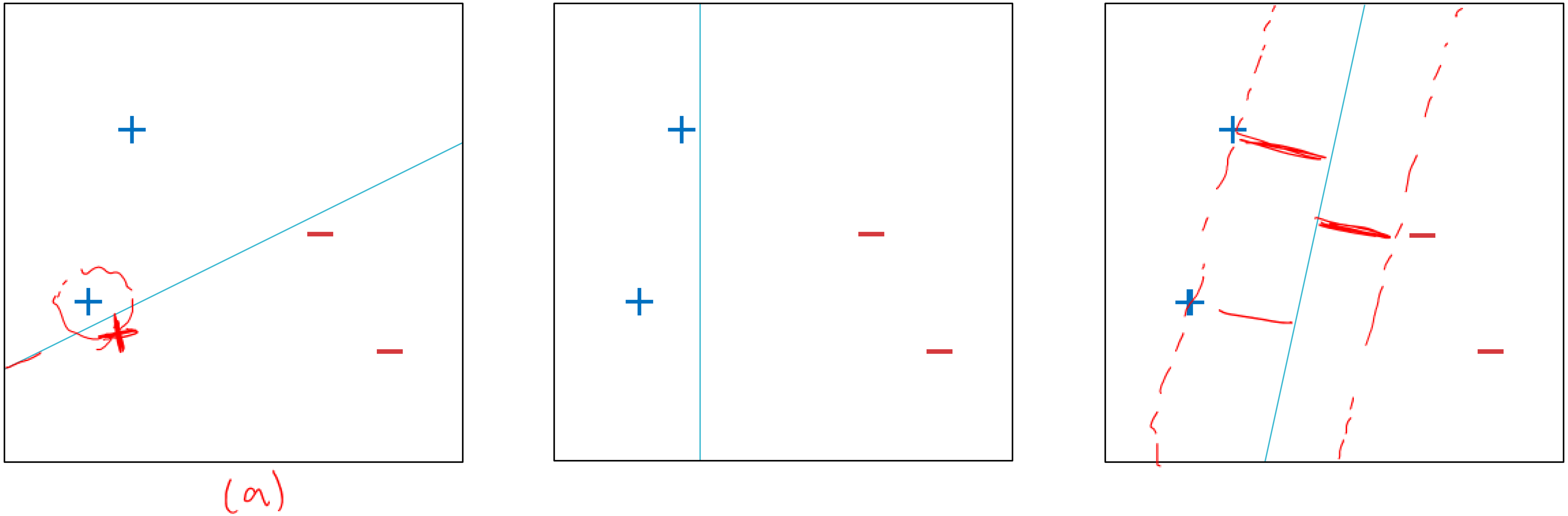
10-701: Introduction to Machine Learning

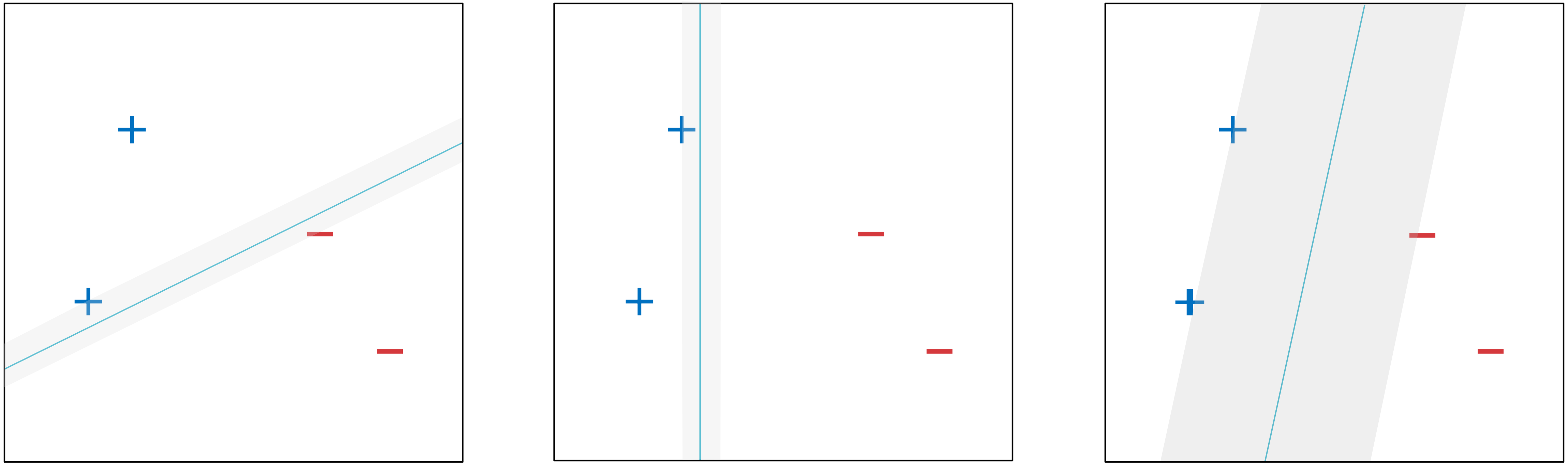
Lecture 23 –Support Vector Machines

Hoda Heidari

* Slides adopted from F24 offering of 10701 by Henry Chai.



Which linear separator is best?



Which linear separator is best?

Maximal Margin Linear Separators

- Standing Assumption: Our training data is linearly separable.

- The **margin** of a linear separator is the distance between it and the nearest training data point.
- Questions:
  1. How can we efficiently find a maximal-margin linear separator?
  2. Why are linear separators with larger margins better?
  3. What can we do if the data is not linearly separable?

Recall: Hyperplanes

- For linear models, decision boundaries are D-dimensional ***hyperplanes*** defined by a weight vector, $[b, \boldsymbol{w}]$

$$c(\boldsymbol{w}^T\boldsymbol{x} + b = 0)$$

w's and b's

- Problem: there are infinitely many weight vectors that describe the same hyperplane
  - $x_1 + 2x_2 + 2 = 0$ is the same line as

$\times 2 \longrightarrow$ $2x_1 + 4x_2 + 4 = 0$, which is the same line as

$\times 1M \longrightarrow$ $1000000x_1 + 2000000x_2 + 2000000 = 0$

- Solution: normalize weight vectors *w.r.t. the training data*

Normalizing Hyperplanes

- Given a dataset $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ where $y \in \{-1, +1\}$,

  = perfect accuracy classifier

  $\hat{y} = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + b)$ is a valid **linear separator** if

  $$y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) > 0 \; \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- For SVMs, we're *only* going to consider **linear separators** in

$$\mathcal{H} = \left\{\hat{y} = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + b) : \min_{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1\right\}$$

- If $\hat{y} = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + b)$ is a linear separator, then

  $\hat{y} = \text{sign}\left(\frac{\boldsymbol{w}^T}{\rho} \boldsymbol{x} + \frac{b}{\rho}\right) \in \mathcal{H}$ where

  $$\rho = \min_{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right)$$

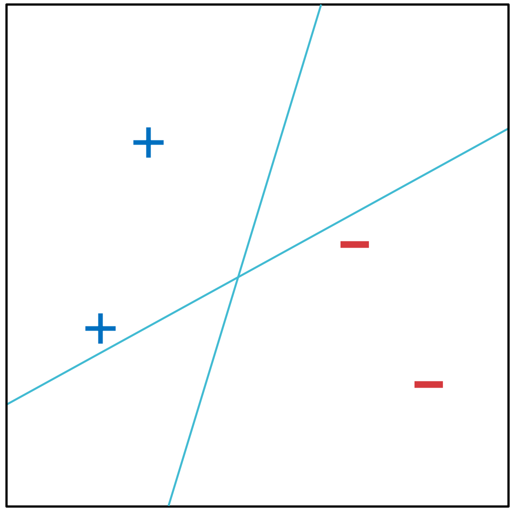
Normalizing Hyperplanes: Example

$\mathcal{H}$

b	w_1	w_2	
-0.2	-0.6	1	$\notin \mathcal{H}$
-0.4	-1.2	2	$\notin \mathcal{H}$
-2	-6	10	$\notin \mathcal{H}$
-10	-30	50	$\in \mathcal{H}$
0.2	-0.6	0.2	$\notin \mathcal{H}$
0.1	-0.3	0.1	$\notin \mathcal{H}$
1	-3	1	$\notin \mathcal{H}$
2	-6	2	$\in \mathcal{H}$

