CSE 417T: Introduction to Machine Learning

Lecture 22: The Kernel Trick

Henry Chai

11/15/18

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

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

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

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

Soft-Margin SVMs

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

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

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

- 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

Primal-Dual Optimization

Primal

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

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

Dual

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

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

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

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. $\alpha_i > 0$

Primal-Dual Optimization

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

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

Primal-Dual Soft-Margin SVMs

Primal

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

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

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

Dual

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

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

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

Soft-Margin SVMs

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

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

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

$$\Updownarrow$$

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

$$L(\vec{a}, \vec{\beta}, \vec{w}, w_0, \vec{\xi}) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\vec{w}^T\vec{x_i} + w_0)) - \sum_{i=1}^{n}\beta_i\xi_i$$

Minimizing the Lagrangian

$$\underset{\vec{w}, w_0, \vec{\xi}}{\text{minimize}} \quad L(\vec{a}, \vec{\beta}, \vec{w}, w_0, \vec{\xi})$$

$$L(\vec{a}, \vec{\beta}, \vec{w}, w_0, \vec{\xi}) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i$$

$$+ \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\vec{w}^T\vec{x_i} + w_0)) - \sum_{i=1}^{n}\beta_i\xi_i$$

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

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

$$\frac{\partial L(\vec{a}, \vec{\beta}, \vec{w}, w_0, \vec{\xi})}{\partial \xi_i} = C - \alpha_i - \beta_i \qquad \rightarrow \beta_i = C - \alpha_i \forall\, i$$

Minimizing the Lagrangian

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

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

$$\beta_i = C - \alpha_i \; \forall \, i$$

$$L\left(\vec{a}, \vec{\beta}, \vec{w}^*, w_0^*, \vec{\xi}^*\right) = \frac{1}{2}\vec{w}^{*T}\vec{w}^* + C\sum_{i=1}^{n}\xi_i^* - \sum_{i=1}^{n}\beta_i\xi_i^*$$

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

$$= \frac{1}{2}\vec{w}^{*T}\vec{w}^* + C\sum_{i=1}^{n}\xi_i^* - \sum_{i=1}^{n}(C - \alpha_i)\xi_i^*$$

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

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

Maximizing the Minimum

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

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

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

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

$$\beta_i = C - \alpha_i \;\forall\, i \in \{1, \ldots, n\}$$

$$\Updownarrow$$

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

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

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

Primal-Dual Soft-Margin SVMs

- 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, $\vec{\alpha}^*$

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

$$w_0^* = ???$$

Complementary Slackness

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

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

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

$$\Updownarrow$$

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

$$L(\vec{a}, \vec{\beta}, \vec{w}, w_0, \vec{\xi}) = \frac{1}{2}\vec{w}^T\vec{w} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i(\vec{w}^T\vec{x_i} + w_0)\right) - \sum_{i=1}^{n}\beta_i\xi_i$$

Complementary Slackness

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

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

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

$$\Updownarrow$$

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

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

Complementary Slackness

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

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

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

$$\Updownarrow$$

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

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

Primal-Dual Soft-Margin SVMs

- 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, $\vec{\alpha}^*$

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

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

  - Support vectors are all $(\vec{x_s}, y_s) \in \mathcal{D}$ s.t. $0 < \alpha_s^* < C$
  - If $\alpha_s^* = C$, then $\xi_s^* > 0 \Rightarrow (\vec{x_s}, y_s)$ is inside the margin or misclassified

Nonlinear SVMs

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

- Find a maximal-margin separating hyperplane in the transformed space, $[\vec{\widetilde{w}}^*, \widetilde{w}_0^*]$, by solving the QP:

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

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

- Return the corresponding predictor in the original space:

  $$g(\vec{x}) = \text{sign}\left(\vec{\widetilde{w}}^{*^T}\Phi(\vec{x}) + \widetilde{w}_0^*\right)$$

