

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

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

\Leftrightarrow

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

SVM

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

Karush-Kuhn-Tucker (KKT) Conditions

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

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

Minimizing the Lagrangian

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

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

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

Maximizing the Minimum

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

\Leftrightarrow

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

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

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

Primal-Dual Optimization

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

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

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

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

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

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, $\hat{\alpha}$

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

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

- 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

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

Primal

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

Dual

Primal-Dual Soft-Margin SVMs

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

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

Recall: Nonlinear Transforms

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

Nonlinear SVMs

- Decide on some transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$
- Find a maximal-margin separating hyperplane in the transformed space, $[\tilde{\mathbf{w}}, \tilde{w}_0]$, by solving the QP:

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

$$\text{subject to } y^{(i)} (\tilde{\mathbf{w}}^T \Phi(\mathbf{x}^{(i)}) + \tilde{w}_0) \geq 1 \quad \forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}$$

- Return the corresponding predictor in the original space:

$$h(\mathbf{x}) = \text{sign}(\tilde{\mathbf{w}}^T \Phi(\mathbf{x}) + \tilde{w}_0)$$

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:

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

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

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

- Return the corresponding predictor in the original space:

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

Efficiency

- Depending on the transformation Φ and the dimensionality of the original input space \mathcal{X} , computing $\Phi(\mathbf{x})$ can be prohibitively computationally expensive
 - Computing $\Phi_2(\mathbf{x}) = [x_1, x_2, \dots, x_D, x_1^2, x_1x_2, \dots, x_D^2]$ requires $D + \binom{D}{2} + D = \frac{D^2 + 3D}{2} = O(D^2)$ time
 - Computing $\Phi_{10}(\mathbf{x})$ 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:

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

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

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

- Return the corresponding predictor in the original space:

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

Nonlinear Dual SVMs

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

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

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

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

- Return the corresponding predictor in the original space:

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

The Kernel Trick

- Approach: instead of computing $\Phi(\mathbf{x})$, find some function K_Φ s.t. $K_\Phi(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}') \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$
 - $K_\Phi(\mathbf{x}, \mathbf{x}')$ should be cheaper to compute than $\Phi(\mathbf{x})$
- Example: $\Phi'_2(\mathbf{x}) = [x_1, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2]$

Common Kernels

- $K_{\Phi'_2}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' + (\mathbf{x}^T \mathbf{x}')^2$

- Implied feature transformation:

$$\Phi'_2(\mathbf{x}) = [x_1, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2]$$

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

- $K_{\Phi_2^{(\gamma)}}(\mathbf{x}, \mathbf{x}') = (1 + \gamma \mathbf{x}^T \mathbf{x}')^2 - 1$

- Implied feature transformation:

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

- γ affects the geometry of the transform

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

Common Kernels

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

- Gaussian-RBF Kernel: $K_{\Phi_r}(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2r}}$

- Implied feature transformation: $\Phi_r(\mathbf{x}) = \left[e^{-\frac{x_1^2}{2r}}, \dots, e^{-\frac{x_D^2}{2r}}, \right.$

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

Common Kernels

- Polynomial Kernel: $K_{\Phi_Q^{(\gamma)}}(\mathbf{x}, \mathbf{x}') = (1 + \gamma \mathbf{x}^T \mathbf{x}')^Q - 1$
 - Implied dimensionality: $O(D^Q)$
 - γ affects the geometry of the transform
- Gaussian-RBF Kernel: $K_{\Phi_r}(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2r}}$
 - Implied feature transformation: $\Phi_r(\mathbf{x}) = \left[\left[e^{-\frac{x_1^2}{2r}} \sqrt{\frac{(x_1^d)^2}{d!r^d}}, \dots, e^{-\frac{x_D^2}{2r}} \sqrt{\frac{(x_1^d)^2}{d!r^d}} \right] : d \in \mathbb{N} \right]$
 - Implied dimensionality: $\infty!$

