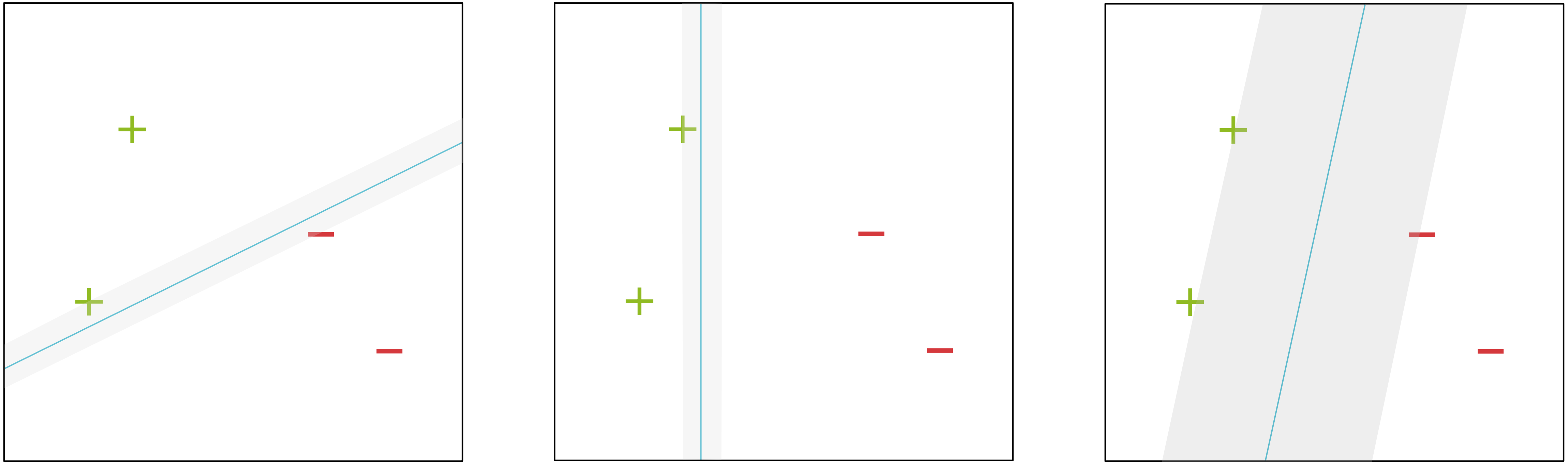
CSE 417T: Introduction to Machine Learning

Lecture 21: Dual SVMs

Henry Chai

11/13/18



Which linear separator is best?

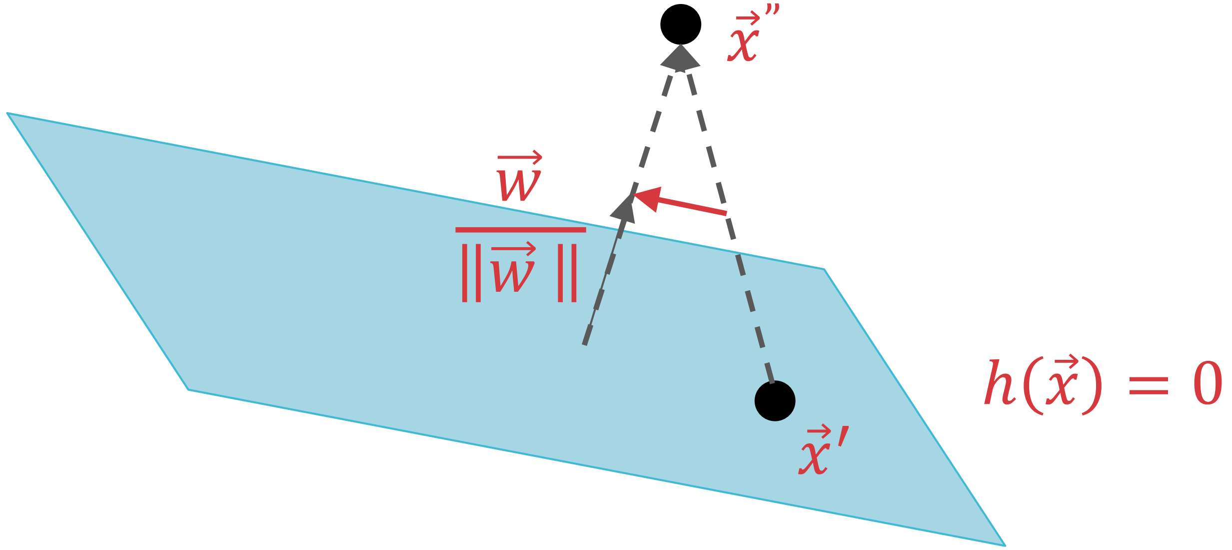
Maximal Margin Linear Separators

- The margin of a separating hyperplane is the distance between the hyperplane and the nearest training point

- Questions:
  - How can we efficiently find a maximal-margin linear separator?
  - Why are linear separators with larger margins better?
  - What can we do if the data is not linearly separable?

