10-701: Introduction to Machine Learning Lecture 25 – Dual SVMs and Kernels

Henry Chai

4/17/24

Front Matter

- Announcements:
  - HW6 released 4/11, due **4/20 (Saturday)** at 11:59 PM
  - Exam 2 on 5/6 **from 1 PM – 3 PM in TEP 1403**
    - You are allowed to bring one letter-/A4-size sheet of notes; you can put *whatever* you want on *both sides*
    - **Pre-midterm material may be referenced but will not be the primary focus of any question**
- Recommended Readings:
  - Murphy, Chapters 14.1-14.5

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!

Primal-Dual Optimization

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

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

Primal

$$\Updownarrow$$

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

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

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

Dual

SVM

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

Karush-Kuhn-Tucker (KKT) Conditions

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

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

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

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

Minimizing the Lagrangian

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

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

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

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

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

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

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

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

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

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

Primal

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

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

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

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

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

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

$$\Updownarrow$$

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

- 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:

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

Dual:

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

Primal-Dual Soft-Margin SVMs

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

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

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

Primal

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

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

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

Dual

Recall: Nonlinear Transforms

- Decide on some transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$
- Given $\mathcal{D} = \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$, learn a hypothesis, $\tilde{h}(\boldsymbol{z})$, using $\widetilde{\mathcal{D}} = \left\{\left(\boldsymbol{z}^{(i)} = \Phi\left(\boldsymbol{x}^{(i)}\right), y^{(i)}\right)\right\}_{i=1}^{N}$
- Return the corresponding predictor in the original space: $h(\boldsymbol{x}) = \tilde{h}\left(\Phi(\boldsymbol{x})\right)$

Nonlinear SVMs

- Decide on some transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$
- Find a maximal-margin separating hyperplane in the transformed space, $\left[\widehat{\widetilde{\boldsymbol{w}}}, \widehat{\widetilde{w_0}}\right]$, by solving the QP:

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

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

- Return the corresponding predictor in the original space:

$$h(\boldsymbol{x}) = \text{sign}\left(\widehat{\widetilde{\boldsymbol{w}}}^T\Phi(\boldsymbol{x}) + \widehat{\widetilde{w_0}}\right)$$

Nonlinear Dual SVMs

- Decide on some transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$

- Find a maximal-margin separating hyperplane in the transformed space by solving the QP:

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

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

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

- Return the corresponding predictor in the original space:

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

Efficiency

- Depending on the transformation Φ and the dimensionality of the original input space $\mathcal{X}$, computing $\Phi(\boldsymbol{x})$ can be prohibitively computationally expensive
  - Computing $\Phi_2(\boldsymbol{x}) = [x_1, x_2, \dots, x_D, x_1^2, x_1 x_2, \dots, x_D^2]$ requires $D + \binom{D}{2} + D = \frac{D^2+3D}{2} = O(D^2)$ time
  - Computing $\Phi_{10}(\boldsymbol{x})$ requires $O(D^{10})$ time
- Tradeoff:
  - High-dimensional transformations can result in good hypotheses (as long as they don't overfit)
  - High-dimensional transformations are expensive

Nonlinear Dual SVMs

- Insight: Φ only appears in inner products!
- Find a maximal-margin separating hyperplane in the transformed space by solving the QP:

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

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

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

- Return the corresponding predictor in the original space:

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

Nonlinear Dual SVMs

- Insight: Φ only appears in inner products!
- Find a maximal-margin separating hyperplane in the transformed space by solving the QP:

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

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

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

- Return the corresponding predictor in the original space:

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

The Kernel Trick

- Approach: instead of computing $\Phi(\boldsymbol{x})$, find some function K_Φ s.t. $K_\Phi(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{x}')\ \forall\ \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$
  - $K_\Phi(\boldsymbol{x}, \boldsymbol{x}')$ should be cheaper to compute than $\Phi(\boldsymbol{x})$

- Example: $\Phi_2'(\boldsymbol{x}) = \left[x_1, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2\right]$

$$\Phi_2'(\boldsymbol{x})^T \Phi_2'(\boldsymbol{x}') = \sum_{i=1}^{D} x_i x_i' + \sum_{i=1}^{D} x_i^2 {x_i'}^2 + \sum_{i=1}^{D} \sum_{j>i} 2 x_i x_i' x_j x_j'$$

$$= \sum_{i=1}^{D} x_i x_i' + \left(\sum_{i=1}^{D} x_i x_i'\right)^2 = \boldsymbol{x}^T \boldsymbol{x}' + (\boldsymbol{x}^T \boldsymbol{x}')^2$$

$$K_{\Phi_2'}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}' + (\boldsymbol{x}^T \boldsymbol{x}')^2$$

