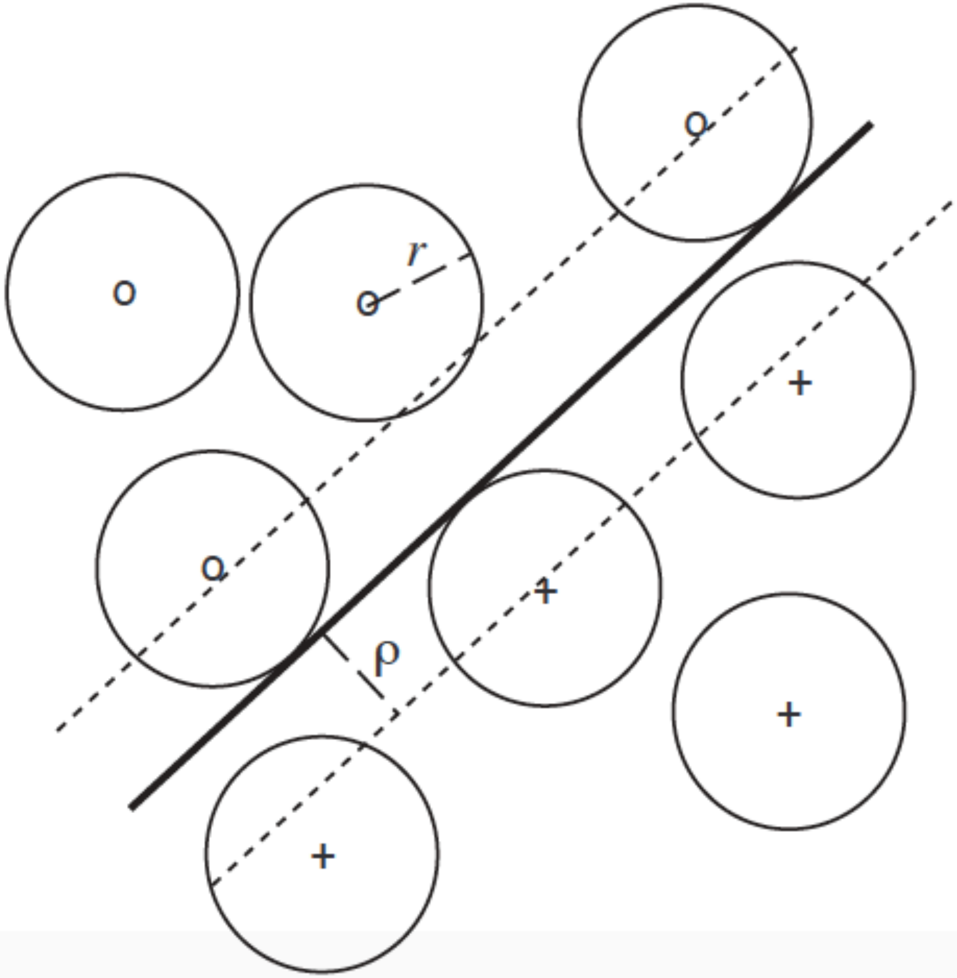


Why large margins?



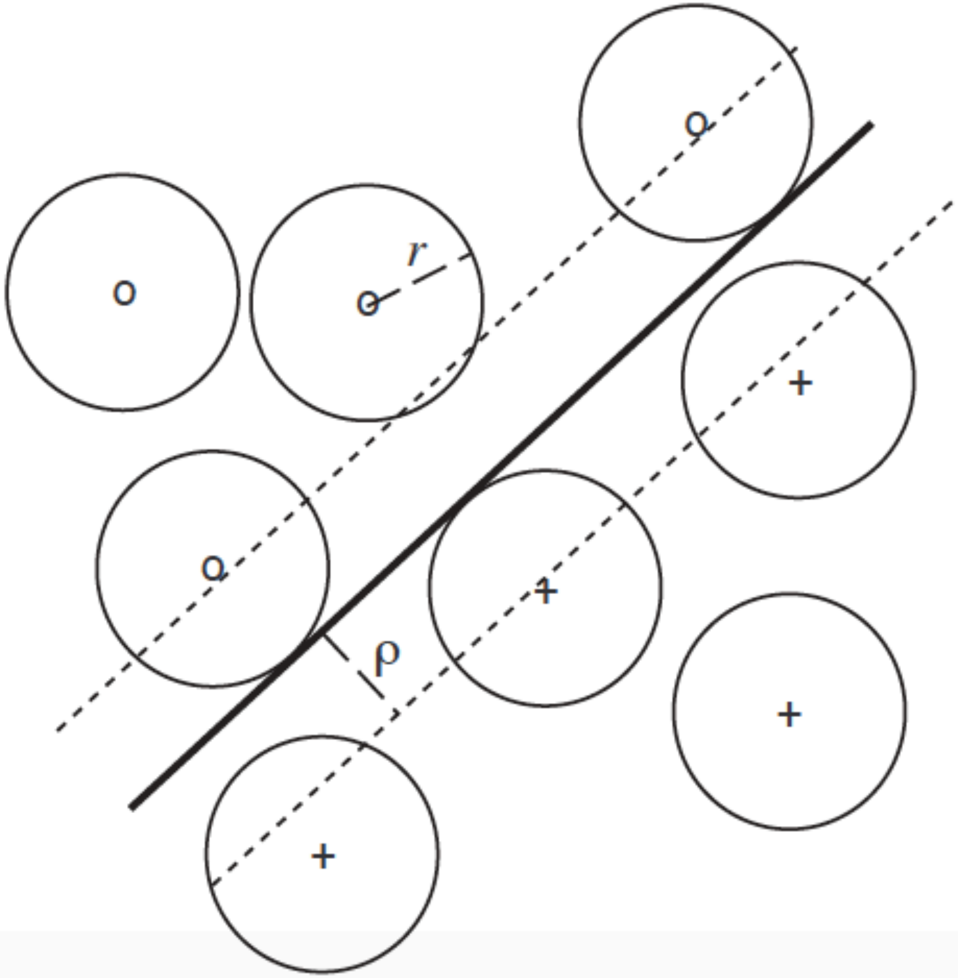
- Maximum robustness relative to uncertainty

Why large margins?



