10-701: Introduction to Machine Learning Lecture 24 - Support Vector Machines

Henry Chai

4/15/24

Front Matter

- Announcements
  - HW6 released 4/11, due **4/20 (Saturday)** at 11:59 PM

Final Exam Logistics

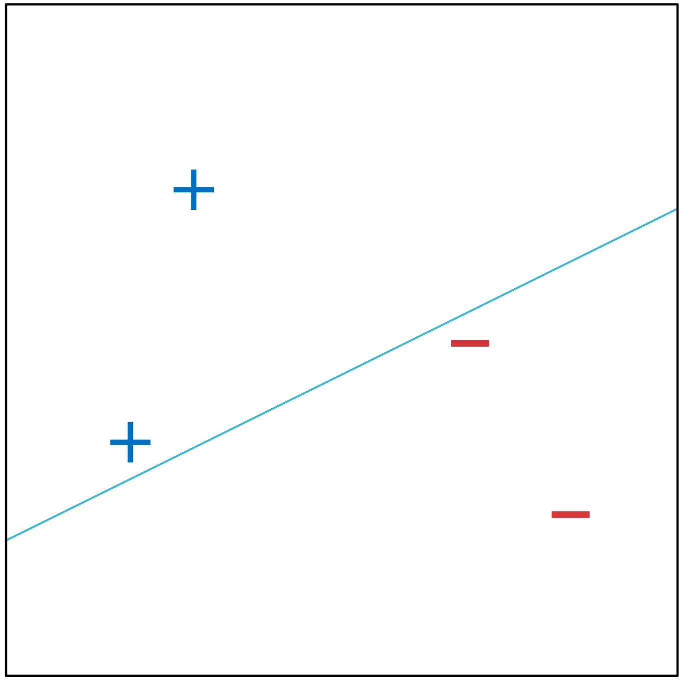
- Format of questions:
  - Multiple choice
  - True / False (with justification)
  - Derivations
  - (*Simple*) Proofs
  - Short answers
  - Drawing & Interpreting figures
  - Implementing algorithms on paper
- No electronic devices (you won't need them!)
- You are allowed to bring one letter-/A4-size sheet of notes; you can put *whatever* you want on *both sides*

Final Exam Topics

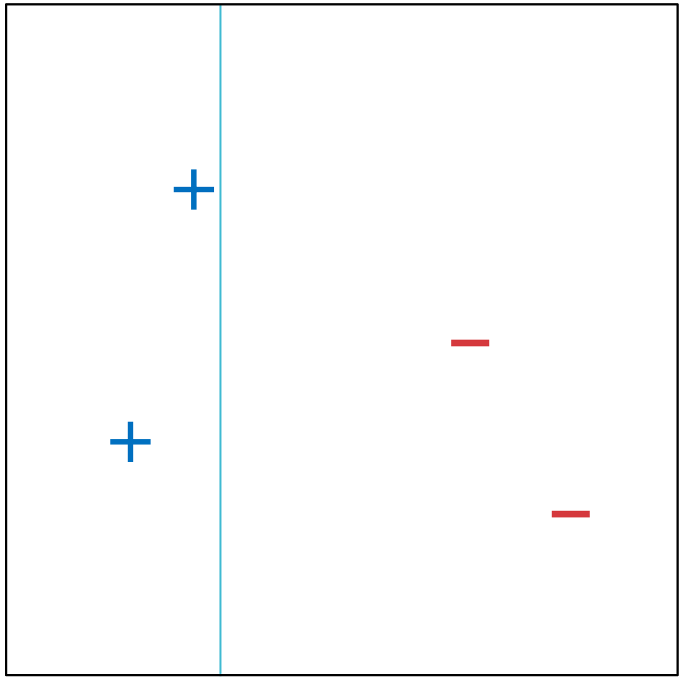
- Covered material: Lectures 14 - 25
  - Unsupervised Learning
  - Reinforcement Learning
  - Pretraining, fine-tuning and in-context learning
  - Algorithmic Bias
  - Learning Theory
  - Ensemble Methods
  - SVMs & Kernels
  - **Pre-midterm material may be referenced but will not be the primary focus of any question**

Final Exam Preparation

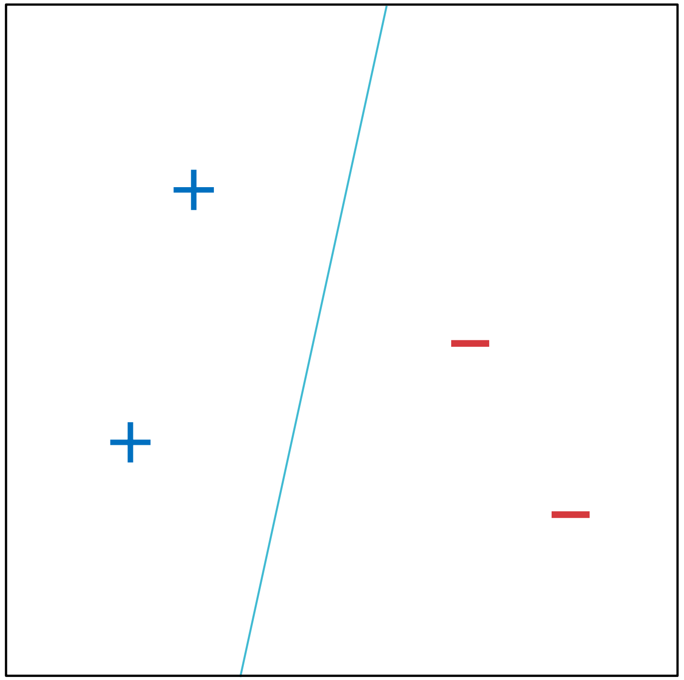
- Review the exam practice problems (to be released on 4/22 to the course website, under the Recitations tab)
- Attend the dedicated final exam review recitation (4/26)
- Review HWs 5 - 6
- Review the key takeaways throughout the lecture slides
- Write your one-page cheat sheet (back and front)