- Computing $\Phi_2'(\boldsymbol{x})^T \Phi_2'(\boldsymbol{x}')$ requires $O(D^2)$ time whereas computing $K_{\Phi_2'}(\boldsymbol{x}, \boldsymbol{x}')$ only takes $O(D)$!

Common Kernels

- $K_{\Phi_2'}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}' + (\boldsymbol{x}^T\boldsymbol{x}')^2$
  - Implied feature transformation:

  $$\Phi_2'(\boldsymbol{x}) = \left[x_1, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2\right]$$

  - Implied dimensionality: $\frac{D^2+3D}{2}$

- $K_{\Phi_2^{(\gamma)}}(\boldsymbol{x}, \boldsymbol{x}') = (1 + \gamma\boldsymbol{x}^T\boldsymbol{x}')^2 - 1$
  - Implied feature transformation:

  $$\Phi_2^{(\gamma)}(\boldsymbol{x}) = \left[\sqrt{2\gamma}x_1, \dots, \sqrt{2\gamma}x_D, \gamma x_1^2, \gamma x_1x_2, \dots, \gamma x_D^2\right]$$

    - γ affects the geometry of the transform
  - Implied dimensionality: $\frac{D^2+3D}{2}$

Common Kernels

- Polynomial Kernel: $K_{\Phi_Q^{(\gamma)}}(\boldsymbol{x}, \boldsymbol{x}') = (1 + \gamma \boldsymbol{x}^T \boldsymbol{x}')^Q - 1$
  - Implied dimensionality: $O(D^Q)$
  - γ affects the geometry of the transform

- Gaussian-RBF Kernel: $K_{\Phi_r}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2r}}$
  - Implied feature transformation: $\Phi_r(\boldsymbol{x}) = \left[e^{-\frac{x_1^2}{2r}}, \dots, e^{-\frac{x_D^2}{2r}}, \right.$

$$\left. e^{-\frac{x_1^2}{2r}} \sqrt{\frac{(x_1)^2}{1!r^1}}, \dots, e^{-\frac{x_D^2}{2r}} \sqrt{\frac{(x_D)^2}{1!r^1}}, e^{-\frac{x_1^2}{2r}} \sqrt{\frac{\left(x_1^2\right)^2}{2!r^2}}, \dots, e^{-\frac{x_D^2}{2r}} \sqrt{\frac{\left(x_D^2\right)^2}{2!r^2}}, \dots \right]$$

Common Kernels

- Polynomial Kernel: $K_{\Phi_Q^{(\gamma)}}(\boldsymbol{x}, \boldsymbol{x}') = (1 + \gamma \boldsymbol{x}^T \boldsymbol{x}')^Q - 1$
  - Implied dimensionality: $O(D^Q)$
  - γ affects the geometry of the transform
- Gaussian-RBF Kernel: $K_{\Phi_r}(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|x - x'\|^2}{2r}}$
  - Implied feature transformation: $\Phi_r(x) =$

$$\left[\left[e^{-\frac{x_1^2}{2r}}\sqrt{\frac{\left(x_1^d\right)^2}{d!r^d}}, \dots, e^{-\frac{x_D^2}{2r}}\sqrt{\frac{\left(x_1^d\right)^2}{d!r^d}}\right] : d \in \mathbb{N}\right]$$

  - Implied dimensionality: ∞!