Nonlinear Dual SVMs

- Decide on a transformation $\Phi: \mathcal{X} \to \mathcal{Z}$

- Find a maximal-margin separating hyperplane in the transformed space, $[\vec{\tilde{w}}^*, \tilde{w}_0^*]$, by solving the QP:

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j \Phi(\vec{x_i})^T\Phi(\vec{x_j}) - \sum_{i=1}^{n}\tilde{\alpha}_i$$

$$\text{subject to} \sum_{i=1}^{n}\tilde{\alpha}_i y_i = 0$$

$$\tilde{\alpha}_i \geq 0 \; \forall \; i \in \{1, \ldots, n\}$$

- Return the corresponding predictor in the original space:

$$g(\vec{x}) = \text{sign}\left(\sum_{i:\alpha_i^* > 0}\tilde{\alpha}_i^* y_i \Phi(\vec{x_i})^T\Phi(\vec{x}) + \tilde{w}_0^*\right)$$

Perceptrons

	Low-Dimensional Input Space	High-Dimensional Input Space
E_{in}	High	Low
Generalization	Good	Bad

SVMs

	Low-Dimensional Input Space	High-Dimensional Input Space
E_{in}	High	Low
Generalization	Good	Okay

$$d_{VC}(\mathcal{H}) = D + 1 \text{ vs. } d_{VC}(\mathcal{H}_\rho) \leq \min\left(D, O\left(\frac{1}{\rho^2}\right)\right) + 1$$

Efficiency

- Depending on the transformation Φ and the dimensionality of the original input space $\mathcal{X}$, computing $\Phi(\vec{x})$ can be computationally expensive
  - Computing $\Phi_2(\vec{x}) = [x_1, x_2, \dots, x_D, x_1^2, x_1 x_2, \dots, x_D^2]$ requires $\widetilde{D} = D + \binom{D}{2} + D = \frac{D^2+3D}{2} = O(D^2)$ time
  - Computing $\Phi_{10}(\vec{x})$ 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: the transformation Φ only appears as the inner product between two transformed points
- Find a maximal-margin separating hyperplane in the transformed space, $[\vec{\tilde{w}}^*, \tilde{w}_0^*]$, by solving the QP:

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j \Phi(\vec{x_i})^T\Phi(\vec{x_j}) - \sum_{i=1}^{n}\tilde{\alpha}_i$$

$$\text{subject to} \quad \sum_{i=1}^{n}\tilde{\alpha}_i y_i = 0$$

$$\tilde{\alpha}_i \geq 0 \;\forall\; i \in \{1, \ldots, n\}$$

- Return the corresponding predictor in the original space:

$$g(\vec{x}) = \text{sign}\left(\sum_{i:\alpha_i^*>0}\tilde{\alpha}_i^* y_i \Phi(\vec{x_i})^T\Phi(\vec{x}) + \tilde{w}_0^*\right)$$

The Kernel Trick

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

$$\Phi_2'(\vec{x}) = \left[x_1, x_2, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2\right]$$

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

$$= \sum_{i=1}^{D} x_i x_i' + \left(\sum_{i=1}^{D} x_i x_i' \right)^2$$

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

The Kernel Trick

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

$$\Phi_2'(\vec{x}) = \left[x_1, x_2, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2\right]$$

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

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

Common Kernels

- $K_{\Phi_2'}(\vec{x}, \vec{x}') = \vec{x}^T\vec{x}' + (\vec{x}^T\vec{x}')^2$
  - Implied feature transformation: $\Phi_2'(\vec{x}) = \left[x_1, x_2, \dots, x_D, x_1^2, \sqrt{2}x_1x_2, \dots, \sqrt{2}x_{D-1}x_D, x_D^2\right]$
  - Implied dimensionality: $\widetilde{D} = \frac{D^2+3D}{2}$
- $K_{\Phi_2^{(\zeta,\gamma)}}(\vec{x}, \vec{x}') = (\zeta + \gamma\vec{x}^T\vec{x}')^2 - \zeta^2$
  - Implied feature transformation: $\Phi_2^{(\zeta,\gamma)}(\vec{x}) = \left[\sqrt{2\zeta\gamma}x_1, \dots, \sqrt{2\zeta\gamma}x_D, \gamma x_1^2, \gamma x_1x_2, \dots, \gamma x_D^2\right]$
  - Implied dimensionality: $\widetilde{D} = \frac{D^2+3D}{2}$
  - γ and ζ affect the geometry of the transform: changing them changes the support vectors / decision boundary
    - Set using validation