Option A

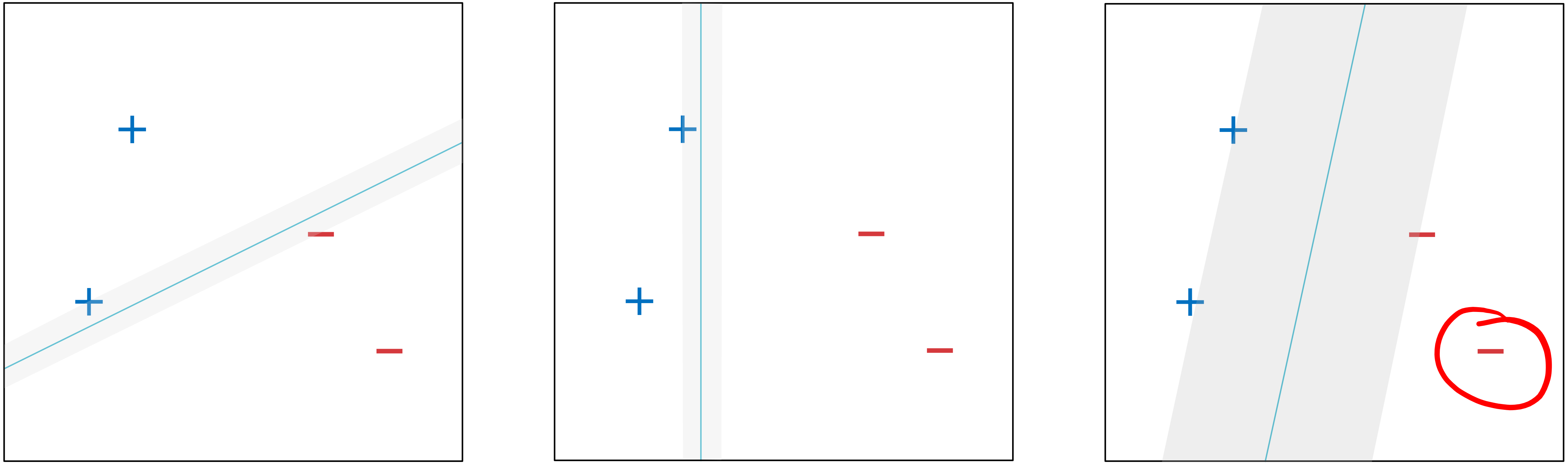


Option B



Option C

Which linear separator is best?



Which linear separator is best?

Maximal Margin Linear Separators

- 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, $[b, \boldsymbol{w}]$

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

- Problem: there are infinitely many weight vectors that describe the same hyperplane
  - $x_1 + 2x_2 + 2 = 0$ is the same line as $2x_1 + 4x_2 + 4 = 0$, which is the same line as $1000000x_1 + 2000000x_2 + 2000000 = 0$
- 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 $y \in \{-1, +1\}$, $\hat{y} = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ 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

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

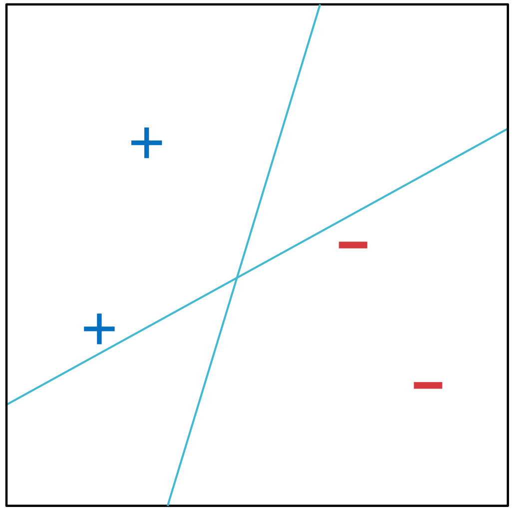
- If $\hat{y} = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ is a linear separator, then ???

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

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

Normalizing Hyperplanes: Example

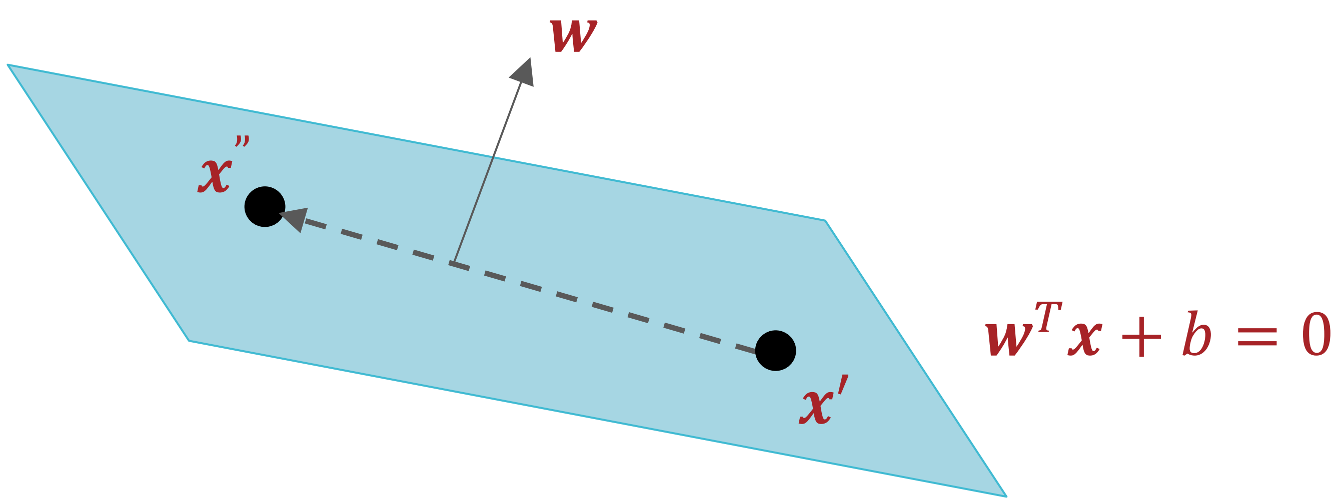
b	w_1	w_2	
-0.2	-0.6	1	$\notin \mathcal{H}$
-0.4	-1.2	2	$\notin \mathcal{H}$
-2	-6	10	$\notin \mathcal{H}$
-10	-30	50	$\in \mathcal{H}$
0.2	-0.6	0.2	$\notin \mathcal{H}$
0.1	-0.3	0.1	$\notin \mathcal{H}$
1	-3	1	$\notin \mathcal{H}$
2	-6	2	$\in \mathcal{H}$



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

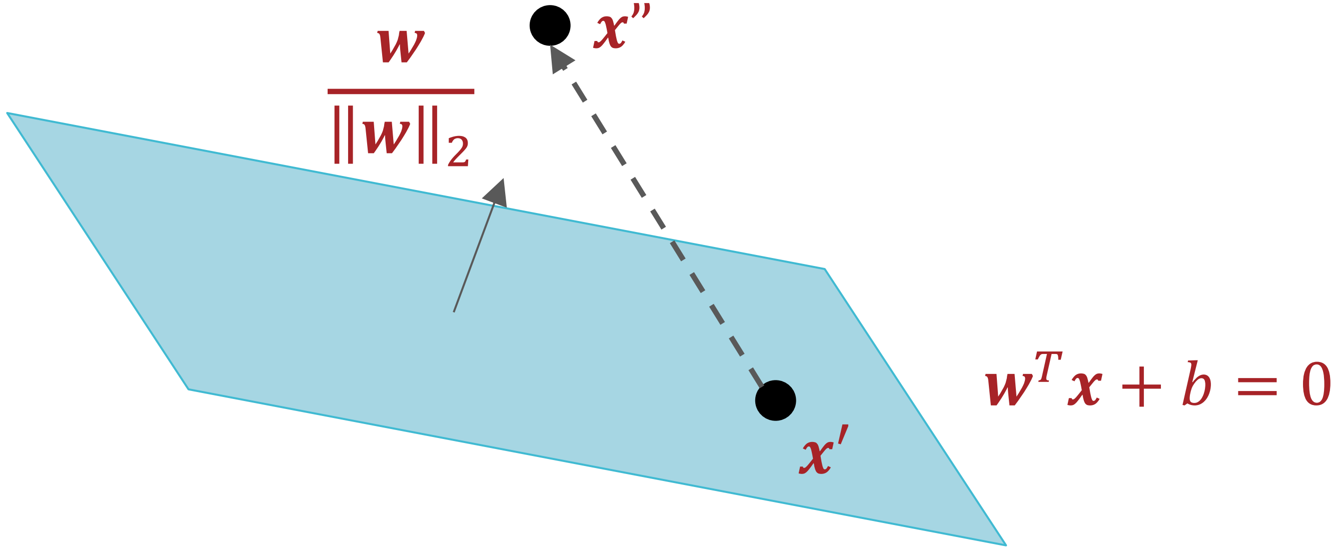
Computing the Margin

- Claim: $\boldsymbol{w}$ is orthogonal to the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ (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 $\boldsymbol{\alpha}^T\boldsymbol{\beta} = 0$



