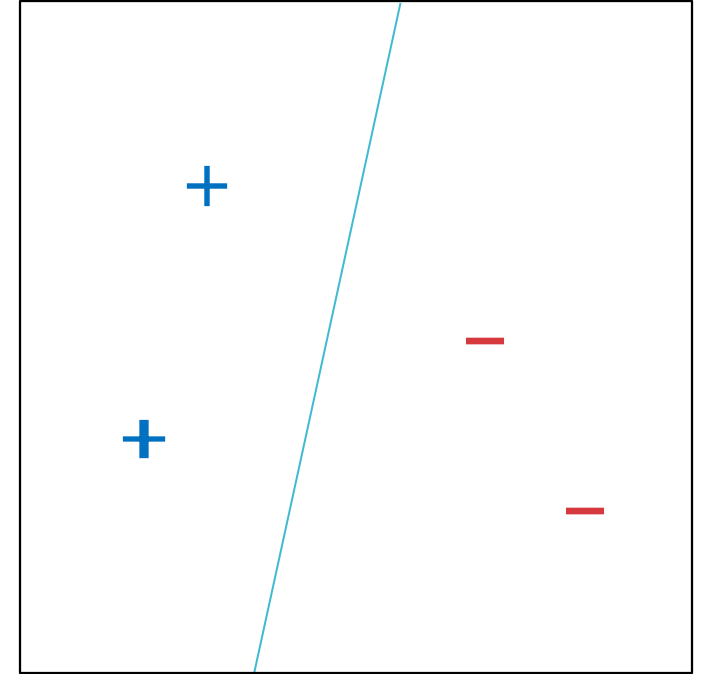
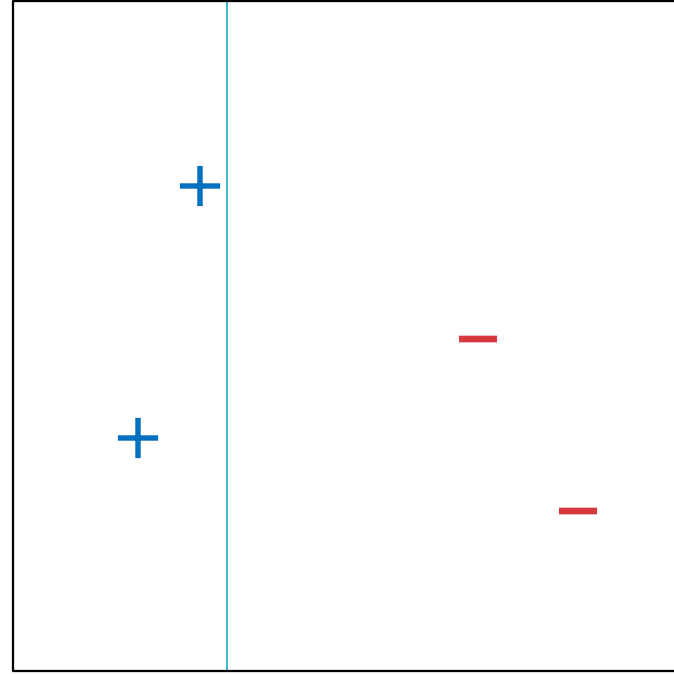
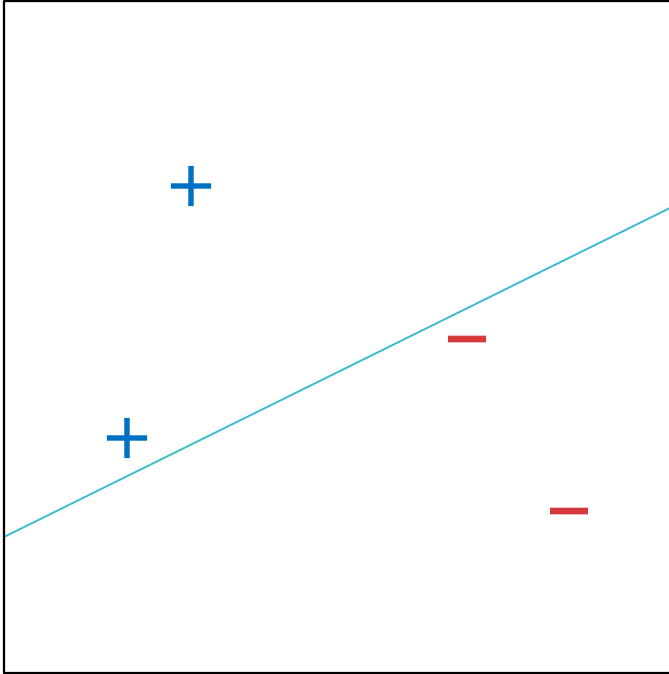


10-701: Introduction to Machine Learning

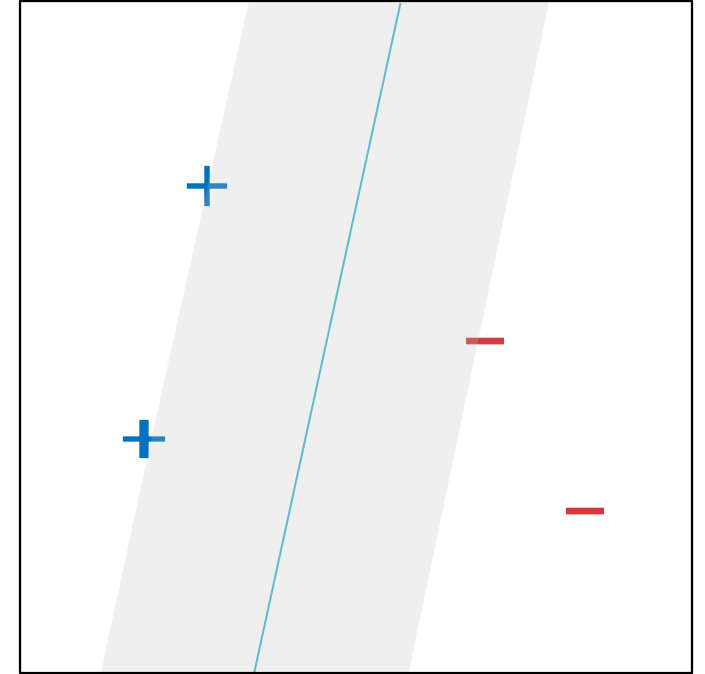
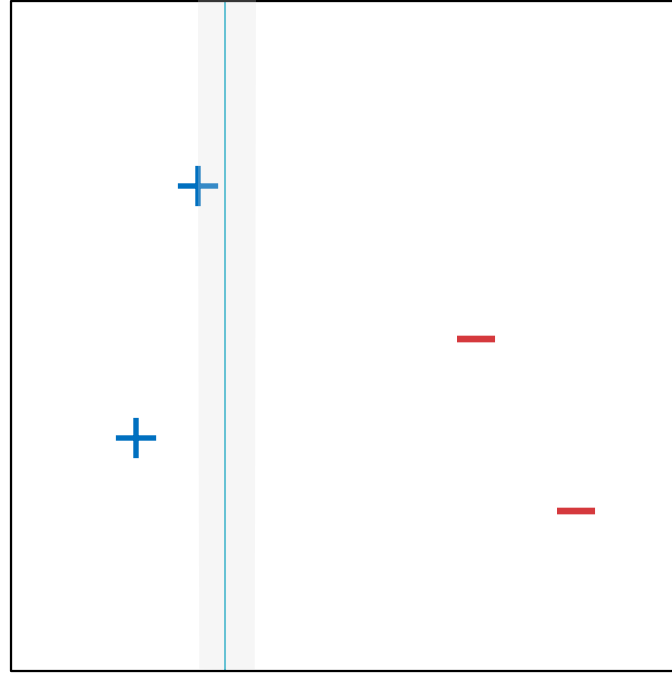
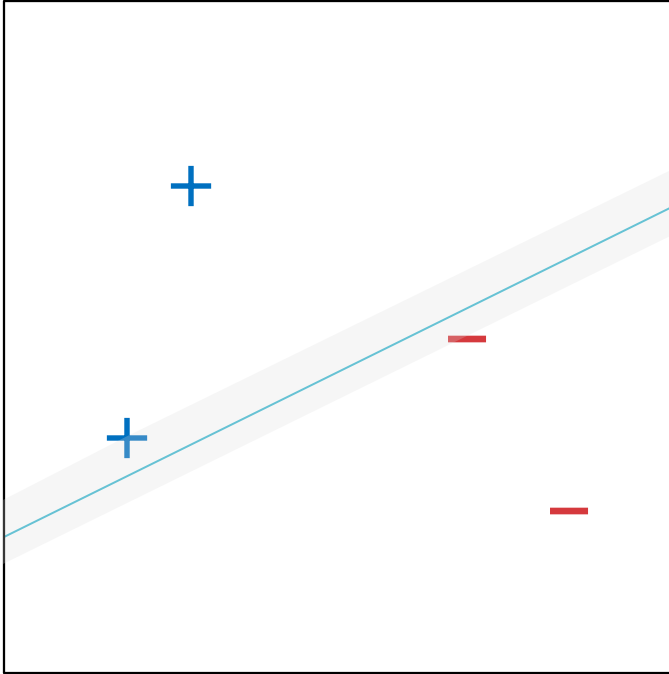
Lecture 23 –Support Vector Machines

