EE269
Signal Processing for Machine Learning

Lecture 11

Instructor : Mert Pilanci

Stanford University

February 13, 2019

Recap: Soft Margin Support Vector Machine

$$\min_{w,b,s_1,...,s_n} \frac{1}{2}||w||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - s_i \ \forall i$$
$$s_i \geq 0 \ \forall i$$

- $s_1, ..., s_n$ are slack variables
- C is a tuning parameter

Recap: Primal SVM vs Dual SVM

- primal problem

$$\min_{w,b,s_1,\ldots,s_n} \frac{1}{2}||w||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - s_i \ \forall i$$
$$s_i \geq 0 \ \forall i$$

- dual problem

$$\max_{\alpha} -\frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$$
$$\text{subject to } \sum_i \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq \frac{C}{n}$$

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

Geometry of SVM

- dual problem

$$\max_{\alpha} -\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \frac{C}{n}$$

- $C \to \infty$ gives hard margin SVM

$$\max_{\alpha} -\frac{1}{2}\left|\left|\sum_i \alpha_i y_i x_i\right|\right|_2^2 + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i$$

Geometry of SVM

- dual problem

$$\max_{\alpha} -\frac{1}{2}\left|\left|\sum_i \alpha_i y_i x_i\right|\right|_2^2 + \sum_i \alpha_i$$
$$\text{subject to } \sum_i \alpha_i y_i = 0$$
$$0 \leq \alpha_i$$

- $C \to \infty$ gives hard margin SVM

$$\max_{\alpha} -\frac{1}{2}\left|\left|\sum_{i\in+} \alpha_i x_i - \sum_{i\in-} \alpha_i x_i\right|\right|_2^2 + \sum_i \alpha_i$$
$$\text{subject to } \sum_{i\in+} \alpha_i = \sum_{i\in-} \alpha_i$$
$$0 \leq \alpha_i$$

Geometry of SVM

$$\max_{\alpha} -\frac{1}{2}\left|\left|\sum_{i\in+}\alpha_i x_i - \sum_{i\in-}\alpha_i x_i\right|\right|_2^2 + \sum_{i\in+}\alpha_i + \sum_{i\in-}\alpha_i$$

$$\text{subject to } \sum_{i\in+}\alpha_i = \sum_{i\in-}\alpha_i$$

$$0 \leq \alpha_i$$

Geometry of SVM

$$\max_{\alpha} -\frac{1}{2}\left|\left|\sum_{i\in+}\alpha_i x_i - \sum_{i\in-}\alpha_i x_i\right|\right|_2^2 + \sum_{i\in+}\alpha_i + \sum_{i\in-}\alpha_i$$

$$\text{subject to } \sum_{i\in+}\alpha_i = \sum_{i\in-}\alpha_i$$

$$0 \leq \alpha_i$$

- impose the constraint $\sum_{i\in+}\alpha_i = \gamma$ for some $\gamma > 0$ and maximize over γ to obtain the same problem

$$\max_{\gamma\geq 0}\max_{\alpha} -\frac{1}{2}\left|\left|\sum_{i\in+}\alpha_i x_i - \sum_{i\in-}\alpha_i x_i\right|\right|_2^2 + 2\gamma$$

$$\text{subject to } \sum_{i\in+}\alpha_i = \sum_{i\in-}\alpha_i = \gamma$$

$$0 \leq \alpha_i$$

$$\max_{\gamma \geq 0} \max_{\alpha} -\frac{1}{2} \left|\left| \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i \right|\right|_2^2 + 2\gamma$$

$$\text{subject to } \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i = \gamma$$

$$0 \leq \alpha_i$$

- variable change $\alpha_i \leftarrow \gamma \alpha_i$

$$\max_{\gamma \geq 0} \max_{\alpha} -\frac{\gamma^2}{2} \left|\left| \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i \right|\right|_2^2 + 2\gamma$$

$$\text{subject to } \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i = 1$$

$$0 \leq \alpha_i$$

$$\max_{\gamma \geq 0} \max_{\alpha} -\frac{\gamma^2}{2} \left|\left| \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i \right|\right|_2^2 + 2\gamma$$

$$\text{subject to } \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i = 1$$

$$0 \leq \alpha_i$$

- optimize over γ: $-\gamma^*(||\sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i||_2^2) + 2 = 0$