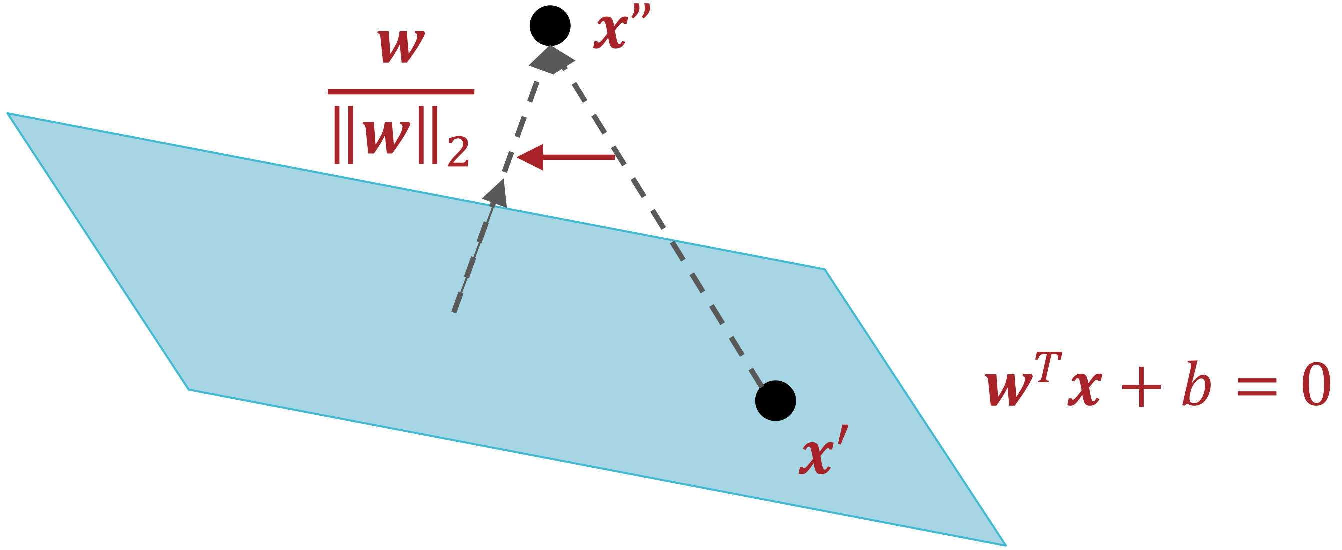
Computing the Margin

- 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



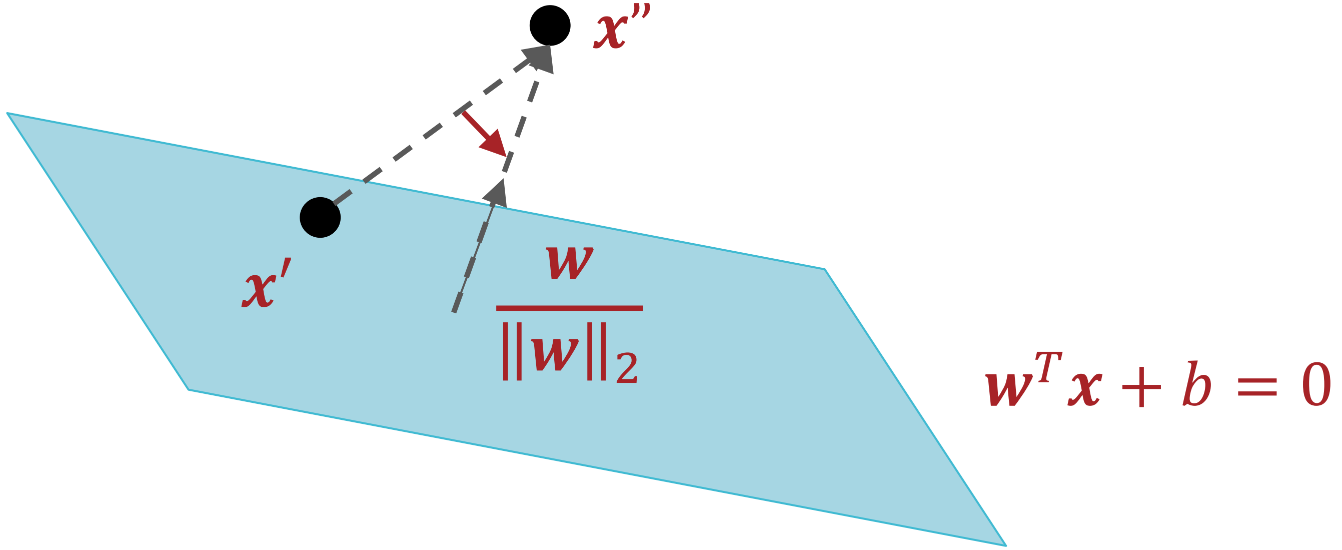
Computing the Margin

- 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 $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane

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

$x' \in h$

$\Rightarrow w^T x' + b = 0$

$\Rightarrow -w^T x' = b$

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

Computing the Margin

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

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

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

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

Maximizing the Margin

$\hat{w}, \hat{b}$

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

$$\Updownarrow$$

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

$$\Updownarrow$$

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{subject to} \quad \min_{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

$$\Updownarrow$$

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

Maximizing the Margin

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

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

- If $[\hat{b}, \hat{\boldsymbol{w}}]$ is the optimal solution, then ∃ at least one training data point $(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}$ s.t $y^{(i)}(\hat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}) = 1$
  - All training data points $(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathcal{D}$ where $y^{(i)}(\hat{\boldsymbol{w}}^T\boldsymbol{x}^{(i)} + \hat{b}) = 1$ 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: $\boldsymbol{w} = [w_1, \dots, w_D]$ and b
- Write down an objective function (with constraints)