1

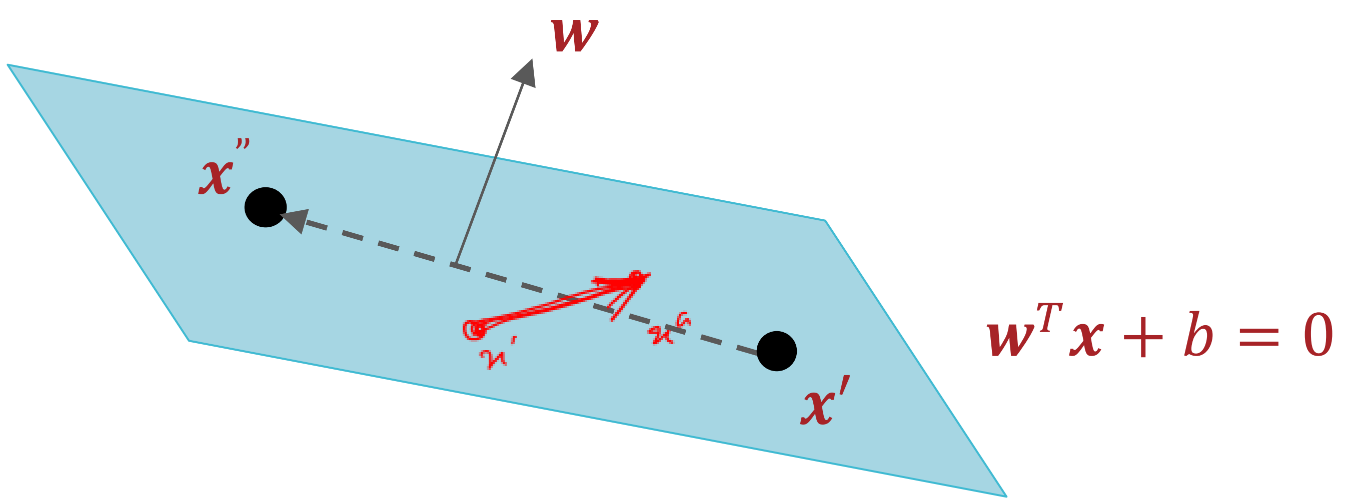
2



x_1	x_2	y	$y(\boldsymbol{w}^T\boldsymbol{x} + b)$
0.2	0.4	+1	1.6
0.3	0.8	+1	1.8
0.7	0.6	-1	1
0.8	0.3	-1	2.2

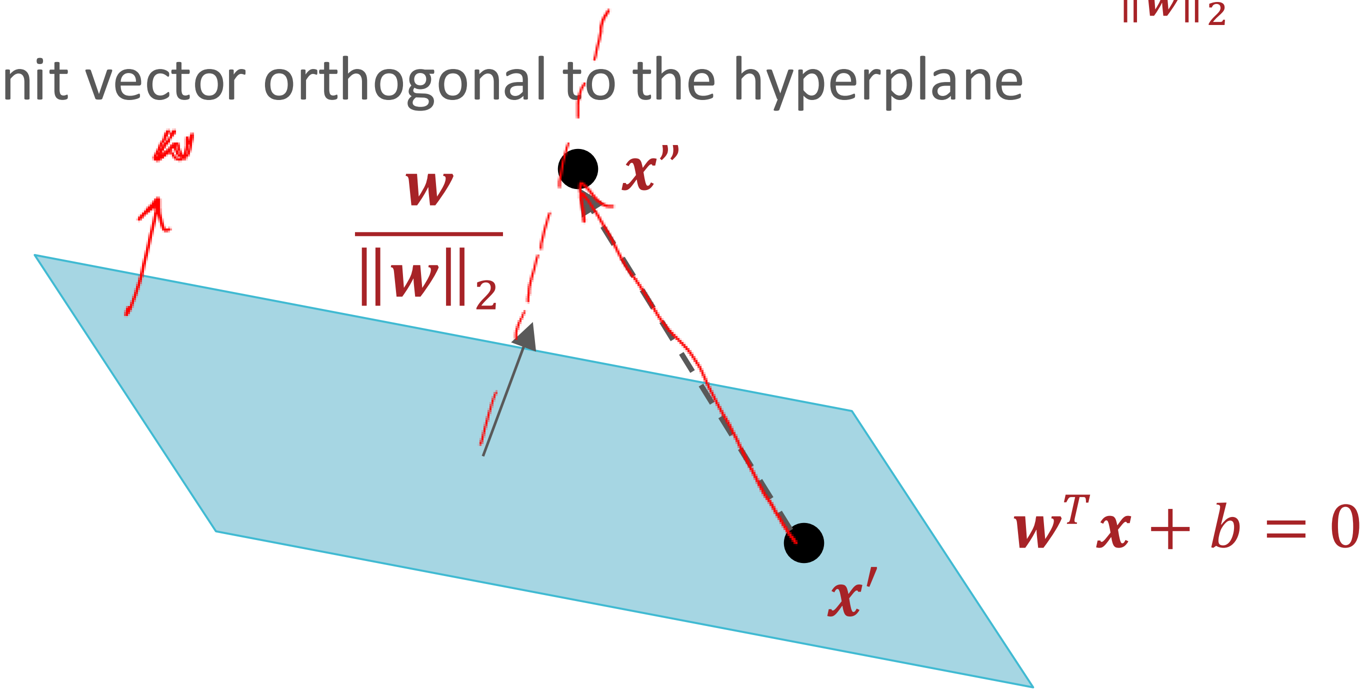
Computing the Margin

- Claim: $\boldsymbol{w}$ is orthogonal to the hyperplane $\underbrace{\boldsymbol{w}^T\boldsymbol{x} + b = 0}$ (the decision boundary)
- A vector is orthogonal to a hyperplane if it is orthogonal to every vector in that hyperplane
- Vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal if $\underbrace{\boldsymbol{\alpha}^T\boldsymbol{\beta}} = 0$



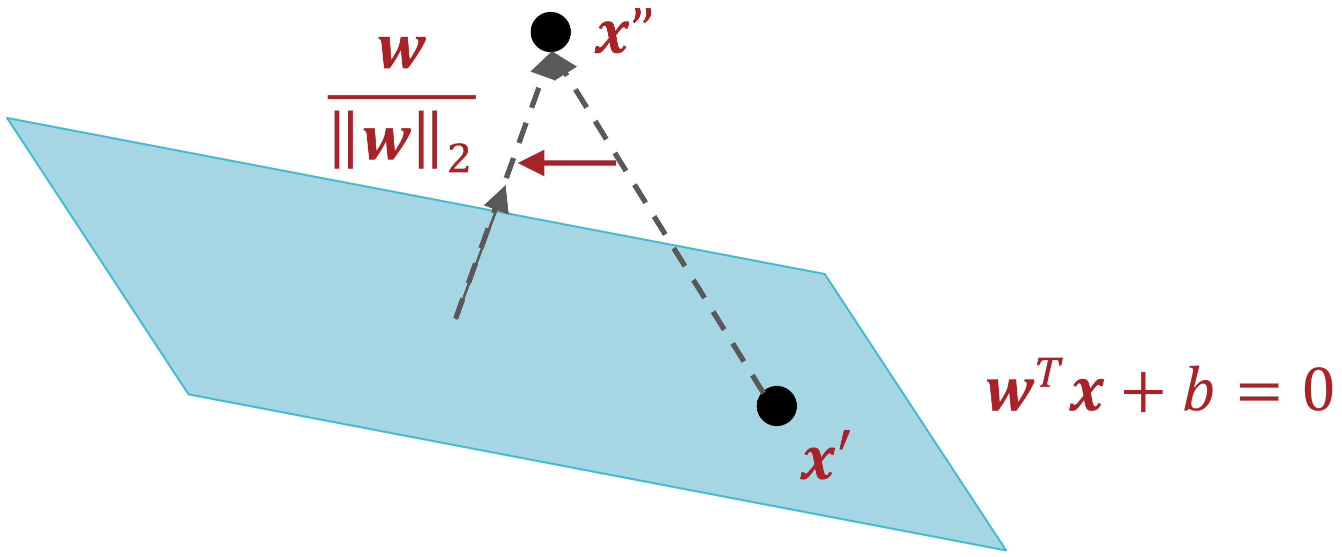
Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