Nonlinear Dual SVMs

- Decide on a (valid) kernel function K_Φ
- Find a maximal-margin separating hyperplane in the transformed space by solving the QP:

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}K_\Phi\left(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\right) - \sum_{i=1}^{N}\alpha^{(i)}$$

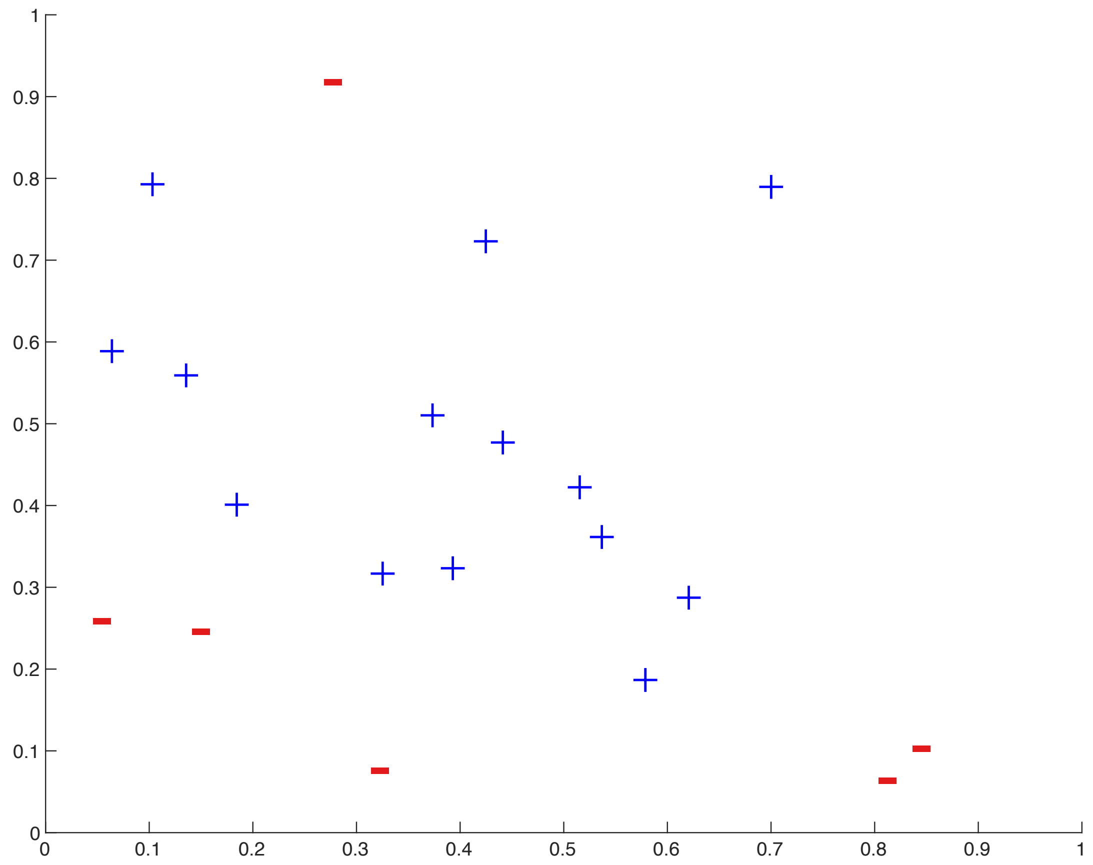
$$\text{subject to} \quad \sum_{i=1}^{N}\alpha^{(i)}y^{(i)} = 0$$