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

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

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

- 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 $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ where $y \in \{-1, +1\}$, $\hat{y} = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$ is a valid **linear separator** if

$$y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) > 0 \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$$

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

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

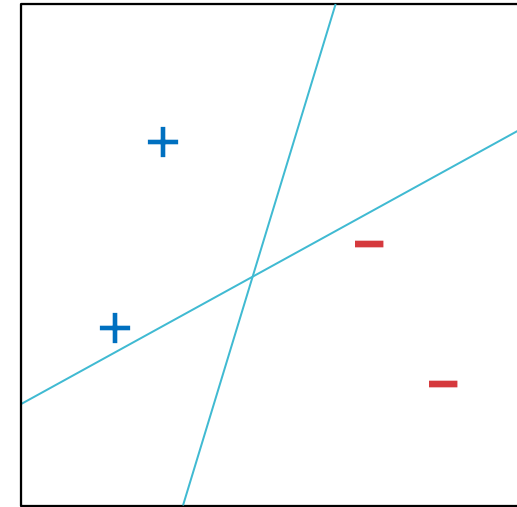
- If $\hat{y} = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$ is a linear separator, then

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

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

Normalizing Hyperplanes: Example

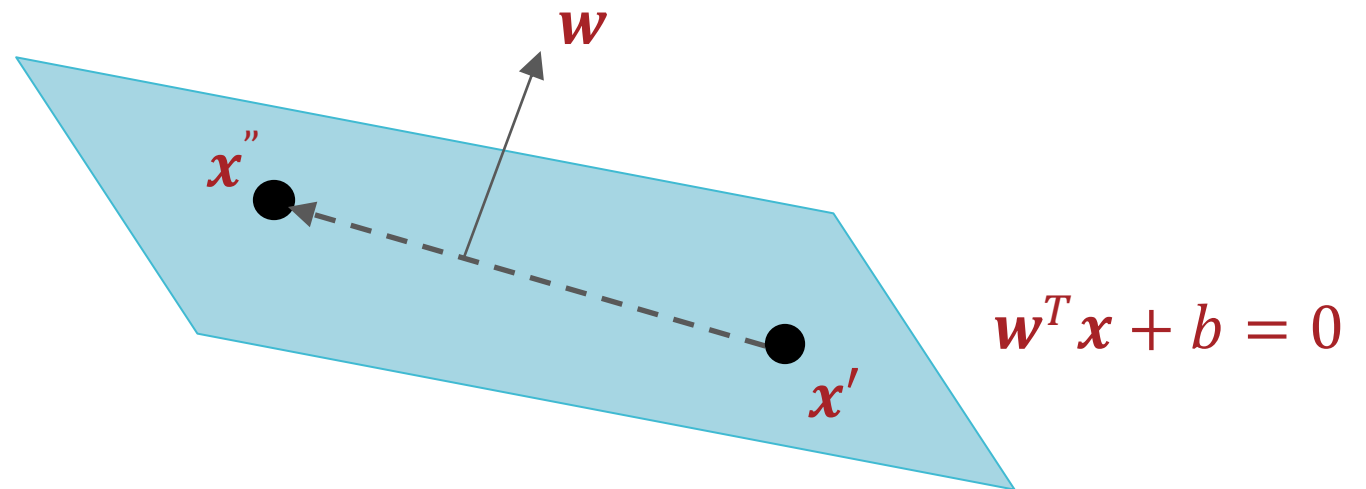
| b | w_1 | w_2 | |
|------|-------|-------|----------------------|
| -0.2 | -0.6 | 1 | $\notin \mathcal{H}$ |
| -0.4 | -1.2 | 2 | $\notin \mathcal{H}$ |
| -2 | -6 | 10 | $\notin \mathcal{H}$ |
| -10 | -30 | 50 | $\in \mathcal{H}$ |
| 0.2 | -0.6 | 0.2 | $\notin \mathcal{H}$ |
| 0.1 | -0.3 | 0.1 | $\notin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



| x_1 | x_2 | y | $y(w^T x + b)$ |
|-------|-------|-----|----------------|
| 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 |