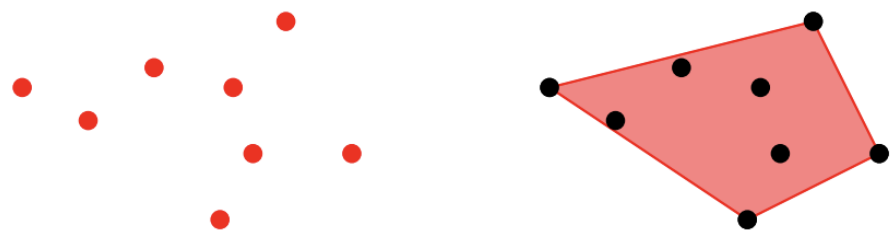
$$\gamma^* = \frac{2}{||\sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i||_2^2}$$

Geometry of SVM

- plug in optimal γ^*

$$\min_{\alpha} \left|\left| \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i \right|\right|_2^2$$
$$\text{subject to } \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i = 1, \quad 0 \leq \alpha_i$$

- **convex hull**: All possible weighted averages of vectors in a set



every point in the shaded region can be reached by a weighted sum

- c. hull$(x_1, \ldots x_n) := \left\{ x \, : \, x = \sum_{i=1}^n \alpha_i x_i \, , \, \alpha \geq 0, \, \sum_{i=1}^n \alpha_i = 1 \right\}$

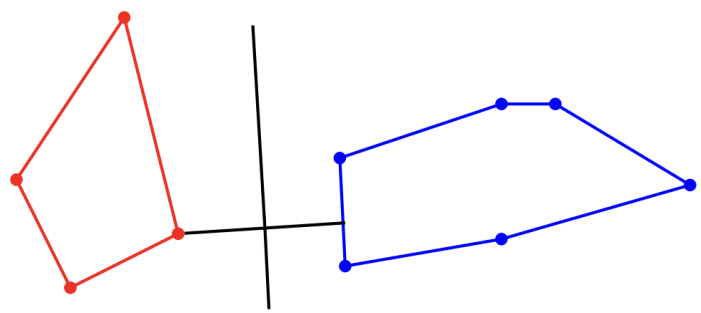
Geometry of SVM

- plug in optimal γ^*

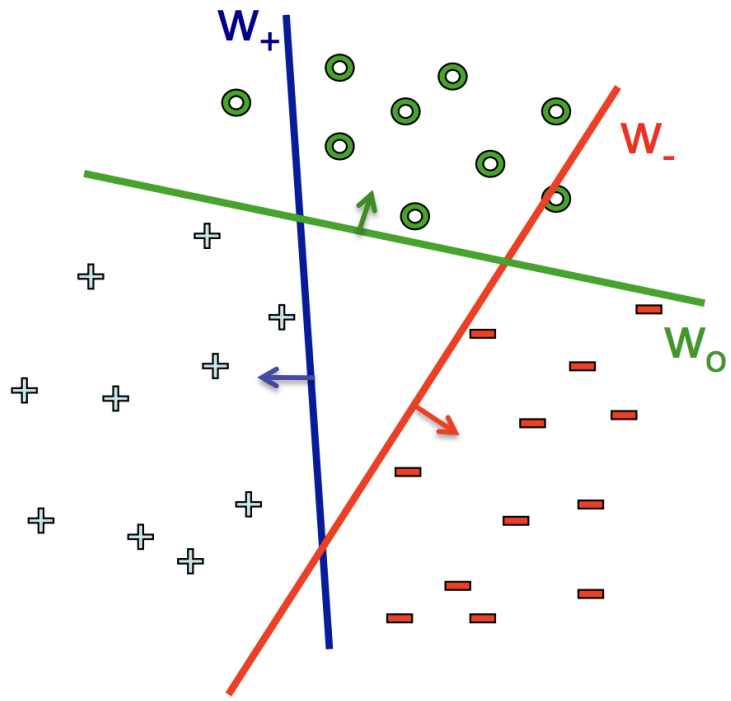
$$\min_{\alpha} \left|\left| \sum_{i \in +} \alpha_i x_i - \sum_{i \in -} \alpha_i x_i \right|\right|_2^2$$
$$\text{subject to } \sum_{i \in +} \alpha_i = \sum_{i \in -} \alpha_i = 1$$
$$0 \leq \alpha_i$$

- minimum distance between convex hulls

$$\min_{\substack{u \in \mathcal{H}(S_1) \\ v \in \mathcal{H}(S_0)}} \|u - v\|^2$$



Multi-class SVM



- Score of the correct class k, i.e., $w_k^T x + b_k$ must be higher

$$w_k^T x_i + b_k > w_{k'}^T x_i + b_{k'} \ , \forall k' = \text{correct class}$$

