Lecture 7 Kernels

1. <continue> Optimal Margin Classifier

$x^{(i)} \in \mathbb{R}$

Suppose $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$ (w can be represented as a linear combination of x)

Why reasonable? <representation theorem>

- Intuition 1: (refer to logistic regression)

$\theta_0 := 0$

(stochastic) gradient descent: $\theta := \theta - \alpha (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$

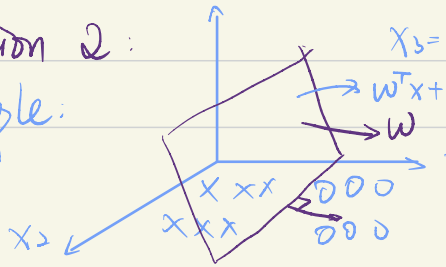
(batch) gradient descent: $\theta := \theta - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$

$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \Rightarrow \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$

No matter iteration, still linear combination

- Intuition 2:

example:



$\Rightarrow$ w should be represented in the span of just feature x_1 and x_2.

$\therefore w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$

$\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, 2, \cdots m$

$\|w\|^2 \rightarrow w^T w$

$$\min \frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T \left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)}\right)$$

$$= \min \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)}$$

$\langle x, z \rangle = x^T z$ is the inner product $\leftarrow \langle x^{(i)}, x^{(j)} \rangle$

$$y^{(i)}\left[\left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)}\right)^T x^{(i)} + b\right] \geq 1$$

$$y^{(i)}\left(\sum_{j=1}^{m} \alpha_j y^{(j)} \langle x^{(j)}, x^{(i)} \rangle + b\right) \geq 1$$

efficient

<a further simplification> : "Dual optimization Problem"

$$\text{Max} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \quad s.t.\ \alpha_i \geq 0, \ \sum_i y^{(i)} \alpha_i = 0$$

To predict:

1) Solve for α_i's; b
2) compute $h_{w,b}(x) = g(w^T x + b)$

$$= g\left(\left(\sum_i \alpha_i y^{(i)} x^{(i)}\right)^T x + b\right) = g\left(\sum_i \alpha_i y^{(i)} \langle x^{(i)}, x\rangle + b\right)$$

2. kernel trick:

① Write algorithm in terms of $\langle x^{(i)}, x^{(j)}\rangle$ (or $\langle x, z\rangle$)

② Let there be mapping from $x \in \mathbb{R}^n \longmapsto \phi(x)$, high-dimensional

③ Find way to compute $K(x, z) = \phi(x)^T \phi(z)$

④ Replace $\langle x, z\rangle$ in algorithm with $K(x, z)$

example:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^n \qquad \phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix} \in \mathbb{R}^{n^2} \qquad \phi(z) = \begin{bmatrix} z_1z_1 \\ z_1z_2 \\ \vdots \\ z_3z_2 \\ z_3z_3 \end{bmatrix}$$

n^2 elements

Need $O(n^2)$ time to compute $\phi(x)$, or $\phi(x)^T\phi(z)$ explicitly

$$K(x, z) = \phi(x)^T \phi(z) \overset{\text{prove}}{=\!=} (x^T z)^2$$

$x^T z \in \mathbb{R}$, $(x^T z)^2 \in \mathbb{R}$, $O(n)$ time

$$\Rightarrow \phi(x)^T \phi(z) = \overbrace{\left(\sum_{i=1}^{n} x_i z_i\right)}^{x^T z} \overbrace{\left(\sum_{j=1}^{n} x_j z_j\right)}^{x^T z}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i z_i x_j z_j = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i x_j)(z_i z_j) = (x^T z)^2$$

3. Optimal Margin Classifier + Kernel trick = SVM

4. How to make kernels?

{ If x, z are "similar", $K(x, z) = \phi(x)^T \phi(z)$ is "large"
  If x, z are "dissimilar", $K(x, z)$ is "small"

= Does the kernel exists ϕ s.t. $K(x, z) = \phi(x)^T \phi(z)$

$K(x, x) = \phi(x)^T \phi(x) \geq 0$

Let $\{x^{(1)}, x^{(2)}, \dots x^{(d)}\}$ be d points

Let $K \in \mathbb{R}^{d \times d}$, $K_{ij} = K(x^{(i)}, x^{(j)})$

↖ kernel matrix

Given any vector z,

$$z^T K z = \sum_i \sum_j z_i k_{ij} z_j = \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j$$

$$= \sum_i \sum_j z_i \sum_k (\phi(x^{(i)}))_k (\phi(x^{(j)}))_k z_j$$

$$= \sum_k \sum_i \sum_j z_i (\phi(x^{(i)}))_k (\phi(x^{(j)}))_k z_j$$

$$= \sum_k \left(\sum_i z_i \phi(x^{(i)})_k \right)^2 \geq 0$$

So $K \geq 0$

Mercer's Theorem: K is a valid kernel function (i.e. $\exists \phi$ s.t. $K(x,z) = \phi(x)^T \phi(z)$) if and only if for any d points $\{x^{(1)}, \dots x^{(d)}\}$, the corresponding kernel matrix $K \geq 0$.

5. Other kernels:

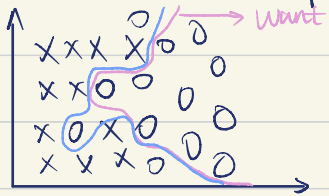
Linear kernel: $k(x,z) = x^T z$; $\phi(x) = x$

Gaussian kernel: $k(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$; $\phi(x) \in \mathbb{R}^\infty$

Other kernel: $k(x,z) = (x^T z + c)^2$

Polynomial kernel: $k(x,z) = (x^T z + c)^d$ e.g. $\begin{bmatrix} x_1 x_3 \\ x_2 x_3 \\ \vdots \\ x_3 x_3 \\ \sqrt{2c}\, x_1 \\ \vdots \\ \sqrt{2c}\, x_3 \end{bmatrix}$

b. L_1 norm soft margin SVM (when you don't want the algorithm to be so correct)



$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \quad i = 1, 2, \cdots m$$

$$\xi_i \geq 0$$

$$\Rightarrow \max_\alpha W(\alpha) = \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t. } \sum_{i=1}^{m} y^{(i)} \alpha_i = 0 \quad , \; 0 \leq \alpha_i \leq C, \; i = 1, 2, \cdots m$$