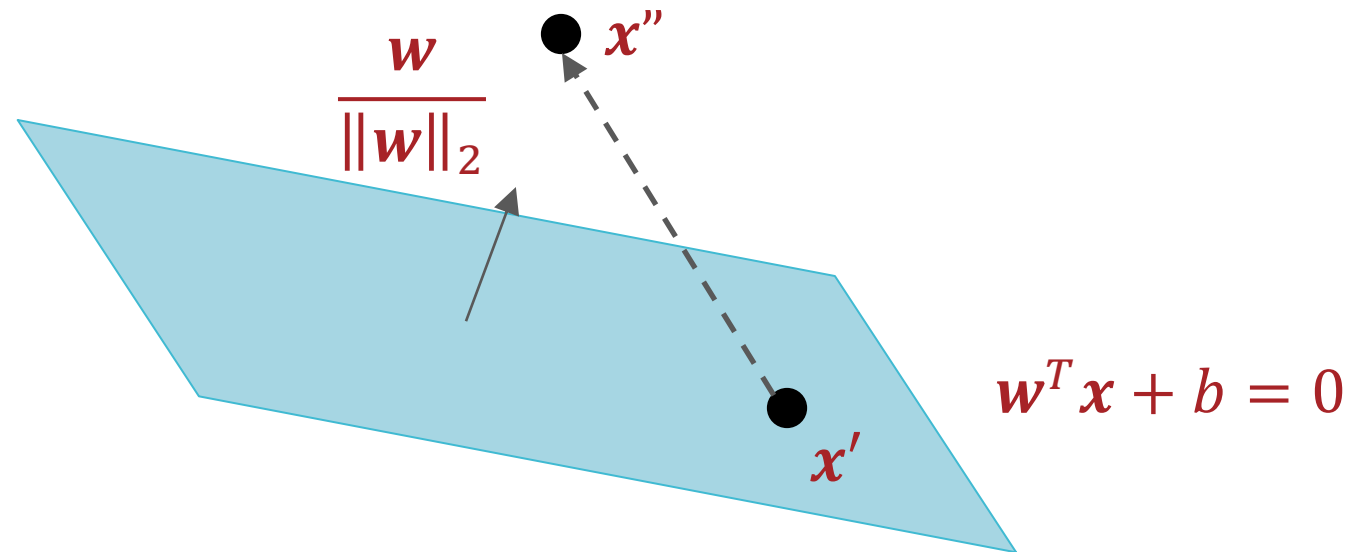
Computing the Margin

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



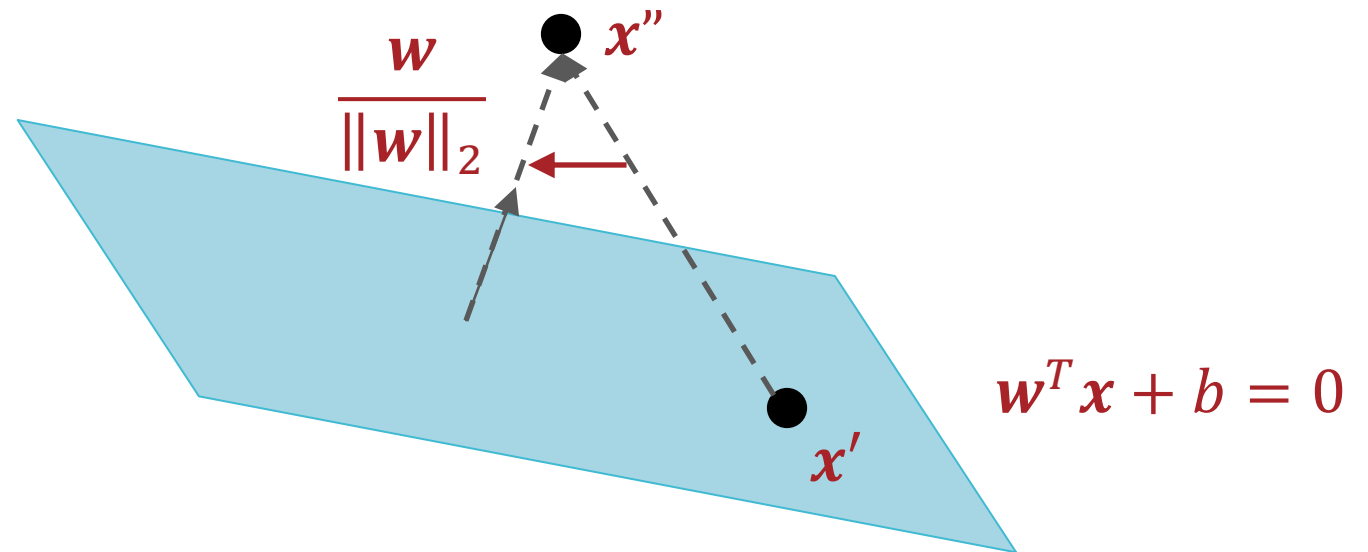
Computing the Margin

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



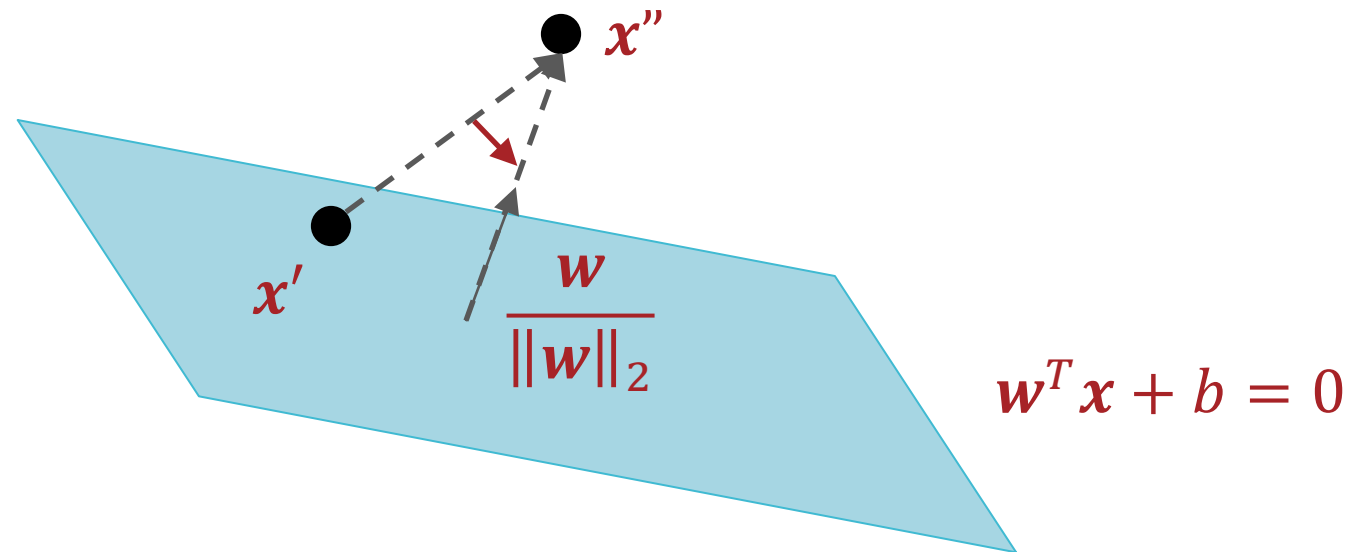
Computing the Margin

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



Computing the Margin

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



Computing the Margin

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

$$\begin{aligned} d(\mathbf{x}'', h) &= \left| \frac{\mathbf{w}^T (\mathbf{x}'' - \mathbf{x}')}{\|\mathbf{w}\|_2} \right| = \frac{|\mathbf{w}^T \mathbf{x}'' - \mathbf{w}^T \mathbf{x}'|}{\|\mathbf{w}\|_2} \\ &= \frac{|\mathbf{w}^T \mathbf{x}'' + b|}{\|\mathbf{w}\|_2} \end{aligned}$$

Computing the Margin

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

$$\begin{aligned} \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} d(\mathbf{x}^{(i)}, h) &= \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{|\mathbf{w}^T \mathbf{x}^{(i)} + b|}{\|\mathbf{w}\|_2} \\ &= \frac{1}{\|\mathbf{w}\|_2} \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} |\mathbf{w}^T \mathbf{x}^{(i)} + b| \\ &= \frac{1}{\|\mathbf{w}\|_2} \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) \\ &= \frac{1}{\|\mathbf{w}\|_2} \end{aligned}$$

Maximizing the Margin

$$\begin{aligned} & \text{maximize} \quad \frac{1}{\|\mathbf{w}\|_2} \\ & \text{subject to} \quad \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) = 1 \\ & \quad \Updownarrow \\ & \text{minimize} \quad \|\mathbf{w}\|_2 \\ & \text{subject to} \quad \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) = 1 \\ & \quad \Updownarrow \\ & \text{minimize} \quad \frac{1}{2} \|\mathbf{w}\|_2^2 \\ & \text{subject to} \quad \min_{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}} y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) = 1 \\ & \quad \Updownarrow \\ & \text{minimize} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ & \text{subject to} \quad y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} \end{aligned}$$

Maximizing the Margin

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

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

- If $[\hat{b}, \hat{\mathbf{w}}]$ is the optimal solution, then \exists at least one training data point $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$ s.t $y^{(i)}(\hat{\mathbf{w}}^T \mathbf{x}^{(i)} + \hat{b}) = 1$
 - All training data points $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$ where $y^{(i)}(\hat{\mathbf{w}}^T \mathbf{x}^{(i)} + \hat{b}) = 1$ 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