Multi-class SVM

- K classes

$$\min_{w,b,s_1,\ldots,s_n} \frac{1}{2}||w_k||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$

$$s.t. \quad w_k^T x_i + b_k \geq w_{k'}^T x_i + b_{k'} + 1 - s_i \ , \forall k' \neq y_i \ \forall i$$

$$s_i \geq 0 \quad \forall i$$

Multi-class SVM: Prediction

- K classes

$$\min_{w,b,s_1,\ldots,s_n} \frac{1}{2}||w_k||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$

$$s.t. \quad w_k^T x_i + b_k \geq w_{k'}^T x_i + b_{k'} + 1 - s_i \ , \forall k' \neq y_i \ \forall i$$

$$s_i \geq 0 \quad \forall i$$

Multi-class SVM via binary classifiers

- K classes
- Alternative 1: Train K binary classifiers

  class k vs **all other classes** for $k = 1, ..., K$

- Alternative 2: Train $\binom{K}{2}$ binary classifiers

  class k vs **class** k' for $k = 1, ..., K$, $k' = 1, ..., K$

Kernels

- Nonlinear feature maps: transform feature vectors via

$$\Phi : \mathbb{R}^d \to \mathbb{R}^m$$

- example: $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$
- consider the nonlinear feature map $\Phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$
- circular classifier $y \to \text{sign}\Big((x_1 - c_1)^2 + (x_2 - c_2)^2 - r^2\Big)$

Recap: Optimal signal classification with different variances

- $x[n] = [x_1, ...x_N] \sim N(0, \sigma_1^2)$ i.i.d. for $y = 1$ (noise)
- $x[n] = [x_1, ...x_N] \sim N(0, \sigma_2^2)$ i.i.d. for $y = 2$ (signal+noise)
- $P(y = 1) = \pi_1$ and $P(y = 2) = \pi_2$

  Assume σ_1 and σ_2 is known

$$g_1(x) := P_{x|y=1} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_i)^2}$$

$$g_2(x) := P_{x|y=2} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2\sigma_2^2}(x_i)^2}$$

Bayes classifier:

$$\text{Classify as class 2 if } \pi_1 g_1(x) < \pi_2 g_2(x)$$

$$\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2}\right) \sum_{i=1}^{N} x_i^2 > \log(\frac{\pi_1}{\pi_2}) - \frac{N}{2} \log(\frac{\sigma_1^2}{\sigma_2^2})$$

Kernels

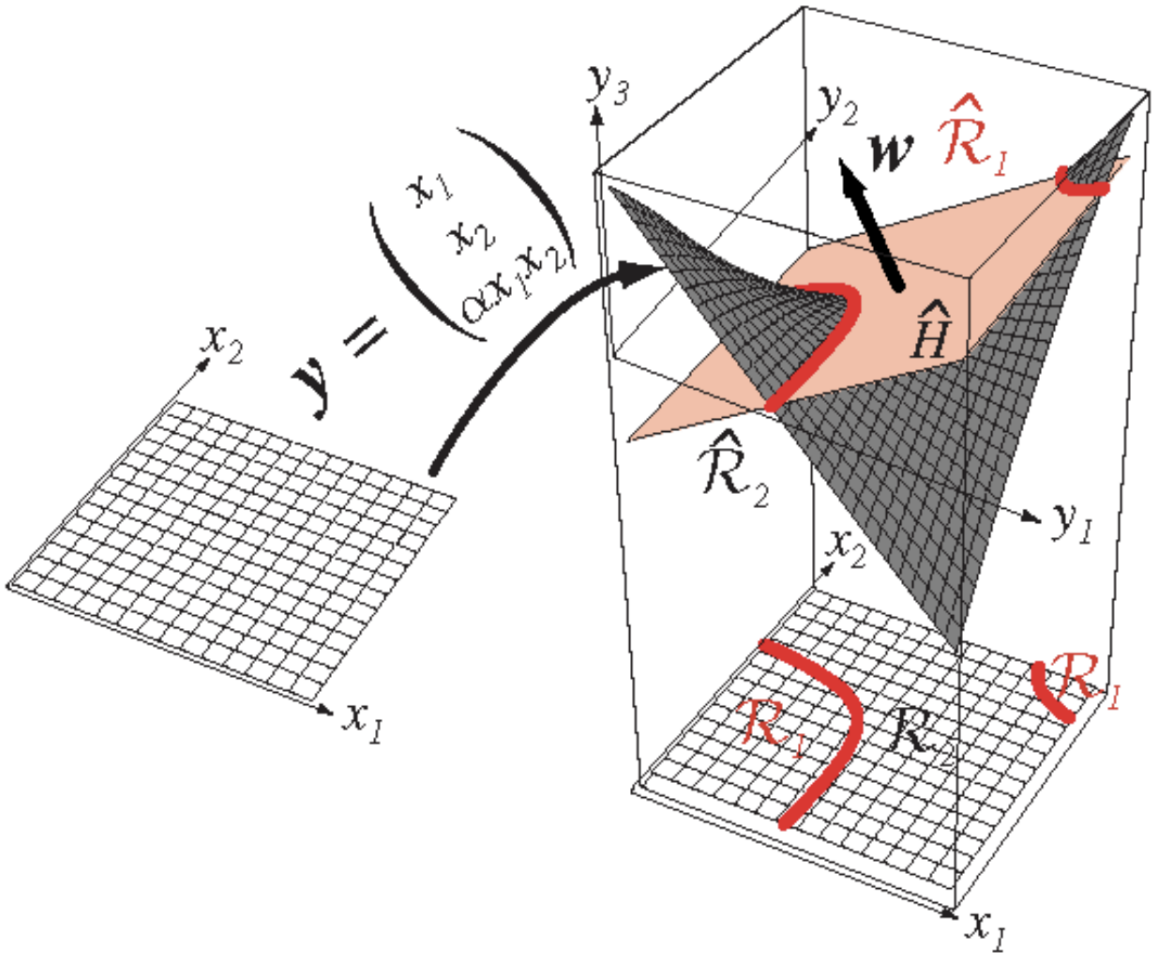
- Nonlinear feature maps: transform feature vectors via

