Outline:

- Kernel Methods
- SVM

Feature map (Recap)

$\phi: \mathbb{R}^d \to \mathbb{R}^p$

$x \to \phi(x)$

fitting $h_\theta(x) = \theta^T \phi(x)$ using gradient descent.

Issue: $\theta, \phi(x) \in \mathbb{R}^p$ can be very high dimensional
runtime per iteration $O(np)$

Goal: improve to $O(n^2)$ per iteration

Kernel Methods

Key observation:

θ can be represented as

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x_i)$$

(β_i: scalar; n variables instead of p)

Proof by induction

At iteration 0

$$\theta = 0 = \sum_{i=1}^{n} 0 \cdot \phi(x^{(i)})$$

Suppose at iteration t, $\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})$

Next iteration

$$\theta := \theta + \alpha \sum (y^{(i)} - \theta^T \phi(x^{(i)})) \phi(x^{(i)})$$

(θ: $\sum \beta_i \phi(x^{(i)})$)

$$\theta := \sum_{i=1}^{n} (\beta_i + \alpha \sum_j (y^{(j)} - \theta^T \phi(x^{(j)})) \phi(x^{(i)})$$

(New β_i)

Represent $\theta \in \mathbb{R}^p$ implicitly by $\beta \in \mathbb{R}^n$
works better when $p \gg n$

Update rule for β

$$\beta_i = \beta_i + \alpha (y^{(i)} - \theta^T (x^{(i)}))$$

$$\theta^T = \left(\sum_{j=1}^{n} \beta_j \phi(x^{(j)})\right)^T = \sum_j \beta_j \phi(x^{(j)})^T$$

$= \beta_i + \alpha\left(y^{(i)} - \left(\sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T\right)\phi(x^{(i)})\right)$

$= \beta_i + \alpha\left(y^{(i)} - \sum_{j=1}^{n} \beta_j \langle \phi(x^{(j)}), \phi(x^{(i)}) \rangle\right)$ Note: $a^T b = \langle a, b \rangle$

Observations

① $\langle \phi(x^{(j)}), \phi(x^{(i)}) \rangle$ $\forall i, j$ can be precomputed

② Often $\langle \phi(x^{(j)}), \phi(x^{(i)}) \rangle$ can be computed faster than $O(p)$ ex: $O(d)$

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1 x_1 \\ \vdots \\ x_d x_d \\ \vdots \\ x_d^3 \end{bmatrix} \qquad p = 1 + d + d^2 + d^3 = O(d^3)$$

$$\langle \phi(x), \phi(z) \rangle = 1 + \sum_{i=1}^{d} x_i z_i + \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j + \sum_{i}\sum_{j}\sum_{k} x_i x_j x_k z_i z_j z_k$$

$$= 1 + \underbrace{\langle x, z \rangle}_{O(d)} + \langle x, z \rangle^2 + \langle x, z \rangle^3$$

Can compute $\langle \phi(x), \phi(z) \rangle$ in $O(d)$ time!

$K(x, z) = \langle \phi(x), \phi(z) \rangle$ Kernel function

$K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

Algorithm for computing Kernel

Compute $K(x^{(i)}, x^{(j)}) = \langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$ for all $i, j \in \{1 \dots n\}$

Set $\beta = 0$

$\beta_i := \beta_i + \alpha\left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)})\right)$

preprocessing: $O(n^2 d)$

Each iteration: $O(n^2)$ per iteration

Prediction: Given x, compute $\theta^T \phi(x)$

$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x)$ $O(nd)$ time

$= \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$

Deeper Observations

The algo only depends on $K(.,.)$, no need to know ϕ

design of features $\longrightarrow$ design of kernel functions

What Kernels are valid?

$\exists\ \phi$ s.t. $K(x, z) = \langle \phi(x), \phi(z) \rangle$

Necessary conditions:

n data points. $x^{(1)}, \dots, x^{(n)}$

Kernel matrix $K \in \mathbb{R}^{n \times n}$ $K_{ij} = K(x^{(i)}, x^{(j)})$

matrix K is positive semidefinite

$K \succeq 0$

$K \succeq 0 \iff z^T K z \geq 0$

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

$\phi(x) = (\phi_1(x), \phi_2(x), \dots)$

$$z^T K z = \sum_i \sum_j z_i k_{ij} z_j$$

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

$$= \sum_i \sum_j z_i \sum_l \phi_l(x^{(i)}) \phi_l(x^{(j)}) z_j$$

$$= \sum_l \left(\sum_i z_i \phi_l(x^{(i)}) \right)^2 \geq 0$$

$\Rightarrow K \succeq 0$

Theorem (Mercer) K is a valid kernel fn (i.e $K(x,z) = \phi(x)^T \phi(z)$) iff for any $n < \infty$ and $x^{(1)} \dots x^{(n)}$ the corresponding Kernel matrix K s.t. $K_{ij} = K(x^{(i)}, x^{(j)})$ is positive semidefinite