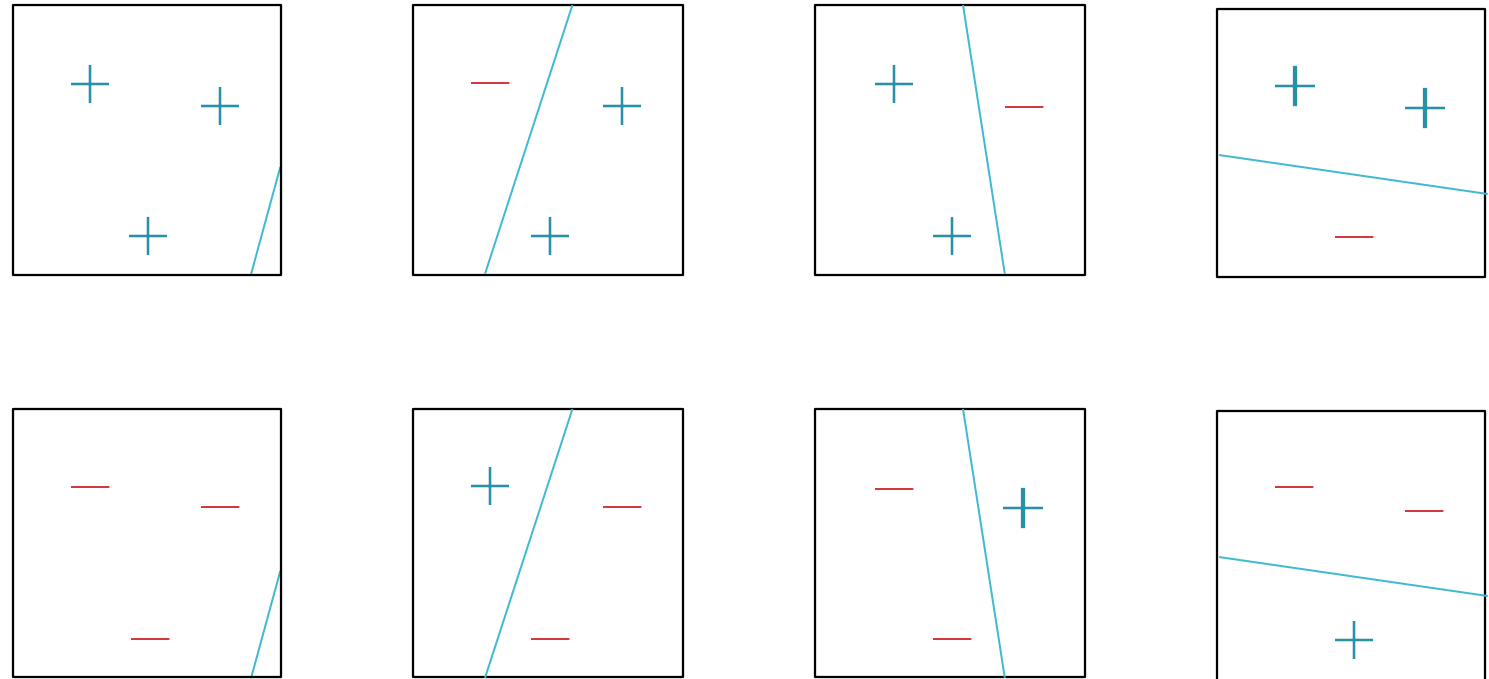
- Define a model and model parameters
 - Assume a linear decision boundary (with normalized weights)

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

- Parameters: $\mathbf{w} = [w_1, \dots, w_D]$ and b
- Write down an objective function (with constraints)
minimize $\frac{1}{2} \mathbf{w}^T \mathbf{w}$
subject to $y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$
- Optimize the objective w.r.t. the model parameters
 - Solve using quadratic programming

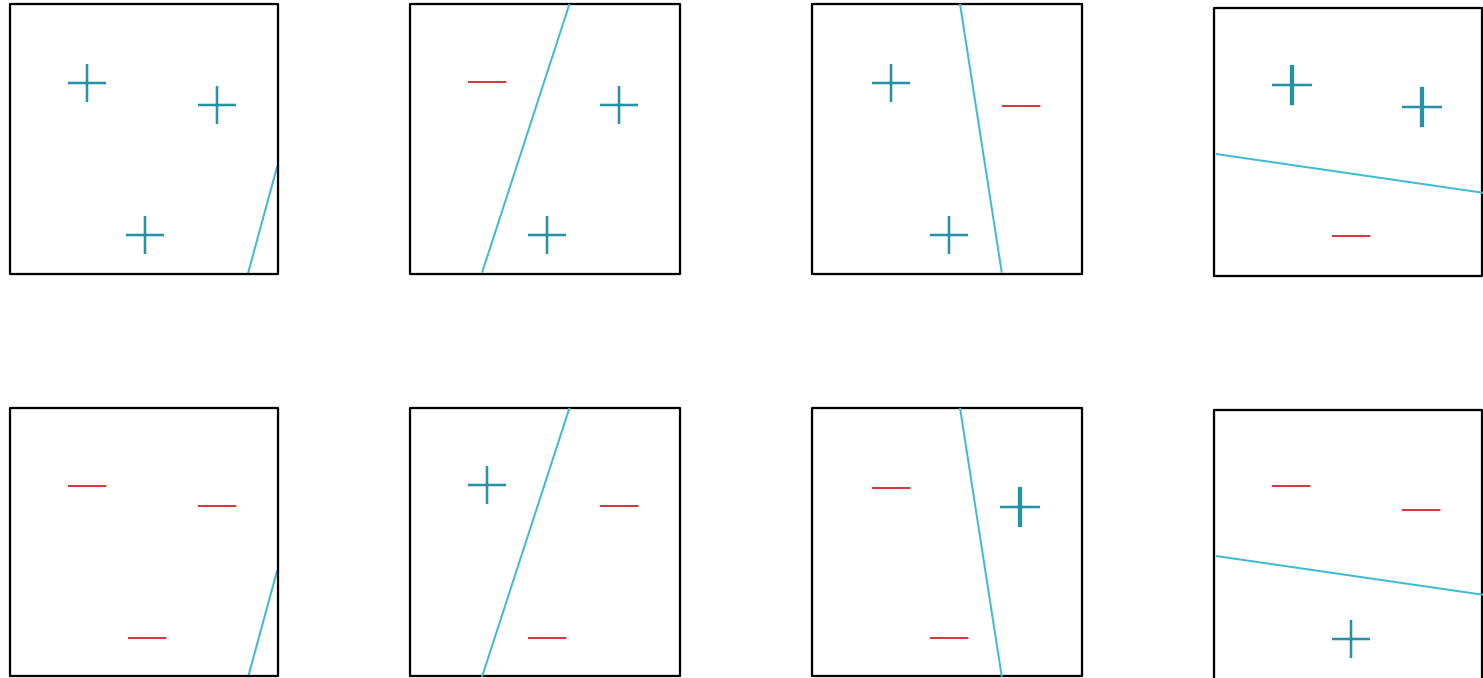
Why Maximal Margins?

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



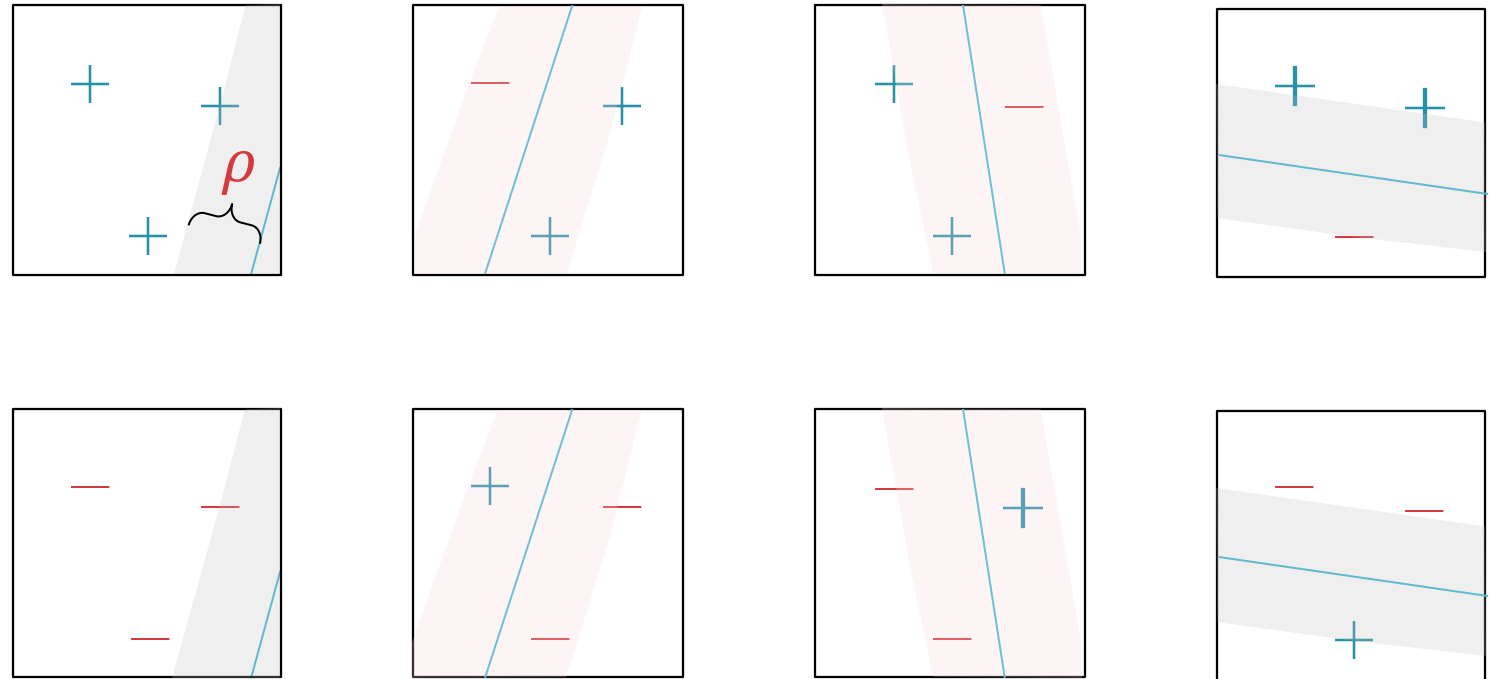
Why Maximal Margins?