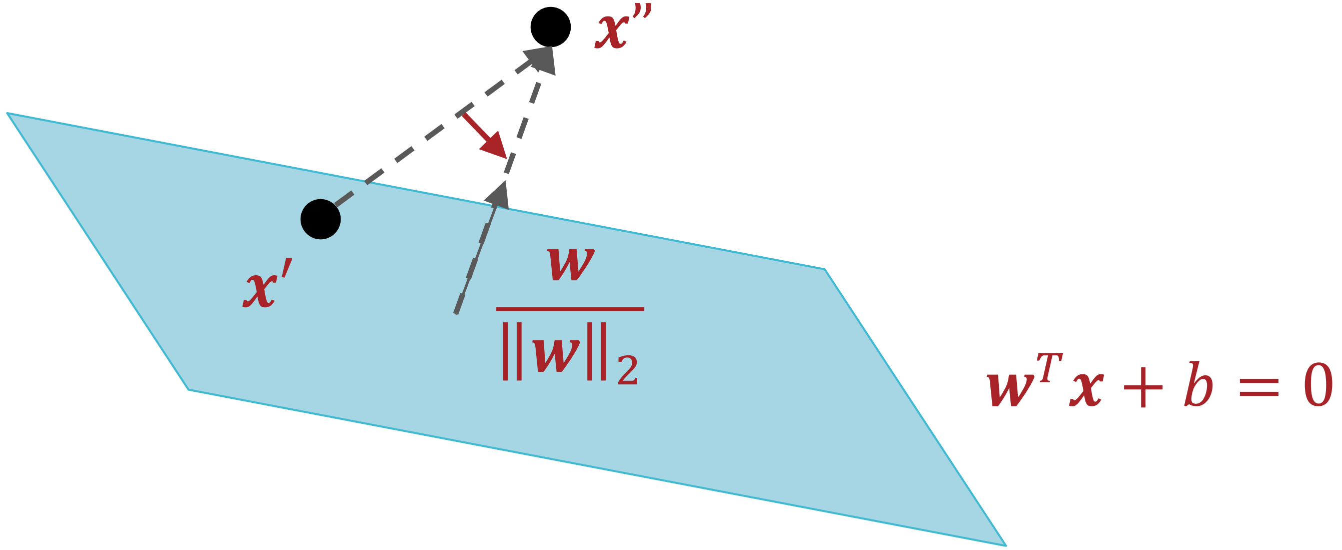
Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boxed{\boldsymbol{x}'' - \boldsymbol{x}'}$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane

$$d(\boldsymbol{x}'', h) = \left|\frac{\boldsymbol{w}^T\boxed{(\boldsymbol{x}'' - \boldsymbol{x}')}}{\|\boldsymbol{w}\|_2}\right| = \frac{|\boldsymbol{w}^T\boldsymbol{x}'' - \cancelto{b}{\boldsymbol{w}^T\boldsymbol{x}'}|}{\|\boldsymbol{w}\|_2}$$

$$\boxed{= \frac{|\boldsymbol{w}^T\boldsymbol{x}'' + b|}{\|\boldsymbol{w}\|_2}}$$

because x' was on hyperplane, $w^Tx' + b = 0 \Rightarrow \underbrace{-w^Tx' = b}$

distance btw x'' and hyperplane $(\vec{w}, b)$

Computing the Margin

- The margin of a linear separator is the distance between it and the nearest training data point

A fixed linear decision boundary defined by $(\boldsymbol{w}^T, b)$

$$\min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} d\left(\boldsymbol{x}^{(i)}, h\right) = \min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{\left|\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right|}{\|\boldsymbol{w}\|_2}$$

$$= \frac{1}{\|\boldsymbol{w}\|_2} \min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} \underbrace{\left|\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right|}_{M}$$

$$= \frac{1}{\|\boldsymbol{w}\|_2} \min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right)$$

$$= \boxed{\frac{1}{\|\boldsymbol{w}\|_2}}$$

Maximizing the Margin

Margin

$$\underset{w,b}{\text{maximize}} \quad \frac{1}{\|\boldsymbol{w}\|_2}$$

$$\text{subject to} \quad \underbrace{\min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1}_{\text{Normalization constraints}}$$

$$\Updownarrow$$

$$\text{minimize} \quad \|\boldsymbol{w}\|_2$$

$$\text{subject to} \quad \min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

$$\Updownarrow$$

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 \quad = \frac{1}{2} w^T w$$

$$\text{subject to} \quad \underbrace{\min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1}$$

$$\Updownarrow$$

$$\boxed{\begin{aligned} &\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T \boldsymbol{w} \\ &\text{subject to} \quad \underbrace{y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right)} \geq 1 \;\forall\, \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D} \end{aligned}}$$

Maximizing the Margin

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$\underbrace{\hat{w}, \hat{b}}_{\text{optimal solution}}$

- If $[\hat{b}, \hat{\boldsymbol{w}}]$ is the optimal solution, then $\exists$ at least one training data point $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ s.t $y^{(i)}\left(\hat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right) = 1$
  - All training data points $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ where $y^{(i)}\left(\hat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right) = 1$ are known as **support vectors**
- Converting the non-linear constraint (involving the $\min$) to N linear constraints means we can use quadratic programming (QP) to solve this problem in $O(D^3)$ time

Recipe for SVMs

- Define a model and model parameters
  - Assume a linear decision boundary (with normalized weights)

$$h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$$

  - Parameters: $\boldsymbol{w} = [w_1, \ldots, w_D]$ and b

- Write down an objective function (with constraints)

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

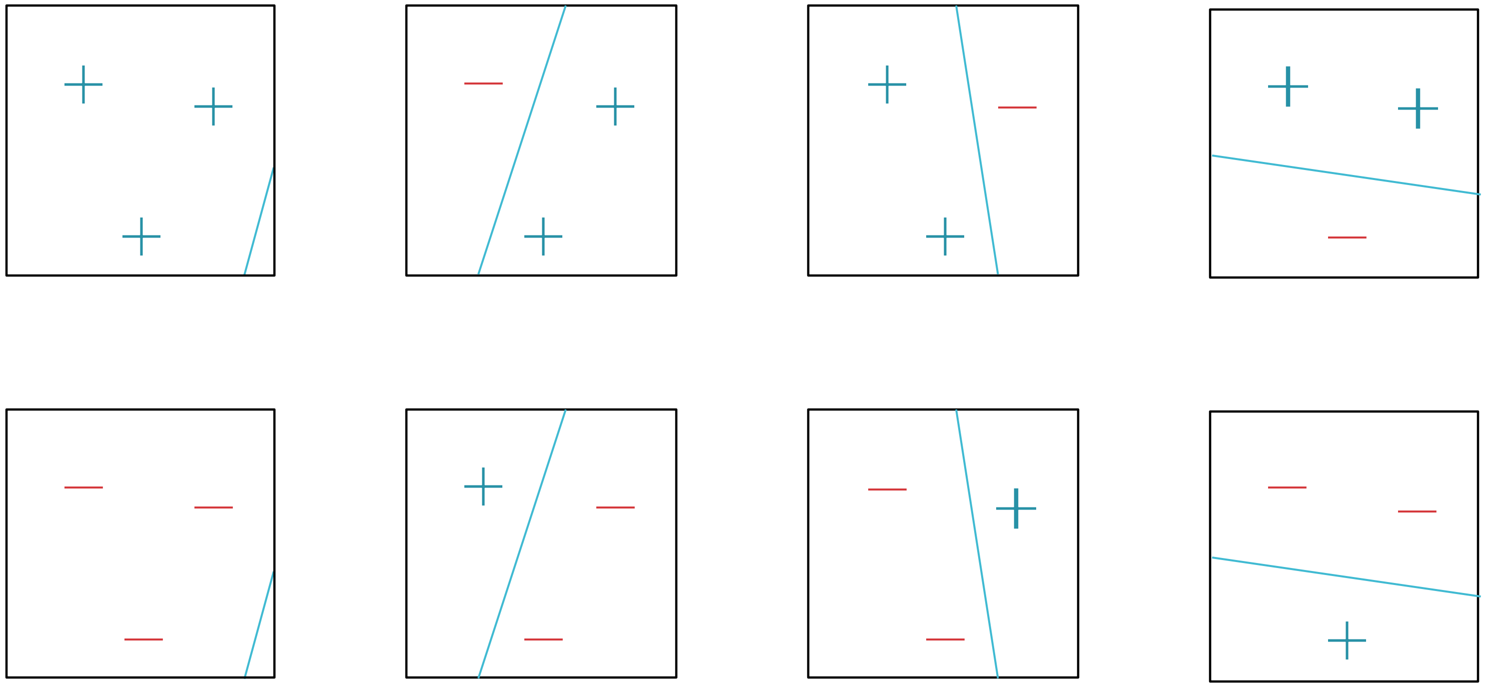
$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \;\forall\; \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- Optimize the objective w.r.t. the model parameters
  - Solve using quadratic programming

Why Maximal Margins?