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:

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

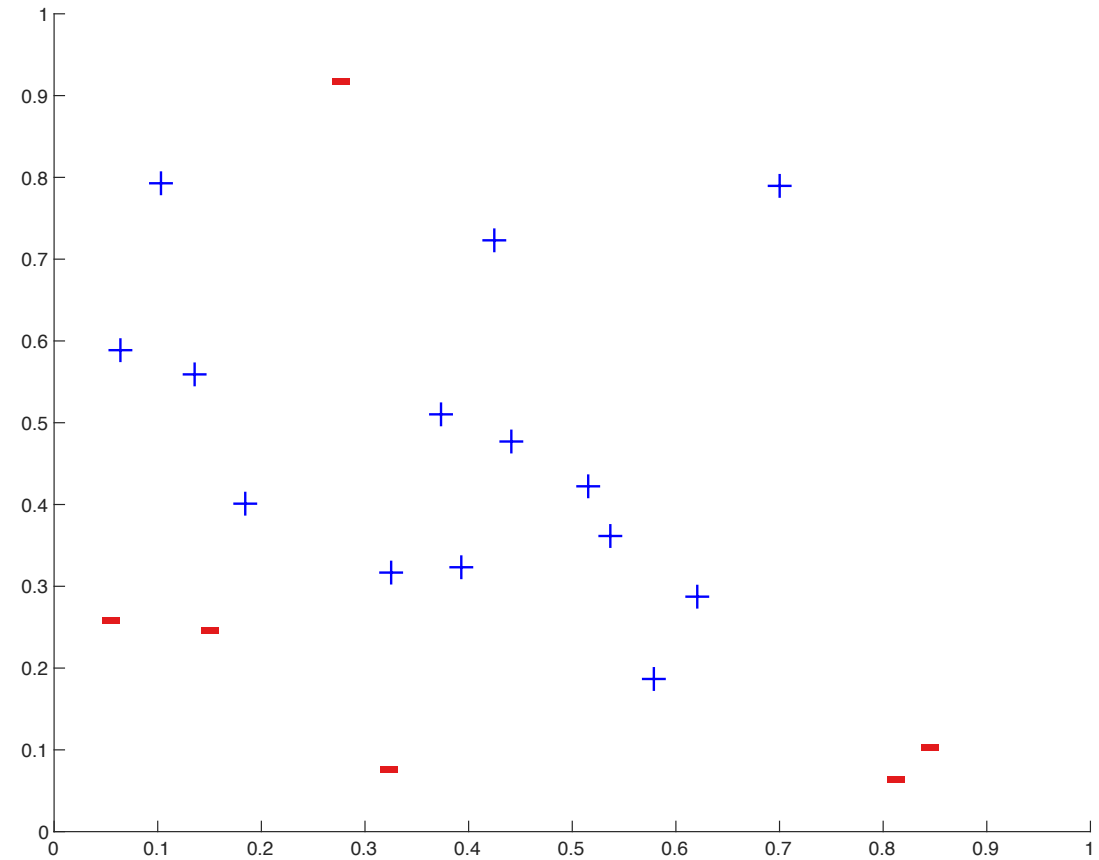
$$\text{subject to } \sum_{i=1}^N \alpha^{(i)} y^{(i)} = 0$$