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

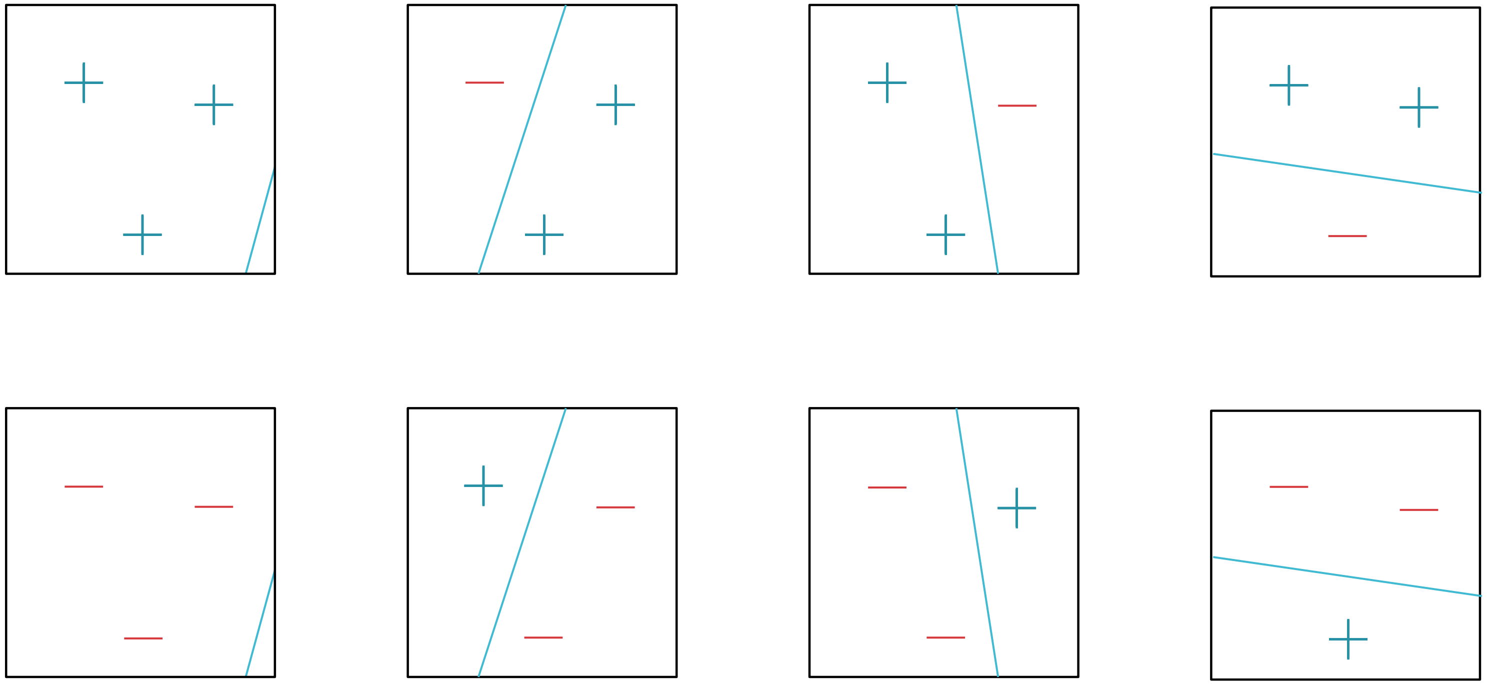
$$\text{subject to} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

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

Why Maximal Margins?

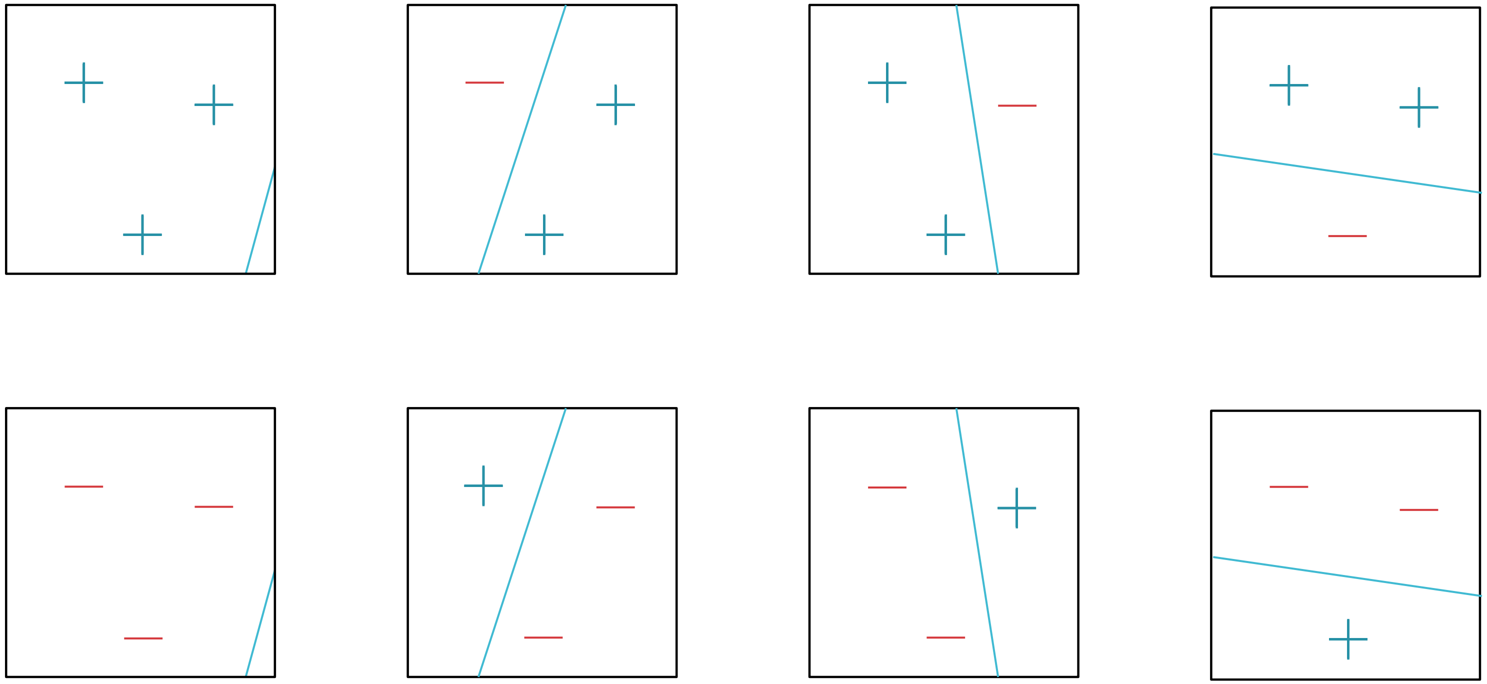
- Consider three binary data points in a **bounded** 2-D space
- Let $\mathcal{H}$ = {all linear separators} and