Common Kernels

- Polynomial Kernel: $K_{\Phi_Q^{(\zeta,\gamma)}}(\vec{x},\vec{x}') = (\zeta + \gamma\vec{x}^T\vec{x}')^Q - \zeta^Q$
  - Implied dimensionality: $\widetilde{D} = O(D^Q)$
  - γ and ζ affect the geometry of the transform: changing them changes the support vectors / decision boundary
    - Set using validation
- Gaussian-RBF Kernel: $K_{\Phi_r}(\vec{x},\vec{x}') = e^{-\frac{\|\vec{x}-\vec{x}'\|^2}{2r}}$
  - Implied feature transformation: $\Phi_r(\vec{x}) = \left[e^{-\frac{x_1^2}{2r}}, \dots, e^{-\frac{x_D^2}{2r}}, e^{-\frac{x_1^2}{2r}}\sqrt{\frac{(x_1)^2}{1!r^1}}, \dots, e^{-\frac{x_D^2}{2r}}\sqrt{\frac{(x_D)^2}{1!r^1}}, e^{-\frac{x_1^2}{2r}}\sqrt{\frac{(x_1^2)^2}{2!r^2}}, \dots, e^{-\frac{x_D^2}{2r}}\sqrt{\frac{(x_D^2)^2}{2!r^2}}, \dots\right]$

Common Kernels

- Polynomial Kernel: $K_{\Phi_Q^{(\zeta,\gamma)}}(\vec{x},\vec{x}') = (\zeta + \gamma\vec{x}^T\vec{x}')^Q - \zeta^Q$
  - Implied dimensionality: $\widetilde{D} = O(D^Q)$
  - γ and ζ affect the geometry of the transform: changing them changes the support vectors / decision boundary
    - Set using validation
- Gaussian-RBF Kernel: $K_{\Phi_r}(\vec{x},\vec{x}') = e^{-\frac{\|\vec{x}-\vec{x}'\|^2}{2r}}$
  - Implied feature transformation: $\Phi_r(\vec{x}) = \left[\left[e^{-\frac{x_1^2}{2r}}\sqrt{\frac{(x_1^d)^2}{d!r^d}}, \ldots, e^{-\frac{x_D^2}{2r}}\sqrt{\frac{(x_1^d)^2}{d!r^d}}\right] : d \in \mathbb{N}\right]$
  - Implied dimensionality: $\widetilde{D} = \infty$!
  - Set r using validation

Valid Kernels

- Any function K is a valid kernel if and only if:
  - $\exists$ a transformation Φ s.t. $K(\vec{x}, \vec{x}') = \Phi(\vec{x})^T \Phi(\vec{x}') \; \forall \; \vec{x}, \vec{x}'$

$$\Updownarrow$$

  - the Gram matrix

$$\mathrm{K} = \begin{bmatrix} K(\overrightarrow{x_1}, \overrightarrow{x_1}) & K(\overrightarrow{x_1}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_1}, \overrightarrow{x_n}) \\ K(\overrightarrow{x_2}, \overrightarrow{x_1}) & K(\overrightarrow{x_2}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_2}, \overrightarrow{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\overrightarrow{x_n}, \overrightarrow{x_1}) & K(\overrightarrow{x_n}, \overrightarrow{x_2}) & \cdots & K(\overrightarrow{x_n}, \overrightarrow{x_n}) \end{bmatrix}$$

  is positive semi-definite $\forall$ sets $\{\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}\}$