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



Why Maximal Margins?

- Consider three binary data points in a **bounded** 2-D space
- $\mathcal{H}_\rho = \{\text{all linear separators with minimum margin } \rho\}$ cannot always correctly classify three non-collinear 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 $\frac{1}{2} \mathbf{w}^T \mathbf{w}$

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

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

Hard-margin SVMs

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

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

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

Soft-margin SVMs

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi^{(i)} \\ &\text{subject to} \quad y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} \\ &\quad \quad \quad \xi^{(i)} \geq 0 \quad \quad \quad \forall i \in \{1, \dots, N\} \end{aligned}$$

Soft-margin SVMs

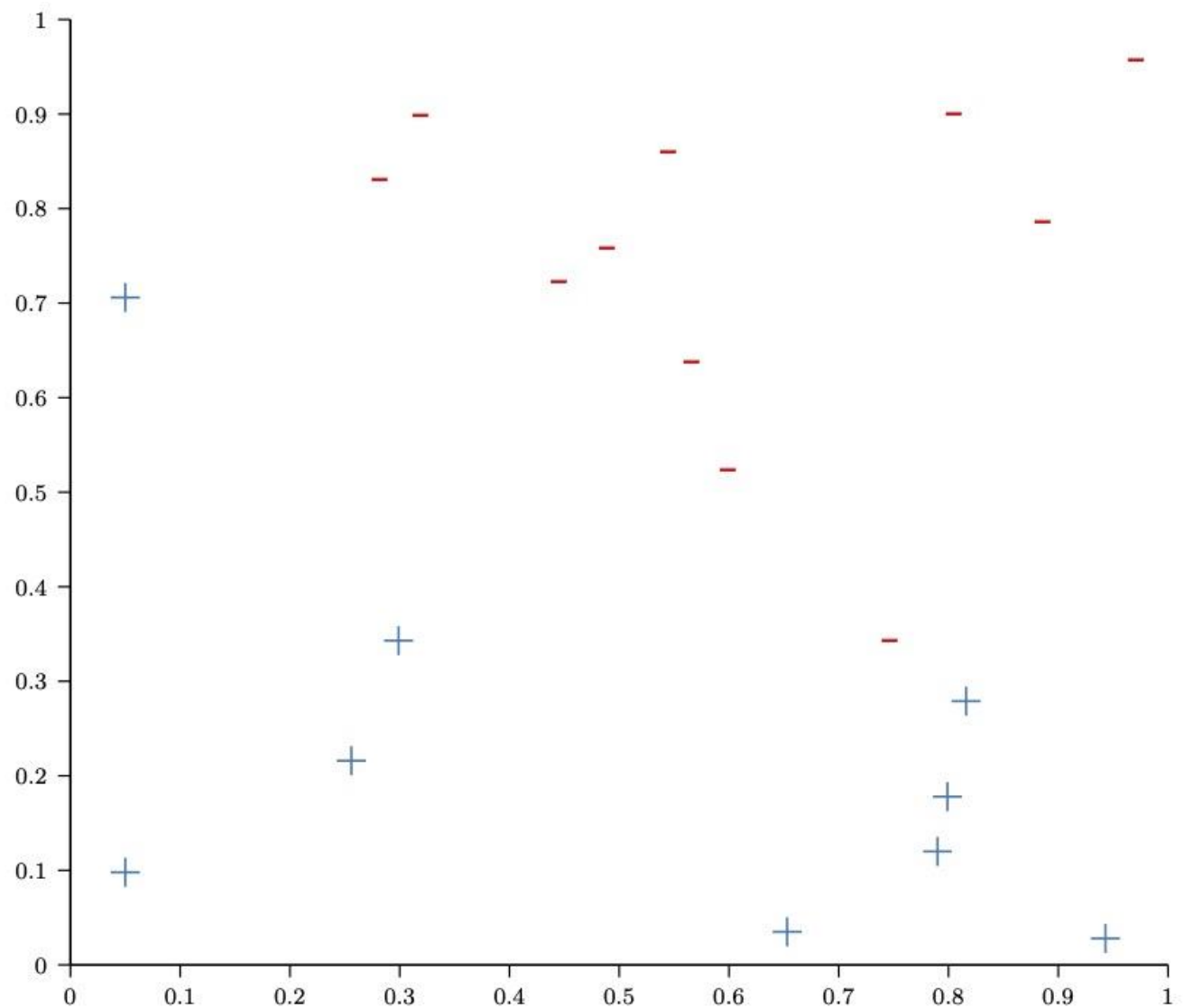
- minimize $\frac{1}{2} \mathbf{w}^T \mathbf{w} + c \sum_{i=1}^N \xi^{(i)}$
- subject to $y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$
- $\xi^{(i)} \geq 0 \quad \forall i \in \{1, \dots, N\}$
- $\xi^{(i)}$ is the “soft” error on the i^{th} training data point
 - If $\xi^{(i)} > 1$, then $y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) < 0 \Rightarrow (\mathbf{x}^{(i)}, y^{(i)})$ is incorrectly classified
 - If $0 < \xi^{(i)} < 1$, then $y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) > 0 \Rightarrow (\mathbf{x}^{(i)}, y^{(i)})$ is correctly classified but inside the margin
 - $\sum_{i=1}^N \xi^{(i)}$ is the “soft” training error