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



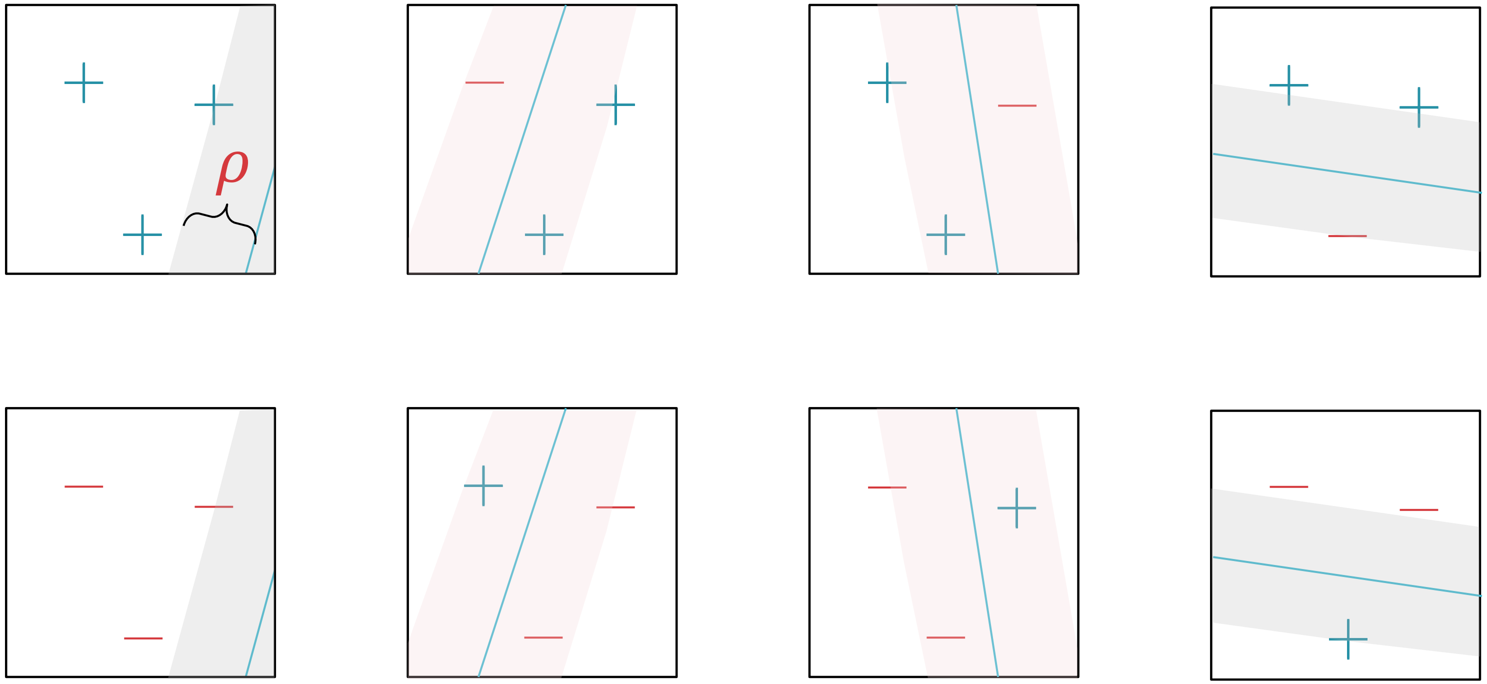
Why Maximal Margins?

- 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

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

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

Hard-margin SVMs

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

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

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

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

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

Soft-margin SVMs

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

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

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

- $\xi^{(i)}$ 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

error

Soft-margin SVMs

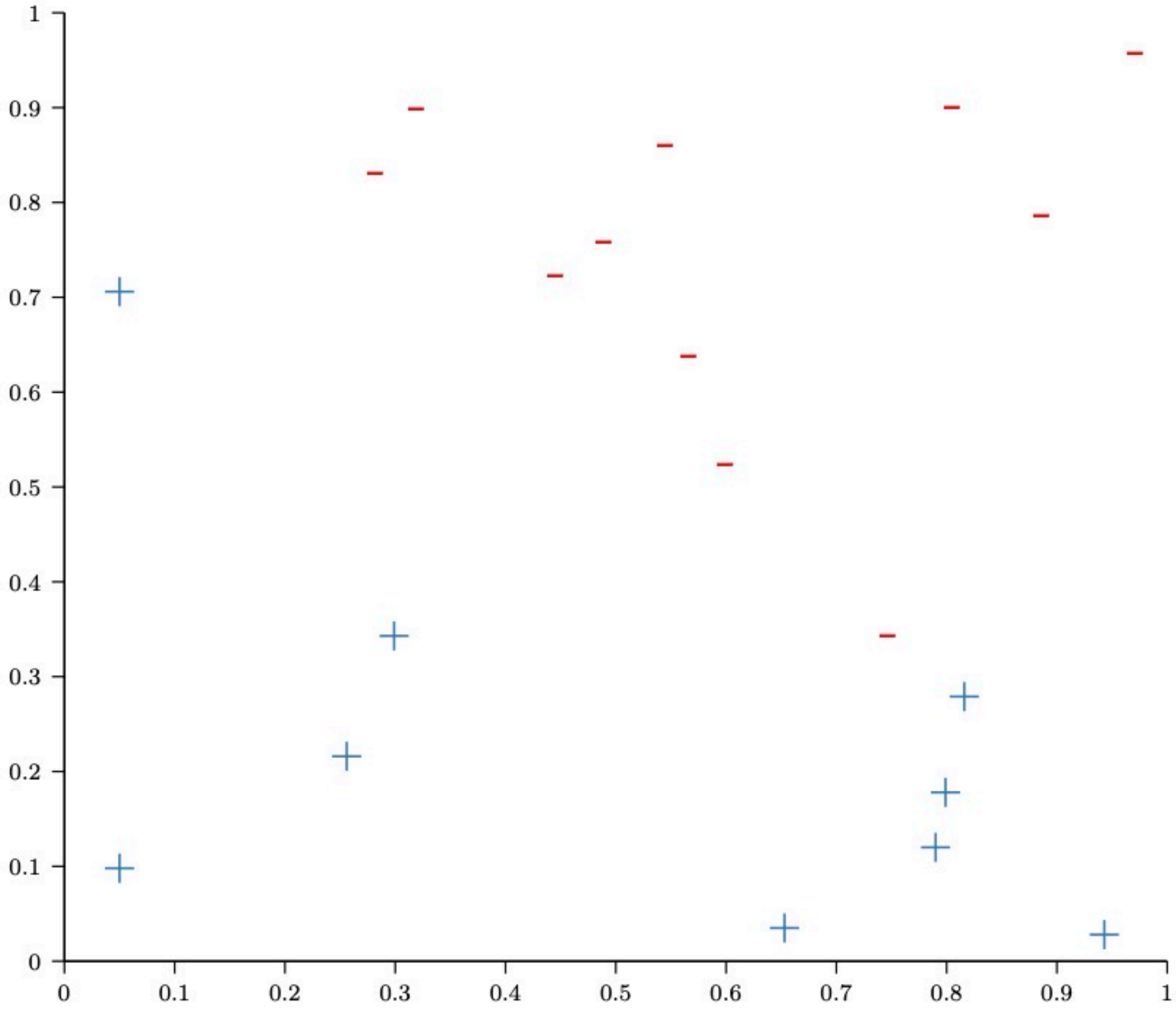
$$\text{minimize } \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