Nonlinear Dual SVMs

- Decide on a (valid) kernel function K_Φ
- Find a maximal-margin separating hyperplane in the transformed space, $[\vec{\tilde{w}}^*, \tilde{w}_0^*]$, by solving the QP:

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j K_\Phi(\vec{x_i}, \vec{x_j}) - \sum_{i=1}^{n}\tilde{\alpha}_i$$

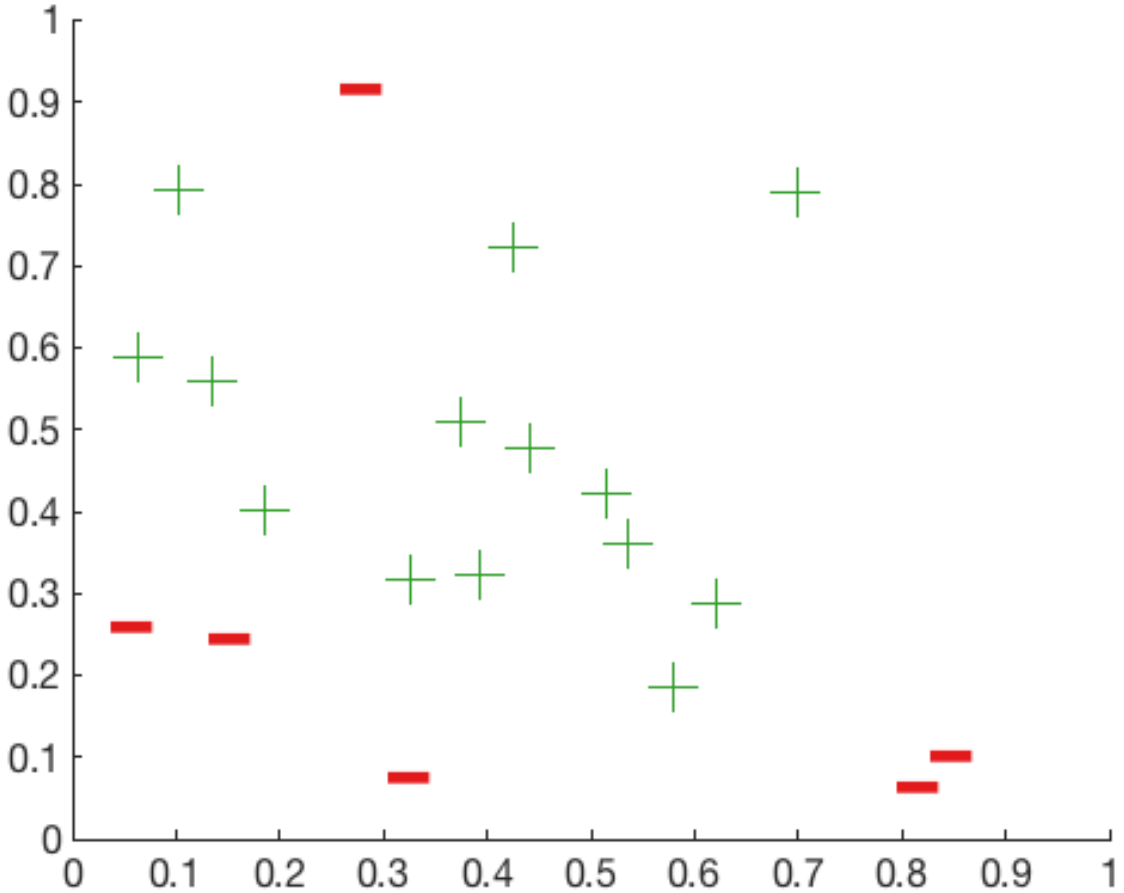
$$\text{subject to} \sum_{i=1}^{n}\tilde{\alpha}_i y_i = 0$$