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

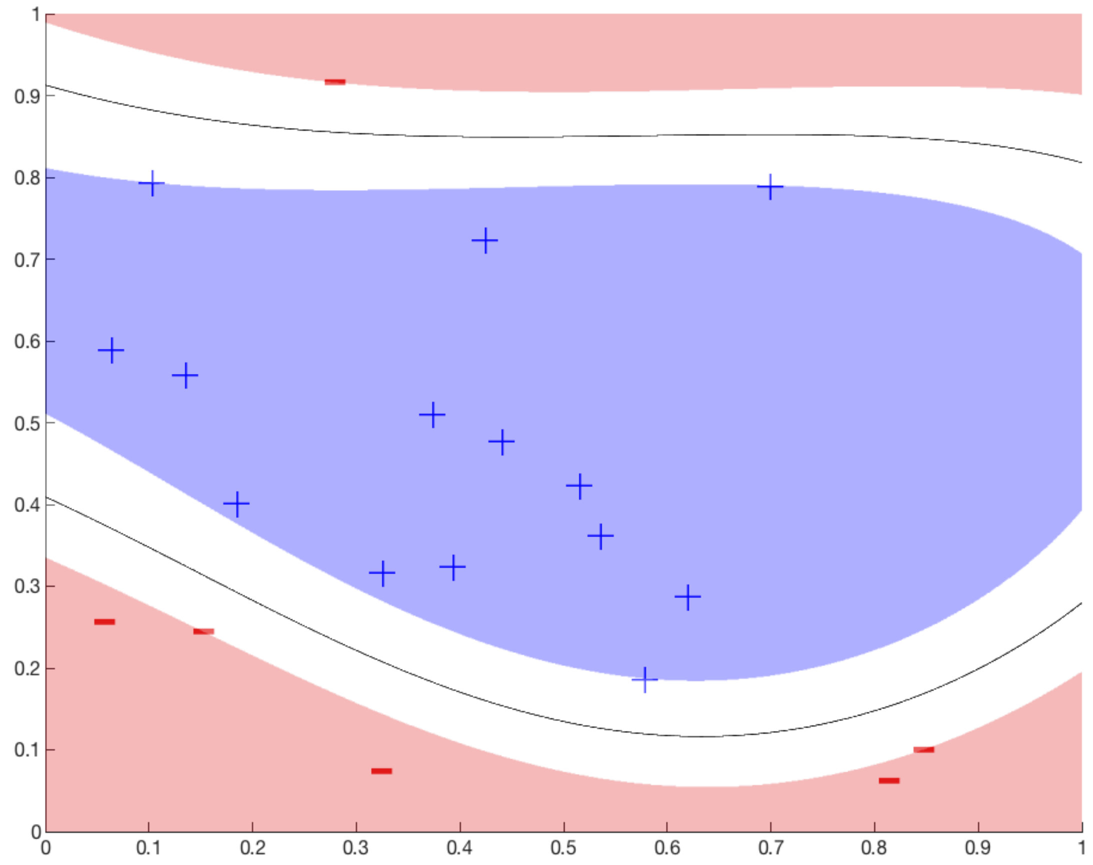
- Return the corresponding predictor in the original space:

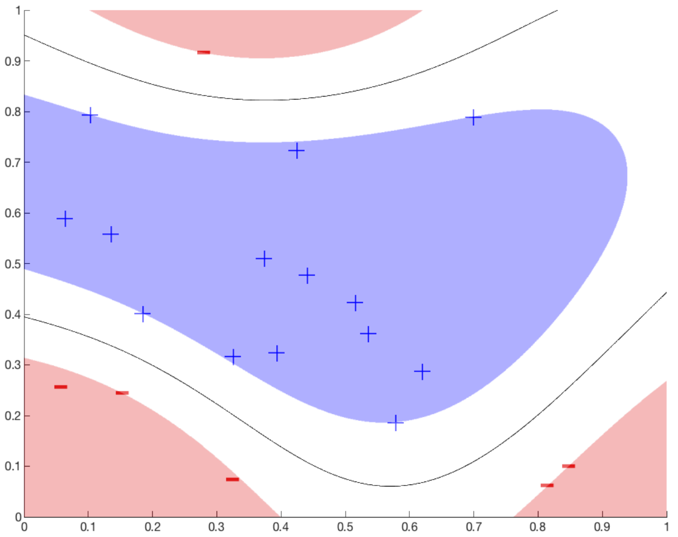
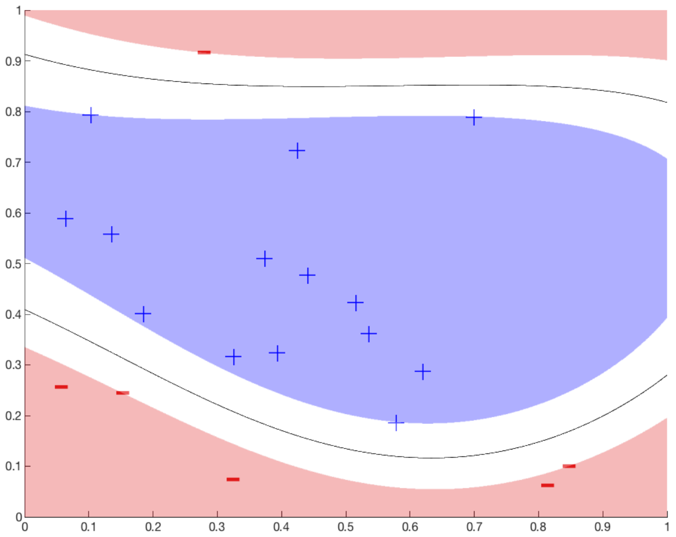
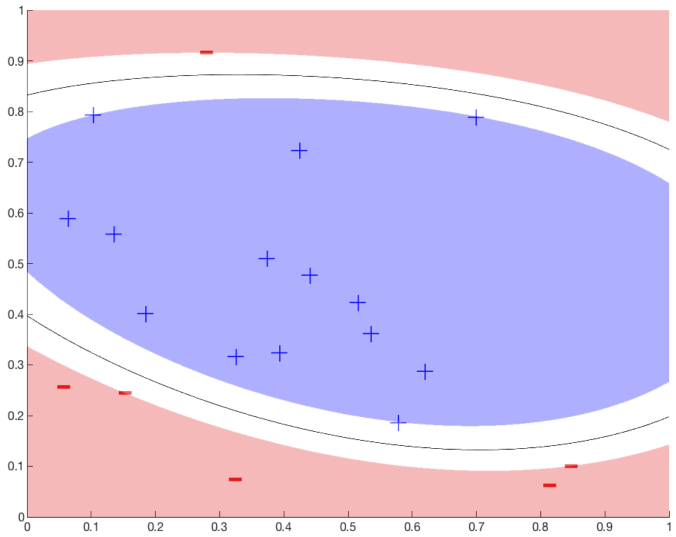
$$h(\boldsymbol{x}) = \text{sign}\left(\sum_{i:\hat{\alpha}^{(i)}>0} \hat{\alpha}^{(i)}y^{(i)}K_\Phi\left(\boldsymbol{x}^{(i)},\boldsymbol{x}\right) + \widehat{\widetilde{w}}_0\right)$$