Computing the Margin

- Let $\vec{x}'$ be an arbitrary point on the hyperplane $h(\vec{x}) = \vec{w}^T\vec{x} + w_0 = 0$ and let $\vec{x}''$ be an arbitrary point
- The distance between $\vec{x}''$ and $h(\vec{x}) = 0$ is equal to the magnitude of the projection of $\vec{x}'' - \vec{x}'$ onto $\frac{\vec{w}}{\|\vec{w}\|}$, the unit vector orthogonal to $h(\vec{x}) = 0$



Computing the Margin

- The margin of a separating hyperplane is the distance between the hyperplane and the nearest training point

$$\min_{(\vec{x_i},y_i)\in\mathcal{D}} d(\vec{x_i},h) = \min_{(\vec{x_i},y_i)\in\mathcal{D}} \frac{|\vec{w}^T\vec{x_i}+w_0|}{||\vec{w}||}$$

$$= \frac{1}{||\vec{w}||}\min_{(\vec{x_i},y_i)\in\mathcal{D}} |\vec{w}^T\vec{x_i}+w_0|$$

$$= \frac{1}{||\vec{w}||}\min_{(\vec{x_i},y_i)\in\mathcal{D}} y_i(\vec{w}^T\vec{x_i}+w_0)$$

$$= \frac{1}{||\vec{w}||} = \frac{1}{\vec{w}^T\vec{w}}$$

Maximizing the Margin

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

subject to $\min_{(\vec{x_i},y_i)\in\mathcal{D}} y_i(\vec{w}^T\vec{x_i} + w_0) = 1$

$\Downarrow$

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$

- If $[w_0^*, \vec{w}^*]$ is the optimal solution, then $\exists$ at least one training point $(\vec{x_i}, y_i) \in \mathcal{D}$ s.t. $y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right) = 1$
  - All points $(\vec{x_i}, y_i) \in \mathcal{D}$ where $y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right) = 1$ are known as support vectors

Maximizing the Margin

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

$$\text{subject to} \quad \min_{(\vec{x_i}, y_i) \in \mathcal{D}} y_i(\vec{w}^T\vec{x_i} + w_0) = 1$$

$$\Downarrow$$

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

$$\text{subject to} \quad y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

- Converting the non-linear constraint (involving the min function) to n linear constraints allows the optimization problem to be solved (approximately) using quadratic programming (QP) in $O(D^3)$ time

VC-dimension of $\mathcal{H}_\rho$

- If the input space, $\mathcal{X}$, is a D-dimensional sphere of radius R, then:

$$d_{VC}(\mathcal{H}_\rho) \leq \min\left(D, \left\lceil \frac{R^2}{\rho^2} \right\rceil\right) + 1$$

Linearly Inseparable Data

- What can we do if the data is not linearly separable?
  - Accept some non-zero in-sample error
    - How much in-sample error should we tolerate?
  - 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?

Hard-Margin SVMs

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \; \forall(\vec{x_i}, y_i) \in \mathcal{D}$

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

Soft-Margin SVMs

minimize $\frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i$

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 - \xi_i \; \forall(\vec{x_i}, y_i) \in \mathcal{D}$