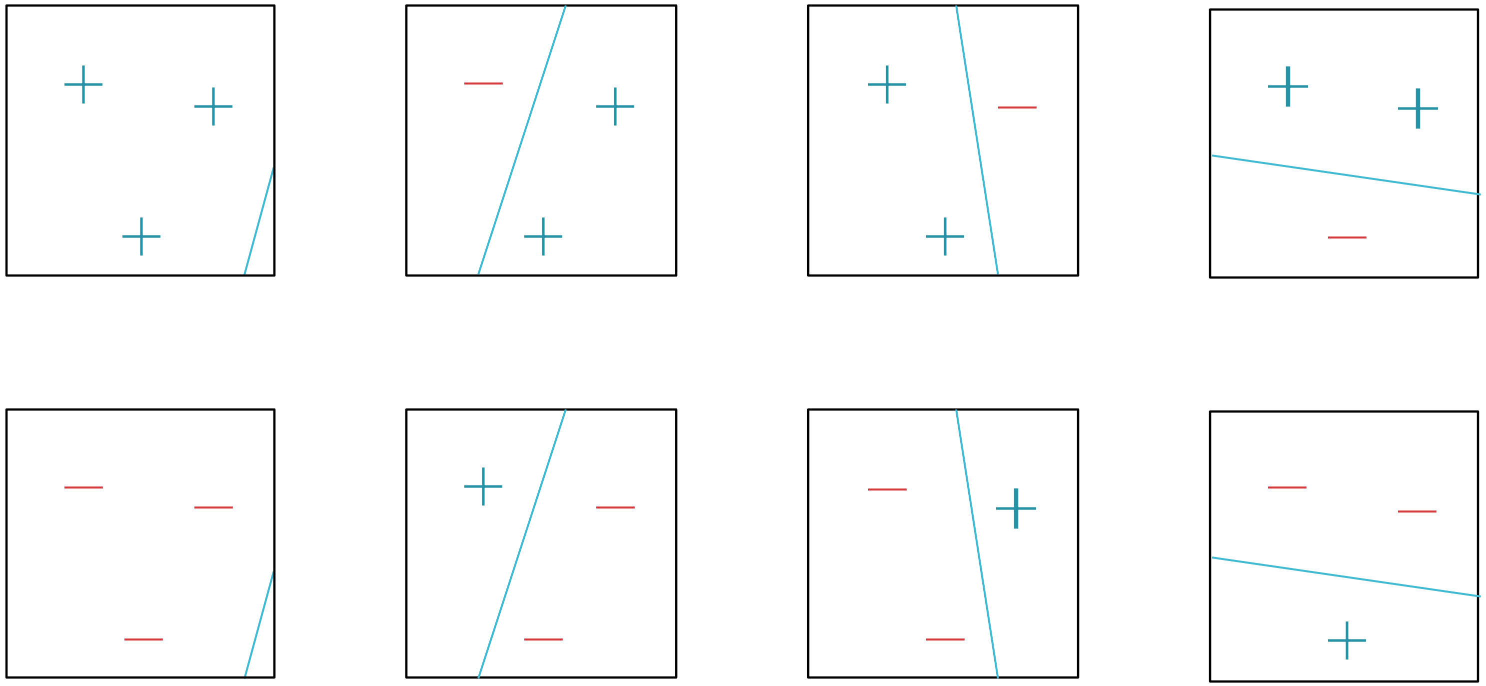
- Consider three binary data points in a **bounded** 2-D space
- Let $\mathcal{H} = \{\text{all linear separators}\}$ and

  $\mathcal{H}_\rho = \{\text{all linear separators with minimum margin } \rho\}$



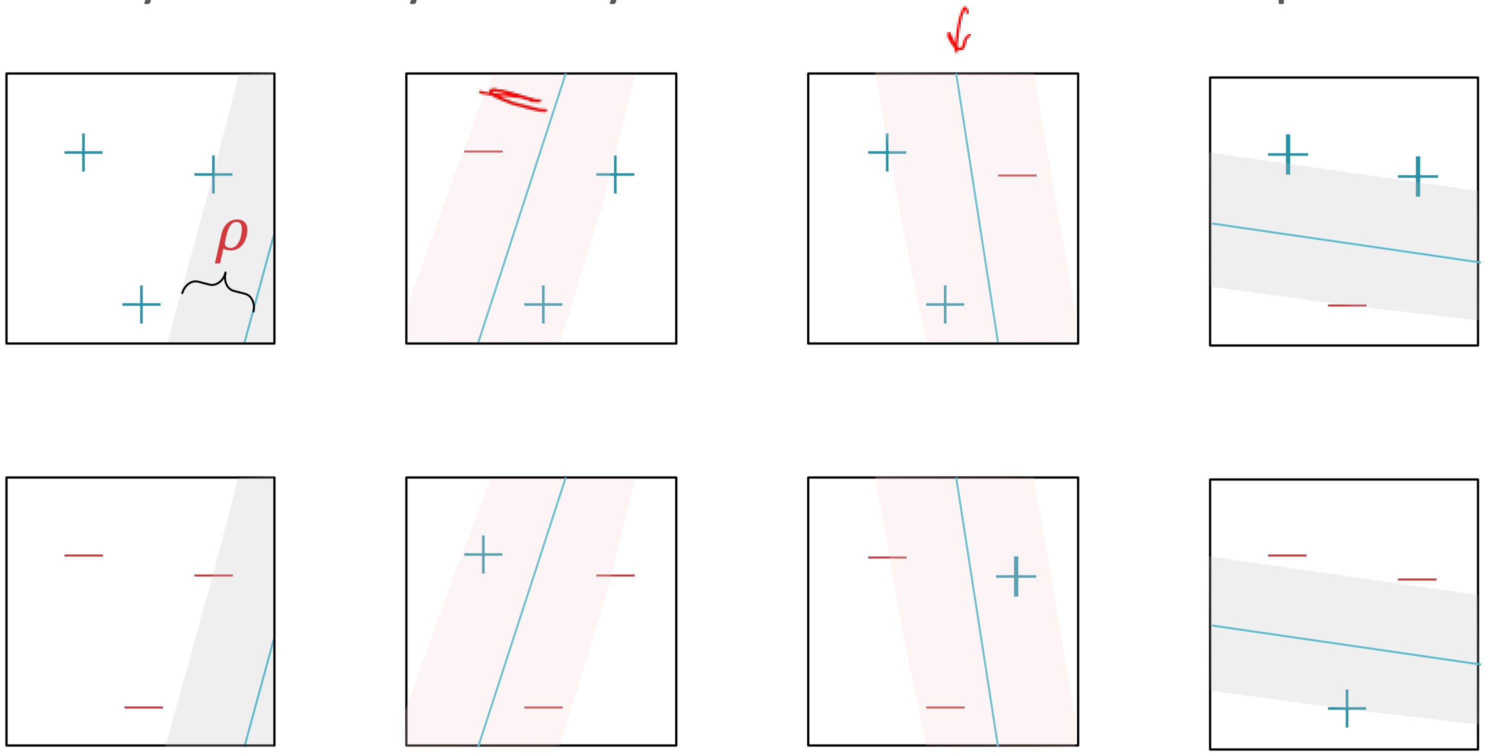
Why Maximal Margins?

- Consider three binary data points in a **bounded** 2-D space
- $\mathcal{H}$ = {all linear separators} can always correctly classify any three (non-colinear) data points in this space



Why Maximal Margins?

- Consider three binary data points in a **bounded** 2-D space
- $\mathcal{H}_\rho$ = {all linear separators with minimum margin ρ} cannot always correctly classify three non-colinear data points



Summary Thus Far

- The margin of a linear separator is the distance between it and the nearest training data point
- Questions:
  1. How can we efficiently find a maximal-margin linear separator? By solving a constrained quadratic optimization problem using quadratic programming
  2. Why are linear separators with larger margins better? They're simpler *waves hands*
  3. What can we do if the data is not linearly separable? Next!

Linearly Inseparable Data

- What can we do if the data is not linearly separable?
  1. Accept some non-zero training error
     - How much training error should we tolerate?
  2. Apply a non-linear transformation that shifts the data into a space where it is linearly separable
     - How can we pick a non-linear transformation?

SVMs

$$\begin{cases} \underset{w,b}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ \text{subject to } \underbrace{y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1}\ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D} \end{cases}$$

- When $\mathcal{D}$ is not linearly separable, there are no feasible solutions to this optimization problem

Hard-margin SVMs

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

- When $\mathcal{D}$ is not linearly separable, there are no feasible solutions to this optimization problem

Soft-margin SVMs

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)} \quad \forall\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\xi^{(i)} \geq 0 \quad \forall\, i \in \{1, \ldots, N\}$$

Soft-margin SVMs

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)} \;\forall\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\xi^{(i)} \geq 0 \qquad \forall\, i \in \{1, \dots, N\}$$