$$\tilde{\alpha}_i \geq 0 \; \forall \; i \in \{1, \ldots, n\}$$

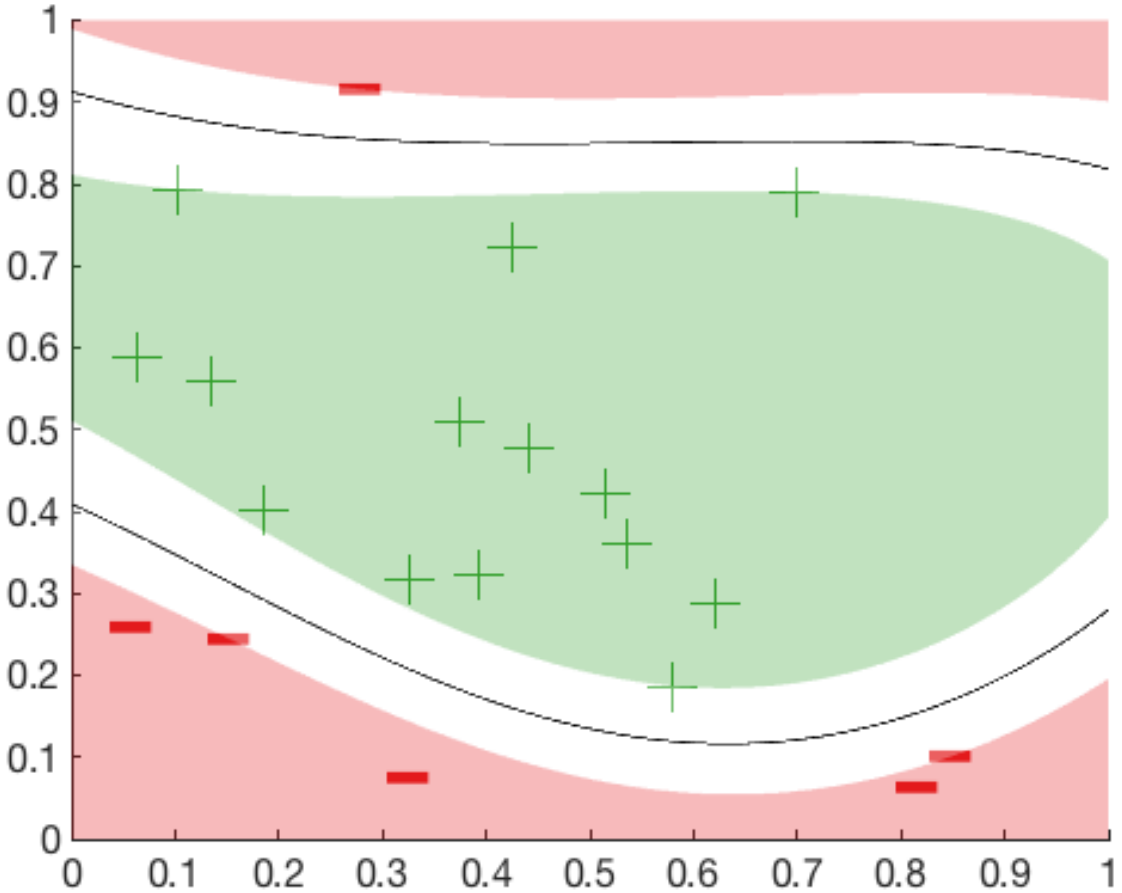
- Return the corresponding predictor in the original space:

$$g(\vec{x}) = \text{sign}\left(\sum_{i:\alpha_i^* > 0}\tilde{\alpha}_i^* y_i K_\Phi(\vec{x_i}, \vec{x}) + \tilde{w}_0^*\right)$$