$\xi_i \geq 0 \qquad \forall\, i \in \{1, \ldots, n\}$

Soft-Margin SVMs

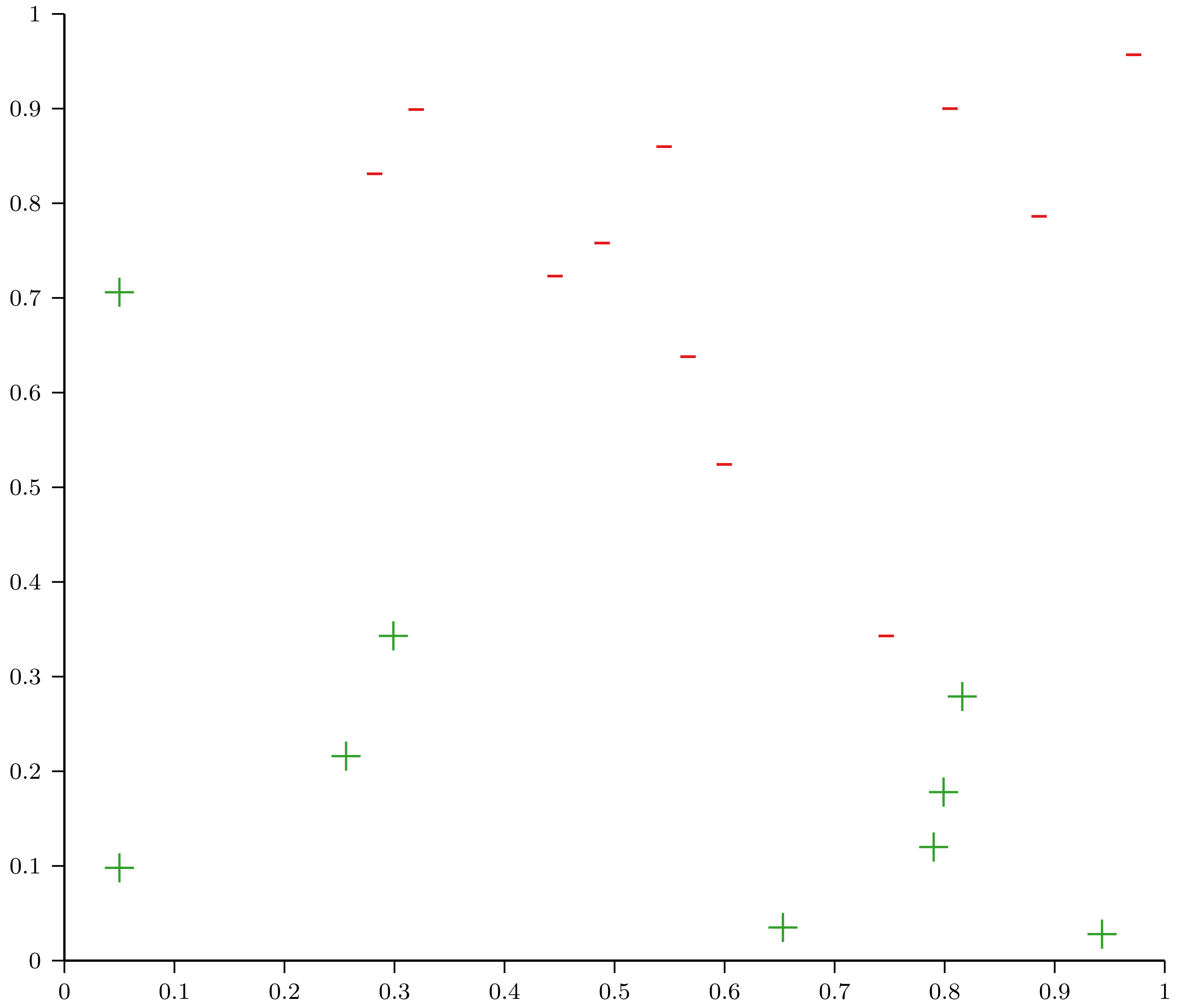
minimize $\frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i$

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 - \xi_i \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$