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

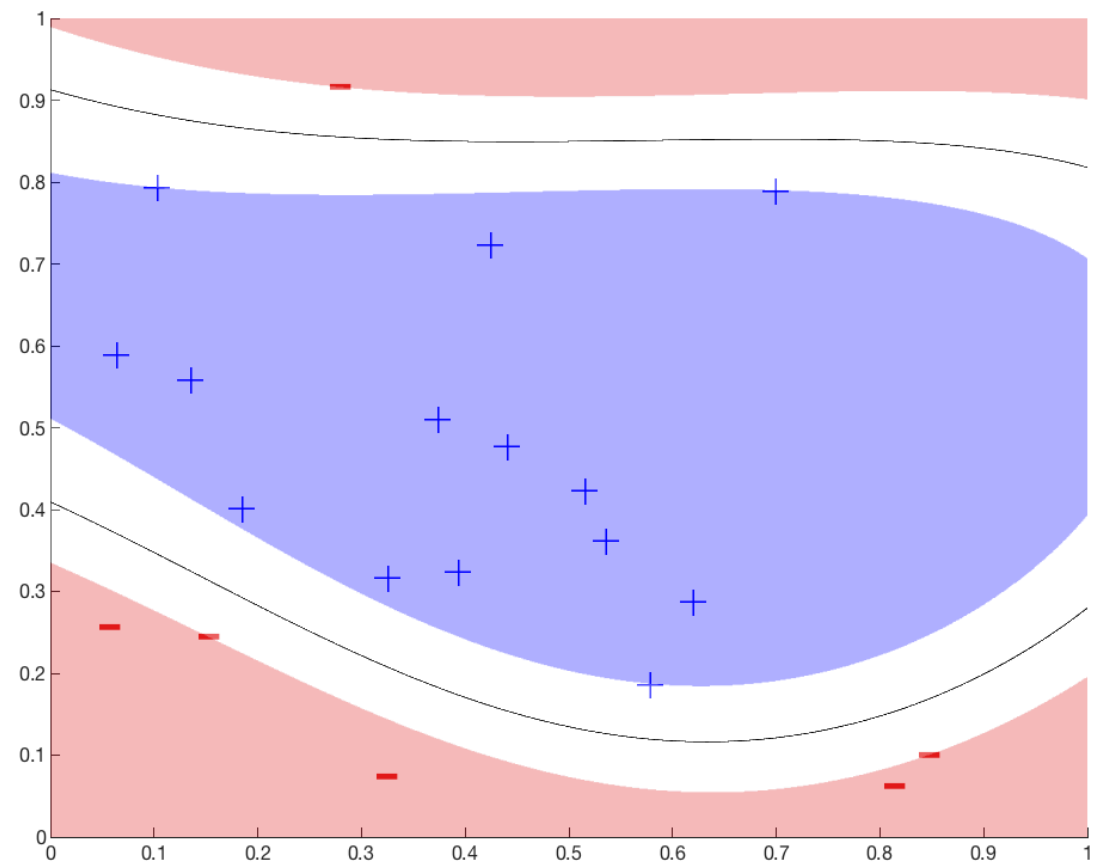
- Return the corresponding predictor in the original space:

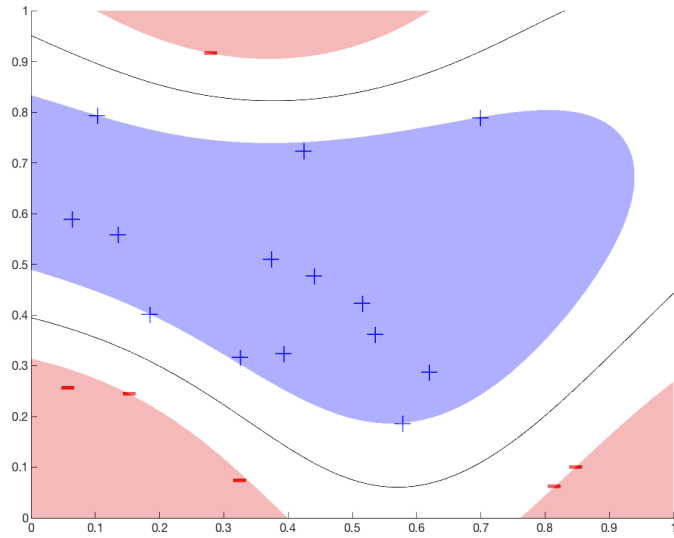
$$h(\mathbf{x}) = \text{sign} \left(\sum_{i: \hat{\alpha}^{(i)} > 0} \hat{\alpha}^{(i)} y^{(i)} K_{\Phi}(\mathbf{x}^{(i)}, \mathbf{x}) + \widetilde{w}_0 \right)$$

