CS 307 Lecture 15

Primal (Hard) SVM: When $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ is linearly separable

$$\left[\theta^T x + \theta_0 \begin{matrix} \geq 0 \\ < 0 \end{matrix} \right]$$

$$\min \ \frac{1}{2} \theta^T \theta$$

$$\text{s.t} \quad y^{(i)} \left(\theta^T x^{(i)} + \theta_0 \right) \geq 1 \quad \forall i \in \{1, 2 \ldots n\}$$

Lagrangian: $\mathcal{L}(\alpha, \theta, \theta_0) = \frac{1}{2} \theta^T \theta + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)} \left(\theta^T x^{(i)} + \theta_0 \right) \right)$

$(\alpha_i \geq 0)$

$$\text{Primal} = \min_{\theta, \theta_0} \sup_{\alpha} \mathcal{L}(\alpha, \theta, \theta_0)$$

Since $\mathcal{L}$ is convex

$$\sup_{\alpha} \underbrace{\min_{\theta, \theta_0} \mathcal{L}(\alpha, \theta, \theta_0)}_{\text{Achived when } \theta = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}} = \min_{\theta, \theta_0} \sup_{\alpha} \mathcal{L}(\alpha, \theta, \theta_0)$$

Dual (Hard) SVM: Find α such that

[Taking $\theta_0 = 0$]

General case $\theta_0 \neq 0$

$$\sup_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \underbrace{\langle x^{(i)}, x^{(j)} \rangle}_{x^{(i)T} x^{(j)}}$$

[Inner prod / dot prod]

$$\text{s.t} \quad \alpha_i \geq 0 \quad \forall i \in \{1, 2 \ldots n\}$$

$$\sum \alpha_i y^{(i)} = 0$$

Prediction on new x: Compute $\text{sign}\left(\sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle \right)$

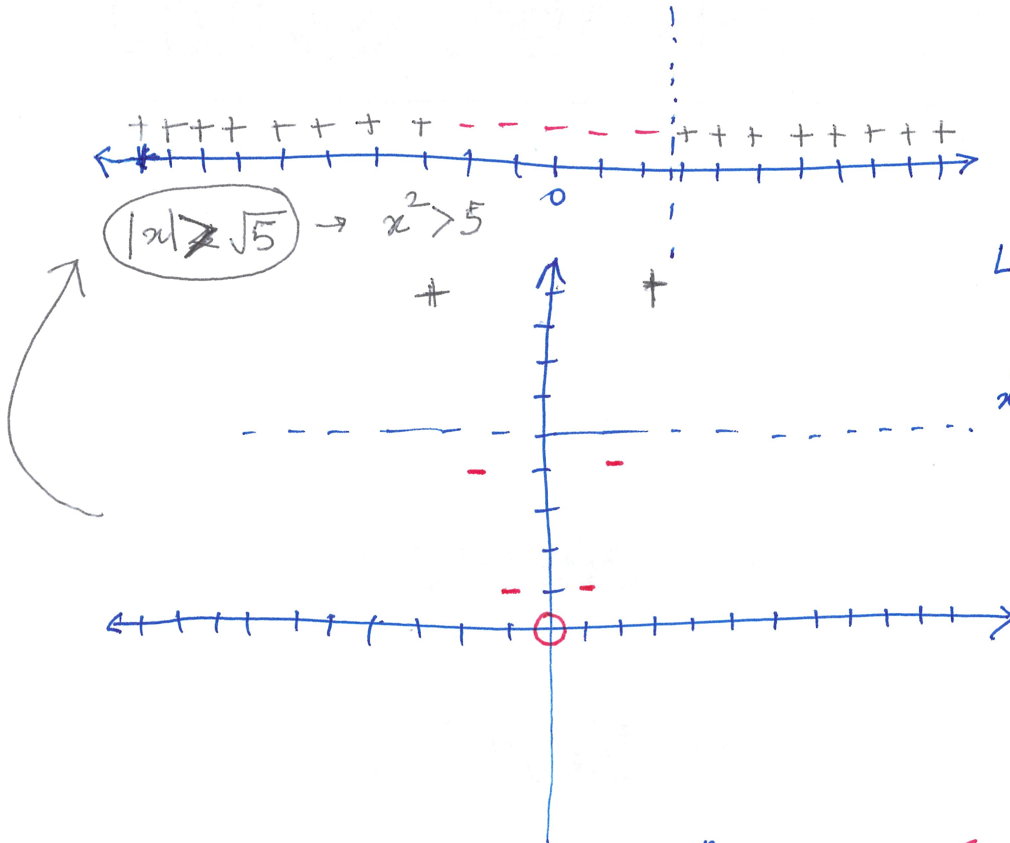
$$\text{sign}(a) = \begin{cases} +1 & \text{when } a > 0 \\ -1 & \text{when } a \leq 0 \end{cases}$$

$\text{sign}(\theta^T x)$

Example: $S = \{(-10,+1), (-9,+1) \dots (-2,-1), (-1,-1), (0,-1) \dots (2,-1), (3,+1) \dots (10,+1)\}$