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

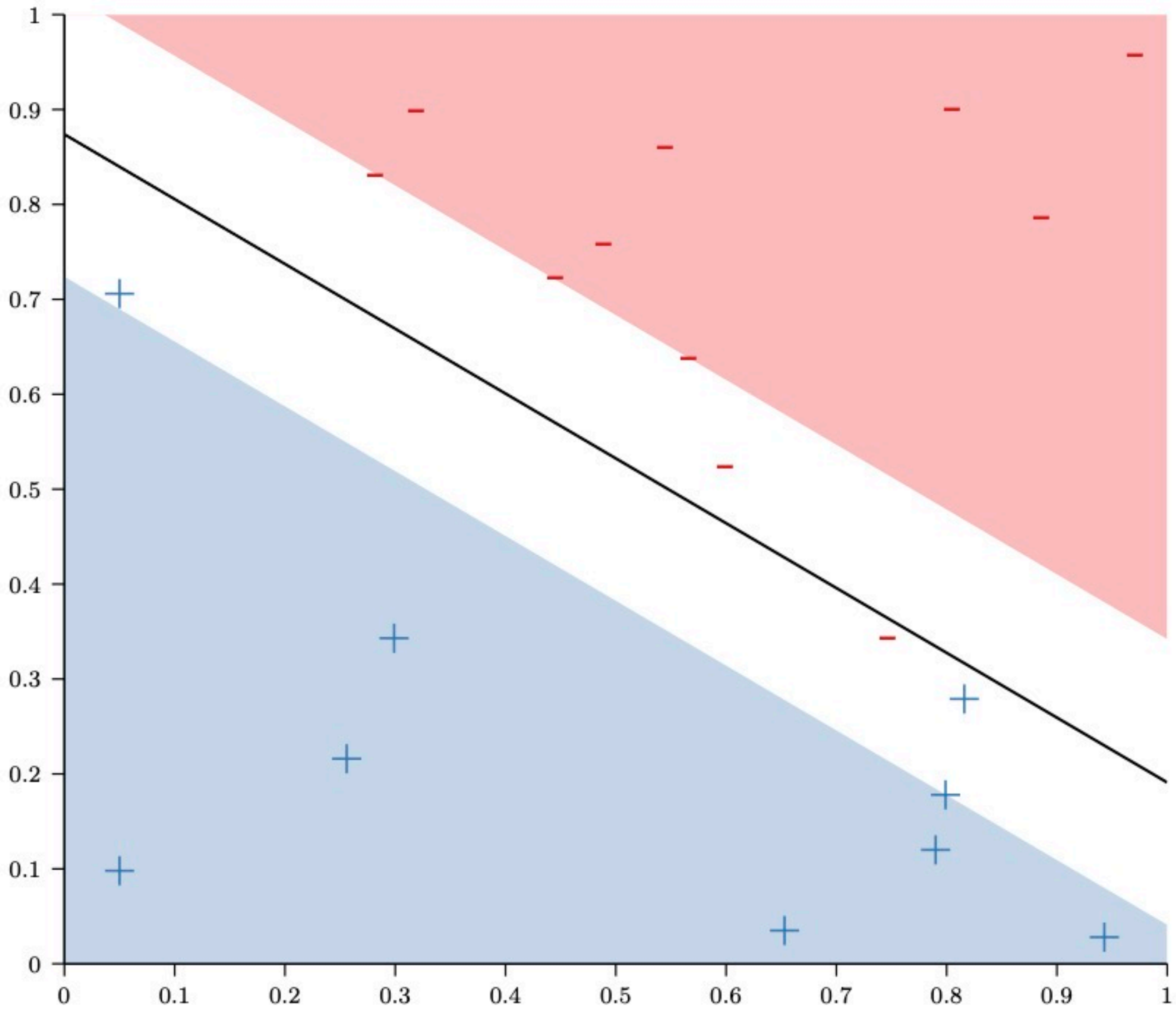
$$\xi^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, N\}$$