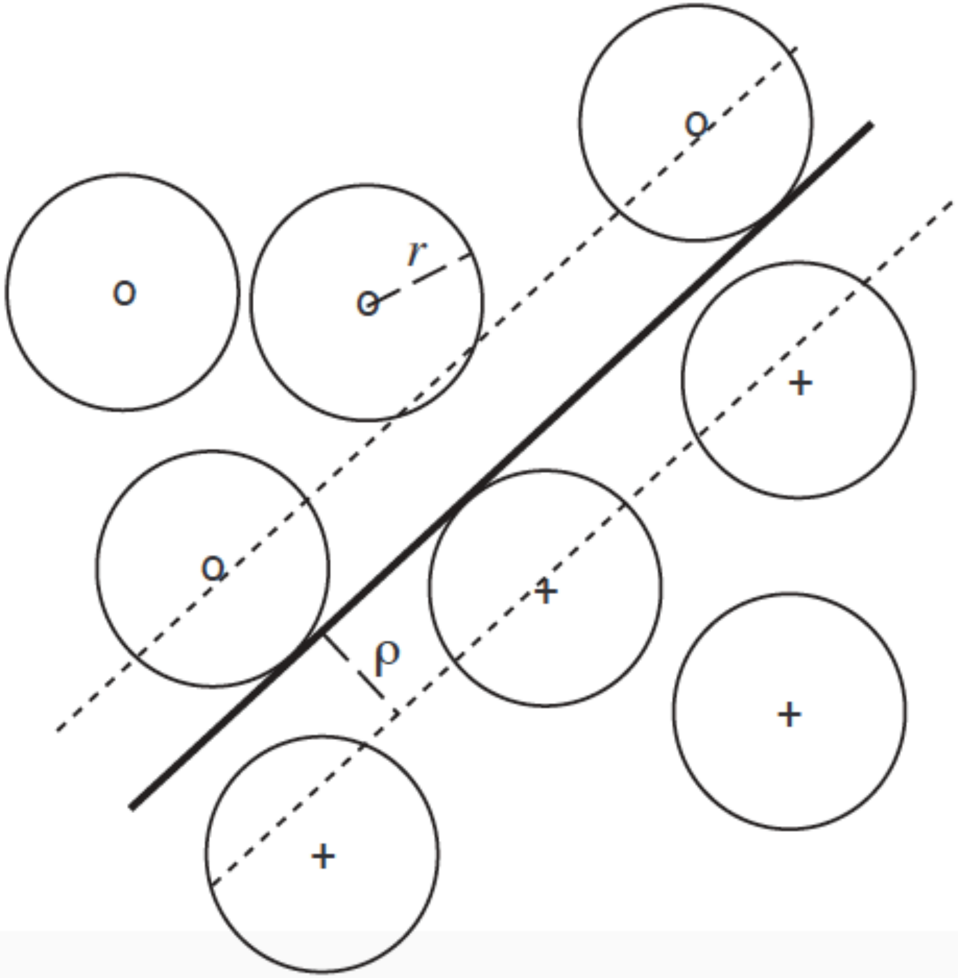
- Maximum robustness relative to uncertainty
- Symmetry breaking

Why large margins?



- Maximum robustness relative to uncertainty
- Symmetry breaking
- Independent of correctly classified instances

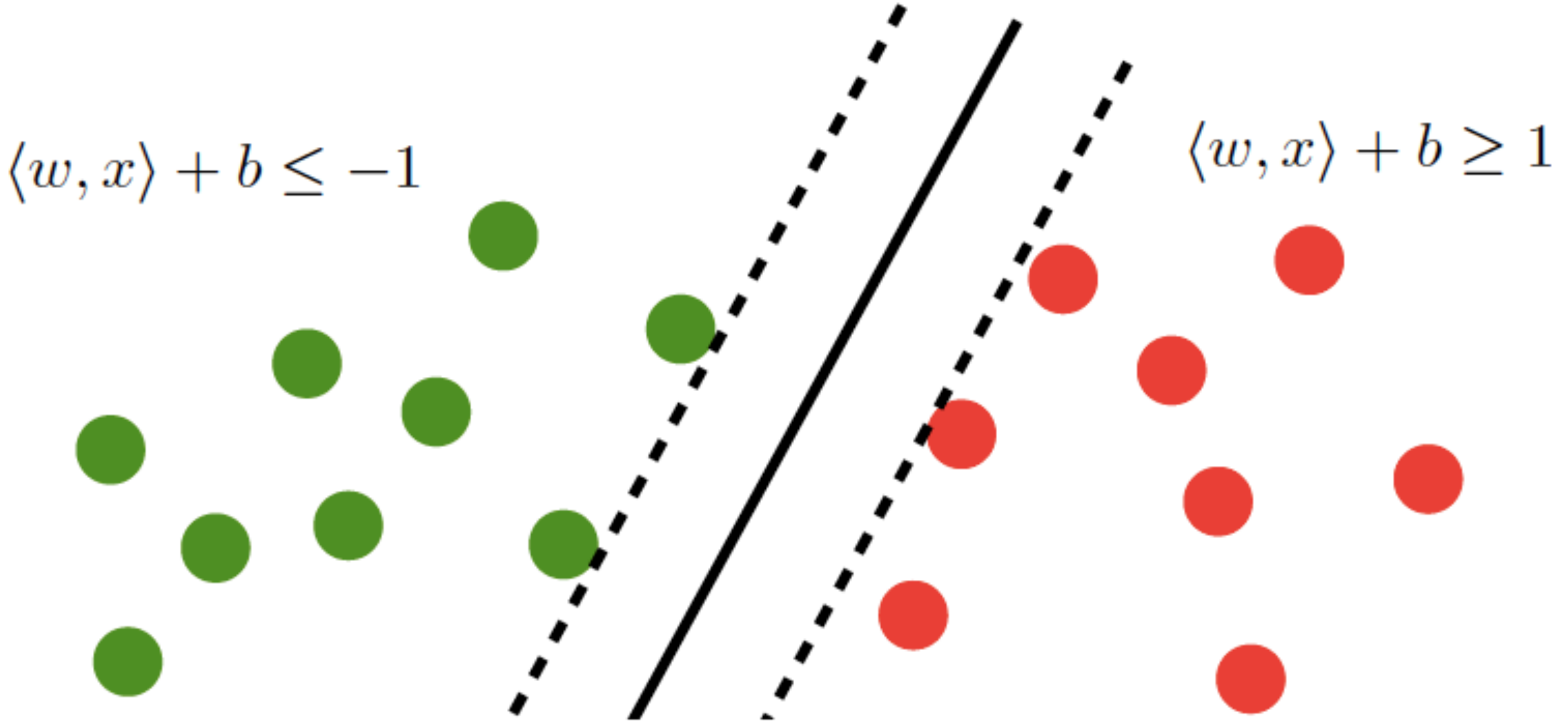
Why large margins?