Soft-margin SVMs

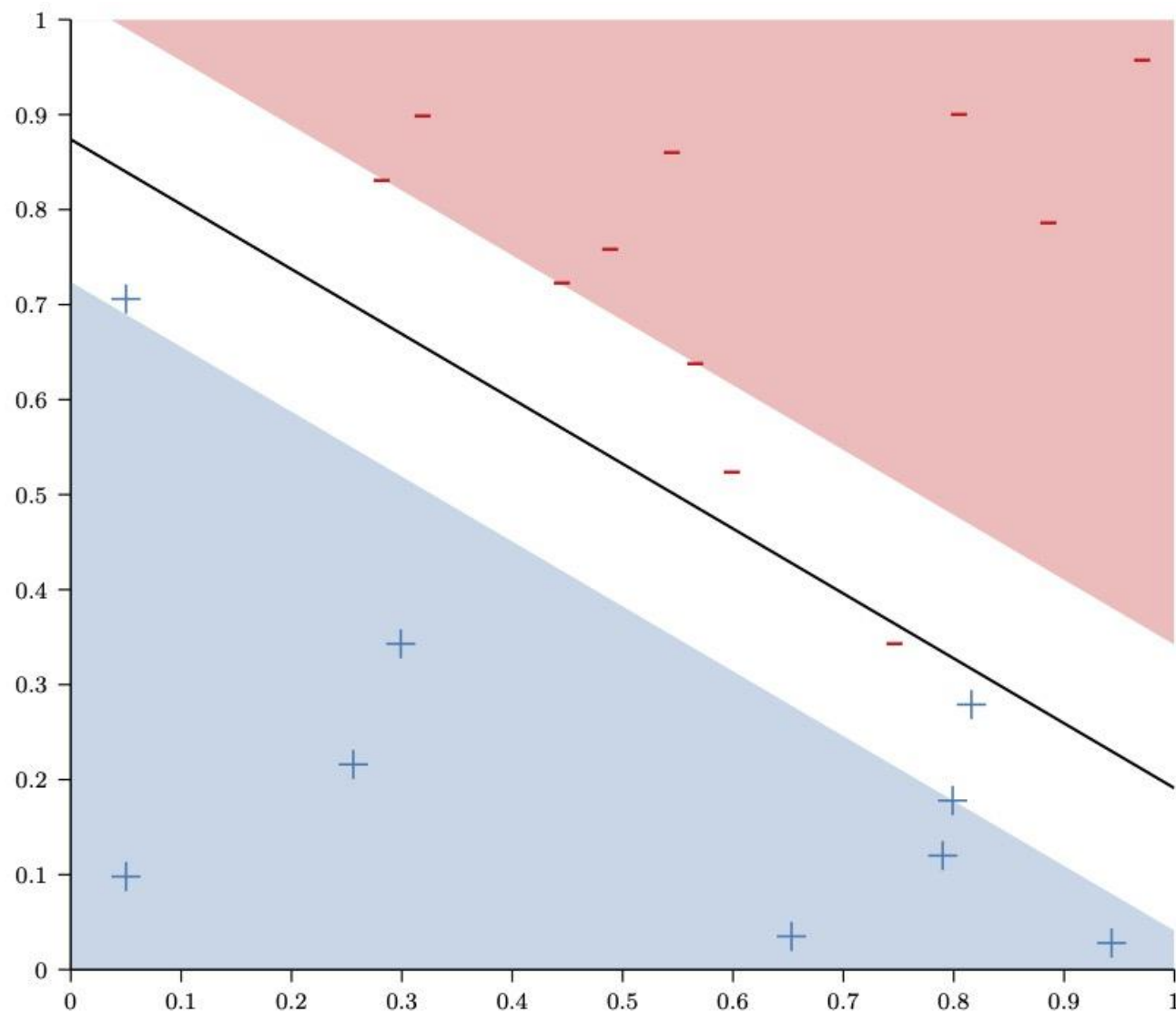
$$\begin{aligned} &\text{minimize } \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi^{(i)} \\ &\text{subject to } y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \xi^{(i)} \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} \\ &\quad \xi^{(i)} \geq 0 \quad \forall i \in \{1, \dots, N\} \end{aligned}$$