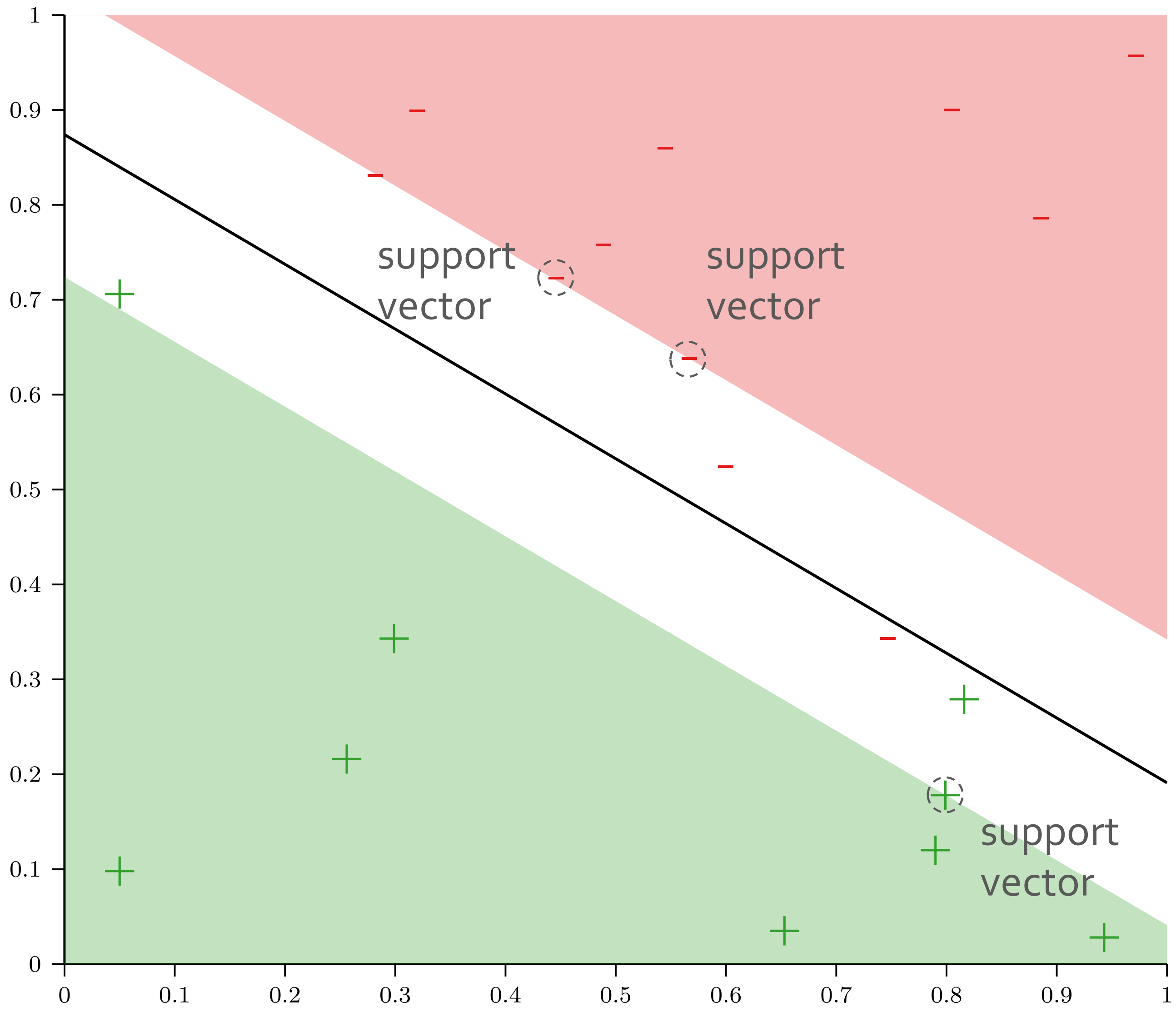
$\xi_i \geq 0 \qquad \forall\, i \in \{1, \dots, n\}$

- ξ_i is the "soft" error on the i^{th} training
  - If $\xi_i > 1$, then $y_i(\vec{w}^T\vec{x_i} + w_0) < 0 \Rightarrow (\vec{x_i}, y_i)$ is incorrectly classified
  - If $0 < \xi_i < 1$, then $y_i(\vec{w}^T\vec{x_i} + w_0) > 0 \Rightarrow (\vec{x_i}, y_i)$ is correctly classified but inside the margin
- $\sum_{i=1}^{n}\xi_i$ is the "soft" in-sample error