Gaussian- RBF Kernel

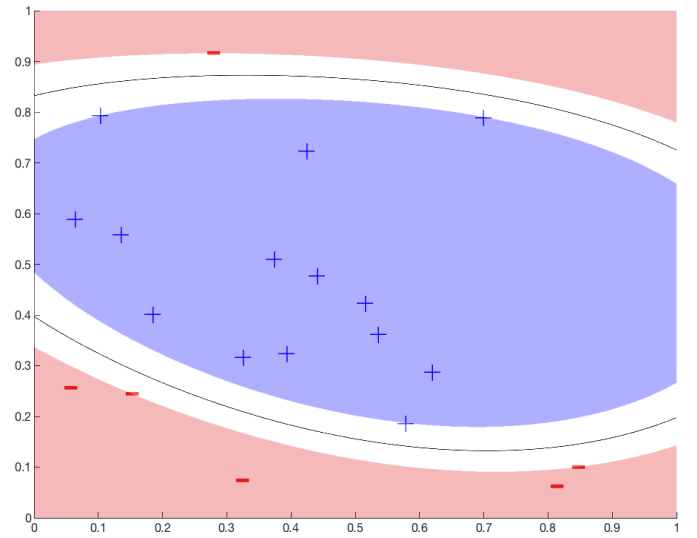
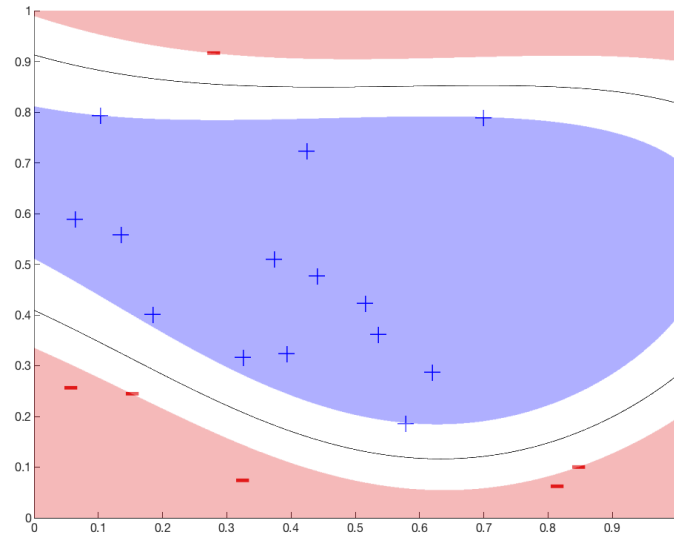


Gaussian- RBF Kernel





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:

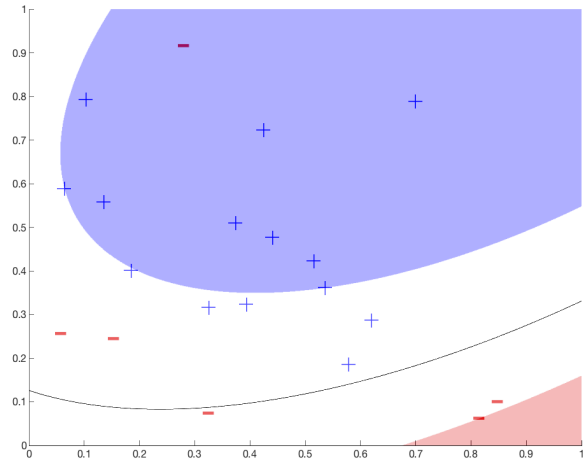
$$\text{minimize } \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} K_{\Phi}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) - \sum_{i=1}^N \alpha^{(i)}$$