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

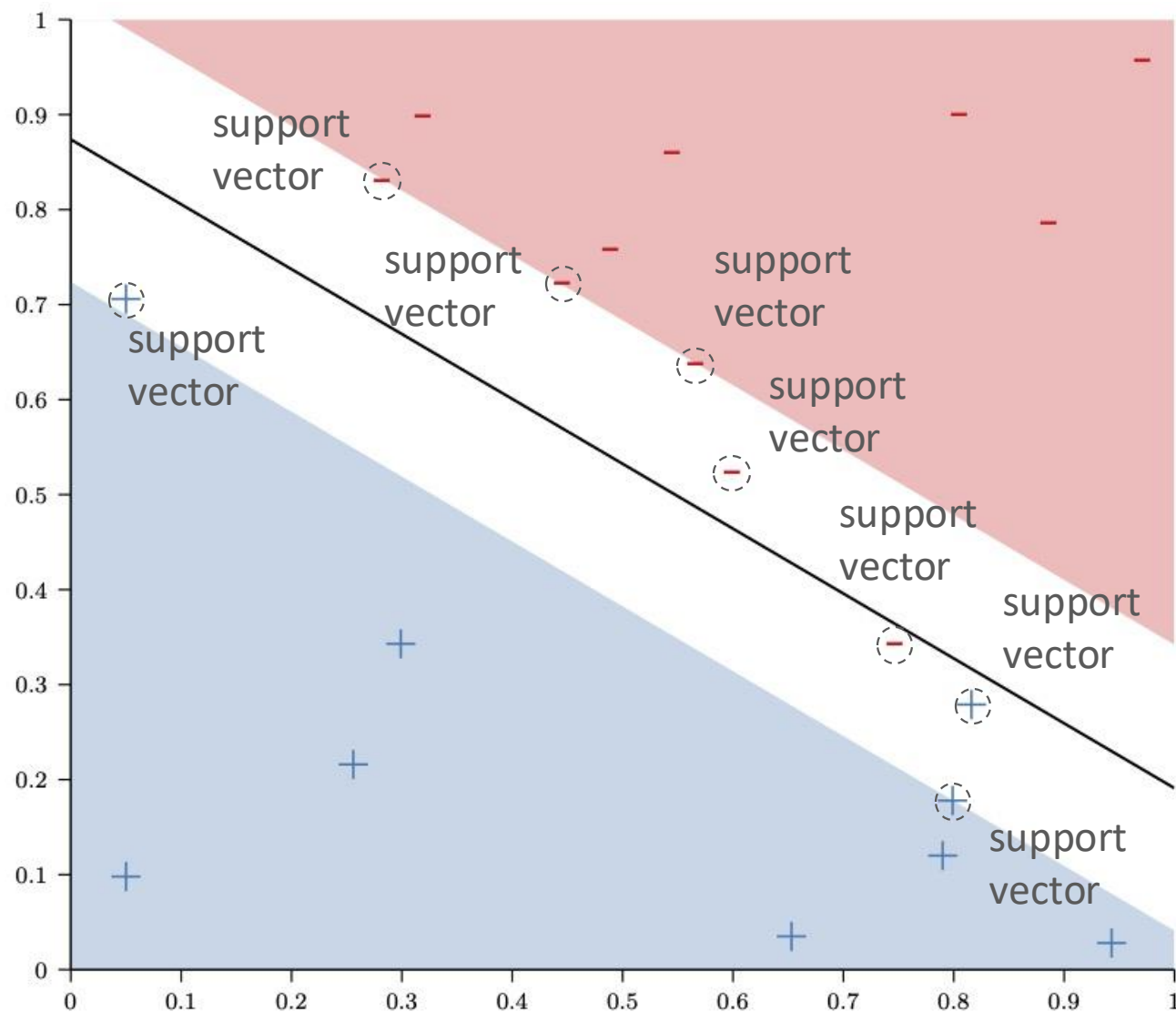
Interpreting $\xi^{(i)}$



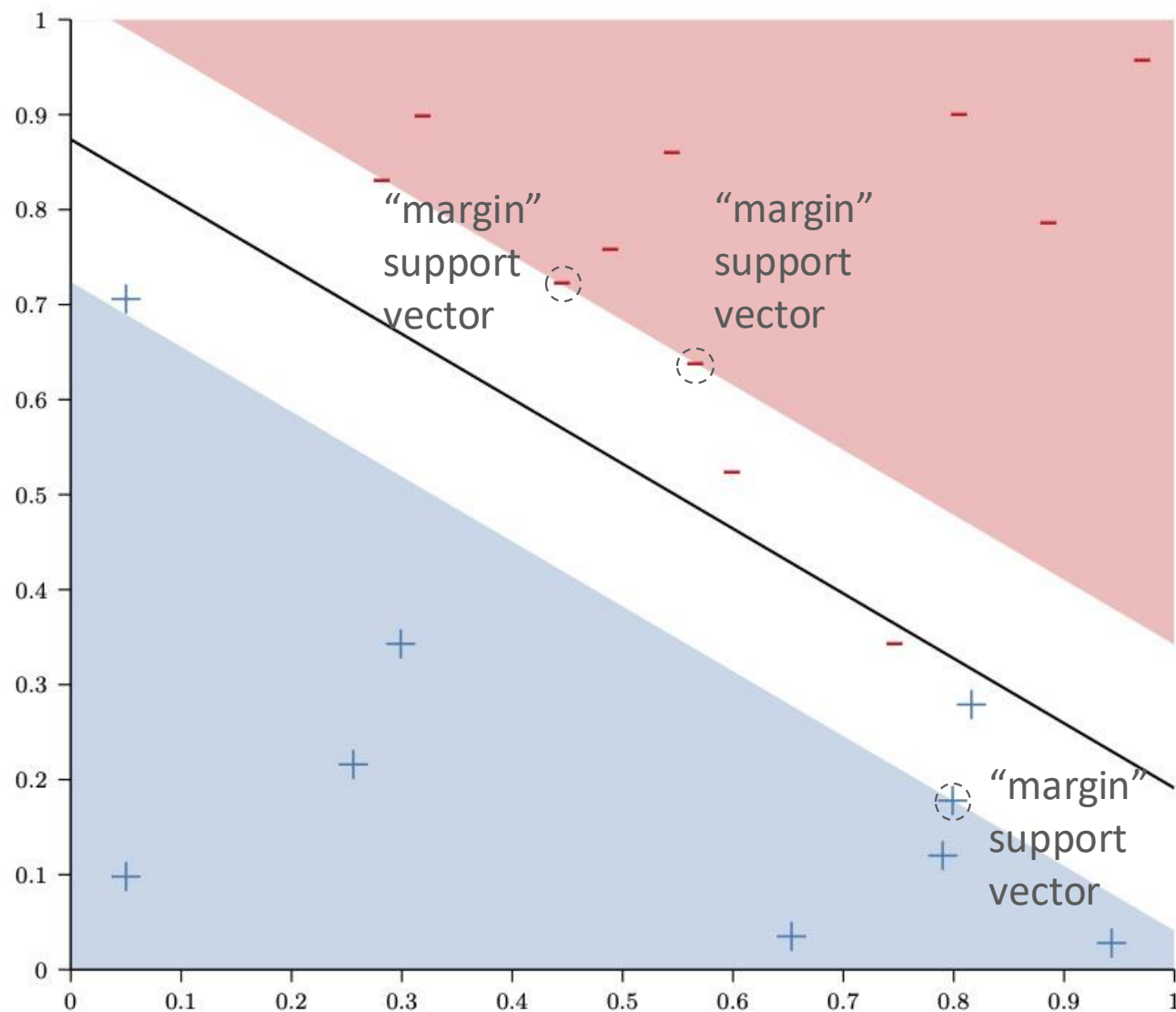
Interpreting $\xi^{(i)}$



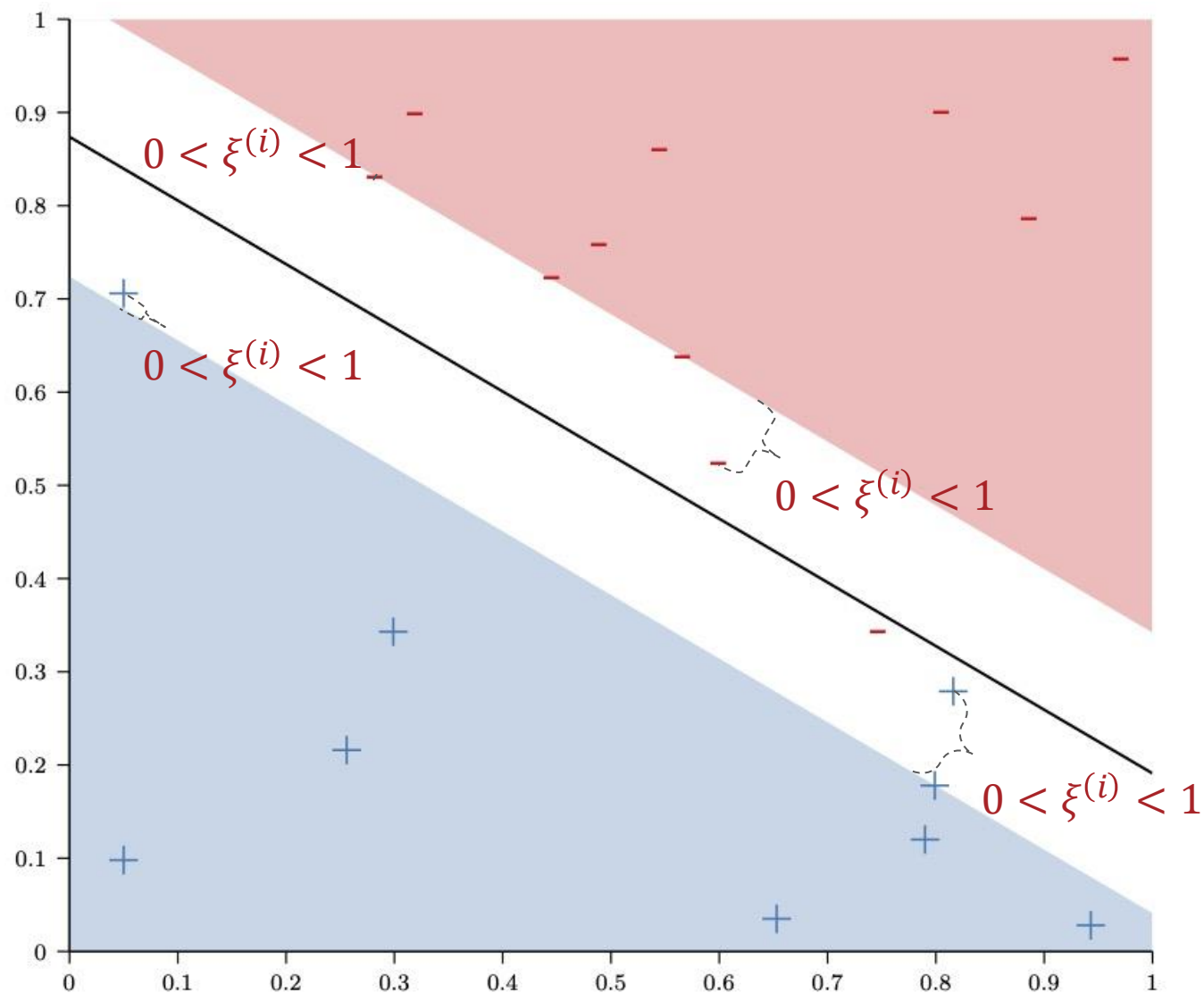
Interpreting $\xi^{(i)}$



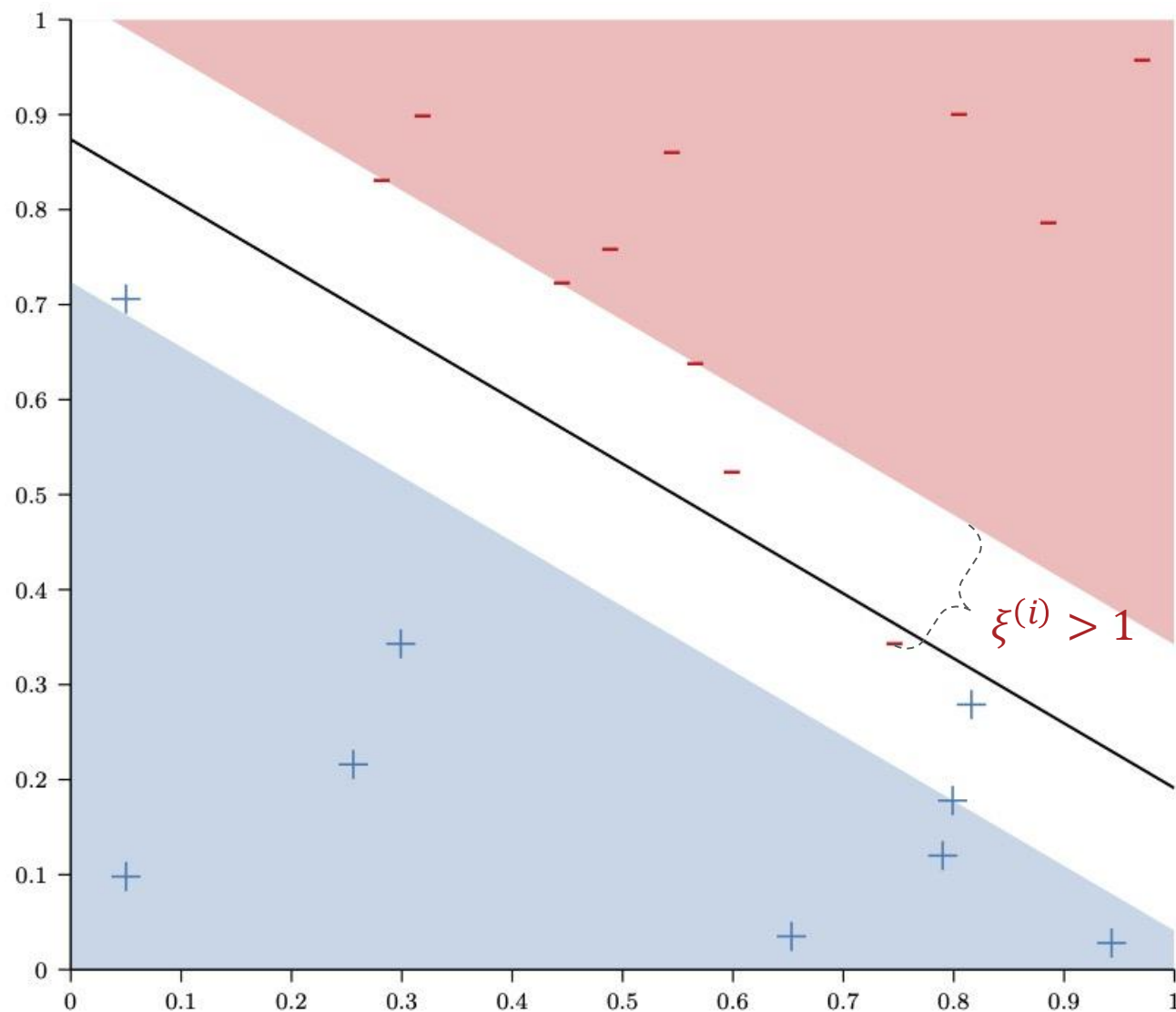
Interpreting $\xi^{(i)}$

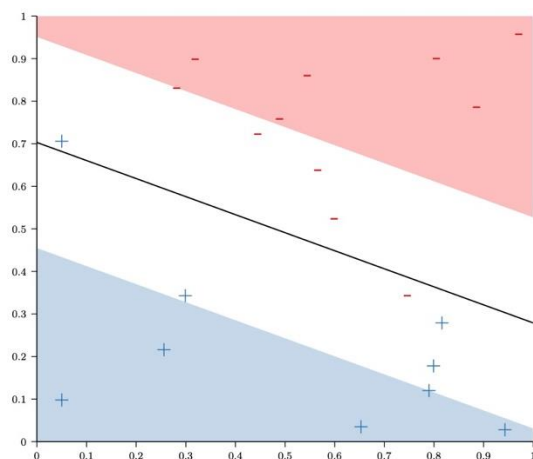


Interpreting $\xi^{(i)}$

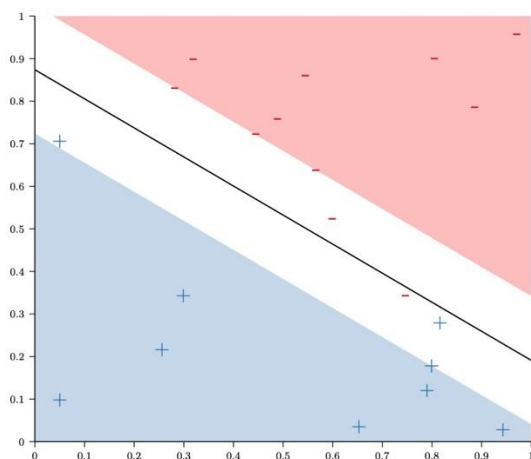


Interpreting $\xi^{(i)}$

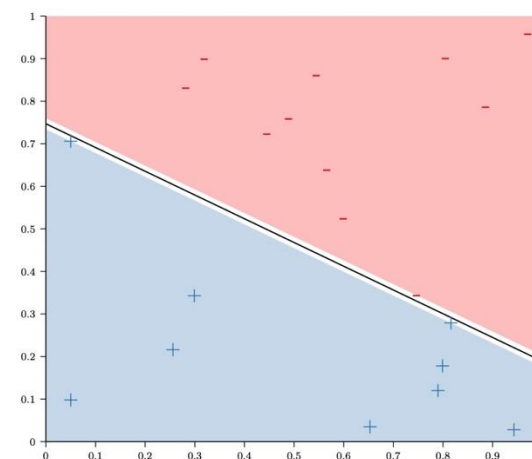




Smaller C



Larger C



Hard Margin

Setting C

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

Hard-margin SVMs

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

$$\begin{array}{ll} \text{minimize} & E_{train} \\ \text{subject to} & \mathbf{w}^T \mathbf{w} \leq C \end{array} \quad \left. \vphantom{\begin{array}{l} \text{minimize} \\ \text{subject to} \end{array}} \right\} \text{Regularization}$$

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

Primal-Dual Optimization

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

\Leftrightarrow

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \mathbf{x}^{(i)T} \mathbf{x}^{(j)} + \sum_{i=1}^N \alpha^{(i)} \\ \text{subject to} & \sum_{i=1}^N \alpha^{(i)} y^{(i)} = 0 \\ & \alpha^{(i)} \geq 0 \quad \forall i \in \{1, \dots, N\} \end{array} \quad \left. \vphantom{\begin{array}{l} \text{maximize} \\ \text{subject to} \end{array}} \right\} \text{Dual}$$

SVM

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

\Leftrightarrow

$$\begin{aligned} &\text{minimize } \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ &\text{subject to } 1 - y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + w_0) \leq 0 \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} \end{aligned}$$

\Leftrightarrow

$$\begin{aligned} &\text{minimize}_{\mathbf{w}, w_0} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \text{maximize}_{\alpha^{(i)} \geq 0} \sum_{i=1}^N \alpha^{(i)} \left(1 - y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} + w_0) \right) \end{aligned}$$

SVM

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

\Leftrightarrow

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

\Leftrightarrow

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

\Leftrightarrow

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

Karush-Kuhn-Tucker (KKT) Conditions

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

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

$$\frac{\partial L(\boldsymbol{\alpha}, \mathbf{w}, w_0)}{\partial \mathbf{w}} =$$

$$\frac{\partial L(\boldsymbol{\alpha}, \mathbf{w}, w_0)}{\partial w_0} =$$

Karush-Kuhn-Tucker (KKT) Conditions

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

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

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

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

Minimizing the Lagrangian

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

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

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

Minimizing the Lagrangian