- Maximum robustness relative to uncertainty
- Symmetry breaking
- Independent of correctly classified instances
- Easy to find for easy problems

Soft Margin Classifiers

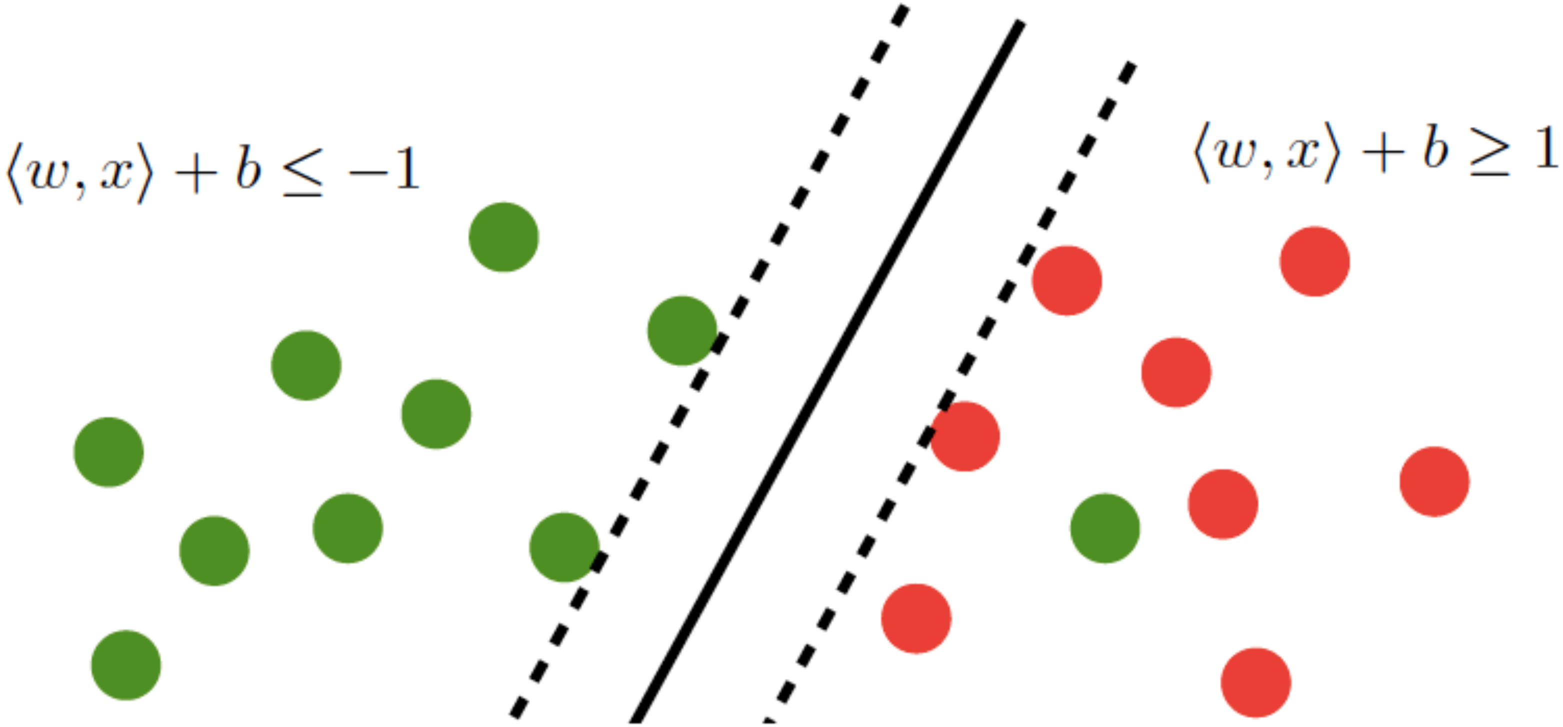
Large Margin Classifier



linear function

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

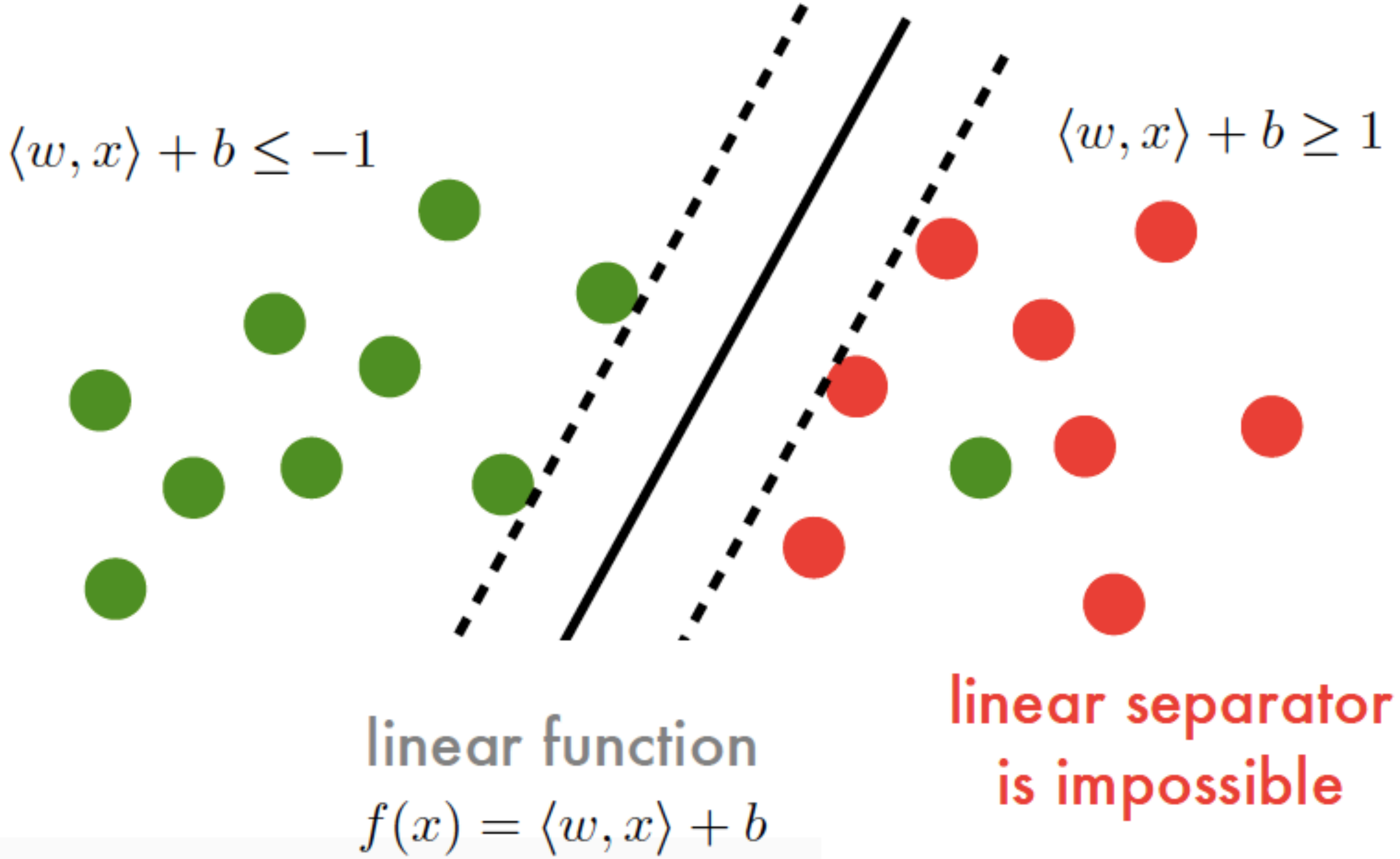
Large Margin Classifier



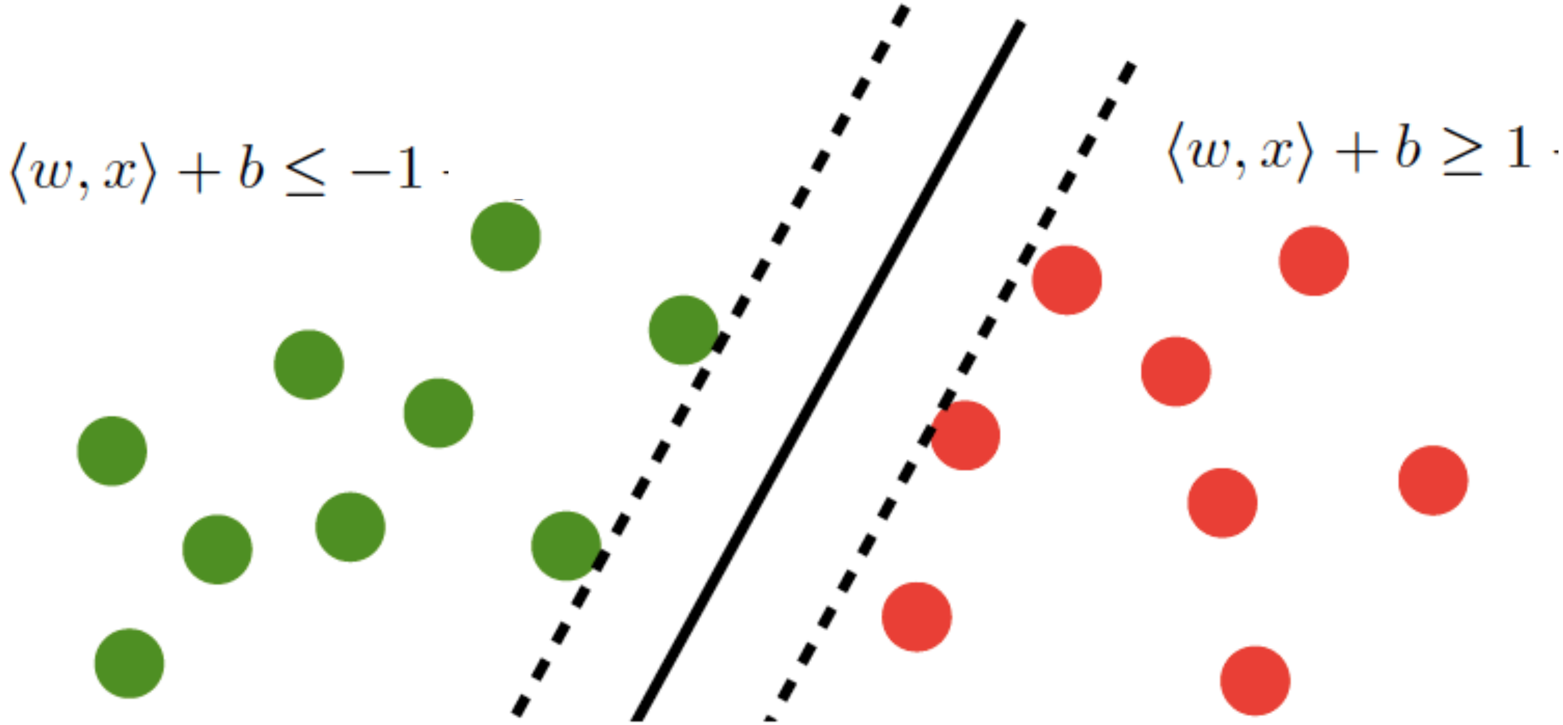
linear function

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

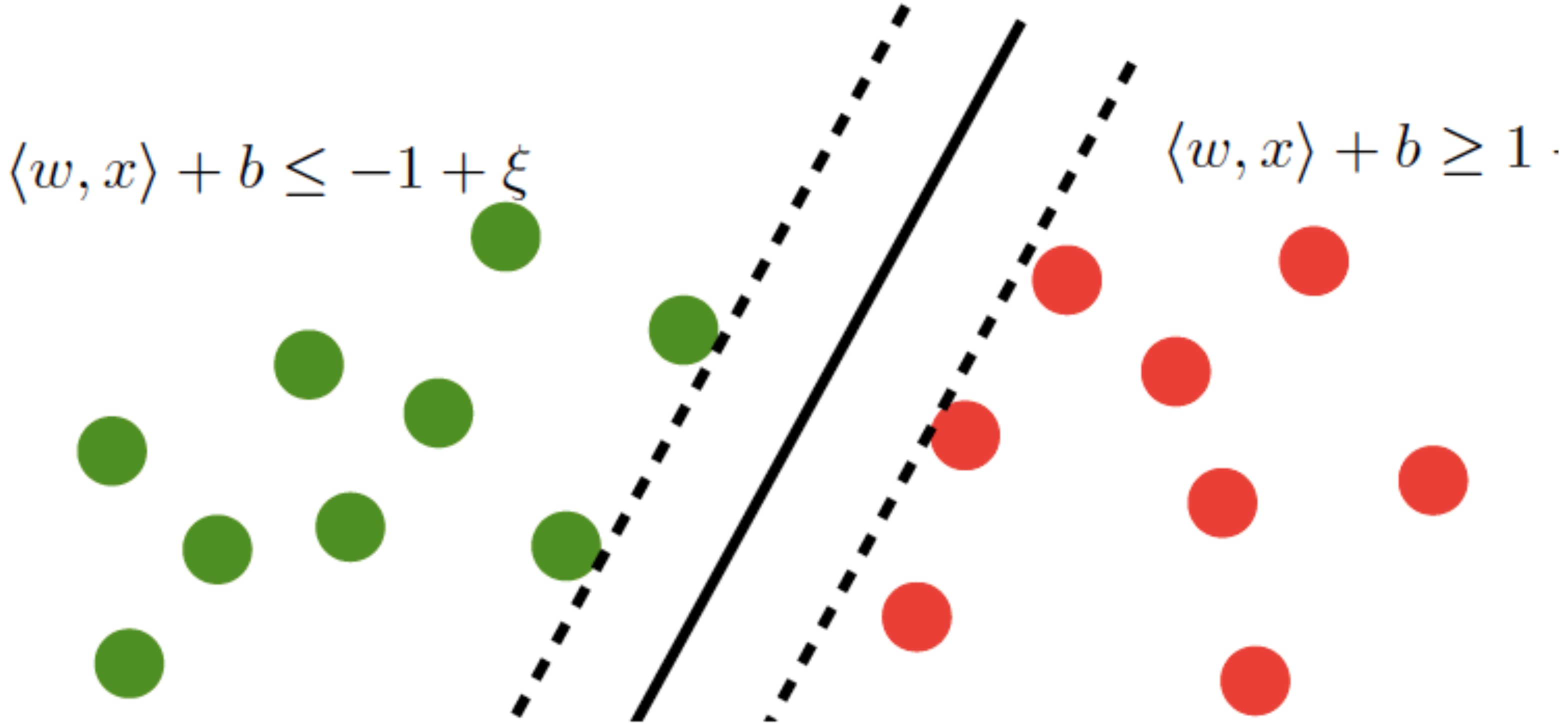
Large Margin Classifier



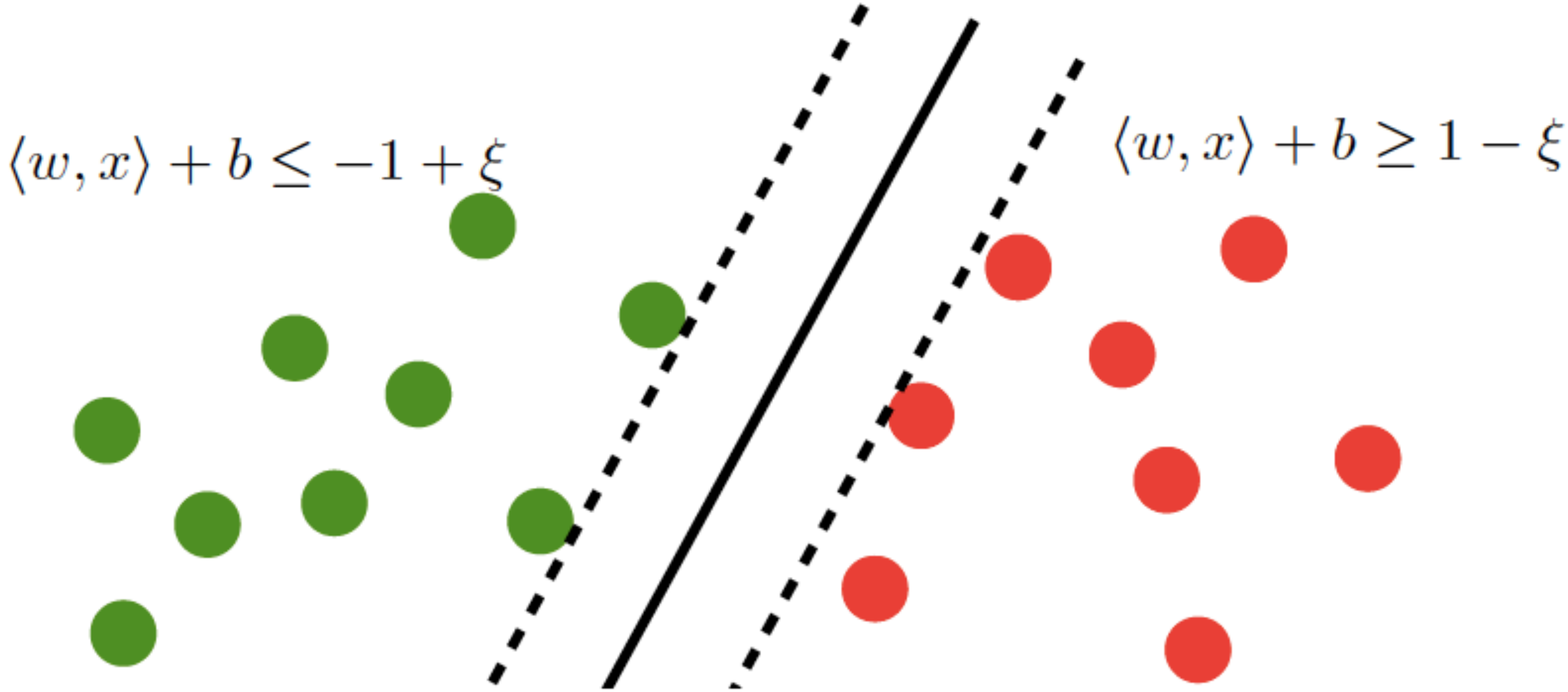
Adding slack variables



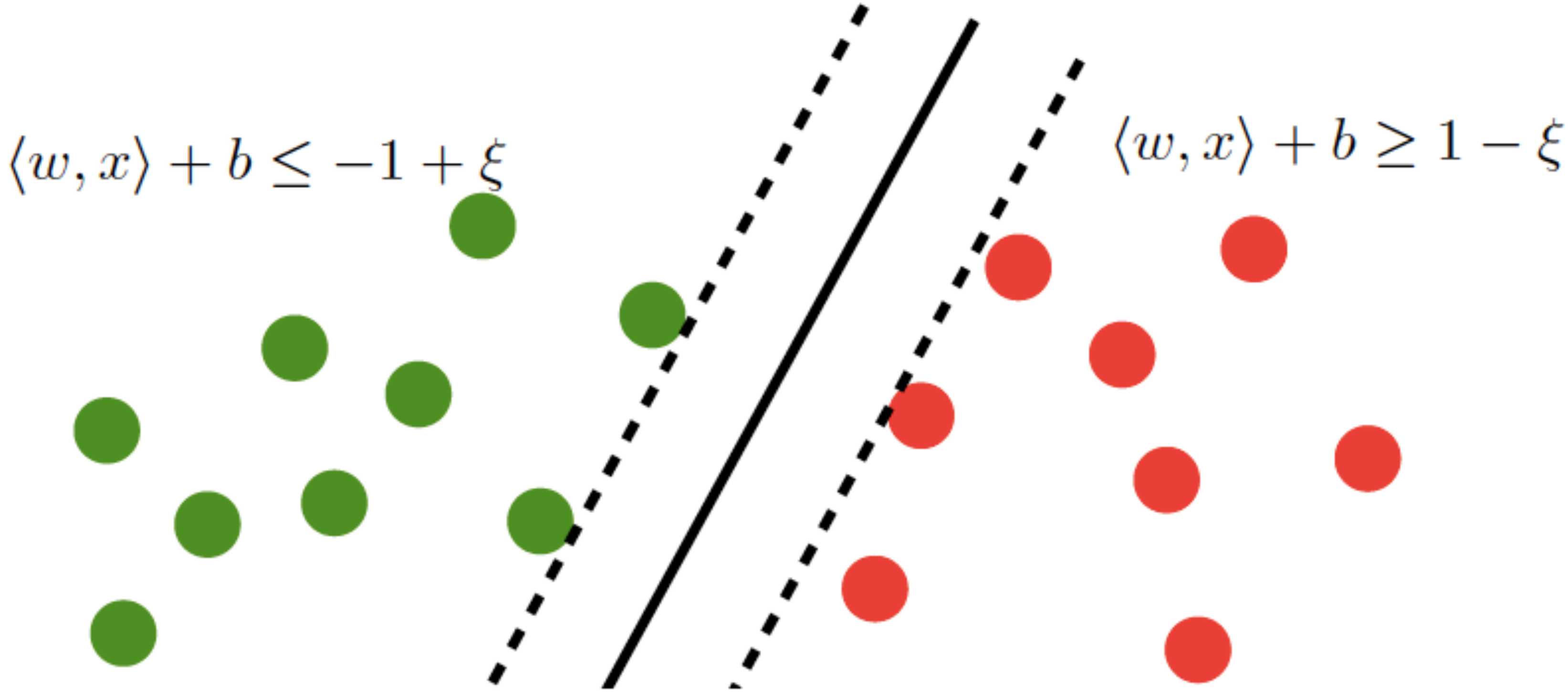
Adding slack variables



Adding slack variables