$$\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^M$$

- we can apply a linear method after the transformation $\Phi(x_1), \Phi(x_2), ...\Phi(x_n)$
- classifier $y \rightarrow \mathrm{sign}(\mathrm{w}^{\mathrm{T}}\mathrm{x} + \mathrm{b})$
- Kernel classifier $y \rightarrow \mathrm{sign}(\mathrm{w}^{\mathrm{T}}\Phi(\mathrm{x}) + \mathrm{b})$

Lifting

Hyperplanes in lifted space



Inner product Kernel

- since $\Phi(x)$ is high dimensional, we may not compute it explicitly
- many learning algorithms depend on $\Phi(x)$ via

$$\langle \Phi(x), \Phi(x') \rangle$$

- **kernel function** $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- can be computed efficiently even if $\Phi(x)$ is infinite dimensional!

Kernel functions

$$\langle \Phi(x), \Phi(x') \rangle$$

- **kernel function** $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- real valued function of two arguments $\kappa(x, x') \in \mathbb{R}$
- typically symmetric $\kappa(x, x') = \kappa(x', x)$ and nonnegative $\kappa(x, x') \geq 0$, and can be interpreted as a measure of similarity

Kernel functions: Examples

- example: Let $x, y \in \mathbb{R}^2$
  $\kappa(x, y) = (x^T y)^2 = \langle x, y \rangle^2$

- corresponding feature map $\Phi = \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}$

Kernel functions: Examples

- example: Let $x, y \in \mathbb{R}^d$
  $\kappa(x, y) = (x^T y)^2 = (\sum_{i=1}^{d} x_i y_i)^2 = \sum_{i=1}^{d} x_i y_i \sum_{j=1}^{d} x_j y_j$

Kernel functions: Examples

- example: Let $x, y \in \mathbb{R}^d$

  $\kappa(x, y) = (x^T y)^2 = (\sum_{i=1}^{d} x_i y_i)^2 = \sum_{i=1}^{d} x_i y_i \sum_{j=1}^{d} x_j y_j$

- corresponding feature map $(d + \binom{d}{2})$ dimensional

$$\Phi = \begin{bmatrix} x_1^2 \\ \vdots \\ x_d^2 \\ x_1 x_d \\ \vdots \\ x_{d-1} x_d \end{bmatrix}$$

Kernel functions: Examples

- exercise: Let $x, y \in \mathbb{R}^d$
  $\kappa(x, y) = (x^T y)^2 = (\sum_{i=1}^{d} x_i y_i)^3$
- find the corresponding feature map $\Phi(x)$

Kernel functions: Examples

- example: $\kappa(x, y) = (x^T y)^p$

Kernel functions: Examples

- example: $\kappa(x, y) = (x^T y)^p$

$$(x^T y)^p = \sum_{\substack{j_1,\ldots,j_d \\ \sum_i j_i = p}} \binom{p}{j_1, \ldots j_d} x_1^{j_1} \ldots x_d^{j_d} y_1^{j_1} \ldots y_1^{j_d}$$

Kernel functions: Examples

- example: $\kappa(x, y) = (x^T y)^p$

$$(x^T y)^p = \sum_{\substack{j_1,\ldots,j_d \\ \sum_i j_i = p}} \binom{p}{j_1, \ldots j_d} x_1^{j_1} \ldots x_d^{j_d} y_1^{j_1} \ldots y_1^{j_d}$$

- corresponding feature map

$$\Phi = \begin{bmatrix} \vdots \\ \sqrt{\binom{p}{j_1,\ldots j_d}} x_1^{j_1} \ldots x_d^{j_d} \\ \vdots \end{bmatrix}$$

all monomials of degree p

When can we find a feature map ?