Other Kernels:

- $K(x,z) = (x^T z + c)^2 = \langle \phi(x), \phi(z) \rangle$

for $\phi(x) = \begin{bmatrix} \frac{c}{\sqrt{2c}} x_1 \\ \vdots \\ \frac{c}{\sqrt{2c}} x_d \\ x_1^2 \\ \vdots \\ x_d^2 \end{bmatrix}$

- $K(x,z) = (x^T z + c)^t$

Gaussian Kernel: $K(x,z) = \exp\left(\frac{-\|x - z\|_2^2}{2\sigma^2} \right)$

$K(x,z) = \langle \phi(x), \phi(z) \rangle \Rightarrow \phi$ is actually infinite dimensional!

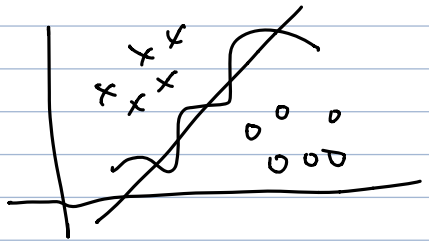
Example: Protein sequence classification

seq of ammoacids (A - T)

$\begin{bmatrix} AAAA \\ AAAB \\ \vdots \\ TTTT \end{bmatrix} \rightarrow 20^4$ dim vector

Histograms: take min and sum up.

SVM for classification



linear: $\{x: w^T x + b = 0\}$

non-linear: $\{x: w^T \phi(x) + b = 0\}$.

SVM

$\phi(x) = x$ (for now)

$y^{(i)} \in \{-1, 1\}$

Warmup

find w, b s.t.

if $y^{(i)} = 1$ $\quad w^T x^{(i)} + b > 0$ ①

$y^{(i)} = -1$ $\quad w^T x^{(i)} + b < 0$ ②

Many such w, b.

New goal: Among all w, b. satisfies ① & ②.

Find w, b s.t. $\max_{w,b} \min_{i \in \{1,\dots n\}}$ dist$(x^{(i)}$, boundary$)$

boundary $= \{x: w^T x + b = 0\}$

$$= \begin{cases} \dfrac{w^T x^{(i)} + b}{\|w\|_2} & \text{if } x^{(i)} \text{ is on the positive side } w^T x^{(i)} + b > 0 \\ -\dfrac{(w^T x^{(i)} + b)}{\|w\|_2} \end{cases}$$

$$\max_{w,b} \min_{i \in \{1,\dots n\}} \frac{y^{(i)}(w^T x^{(i)} + b)}{\|w\|_2}$$

Scale invariant $(w, b) \to (100w, 100b)$

we want to find w, b

s.t. $\min_{i \in \{1,\dots n\}} y^{(i)}(w^T x^{(i)} + b) = 1$ ③

$$\max \frac{1}{\|w\|_2} \equiv \min \|w\|_2$$

$\min \|w\|_2$

s.t. $\forall i \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$ ⇔ equivalant

$\min \|w\|_2 \Rightarrow \min \frac{1}{2}\|w\|_2^2$ (Convex!)

s.t. $\forall i,\ y^{(i)}(w^T x^{(i)} + b) \geq 1$

Facts: (i) optimal solution $w^* = \sum_{i=1}^{h} \alpha_i y^{(i)} x^{(i)}$ for some $\alpha_i \geq 0$ —④ (linear combination)

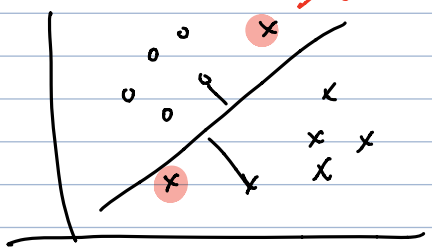
(ii) The α in ④ is the optimizer of

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

s.t. $\alpha_i \geq 0$, $\sum_i^n \alpha_i y^{(i)} = 0$

Kernelize: replace $\langle x^{(i)}, x^{(j)} \rangle$ with $K(x^{(i)}, x^{(j)})$

n variables.



$y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$ (relax)

$\min \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{n} \xi_i$

C pay some penalty