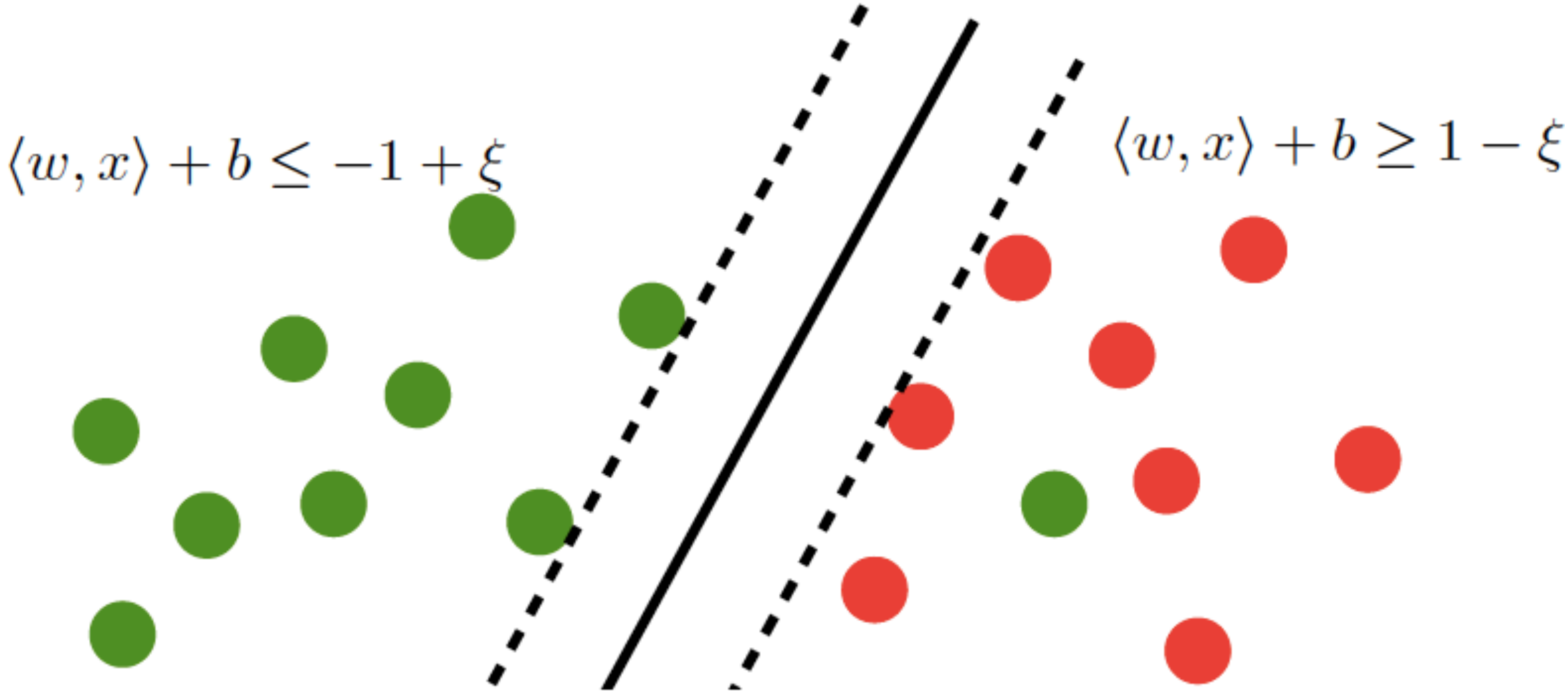


Adding slack variables



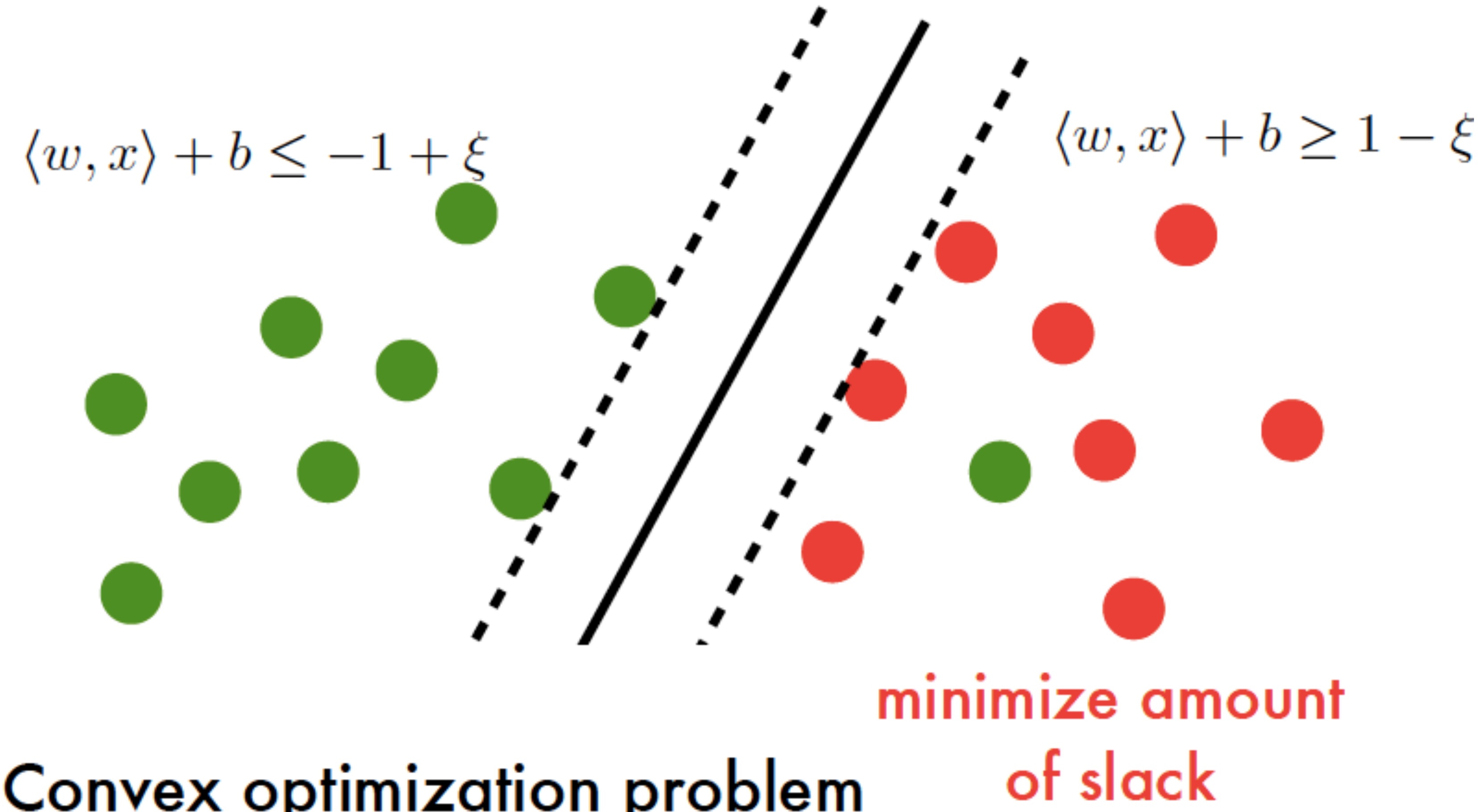
Convex optimization problem

Adding slack variables



Convex optimization problem

Adding slack variables



Convex optimization problem

Adding slack variables

Adding slack variables

- **Hard margin problem**

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \ \text{ subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1$$

Adding slack variables

- Hard margin problem

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1$$

- With slack variables

$$\underset{w,b}{\text{minimize}}\ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Adding slack variables

- Hard margin problem