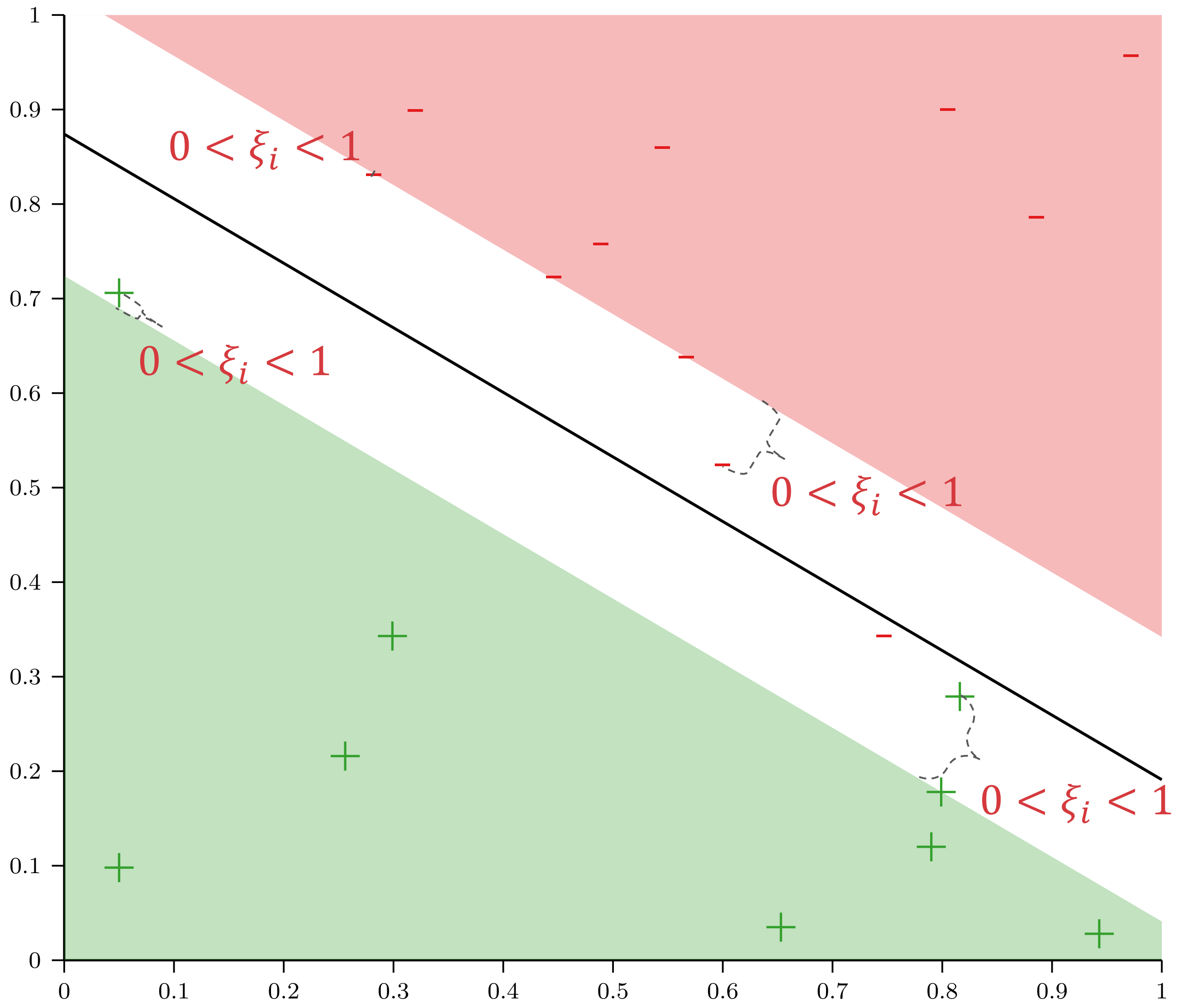
Interpreting ξ_i



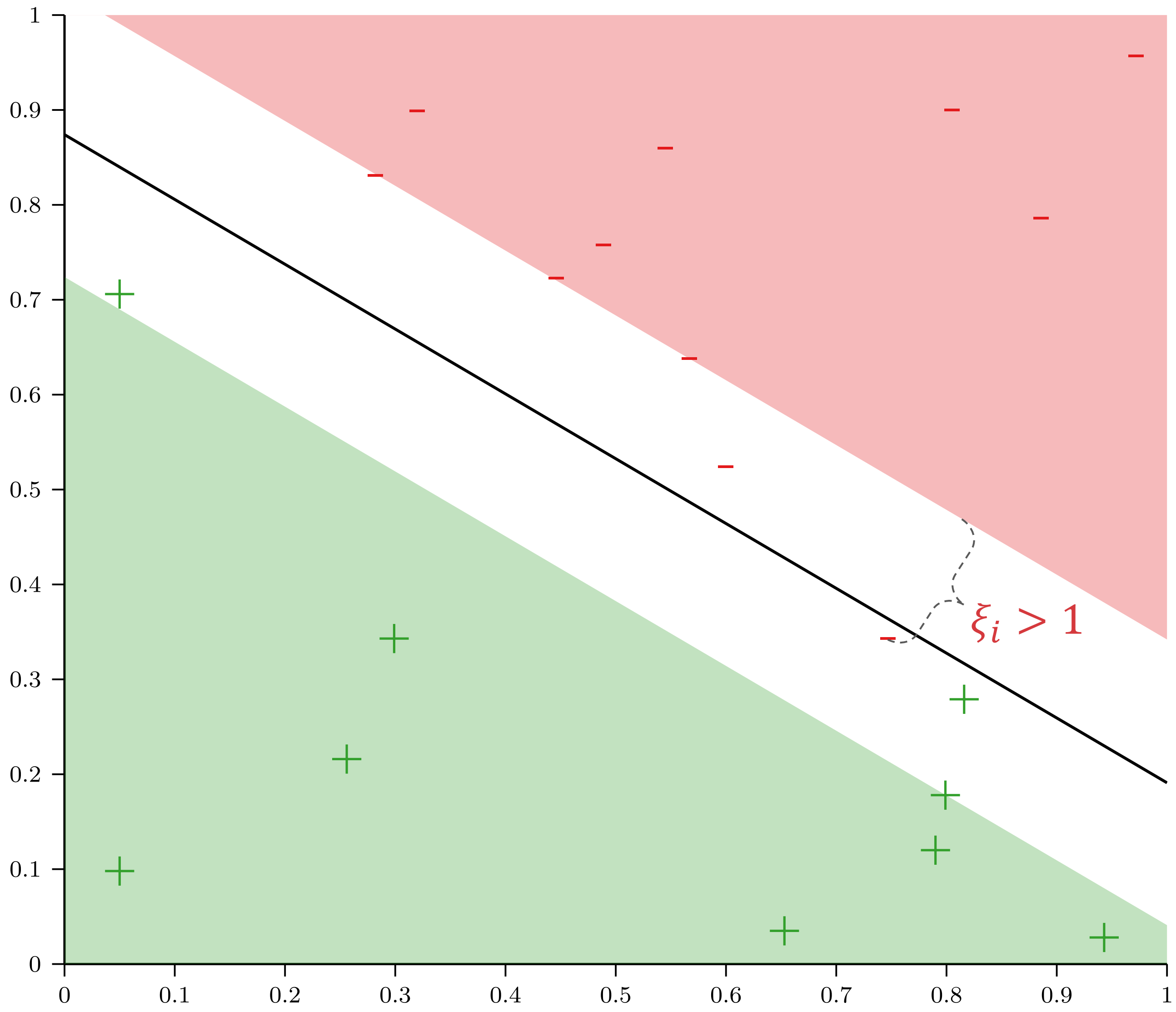
Interpreting ξ_i

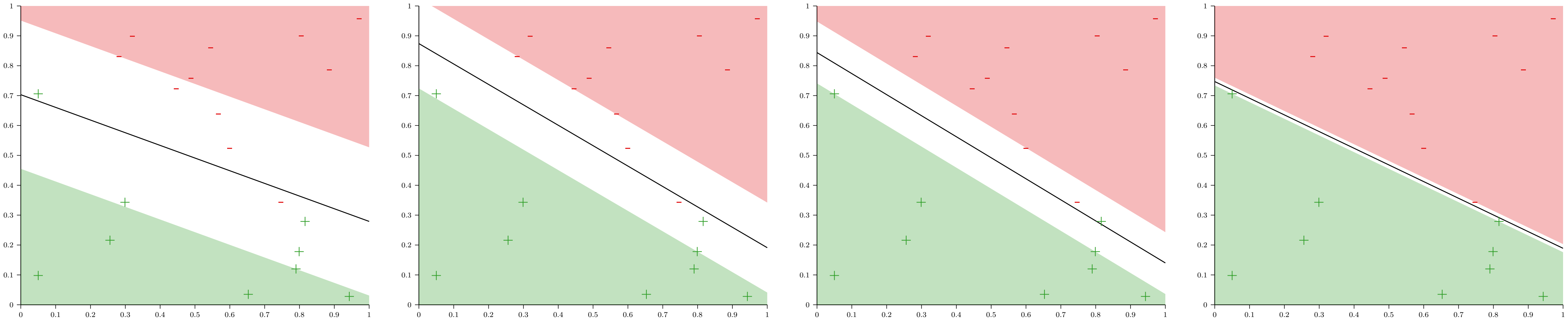


Interpreting ξ_i



Interpreting ξ_i





Smaller C | Larger C | Hard Margin

Setting C

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

Use validation!

Regularization

minimize $E_{in}(\vec{w}) = \frac{1}{n}(X\vec{w} - \vec{y})^T(X\vec{w} - \vec{y})$

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

$$\Updownarrow$$

minimize $E_{aug}(\vec{w}, \lambda_C) = E_{in}(\vec{w}) + \frac{\lambda_C}{n}\Omega(\vec{w})$

$$= \frac{1}{n}(X\vec{w} - \vec{y})^T(X\vec{w} - \vec{y}) + \frac{\lambda_C}{n}\vec{w}^T\vec{w}$$

Primal-Dual Optimization

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \; \forall (\vec{x_i}, y_i) \in \mathcal{D}$

Primal

maximize $-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_i}^T \vec{x_j} + \sum_{i=1}^{n} \alpha_i$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$

$\alpha_i \geq 0 \; \forall \, i \in \{1, \dots, n\}$

Dual

SVM

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

$$\text{subject to} \quad y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

$$\Updownarrow$$

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

$$\text{subject to} \quad 1 - y_i(\vec{w}^T\vec{x_i} + w_0) \leq 0 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

$$\Updownarrow$$

$$\text{minimize} \quad \frac{1}{2}\vec{w}^T\vec{w} + \max\left(\sum_{i=1}^{n} \alpha_i\left(1 - y_i(\vec{w}^T\vec{x_i} + w_0)\right)\right)$$

$$\text{subject to} \quad \alpha_i \geq 0 \ \forall\, i \in \{1, \ldots, n\}$$

SVM

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

$$\text{subject to } y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

$$\Updownarrow$$

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

$$\text{subject to } 1 - y_i(\vec{w}^T\vec{x_i} + w_0) \leq 0 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

$$\Updownarrow$$

$$\underset{\vec{w}, w_0}{\text{minimize}} \ \frac{1}{2}\vec{w}^T\vec{w} + \underset{\alpha_i \geq 0}{\text{maximize}} \left(\sum_{i=1}^{n} \alpha_i\big(1 - y_i(\vec{w}^T\vec{x_i} + w_0)\big)\right)$$

