Outline:

Kernel Methods

SVM

Feature map

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

$x \longrightarrow \phi(x)$

fitting $h_\theta(x) = \theta^T \phi(x)$ using gradient descent with $\phi(x^{(i)})$ as features

$\theta = 0$

Loop

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

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

Key observation:

$\theta$ can be represented as

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

($\beta_i$: 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:

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

$\theta \downarrow \sum \beta_i \phi(x^{(i)})$

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

↑ new $\beta_i$

represent $\theta \in \mathbb{R}^p$ implicitly by $\beta \in \mathbb{R}^n$

works better $p >> n$

Update rule for $\beta$

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

$$\begin{aligned}
\beta_i &= \beta_i + \alpha\left(y^{(i)} - \theta^T \phi(x^{(i)})\right) \\
&= \beta_i + \alpha\left(y^{(i)} - \left(\sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T\right) \phi(x^{(i)})\right) \\
&= \beta_i + \alpha\left(y^{(i)} - \sum_{j=1}^{n} \beta_j \langle \phi(x^{(j)}), \phi(x^{(i)}) \rangle\right)
\end{aligned}$$

$$a^T b = \langle a, b \rangle$$

Observations

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

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

$$\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_d^2 \\ x_1^3 \\ x_1^2 x_2 \\ \vdots \\ x_d^3 \end{bmatrix}$$

$$p = 1 + d + d^2 + d^3 = O(d^3)$$

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

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

$$= 1 + \langle x, z \rangle + \langle x, z \rangle^2 + \langle x, z \rangle^3$$

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

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

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

Algo:

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

Set $\beta = 0$

Loop

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

preprocessing: $O(n^2 d)$

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

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

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

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

$O(nd)$ time

$x^{(i)}$: $i^{th}$ training example

$x$: new example

Deeper Observation

The algo only depends on $K(\cdot, \cdot)$

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 condition:
$n$ data points $x^{(1)} \ldots x^{(n)}$

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

matrix $K$ is positive semidefinite

$K \succeq 0$

$K_{ij}$

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

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

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

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

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

$$= \sum_i \sum_j z_i \sum_\ell \phi_\ell(x^{(i)}) \phi_\ell(x^{(j)}) \, z_j$$

$$= \sum_\ell \left( \sum_i z_i \, \phi_\ell(x^{(i)}) \right)^2 \geq 0$$

$\Rightarrow K \succeq 0$

Thm (Mercer) $K$ is a valid kernel fn (i.e. $K(x,z) = \phi(x)^T \phi(z)$) iff for any $n < \infty$ and any $x^{(1)} \dots x^{(n)}$ 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 $\phi(x) =$

$$\begin{bmatrix} c \\ \sqrt{2c}\, x_1 \\ \vdots \\ \sqrt{2c}\, x_d \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_d^2 \end{bmatrix}$$

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

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

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

$\phi$: infinite dimensional!

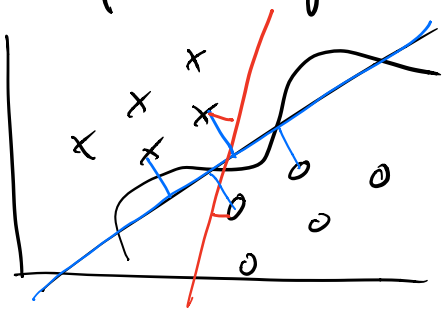
$x_1, x_1x_2 \dots x_1^{16} x_2^{27} \dots$

Protein sequence classification
sequences of amino acids (A-T)

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

Histograms: take min and sum up

SVMs 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 $y^{(i)} = 1$ $w^T x^{(i)} + b > 0$ ①
$y^{(i)} = -1$ $w^T x^{(i)} + b < 0$ ②
Many such $w, b$

New goal: Among all $w, b$ satisfying ① & ②

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

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

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

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

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

wlog. 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. $\forall i \quad y^{(i)}(w^T x^{(i)} + b) \geq 1$

$\min \|w\|_2 \Rightarrow \min \frac{1}{2}\|w\|_2^2$

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

Facts: ① optimal sol$^n$ $w^* = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$ for some $\alpha_i \geq 0$ ④

② The $\alpha$ in ④ is the optimizer of

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

$$s.t. \ \alpha_i \geq 0 \qquad \sum \alpha_i y^{(i)} = 0$$

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

n variables

Fixing linearly separable assumption:

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

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