$$\underset{w,b}{\text{minimize}} \frac{1}{2}\|w\|^2 \text{ subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1$$

- With slack variables

$$\underset{w,b}{\text{minimize}} \; \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Problem is always feasible.

Dual Problem

Dual Problem

- Primal optimization problem

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i \left[\langle w, x_i \rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Dual Problem

- **Primal optimization problem**

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

- **Lagrange function**

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\left[y_i\left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i\xi_i$$

Dual Problem

- Primal optimization problem

$$\underset{w,b}{\text{minimize}} \;\; \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to } y_i\left[\langle w, x_i\rangle + b\right] \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

- Lagrange function

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i\left[y_i\left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

Optimality in w,b,ξ is at saddle point with α,η

- Derivatives in w,b,ξ need to vanish

Dual Problem

Dual Problem

- **Lagrange function**

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i \left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

Dual Problem

- **Lagrange function**

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i\left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

Dual Problem

- **Lagrange function**

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i\left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

Dual Problem

- **Lagrange function**

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i \left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

- **Derivatives in w, b need to vanish**

$$\partial_w L(w,b,\xi,\alpha,\eta) = w - \sum_i \alpha_i y_i x_i = 0$$

$$\partial_b L(w,b,\xi,\alpha,\eta) = \sum_i \alpha_i y_i = 0$$

$$\partial_{\xi_i} L(w,b,\xi,\alpha,\eta) = C - \alpha_i - \eta_i = 0$$

Dual Problem

- Lagrange function

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \left[y_i \left[\langle x_i, w\rangle + b\right] + \xi_i - 1\right] - \sum_i \eta_i \xi_i$$