SVM

$$\underset{\vec{w},w_0}{\text{minimize}}\ \frac{1}{2}\vec{w}^T\vec{w} + \underset{\alpha_i \geq 0}{\text{maximize}}\left(\sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(\vec{w}^T\vec{x_i} + w_0\right)\right)\right)$$

$$\Updownarrow$$

$$\underset{\vec{w},w_0}{\text{minimize}}\left(\underset{\alpha_i \geq 0}{\text{maximize}}\ \frac{1}{2}\vec{w}^T\vec{w} + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(\vec{w}^T\vec{x_i} + w_0\right)\right)\right)$$

$$\Updownarrow$$

$$\underset{\alpha_i \geq 0}{\text{maximize}}\left(\underset{\vec{w},w_0}{\text{minimize}}\ \frac{1}{2}\vec{w}^T\vec{w} + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(\vec{w}^T\vec{x_i} + w_0\right)\right)\right)$$

$$\Updownarrow$$

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

Minimizing the Lagrangian

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

$$L(\vec{\alpha}, \vec{w}, w_0) = \frac{1}{2}\vec{w}^T\vec{w} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\vec{w}^T\vec{x_i} + w_0)\right)$$

$$\frac{\partial L(\vec{\alpha}, \vec{w}, w_0)}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^{n} \alpha_i y_i \vec{x_i} \rightarrow \vec{w}^* = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i}$$

$$\frac{\partial L(\vec{\alpha}, \vec{w}, w_0)}{\partial w_0} = -\sum_{i=1}^{n} \alpha_i y_i \quad \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

Minimizing the Lagrangian

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$L(\vec{\alpha}, \vec{w}^*, w_0^*) = \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i \left(1 - y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right)\right)$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^* \sum_{i=1}^{n} \alpha_i y_i$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^*(0)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i y_i \vec{x_i}^T\right)\left(\sum_{j=1}^{n} \alpha_j y_j \vec{x_j}\right) + \sum_{i=1}^{n} \alpha_i$$

$$- \sum_{i=1}^{n} \alpha_i y_i \left(\sum_{j=1}^{n} \alpha_j y_j \vec{x_j}^T\right)\vec{x_i}$$

Minimizing the Lagrangian

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$L(\vec{\alpha}, \vec{w}^*, w_0^*) = \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i \left(1 - y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right)\right)$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^* \sum_{i=1}^{n} \alpha_i y_i$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^*(0)$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_i}^T\vec{x_j} + \sum_{i=1}^{n} \alpha_i$$

$$- \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_j}^T\vec{x_i}$$

Minimizing the Lagrangian

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$L(\vec{\alpha}, \vec{w}^*, w_0^*) = \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i \left(1 - y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right)\right)$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^* \sum_{i=1}^{n} \alpha_i y_i$$

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \vec{w}^{*T}\vec{x_i} - w_0^*(0)$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_i}^T \vec{x_j} + \sum_{i=1}^{n} \alpha_i$$

Maximizing the Minimum

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

$$\Updownarrow$$

$$\underset{\vec{\alpha} \geq 0}{\text{maximize}} \; L(\vec{\alpha}, \vec{w}^*, w_0^*)$$

$$\text{subject to} \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\Updownarrow$$

$$\underset{\vec{\alpha} \geq 0}{\text{maximize}} \; -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_i}^T \vec{x_j} + \sum_{i=1}^{n} \alpha_i$$

$$\text{subject to} \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

Maximizing the Minimum

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

$$\Updownarrow$$

$$\underset{\vec{\alpha} \geq 0}{\text{maximize}} \; L(\vec{\alpha}, \vec{w}^*, w_0^*)$$

$$\text{subject to} \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\Updownarrow$$

$$\text{maximize} \; -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x_i}^T \vec{x_j} + \sum_{i=1}^{n} \alpha_i$$

$$\text{subject to} \; \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \; \forall \, i \in \{1, \ldots, n\}$$

Primal-Dual Optimization

Primal

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \;\forall (\vec{x_i}, y_i) \in \mathcal{D}$

Dual

maximize $-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x_i}^T\vec{x_j} + \sum_{i=1}^{n}\alpha_i$

subject to $\sum_{i=1}^{n}\alpha_i y_i = 0$

$\alpha_i \geq 0 \;\forall\, i \in \{1, \dots, n\}$

Primal-Dual Optimization

Primal

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \; \forall (\vec{x_i}, y_i) \in \mathcal{D}$