Gaussian-RBF Kernel



Gaussian-RBF Kernel



Smaller r

Larger r

Gaussian-RBF Kernel

Nonlinear Dual Soft-Margin SVMs

- Decide on a (valid) kernel function K_Φ
- Find a maximal-margin separating hyperplane in the transformed space by solving the QP:

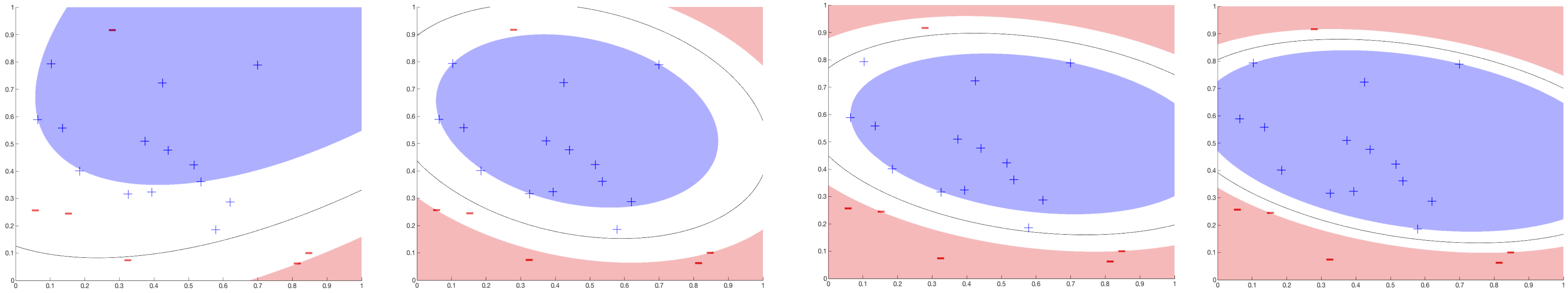
$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}K_\Phi\left(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\right)-\sum_{i=1}^{N}\alpha^{(i)}$$

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

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

- Return the corresponding predictor in the original space:

$$h(\boldsymbol{x})=\operatorname{sign}\left(\sum_{i:\hat{\alpha}^{(i)}>0}\hat{\alpha}^{(i)}y^{(i)}K_\Phi\left(\boldsymbol{x}^{(i)},\boldsymbol{x}\right)+\widetilde{\widehat{w}_0}\right)$$



Smaller C | Larger C | Hard Margin

2^{nd}-Degree Polynomial Kernel

C is a trade-off parameter between the size of the margin and the soft training error

Valid Kernels

- Any function K is a valid kernel if and only if:
  - $\exists$ a transformation Φ s.t.

$$K(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{x}') \ \forall \ \boldsymbol{x}, \boldsymbol{x}'$$

$$\Updownarrow$$

  - the Gram matrix

$$\mathrm{K} = \begin{bmatrix} K\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(1)}\right) & K\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) & \cdots & K\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(N)}\right) \\ K\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}\right) & K\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(2)}\right) & \cdots & K\left(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(N)}\right) \\ \vdots & \vdots & \ddots & \vdots \\ K\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(1)}\right) & K\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(2)}\right) & \cdots & K\left(\boldsymbol{x}^{(N)}, \boldsymbol{x}^{(N)}\right) \end{bmatrix}$$

  is symmetric and positive semi-definite $\forall$ sets $\left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\right\}$

Kernels Everywhere!