- Derivatives in w, b need to vanish

$$\partial_w L(w, b, \xi, \alpha, \eta) = w - \sum_i \alpha_i y_i x_i = 0$$

$$\partial_b L(w, b, \xi, \alpha, \eta) = \sum_i \alpha_i y_i = 0$$

$$\partial_{\xi_i} L(w, b, \xi, \alpha, \eta) = C - \alpha_i - \eta_i = 0$$

- Plugging terms back into L yields

$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]$$

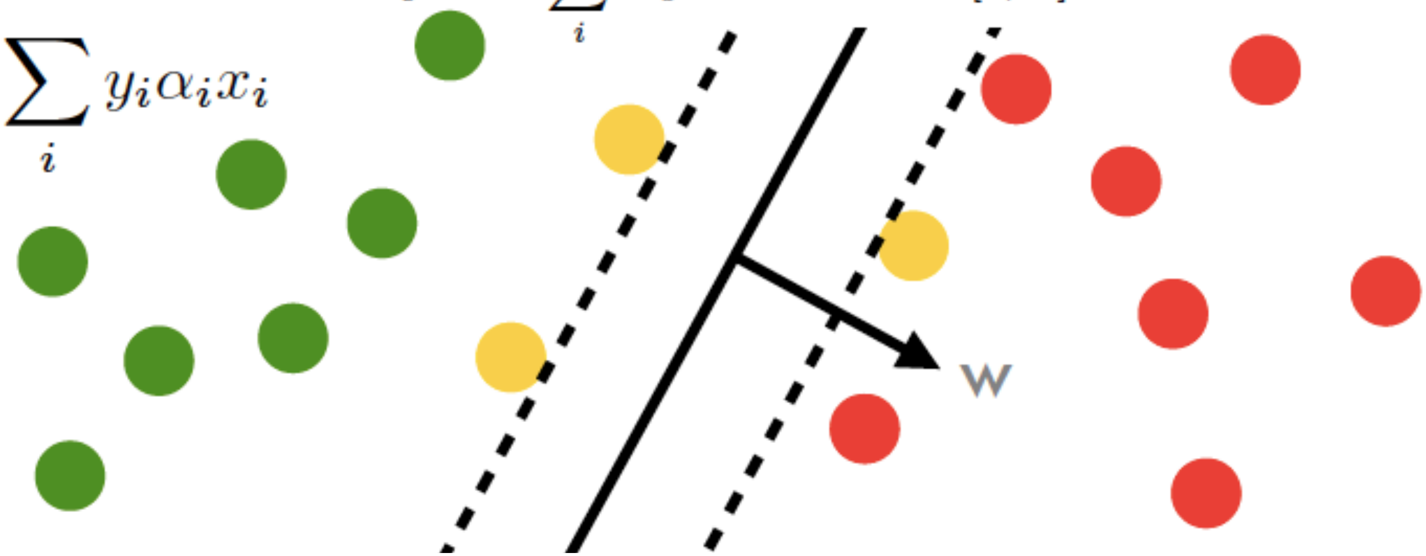
bound influence

Karush Kuhn Tucker Conditions

$$\underset{\alpha}{\text{maximize}} - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]$$

$$w = \sum_i y_i \alpha_i x_i$$

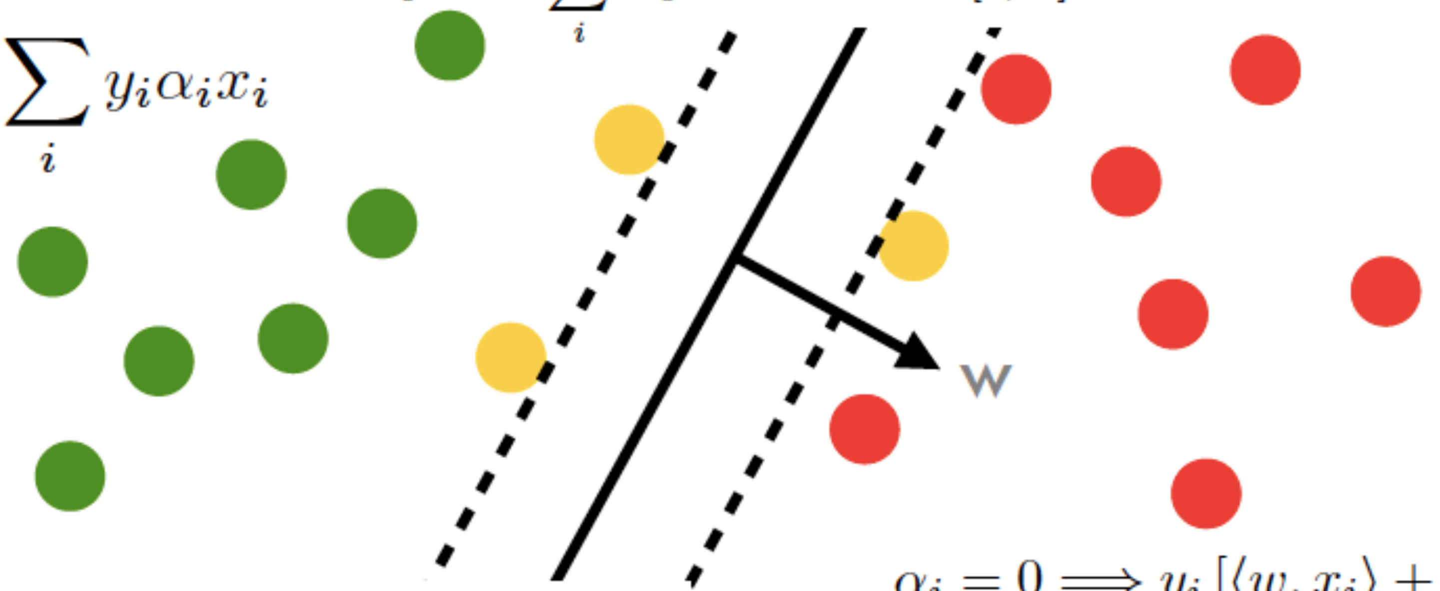


Karush Kuhn Tucker Conditions

$$\underset{\alpha}{\text{maximize}} - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i$$

$$\text{subject to} \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \in [0, C]$$

$$w = \sum_i y_i \alpha_i x_i$$



$$\alpha_i = 0 \implies y_i [\langle w, x_i \rangle + b] \geq 1$$

$$\alpha_i [y_i [\langle w, x_i \rangle + b] + \xi_i - 1] = 0$$

$$\eta_i \xi_i = 0$$

$$0 < \alpha_i < C \implies y_i [\langle w, x_i \rangle + b] = 1$$

$$\alpha_i = C \implies y_i [\langle w, x_i \rangle + b] \leq 1$$

That’s all!