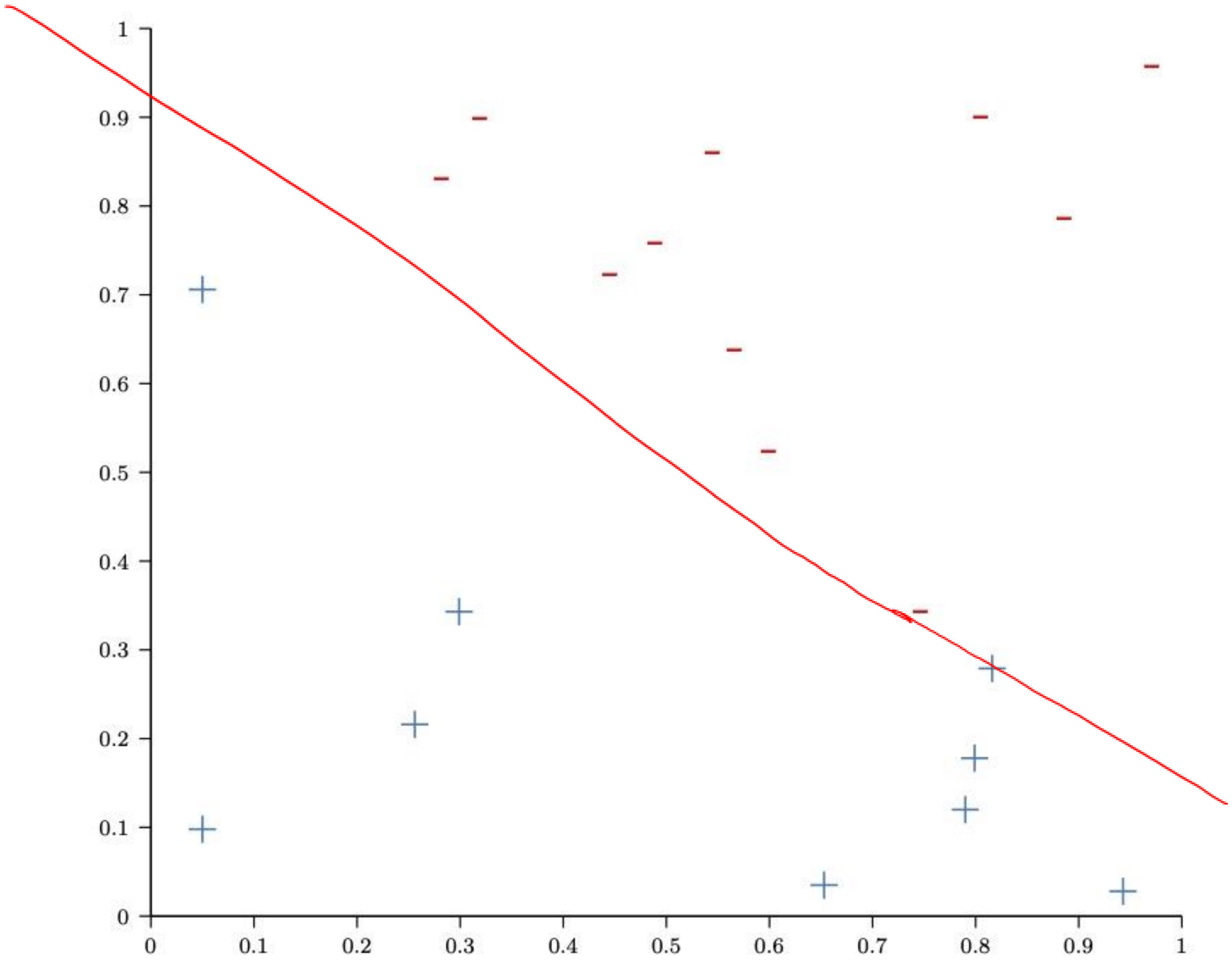
- $\xi^{(i)}$ is the "soft" error on the i^{th} training data point
  - If $\xi^{(i)} > 1$, then $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) < 0 \Rightarrow \left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ is incorrectly classified
  - If $0 < \xi^{(i)} < 1$, then $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) > 0 \Rightarrow \left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ is correctly classified but inside the margin
- $\sum_{i=1}^{N}\xi^{(i)}$ is the "soft" training error

Soft-margin SVMs

$$\begin{cases} \underset{\omega, b, \xi}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \boxed{C}\sum_{i=1}^{N}\xi^{(i)} \\ \text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi^{(i)} \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D} \\ \xi^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, N\} \end{cases}$$

- Still solvable using quadratic programming
- All training data points $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ where $\underbrace{y^{(i)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}\right) \leq 1}$ are known as **support vectors**