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

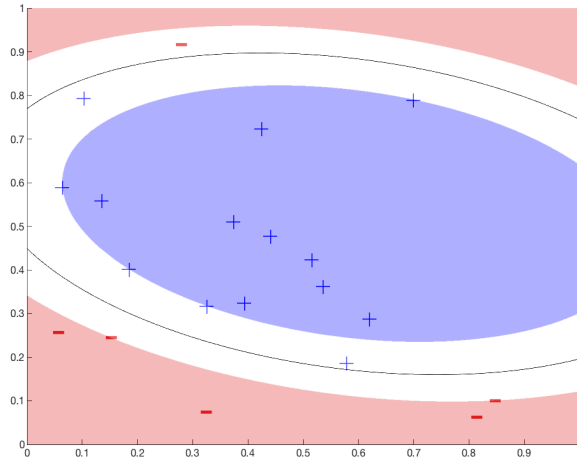
$$0 \leq \alpha^{(i)} \leq C \quad \forall i \in \{1, \dots, N\}$$

- Return the corresponding predictor in the original space:

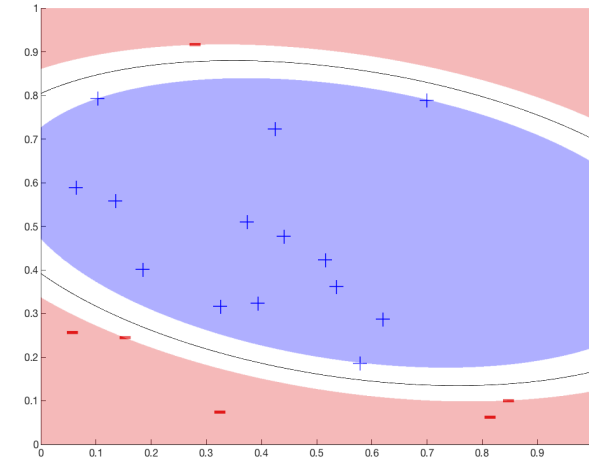
$$h(\mathbf{x}) = \text{sign} \left(\sum_{i: \hat{\alpha}^{(i)} > 0} \hat{\alpha}^{(i)} y^{(i)} K_{\Phi}(\mathbf{x}^{(i)}, \mathbf{x}) + \widetilde{w}_0 \right)$$



Smaller C



Larger C



Hard Margin

2nd-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.

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



- the Gram matrix

$$K = \begin{bmatrix} K(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & K(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) & \dots & K(\mathbf{x}^{(1)}, \mathbf{x}^{(N)}) \\ K(\mathbf{x}^{(2)}, \mathbf{x}^{(1)}) & K(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) & \dots & K(\mathbf{x}^{(2)}, \mathbf{x}^{(N)}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}^{(N)}, \mathbf{x}^{(1)}) & K(\mathbf{x}^{(N)}, \mathbf{x}^{(2)}) & \dots & K(\mathbf{x}^{(N)}, \mathbf{x}^{(N)}) \end{bmatrix}$$

is symmetric and positive semi-definite \forall sets

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$$

Kernels Everywhere!

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

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

} Primal

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

} Dual

Soft-Margin SVMs

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

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

\Leftrightarrow

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

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

Karush-Kuhn-Tucker (KKT) Conditions

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

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

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

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

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

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

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

Minimizing the Lagrangian

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

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

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

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

Maximizing the Minimum

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

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

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

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

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

\Leftrightarrow

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

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

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

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, $\hat{\alpha}$

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

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

Complementary Slackness

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

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

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

\Leftrightarrow

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

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

Complementary Slackness

maximize minimize
 $\alpha, \beta \geq 0$ \mathbf{w}, w_0, ξ $L(\alpha, \mathbf{w}, w_0, \beta, \xi)$

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

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

Primal-Dual Soft-Margin SVMs

- Dual

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

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

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

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