$\boxed{ax+b \geq 0 \quad \leq 0}$

$x \geq b$
$\leq b$



Look at x, x^2

Enhance Features: $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^d$

Mapping $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_1}$ $\quad (d_1 \geq d)$

Example
$\phi : \mathbb{R} \to \mathbb{R}^2$
$\phi(x) = (x, x^2)$

(2). Construct $\phi(S) = \{(\underbrace{\phi(x^{(i)})}_{\in \mathbb{R}^{d_1}}, y^{(i)})\}_{i=1}^{n}$

(3) Train a linear classifier on $\phi(S)$. Find $\hat{\theta} \in \mathbb{R}^{d_1}$, $\theta_0 \in \mathbb{R}$.

(4) Prediction on x : $\mathrm{sign}(\hat{\theta}^T \phi(x) + \theta_0)$

Quadratic Classifier: $x^T = (x_1, x_2 \dots x_d)$

$\phi(x)^T = (1, x_1, \underbrace{x_2, \dots x_d, x_1x_1, x_1x_2, x_1x_3}_{d^2+d+1} \dots x_1x_d, x_2x_1 \dots x_dx_d)$

~~Kernel Trick:~~

~~- Fixet~~

SVM in high dimensions: Suppose $\phi : \mathbb{R}^d \to \mathbb{R}^{d_1}$

Find $\theta \in \mathbb{R}^{d_1}$, $\theta_0 \in \mathbb{R}$

$$\min \ \frac{1}{2}\theta^T\theta$$

$$\text{s.t} \quad y^{(i)}\left(\theta^T \phi(x^{(i)}) + \theta_0\right) \geq 1$$

Optimal $\theta = \sum_{i=1}^{n} \alpha_i y^{(i)} \phi(x^{(i)})$

$$\sup_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$$

$$\text{s.t} \quad \alpha_i \geq 0$$

$$\sum \alpha_i y^{(i)} = 0$$

[Assume $\theta_0 = 0$]

Prediction x: $\operatorname{sign}\left(\sum \alpha_i y^{(i)} \langle \phi(x^{(i)}), \phi(x) \rangle\right)$

Goal ϕ:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \to \phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix} \qquad z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \to \phi(z) = \begin{bmatrix} z_1z_1 \\ z_1z_2 \\ z_1z_3 \\ z_2z_1 \\ z_2z_2 \\ z_2z_3 \\ z_3z_1 \\ z_3z_2 \\ z_3z_3 \end{bmatrix}$$

Compute $\langle \phi(x), \phi(z) \rangle = \left(\langle x, z \rangle\right)^2$

$$\left(\langle x, z \rangle\right)^2 = \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) = \sum_{i=1}^{d}\sum_{j=1}^{d} x_i z_i x_j z_j = \sum_{i=1}\sum_{j=1} \underline{(x_i x_j)(z_i z_j)}$$

$$= \langle \phi(x), \phi(z) \rangle$$

$$\phi_1 : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ \vdots \\ x_3 x_3 \\ x_1 \\ x_2 \\ \vdots \\ x_3 \\ 1 \end{bmatrix}$$

$$\langle \phi_1(x), \phi_1(z) \rangle = (1 + \langle x, z \rangle)^2$$

Instead of quadratic features, we want all k-ary monomials

$$\phi_k : \mathbb{R}^d \rightarrow \mathbb{R}^{d_1}$$

$$\langle \phi_k(x), \phi_k(z) \rangle = (1 + \langle x, z \rangle)^k$$

Kernel: $K(x, z) = \langle \phi(x), \phi(z) \rangle$ (for ϕ)

Compute $\sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y^{(i)} y^{(j)} \cancel{K(x^{(i)}, z)} K(x^{(i)}, x^{(j)})$

$$\text{sign}\left(\sum \alpha_i y^{(i)} K(x^{(i)}, x) \right)$$

Gram matrix: Given $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, G is $n \times n$ matrix

$$G_{ij} = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$$

Kernel: Function that measures the similarity between two vectors

$$K(x, z) \doteq \begin{cases} \text{large} & \text{when } x, z \text{ are similar} \\ \text{small} & \text{when } x, z \text{ are dissimilar} \end{cases}.$$

RBF or Gaussian Kernel:

$$K(x, z) = e^{\frac{-\|x - z\|}{2\sigma^2}}$$

A function K is "kernel" if $\exists\, \phi$ s.t. $K(x, z) = \langle \phi(x), \phi(z) \rangle$

is a kernel ϕ maps d vector to an ∞-dimensional space.

Kernels: "linear": $\phi(x) = x$. $K(x,z) = x^T z$

"polynomial": $\phi(x)$ – all monomials upto k. $K(x,z) = (1+x^T z)^k$

"RBF" or "gaussian": $K(x,z) = e^{-\frac{\|x-z\|}{2\sigma^2}}$

Kernel Trick: ① Write your learning algorithm in terms of inner products.

② Replace inner products with Kernels.