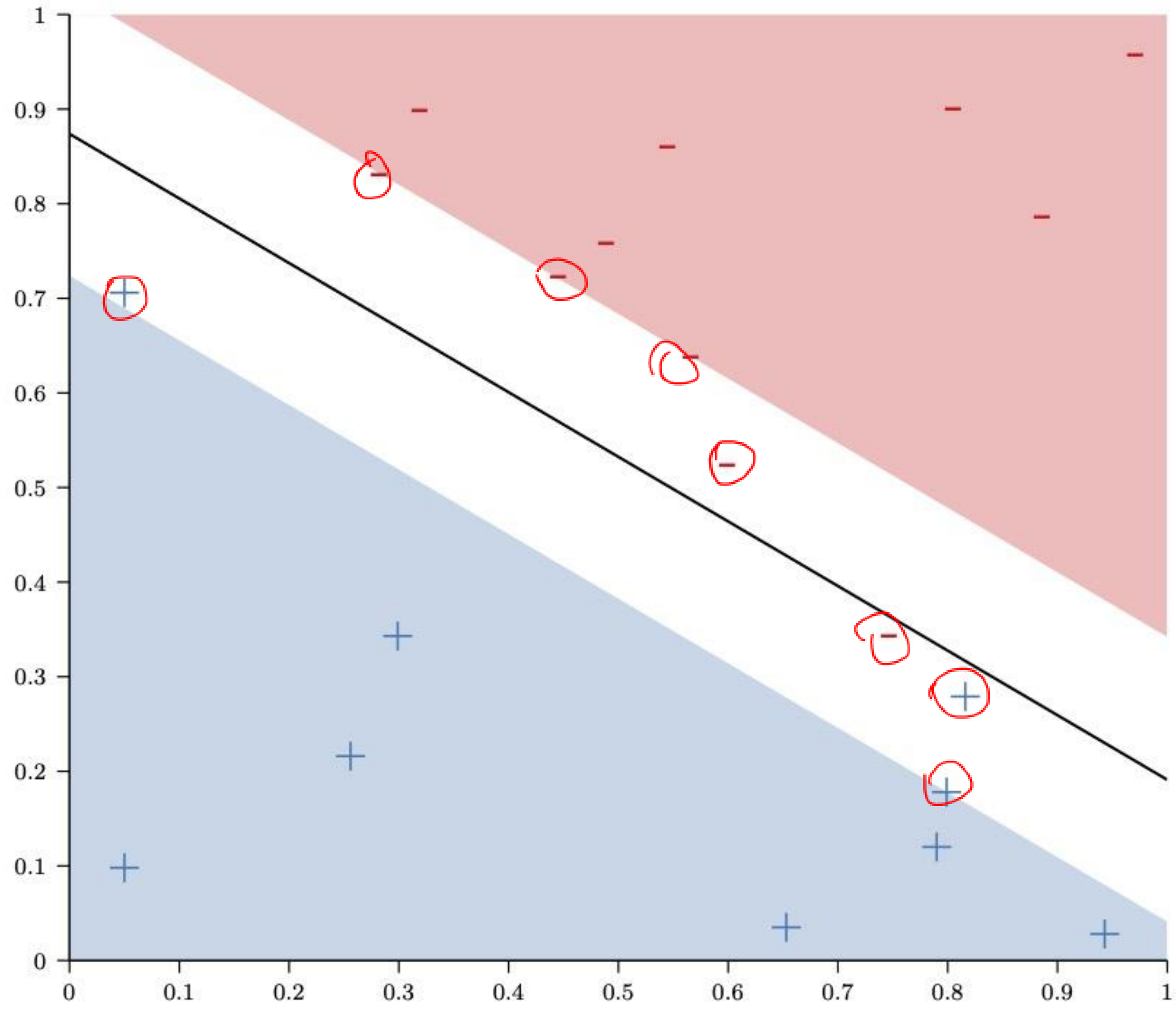
Interpreting $\xi^{(i)}$



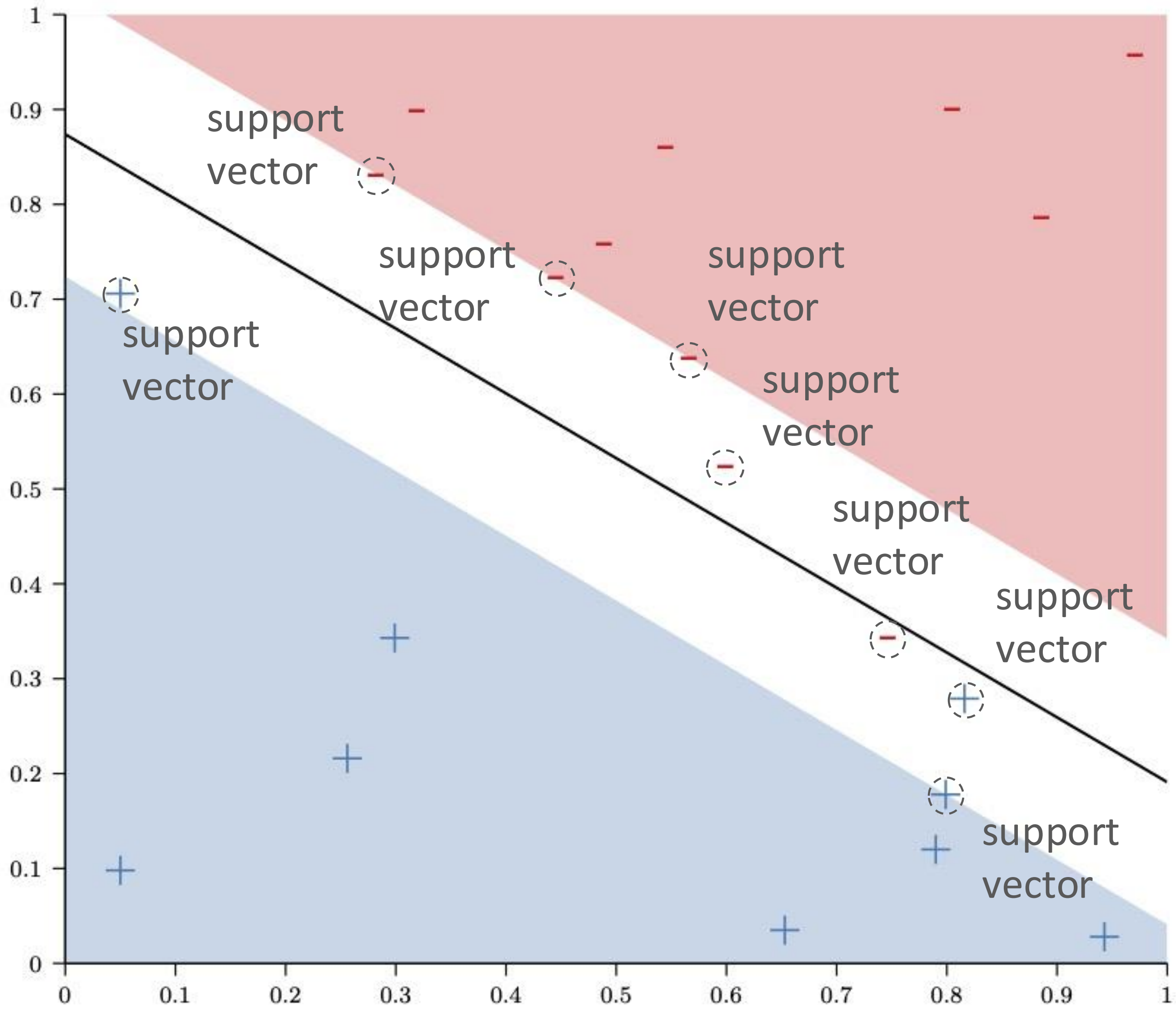
Interpreting $\xi^{(i)}$

C small

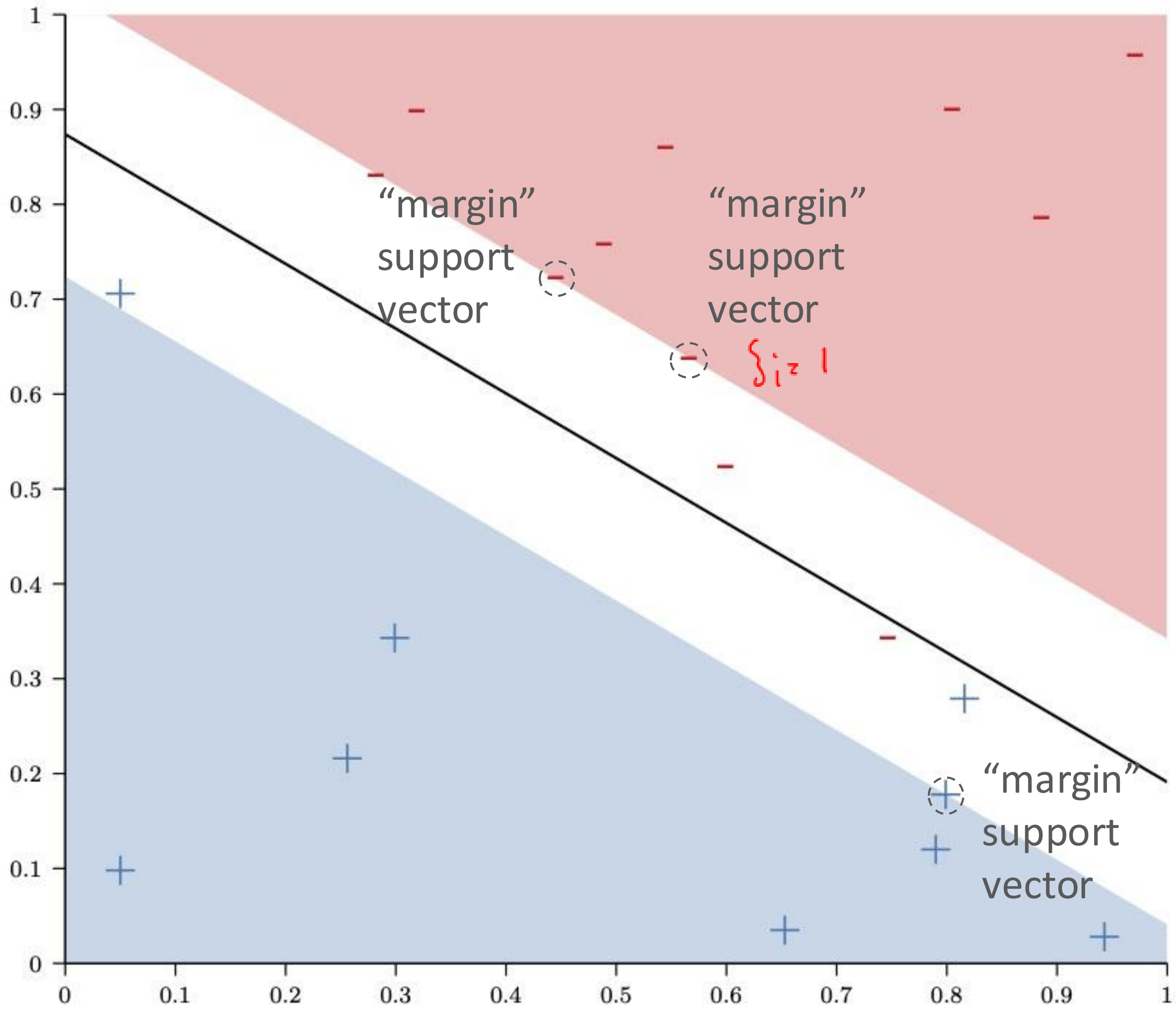
$C \sum_{i=1}^{N} \xi_i$



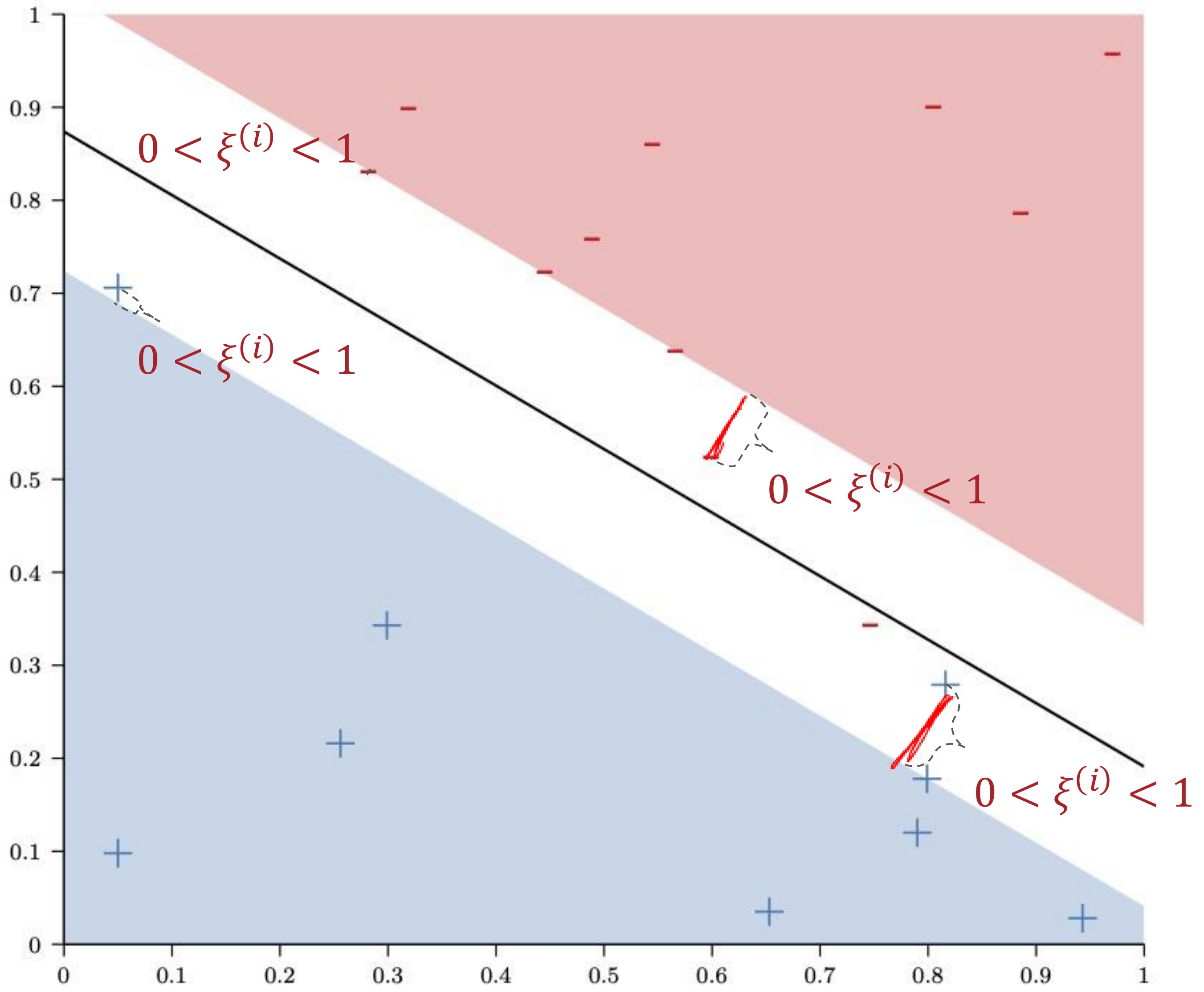
Interpreting $\xi^{(i)}$



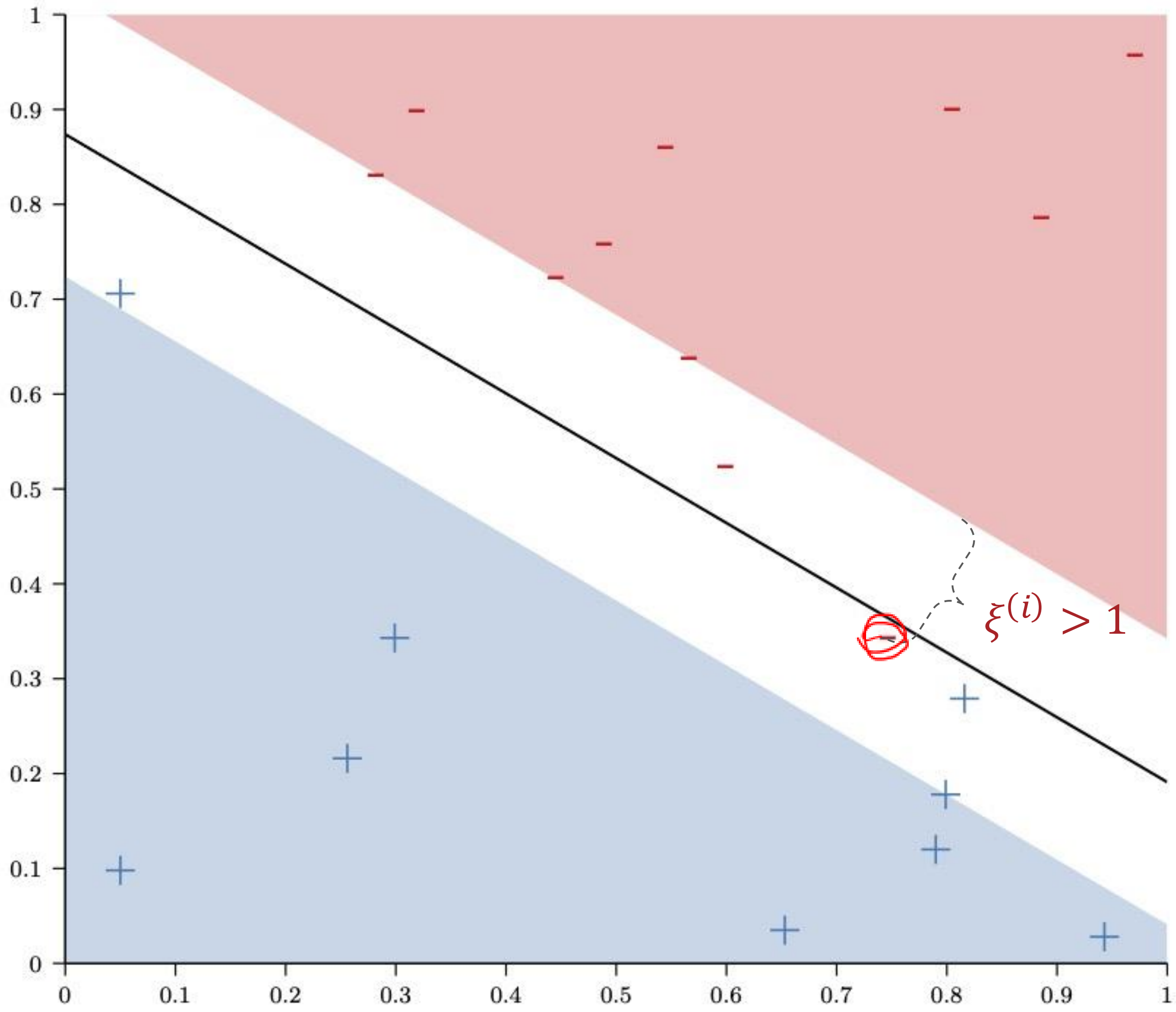
Interpreting $\xi^{(i)}$

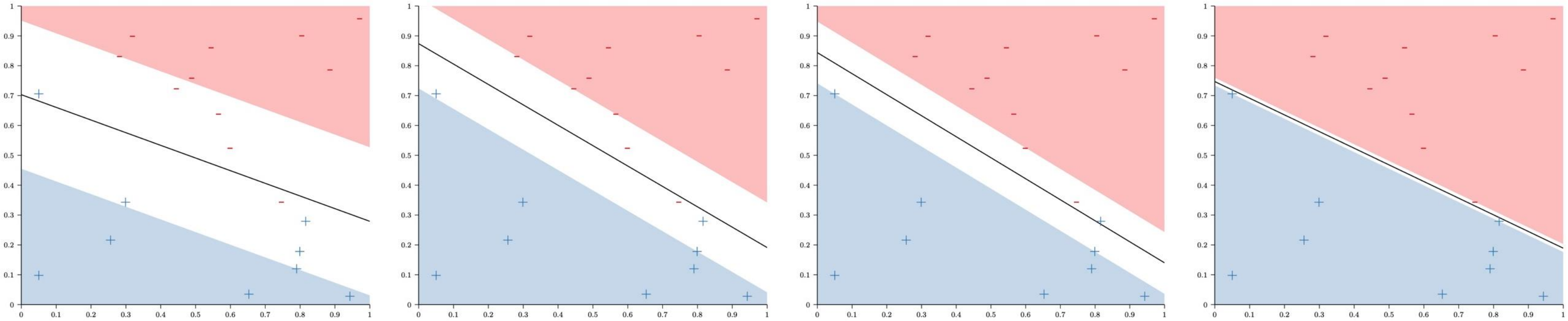


Interpreting $\xi^{(i)}$



Interpreting $\xi^{(i)}$





Smaller C Larger C Hard Margin

Setting C

C is a tradeoff parameter (much like the tradeoff parameter in regularization)

Hard-margin SVMs

$$\left.\begin{aligned}&\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\\&\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)}+b\right)\geq 1\ \forall\left(\boldsymbol{x}^{(i)},y^{(i)}\right)\in\mathcal{D}\end{aligned}\right\}\text{SVMs}$$

$$\left.\begin{aligned}&\text{minimize } E_{train}\\&\text{subject to } \boldsymbol{w}^T\boldsymbol{w}\leq C\end{aligned}\right\}\text{Regularization}$$

	SVM	Regularization
minimize	$\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$	E_{train}
subject to	$E_{train}=0$	$\boldsymbol{w}^T\boldsymbol{w}\leq C$

Hard_margin SVM

Primal-Dual Optimization

primal

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \; \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

Primal

$$\Updownarrow$$

$$\text{maximize } -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} + \sum_{i=1}^{N}\alpha^{(i)}$$

$$\text{subject to } \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0$$

$$\alpha^{(i)} \geq 0 \; \forall \, i \in \{1, \dots, N\}$$

Dual

SVM

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + \overset{b}{w_0}\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\Updownarrow$$

(*)

$$\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to} \quad 1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \leq 0 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\Updownarrow$$

(**)

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \left[\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \boxed{\underset{\alpha^{(i)} \geq 0}{\text{maximize}}} \sum_{i=1}^{N} \alpha^{(i)}\left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right)\right]$$

adversary

SVM

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \left[\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \underset{\alpha^{(i)} \geq 0}{\text{maximize}} \sum_{i=1}^{N} \alpha^{(i)} \left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right) \right]$$

$$\Updownarrow$$

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \left[\underset{\alpha^{(i)} \geq 0}{\text{maximize}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{N} \alpha^{(i)} \left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right) \right]$$

$$\Updownarrow$$

$$\underset{\alpha^{(i)} \geq 0}{\text{maximize}} \left[\underset{\boldsymbol{w}, w_0}{\text{minimize}} \underbrace{\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{N} \alpha^{(i)} \left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right)} \right]$$

$$\Updownarrow$$

$$\underset{\boldsymbol{\alpha} \geq 0}{\text{maximize}} \; \underset{\boldsymbol{w}, w_0}{\text{minimize}} \; \underline{L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)}$$