- Any method that only depends on the Euclidean distance between data points is an inner product method:

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^T(\boldsymbol{x} - \boldsymbol{x}')} = \sqrt{\boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{x}' + \boldsymbol{x}'^T\boldsymbol{x}'}$$

  - We can kernelize kNN!

- We can also kernelize linear/ridge regression!
  - See Murphy, Chapter 14.4

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
- Kernels and the “kernel trick” allow for efficient use of feature transformations for inner product methods
  - Definition of valid kernels
  - Common kernels and combining kernels

Primal-Dual Soft-Margin SVMs

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

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

Soft-Margin SVMs

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

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

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

$$\Updownarrow$$

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

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0, \boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$+\sum_{i=1}^{N}\alpha^{(i)}\left(1 - \xi^{(i)} - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right) - \sum_{i=1}^{N}\beta^{(i)}\xi^{(i)}$$

Karush-Kuhn-Tucker (KKT) Conditions

$$\underset{\boldsymbol{w}, w_0, \xi}{\text{minimize}} \; L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0, \boldsymbol{\beta}, \xi)$$

$$L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0, \boldsymbol{\beta}, \xi) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi^{(i)}$$

$$+\sum_{i=1}^{N}\alpha^{(i)}\left(1 - \xi^{(i)} - y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right)\right) - \sum_{i=1}^{N}\beta^{(i)}\xi^{(i)}$$

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

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

$$\frac{\partial L(\boldsymbol{\alpha}, \boldsymbol{w}, w_0, \boldsymbol{\beta}, \xi)}{\partial \xi^{(i)}} = C - \alpha^{(i)} - \beta^{(i)} \qquad \rightarrow \beta^{(i)} = C - \alpha^{(i)}$$

$$\forall\, i \in \{1, \dots, N\}$$

Minimizing the Lagrangian

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

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

$$\beta^{(i)} = C - \alpha^{(i)} \; \forall\, i$$

$$L(\boldsymbol{\alpha}, \hat{\boldsymbol{w}}, \hat{w}_0, \boldsymbol{\beta}, \hat{\boldsymbol{\xi}}) = \frac{1}{2}\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} + C\sum_{i=1}^{N} \hat{\xi}^{(i)}$$

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

$$= \frac{1}{2}\hat{\boldsymbol{w}}^T\hat{\boldsymbol{w}} + C\sum_{i=1}^{N} \hat{\xi}^{(i)} - \sum_{i=1}^{n} (C - \alpha^{(i)})\hat{\xi}^{(i)}$$

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

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

Maximizing the Minimum

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

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

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

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

$$\beta^{(i)} = C - \alpha^{(i)} \quad \forall\, i \in \{1, \dots, N\}$$

$$\Updownarrow$$

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

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

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

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, $\hat{\boldsymbol{\alpha}}$

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

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

Complementary Slackness

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

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

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

$$\Updownarrow$$

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

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

Complementary Slackness

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

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

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

Primal-Dual Soft-Margin SVMs

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

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

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

  - Support vectors are all $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right) \in \mathcal{D}$ s.t. $0 < \hat{\alpha}^{(s)}$
  - If $0 < \hat{\alpha}^{(s)} < C$, then $y^{(i)}\left(\hat{\boldsymbol{w}}^T \boldsymbol{x}^{(i)} + \hat{w}_0\right) = 1 \Rightarrow$ $\left(\boldsymbol{x}^{(s)}, y^{(s)}\right)$ defines the margin
  - If $\hat{\alpha}^{(s)} = C$, then $\hat{\xi}^{(s)} > 0 \Rightarrow \left(\boldsymbol{x}^{(s)}, y^{(s)}\right)$ is inside the margin or misclassified.