$$\hat{\mathbf{w}} = \sum_{i=1}^N \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)}$$
$$\sum_{i=1}^N \alpha^{(i)} y^{(i)} = 0$$

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

Maximizing the Minimum

$$\begin{array}{ll} \text{maximize} & \text{minimize} \\ \boldsymbol{\alpha} \geq 0 & \boldsymbol{w}, w_0 \end{array} L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0)$$

\Updownarrow

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} \boldsymbol{x}^{(i)T} \boldsymbol{x}^{(j)} + \sum_{i=1}^N \alpha^{(i)} \\ \text{subject to} & \sum_{i=1}^N \alpha^{(i)} y^{(i)} = 0 \\ & \alpha^{(i)} \geq 0 \quad \forall i \in \{1, \dots, N\} \end{array}$$

Primal-Dual Optimization

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

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

Primal-Dual Optimization

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

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

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

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

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

Complementary Slackness

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

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

\Leftrightarrow

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

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

Computing \hat{w}_0

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

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

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

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

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

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

Primal-Dual Optimization

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

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

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

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

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

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

Primal-Dual Optimization

- Primal

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

- Dual

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

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

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

Primal-Dual Soft-Margin SVMs

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

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

Primal-Dual Soft-Margin SVMs

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

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

Primal-Dual Soft-Margin SVMs

- Primal

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

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

- Dual

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

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

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

- Support vectors are all $(\mathbf{x}^{(s)}, y^{(s)}) \in \mathcal{D}$ s.t. $0 < \hat{\alpha}^{(s)} < C$
- If $\hat{\alpha}^{(s)} = C$, then $\hat{\xi}^{(s)} > 0 \Rightarrow (\mathbf{x}^{(s)}, y^{(s)})$ 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