Karush-Kuhn-Tucker (KKT) Conditions

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \quad L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)$$

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{N} \alpha^{(i)}\left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right)$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} = w + \sum_{i=1}^{N} \alpha^{(i)}\left(-y^{(i)} x^{(i)}\right) = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x^{(i)}$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)}{\partial w_0} = \sum_{i=1}^{N} \alpha^{(i)}(-y^{(i)}) = 0 \Rightarrow \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0$$

Karush-Kuhn-Tucker (KKT) Conditions

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \quad L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)$$

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{N} \alpha^{(i)}\left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right)$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)} \rightarrow \widehat{\boldsymbol{w}} = \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)}{\partial w_0} = -\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \quad \rightarrow \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0$$

Minimizing the Lagrangian

$$\boxed{\widehat{\boldsymbol{w}} = \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}}$$

$$\boxed{\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0}$$

$$\underbrace{L(\boldsymbol{\alpha}, \widehat{\boldsymbol{w}}, \widehat{w}_0) = \frac{1}{2}\widehat{\boldsymbol{w}}^T \widehat{\boldsymbol{w}} + \sum_{i=1}^{N} \alpha^{(i)} \left(1 - y^{(i)}\left(\widehat{\boldsymbol{w}}^T \boldsymbol{x}^{(i)} + \widehat{w}_0\right)\right)}$$

$$= \frac{1}{2}\widehat{\boldsymbol{w}}^T \widehat{\boldsymbol{w}}$$

$$+ \sum_{i=1}^{N} \alpha^{(i)} - \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \widehat{\boldsymbol{w}}^T \boldsymbol{x}^{(i)} - \widehat{w}_0 \overbrace{\sum_{i=1}^{N} \alpha^{(i)} y^{(i)}}^{=0}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}\right)\left(\sum_{j=1}^{N} \alpha^{(j)} y^{(j)} \boldsymbol{x}^{(j)}\right)$$

$$+ \sum_{i=1}^{N} \alpha^{(i)} - \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \left(\sum_{j=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}\right)^T \boldsymbol{x}^{(i)}$$

Minimizing the Lagrangian

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

$$\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0$$

$$L(\boldsymbol{\alpha}, \hat{\boldsymbol{w}}, \hat{w}_0) = \frac{1}{2}\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} + \sum_{i=1}^{N} \alpha^{(i)}\left(1 - y^{(i)}\left(\hat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{w}_0\right)\right)$$

$$= \frac{1}{2}\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}}$$

$$+ \sum_{i=1}^{N} \alpha^{(i)} - \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \hat{\boldsymbol{w}}^T \boldsymbol{x}^{(i)} - \hat{w}_0 \sum_{i=1}^{N} \alpha^{(i)} y^{(i)}$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha^{(i)}\alpha^{(j)} y^{(i)} y^{(j)} \boldsymbol{x}^{(i)T}\boldsymbol{x}^{(j)} + \sum_{i=1}^{N} \alpha^{(i)}$$

Maximizing the Minimum

$$\underset{\boldsymbol{\alpha} \geq 0}{\text{maximize}} \; \underset{\boldsymbol{w}, w_0}{\text{minimize}} \; L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)$$

$$\Updownarrow$$

$$\text{maximize} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} + \sum_{i=1}^{N}\alpha^{(i)}$$

$$\text{subject to} \quad \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0$$

$$\alpha^{(i)} \geq 0 \; \forall \, i \in \{1, \dots, N\}$$

Primal-Dual Optimization

$$\left.\begin{aligned}&\text{minimize} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\\&\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}\end{aligned}\right\} \text{Primal}$$

$$\left.\begin{aligned}&\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} - \sum_{i=1}^{N}\alpha^{(i)}\\&\text{subject to} \quad \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0\\&\qquad\qquad\quad \alpha^{(i)} \geq 0 \ \forall \ i \in \{1, \dots, N\}\end{aligned}\right\} \text{Dual}$$

Primal-Dual Optimization

- Primal
  - Directly returns the weights, $[\hat{w}_0, \hat{\boldsymbol{w}}]$
  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t.

$$y^{(s)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = 1$$

- Dual
  - Returns the vector, $\hat{\boldsymbol{\alpha}}$

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

$$\hat{w}_0 = ???$$

Complementary Slackness

$$\underset{}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{subject to } 1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 0 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\Updownarrow$$

$$\underset{\boldsymbol{w}, w_0}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \underset{\alpha^{(i)} \geq 0}{\text{maximize}} \sum_{i=1}^{N} \alpha^{(i)}\left(1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right)$$

- Theorem: $\hat{\alpha}^{(i)}\left(1 - y^{(i)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \widehat{w}_0\right)\right) = 0 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$
  - If $\hat{\alpha}^{(s)} > 0$, then $1 - y^{(s)}\left(\widehat{\boldsymbol{w}}^T\boldsymbol{x}^{(s)} + \widehat{w}_0\right) = 0$

Computing $\hat{w}_0$

$$\hat{\alpha}^{(i)}\left(1 - y^{(i)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(i)} + \hat{w}_0\right)\right) = 0 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

If $\hat{\alpha}^{(s)} > 0 \rightarrow 1 - y^{(s)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = 0$

$$\rightarrow y^{(s)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = 1$$

$$\rightarrow {y^{(s)}}^2\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = y^{(s)}$$

$$\rightarrow \hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0 = y^{(s)}$$

$$\rightarrow \hat{w}_0 = y^{(s)} - \hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)}$$

Primal-Dual Optimization

- Primal
  - Directly returns the weights, $[\hat{w}_0, \hat{\boldsymbol{w}}]$
  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t.

$$y^{(s)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = 1$$

- Dual
  - Returns the vector, $\hat{\boldsymbol{\alpha}}$

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

$$\hat{w}_0 = y^{(s)} - \hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} \text{ for any } s \text{ s.t. } \hat{\alpha}^{(s)} > 0$$

  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t. $\hat{\alpha}^{(s)} > 0$

Primal-Dual Optimization

- Primal
  - $\hat{y} = \text{sign}(\widehat{\boldsymbol{w}}^T \vec{x} + \widehat{w}_0)$

- Dual
  - $\hat{y} = \text{sign}(\widehat{\boldsymbol{w}}^T \vec{x} + \widehat{w}_0)$

$$= \text{sign}\left(\left(\sum_{i=1}^{N} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)}\right)^T \boldsymbol{x} + \widehat{w}_0\right)$$

$$= \text{sign}\left(\sum_{i\,:\,\widehat{\alpha}^{(i)} > 0} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)T} \boldsymbol{x} + \widehat{w}_0\right)$$

Primal-Dual Soft-Margin SVMs

Primal

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)} \\ \text{subject to} \quad & y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 - \xi^{(i)} \; \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D} \\ & \xi^{(i)} \geq 0 \qquad \forall\, i \in \{1, \ldots, N\}\end{aligned}$$

Dual

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} - \sum_{i=1}^{N}\alpha^{(i)} \\ \text{subject to} \quad & \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0 \\ & 0 \leq \alpha^{(i)} \leq C \; \forall\, i \in \{1, \ldots, N\}\end{aligned}$$

Primal-Dual Soft-Margin SVMs

Primal

$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$\text{subject to } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 - \xi^{(i)} \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\xi^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, N\}$$

Dual

$$\text{minimize } \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\boldsymbol{x}^{(i)^T}\boldsymbol{x}^{(j)} - \sum_{i=1}^{N}\alpha^{(i)}$$

$$\text{subject to } \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0$$

$$0 \leq \alpha^{(i)} \leq C \ \forall\, i \in \{1, \dots, N\}$$

Primal-Dual Soft-Margin SVMs

- Primal
  - Directly returns the weights, $[\hat{w}_0, \hat{\boldsymbol{w}}]$
  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t.

$$y^{(s)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)} + \hat{w}_0\right) = 1$$

- Dual
  - Returns the vector, $\hat{\boldsymbol{\alpha}}$

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$$

$\hat{w}_0 = y^{(s)} - \hat{\boldsymbol{w}}^T \boldsymbol{x}^{(s)}$ for any s s.t. $0 < \hat{\alpha}^{(s)} < C$

  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t. $0 < \hat{\alpha}^{(s)} < C$
  - If $\hat{\alpha}^{(s)} = C$, then $\hat{\xi}^{(s)} > 0 \Rightarrow \left(\boldsymbol{x}^{(s)}, y^{(s)}\right)$ is inside the margin or misclassified

Key Takeaways

- SVMs provide a principled way of finding linear decision boundaries with maximal margins
  - Larger margins can lead to better generalization
  - Defined as a constrained optimization problem
  - Interpretation of solution and definition of support vectors
- Soft margins for linearly inseparable data
- Dual formulations
  - Interpretation of solution and definition of support vectors