Dual

minimize $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x_i}^T\vec{x_j} - \sum_{i=1}^{n}\alpha_i$

subject to $\sum_{i=1}^{n}\alpha_i y_i = 0$

$\alpha_i \geq 0 \; \forall \; i \in \{1, \dots, n\}$

Primal-Dual Optimization

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

subject to $y_i(\vec{w}^T\vec{x_i} + w_0) \geq 1 \; \forall (\vec{x_i}, y_i) \in \mathcal{D}$

Primal

minimize $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j(y_i\vec{x_i})^T(y_j\vec{x_j}) - \sum_{i=1}^{n}\alpha_i$

subject to $\sum_{i=1}^{n}\alpha_i y_i = 0$

$\alpha_i \geq 0 \; \forall \, i \in \{1, \ldots, n\}$

Dual

Primal-Dual Optimization

- Primal
  - Directly returns the hyperplane, $[w_0^*, \vec{w}^*]$
  - Support vectors are all $(\vec{x_s}, y_s) \in \mathcal{D}$ s.t. $y_s\left(\vec{w}^{*T}\vec{x_s} + w_0^*\right) = 1$
- Dual
  - Returns the vector, $\vec{\alpha}^*$

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \vec{x_i}$$

$$w_0^* = ???$$

SVM

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

$$\text{subject to} \ 1 - y_i(\vec{w}^T\vec{x_i} + w_0) \leq 0 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$$

$$\Updownarrow$$

$$\underset{\vec{w}, w_0}{\text{minimize}} \ \frac{1}{2}\vec{w}^T\vec{w} + \underset{\alpha_i \geq 0}{\text{maximize}} \left(\sum_{i=1}^{n} \alpha_i \left(1 - y_i(\vec{w}^T\vec{x_i} + w_0)\right) \right)$$

- Theorem: $\alpha_i^* \left(1 - y_i\left(\vec{w}^{*T}\vec{x_i} + w_0^*\right)\right) = 0 \ \forall(\vec{x_i}, y_i) \in \mathcal{D}$
  - If $\alpha_s^* > 0$, then $1 - y_s\left(\vec{w}^{*T}\vec{x_s} + w_0^*\right) = 0$

Computing w_0^*

$$\alpha_i^* \left(1 - y_i\left(\overrightarrow{w^*}^T \overrightarrow{x_i} + w_0^*\right)\right) = 0 \; \forall (\overrightarrow{x_i}, y_i) \in \mathcal{D}$$

$$\downarrow$$

If $\alpha_s^* > 0$, then $1 - y_s\left(\overrightarrow{w^*}^T \overrightarrow{x_s} + w_0^*\right) = 0$

then $y_s\left(\overrightarrow{w^*}^T \overrightarrow{x_s} + w_0^*\right) = 1$

then $y_s^2\left(\overrightarrow{w^*}^T \overrightarrow{x_s} + w_0^*\right) = y_s$

then $\overrightarrow{w^*}^T \overrightarrow{x_s} + w_0^* = y_s$

then $w_0^* = y_s - \overrightarrow{w^*}^T \overrightarrow{x_s}$

Primal-Dual Optimization

- Primal
  - Directly returns the hyperplane, $[w_0^*, \vec{w}^*]$
  - Support vectors are all $(\vec{x_s}, y_s) \in \mathcal{D}$ s.t. $y_s\left({\vec{w}^*}^T\vec{x_s} + w_0^*\right) = 1$
- Dual
  - Returns the vector, $\vec{\alpha}^*$

$$\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \vec{x_i}$$

$$w_0^* = y_s - {\vec{w}^*}^T\vec{x_s} \text{ where } \alpha_s^* > 0$$

  - Support vectors are all $(\vec{x_s}, y_s) \in \mathcal{D}$ s.t. $\alpha_s^* > 0$

Primal-Dual Optimization

- Primal
  - $g(\vec{x}) = \text{sign}\left({\vec{w}^*}^T \vec{x} + w_0^*\right)$
- Dual
  - $g(\vec{x}) = \text{sign}\left({\vec{w}^*}^T \vec{x} + w_0^*\right)$

$$= \text{sign}\left(\left(\sum_{i=1}^{n} \alpha_i^* y_i {\vec{x_i}}^T\right) \vec{x} + w_0^*\right)$$

Primal-Dual Optimization

- Primal
  - $g(\vec{x}) = \text{sign}\left(\vec{w}^{*T}\vec{x} + w_0^*\right)$
- Dual
  - $g(\vec{x}) = \text{sign}\left(\vec{w}^{*T}\vec{x} + w_0^*\right)$

$$= \text{sign}\left(\left(\sum_{i:\alpha_i^* > 0} \alpha_i^* y_i \vec{x_i}^T\right)\vec{x} + w_0^*\right)$$