- when does $\kappa(x, y) = \langle \Phi(x), \Phi(y) \rangle$ hold ?

When can we find a feature map ?

- when does $\kappa(x, y) = \langle \Phi(x), \Phi(y) \rangle$ hold ?
- **Positive Definite Kernels**

  the Kernel matrix $K_{ij} = \kappa(x_i, x_j)$ is a symmetric positive semidefinite matrix for all x_i, x_j

Example kernel functions

1. Homogenous polynomial kernel $\kappa(x, y) = (x^T y)^p$
2. Inhomogenous polynomial kernel $\kappa(x, y) = (1 + x^T y)^p$
3. Gaussian (Radial Basis Function) Kernel
   $\kappa(x, y) = e^{-\frac{1}{2\sigma^2}||x-y||_2^2}$

Kernel trick

1. select an inner product kernel $\kappa(x, x')$
2. formulate a linear learning method that only depends on inner products $\langle x, x' \rangle$
3. replace $\langle x, x' \rangle$ with $\kappa(x, x')$

- The matrix XX^T is replaced by the kernel matrix K

Primal SVM vs Dual SVM

- primal problem

$$\min_{w,b,s_1,\ldots,s_n} \frac{1}{2}||w||_2^2 + C\frac{1}{n}\sum_{i=1}^{n} s_i$$
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - s_i \ \forall i$$
$$s_i \geq 0 \ \forall i$$

- dual problem

$$\max_{\alpha} -\frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$$
$$\text{subject to } \sum_i \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq \frac{C}{n}$$

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

Dual SVM

- dual problem

$$\max_{\alpha} -\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \frac{C}{n}$$

- dual problem in matrix form

$$\max_{\alpha} -\frac{1}{2}\sum_{ij}\alpha^T \text{Diag}(y) X X^T \text{Diag}(y)\alpha + \sum_i \alpha_i$$

$$\text{subject to } \alpha^T y = 0$$

$$0 \leq \alpha_i \leq \frac{C}{n}$$

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

Lifting and SVM Duality

- SVM decision region $\text{sign}(x^T w + b) = \text{sign}(\tilde{x}^T \tilde{w})$
- SVM primal and dual solutions satisfy:
  $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$

- lifting $x \to \Phi(x)$ gives
  $w^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i)$ in the lifted space

Lifting and SVM Duality

- SVM decision region $\text{sign}(x^T w + b) = \text{sign}(\tilde{x}^T \tilde{w})$
- SVM primal and dual solutions satisfy:
  $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$
- lifting $x \to \Phi(x)$ gives
  $w^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i)$ in the lifted space
- decision region for the test sample $x \to \Phi(x)$

$$\text{sign}\Big(\Phi(x)^T w^*\Big) = \text{sign}\Big(\Phi(x)^T \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i)\Big)$$
$$= \text{sign}\Big(\sum_{i=1}^n \alpha_i^* y_i \Phi(x)^T \Phi(x_i)\Big)$$

- $\Phi(x)^T \Phi(x_i)$ is an inner-product in the lifted space

Kernel SVM (Kernel Machine)

- dual problem with lifted features $\Phi(x)$

$$\max_{\alpha} -\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \Phi(x_i)^T\Phi(x_j) + \sum_i \alpha_i$$

$$\text{subject to } \sum_i \alpha_i y_i = 0$$

$$0 \le \alpha_i \le \frac{C}{n}$$

- dual problem in matrix form

$$\max_{\alpha} -\frac{1}{2}\sum_{ij}\alpha^T \text{Diag}(y) K \text{Diag}(y)\alpha + \sum_i \alpha_i$$

$$\text{subject to } \alpha^T y = 0$$

$$0 \le \alpha_i \le \frac{C}{n}$$

- $w^* = \sum_{i=1}^n \alpha_i^* y_i \Phi(x_i)$
- Kernel matrix $K_{ij} = \langle \Phi(x_i), \Phi(x_j)\rangle = \kappa(x_i, x_j)$
- Decision region

$$\text{sign}\Big(\sum_{i=1}^n \alpha_i^* y_i \Phi(x)^T\Phi(x_i)\Big) = \text{sign}\Big(\sum_{i=1}^n \alpha_i^* y_i \kappa(x, x_i)\Big)$$