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

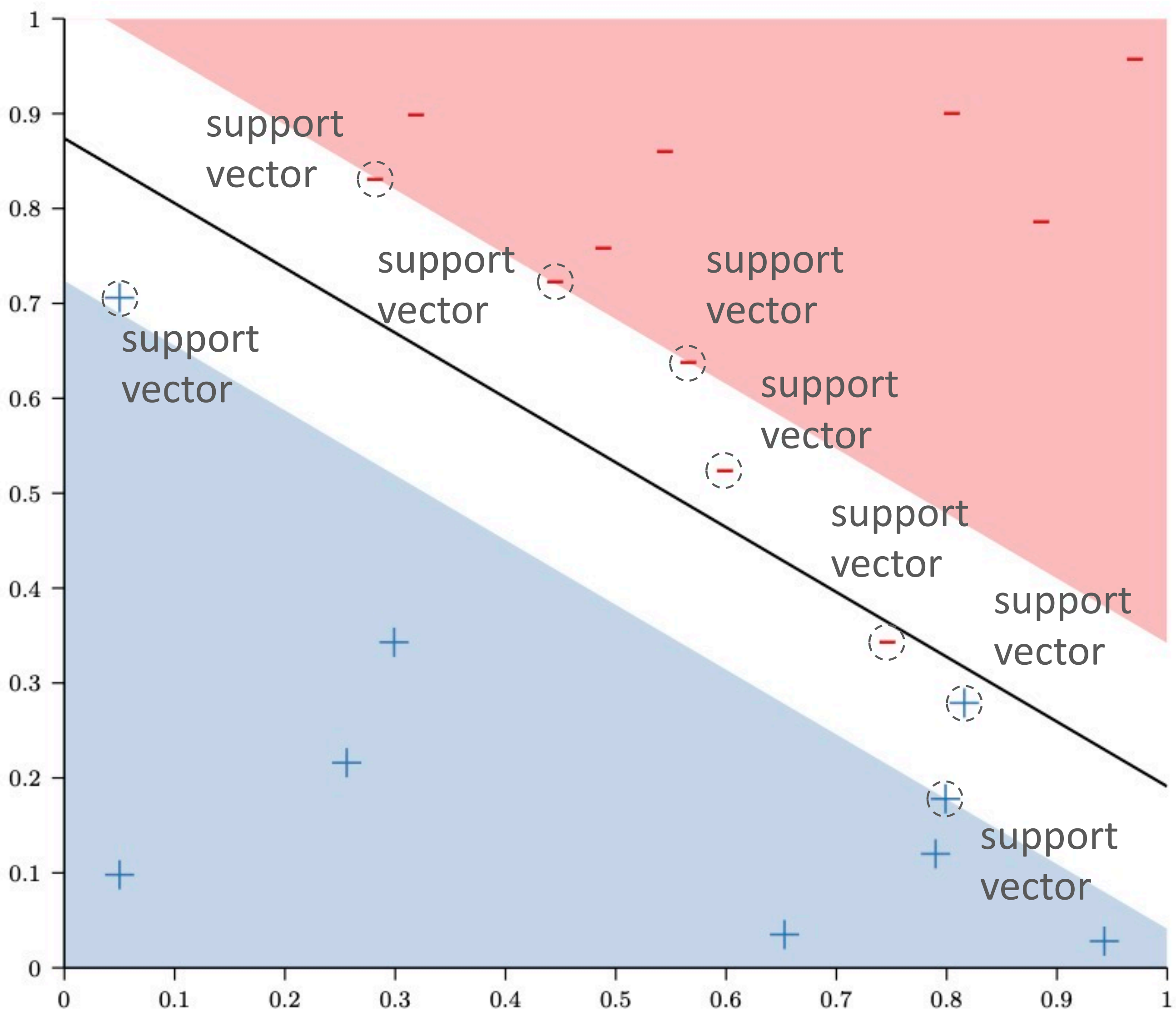
Interpreting $\xi^{(i)}$



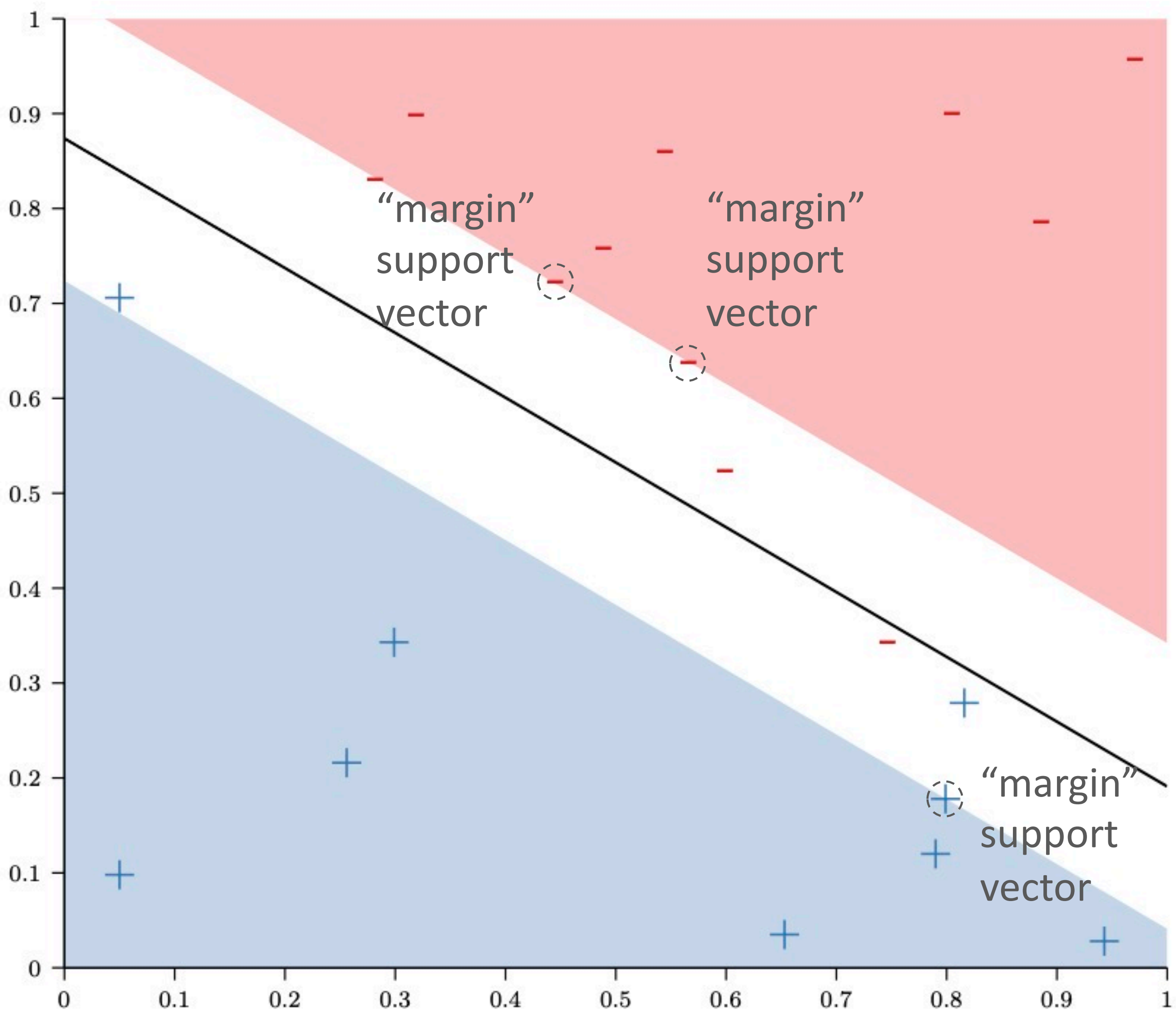
Interpreting $\xi^{(i)}$



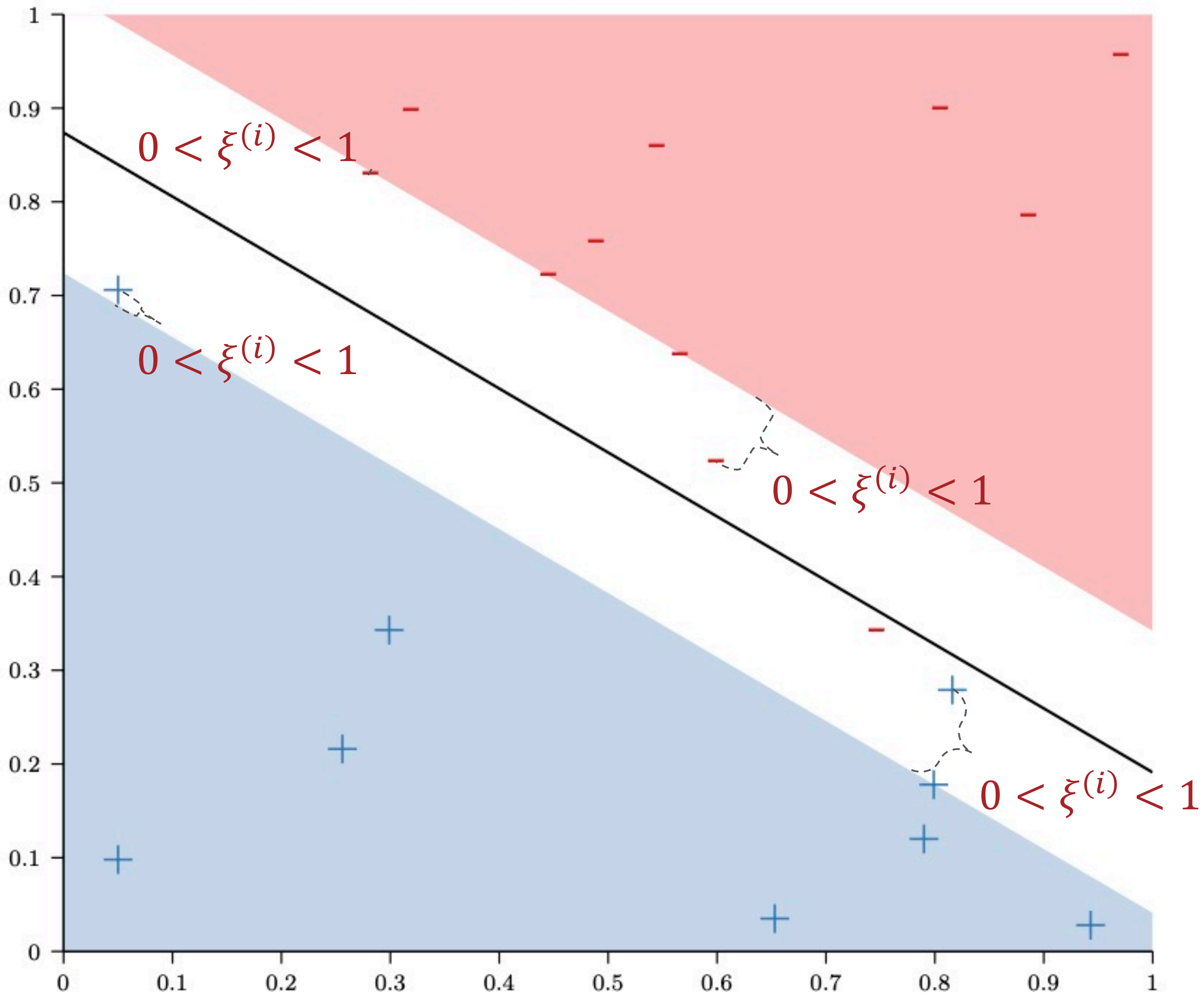
Interpreting $\xi^{(i)}$



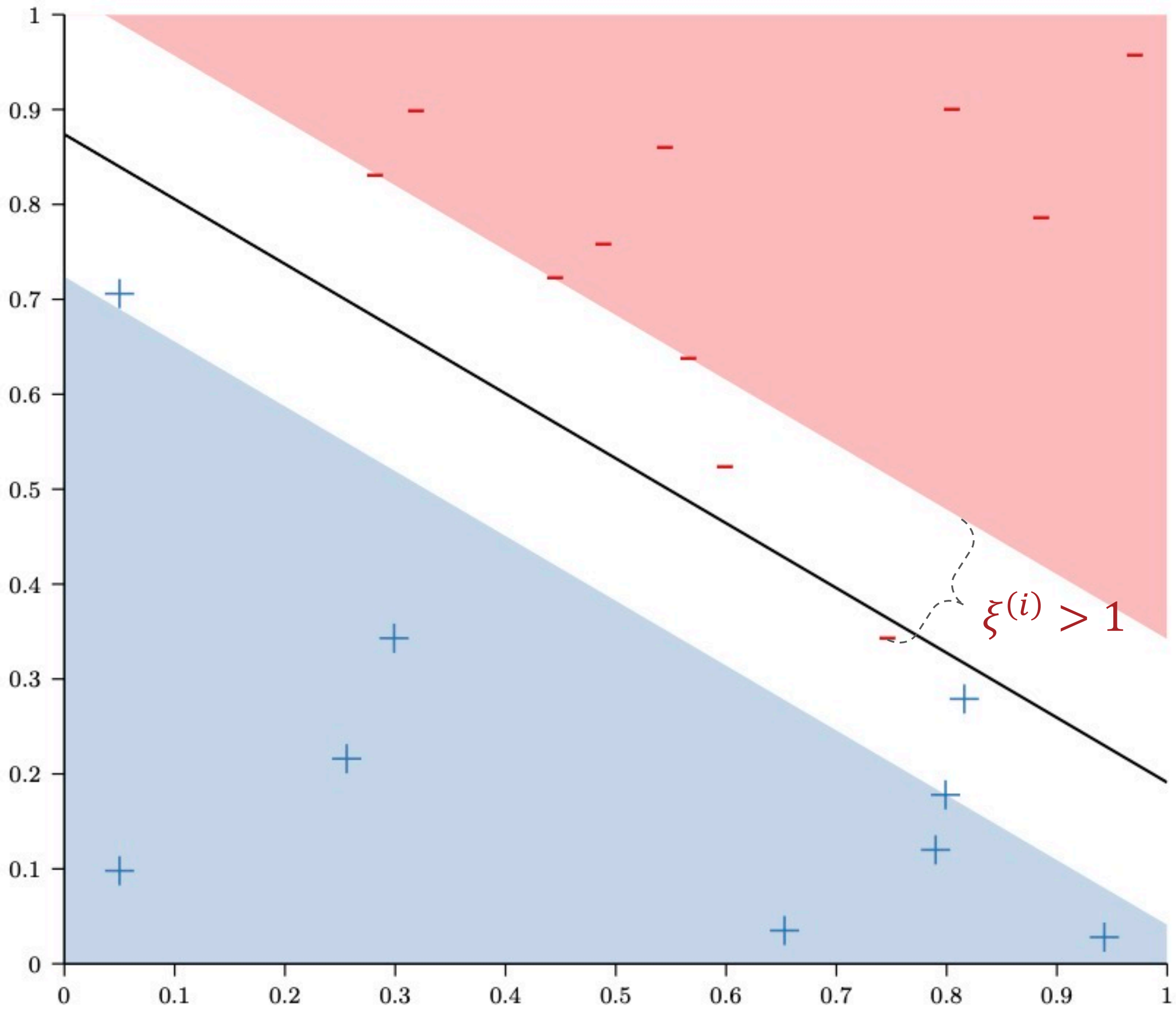
Interpreting $\xi^{(i)}$

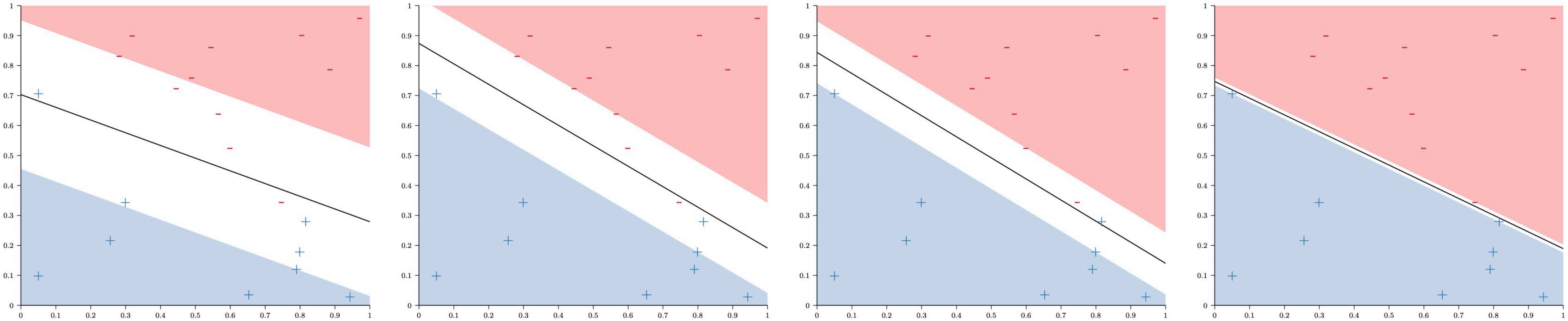


Interpreting $\xi^{(i)}$



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

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

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

} SVMs

minimize E_{train}

subject to $\boldsymbol{w}^T\boldsymbol{w} \leq C$

} Regularization

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

Primal-Dual Optimization

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

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

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

$$\underset{}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$
$$\text{subject to} \ y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\Updownarrow$$

$$\text{minimize} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$
$$\text{subject to} \ 1 - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \leq 0 \ \forall \left(\boldsymbol{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$$

$$\Updownarrow$$

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

SVM

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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