Gaussian-RBF Kernel



Gaussian-RBF Kernel



Nonlinear Soft-Margin Dual SVMs

- Decide on a (valid) kernel function K_Φ
- Find a maximal-margin separating hyperplane in the transformed space, $[\vec{\tilde{w}}^*, \tilde{w}_0^*]$, by solving the QP:

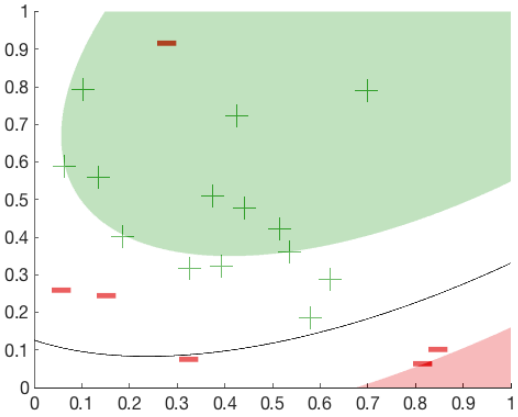
$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j K_\Phi(\vec{x_i}, \vec{x_j}) - \sum_{i=1}^{n}\tilde{\alpha}_i$$

$$\text{subject to} \sum_{i=1}^{n}\tilde{\alpha}_i y_i = 0$$

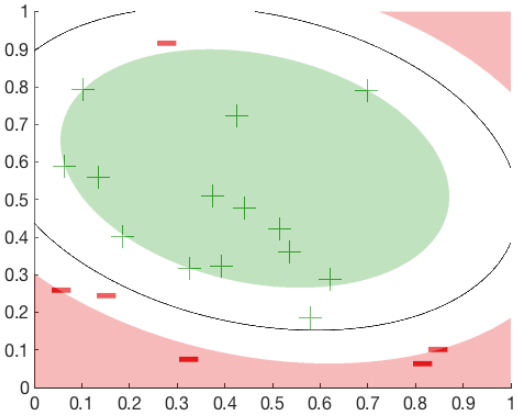
$$0 \le \tilde{\alpha}_i \le C \;\forall\; i \in \{1, \ldots, n\}$$

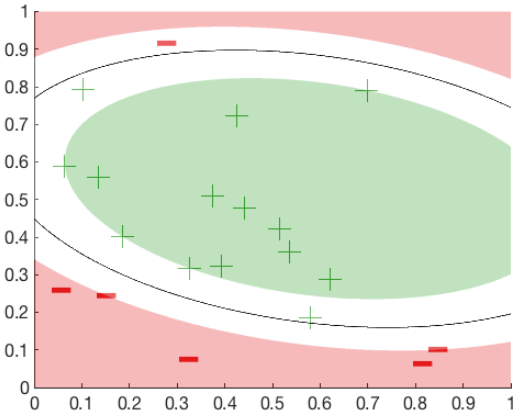
- Return the corresponding predictor in the original space:

$$g(\vec{x}) = \text{sign}\left(\sum_{i:\alpha_i^* > 0}\tilde{\alpha}_i^* y_i K_\Phi(\vec{x_i}, \vec{x}) + \tilde{w}_0^*\right)$$

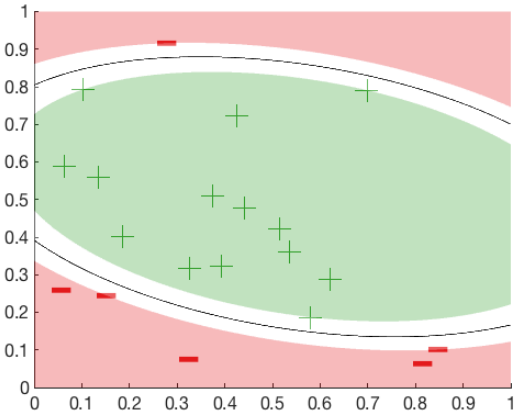


Smaller C





Larger C



Hard Margin

2^{nd}-Degree Polynomial Kernel

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

Use validation!