SVM
- optimization problem
- kernel
- representer theorem

optimal margin classifier

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

scale $\|w\| = \frac{1}{\gamma}$

Then condition becomes $\max \frac{1}{\|w\|}$

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

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

Then becomes $\min \frac{1}{2}\|w\|^2$

s.t. $y^{(i)}(w^T x^{(i)}+b) \geq 1$ — Key idea: scaling w & b by common factor doesn't affect decision boundary

The above all are for $x^{(i)} \in R^n$, n is finite

Now for R^∞

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

Why? This assumption works?

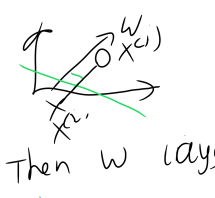
Ans 1 Logistic regression

$\theta = 0$, Gradient descent $\theta = \theta - \alpha(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$

Thus θ optimal must be lin combination of $x^{(i)}$

Batch gradient descent as well $\theta = \theta - \alpha\sum^{m}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$

Ans 2 $g(\underbrace{[\,^2_1\,]}_{w}x - 2)$ suppose



Then w lays in span of $x^{(2)}$ and $x^{(1)}$

Thus $w = \sum \alpha_i y^{(i)} x^{(i)}$

$\min \frac{1}{2}(\sum \quad)^T(\sum)$

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

Condition becomes $y^{(i)}((\sum \alpha_j y^{(j)} x^{(j)})^T x^{(i)} + b) \geq 1$

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

Again simplify problem to

$\sum \alpha_i - \frac{1}{2}\sum_i\sum_j y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$

s.t. $\alpha_i \geq 0$

$\sum y^{(i)}\alpha_i = 0$ — "Dual optimization problem"

1) Solve for α_i's, b

2) to make a prediction

$h_{w,b}(x) = g(w^T x + b)$

$= g(\sum \alpha_i y^{(i)} \langle x^{(i)}, x\rangle + b)$

Kernel trick

1. Write algorithm for $\langle x^{(i)}, x^{(j)}\rangle$ (or $\langle x, z\rangle$)
2. Let there be mapping from $x \to \phi(x)$ (higher or ∞ dim)
3. Find a way to compute $K(x,z) = \phi(x)^T\phi(z)$
4. Replace $\langle x, z\rangle$ with $K(x,z)$

Free lunch theorem

eg Suppose $x \in R^n$, $\phi(x) \in R^{n^2}$ $\quad \phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ x_3x_3 \end{bmatrix}$ similarly for $\phi(z)$

$\Rightarrow O(n^2)$ to compute $\phi(x)^T\phi(z)$

$K(x,z) = \phi(x)^T\phi(z) = (x^Tz)^2$

bcs $\underbrace{(\sum x_i z_i)}_{x^T z}(\sum x_j z_j)$

$= \sum\sum(x_i z_i)(x_j z_j)$

$= \phi^T(x)\phi(z)$

So $O(n)$ calculation

Another eg if $K(x,z) = (x^T z + c)^2$

then $\phi(x) = \begin{bmatrix} x_1x_2 \\ \vdots \\ x_1 \\ x_3 \\ \sqrt{2c}\,x_1 \\ \sqrt{2c}\,x_2 \\ \sqrt{2c}\,x_3 \end{bmatrix}$

Another eg $\phi(x)$ has all feature $\binom{n+d}{d}$ of monomial to order d

SVM = optimal margin boundary + kernel trick

as computation only scales linearly, not that bad

How to make kernels?

Criteria: If x,z "similar" $K(x,z) = \phi^T(x)\phi(z)$ as "large" — for inner product reasons

x,z "not similar" $K(x,z)$ "small"

Does there exist $K(x,z) = \phi(x)^T\phi(z)$

Valid kernel theorem: Let $\{x^{(1)}, \dots, x^{(d)}\}$ be d points

Let $K \in R^{d\times d}$ kernel matrix

$K_{ij} = K(x^{(i)}, x^{(j)})$

Given any vector z

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

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

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

$= \sum_k\sum_i\sum_j \dots$

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

So $K \geq 0 \Rightarrow$ positive-semidefinite

Mercer Theorem: K is valid kernel function iff for any d point $\{x^{(1)}, \dots, x^{(d)}\}$ the kernel matrix $K \geq 0$

Linear kernel $K(x,z) = x^T z$

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

using all monomial features

x_1

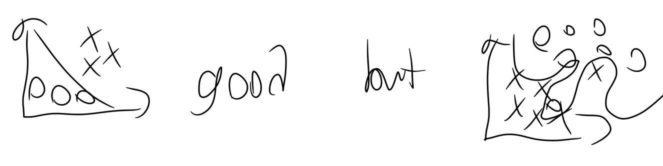
$x_1 x_2$

$x_1^2 x_2$

$x_1^{100} x_2^{1701}$

Kernel trick can apply to all the algorithm so far learned when is just efficient computation of inner inner product

Modification of linearly seperable decision boundary assumption



good but if noisy then bad

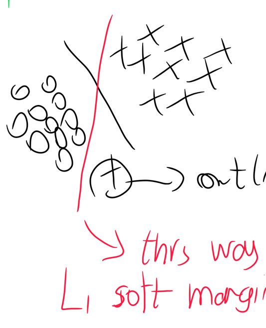
L_1 norm soft margin SVM

$\min \frac{1}{2}\|w\|^2 + C\sum^{m}\xi_i$

s.t. $\underbrace{y^{(i)}(w^T x^{(i)}+b)}_{\text{function margin}} \geq 1 - \xi_i \quad i=1\dots m$

$\xi_i \geq 0$, don't want ξ_i too big so modify optimization problem

Why use L_1 norm soft SVM



outlier

this way too huge impact

L_1 soft margin can prevent this

$\max \sum \alpha_i - \sum\sum y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$

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

$0 \leq \alpha_i \leq C, \; i = 1 \dots m$

How to choose C?

eg $K(x,z) = (x^T z)^d$

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

Protein sequence classifier

A ... Z

Proteins are sequence of amino acid

$\phi(x) = ?$

AAAA

$\vdots$

ZZZZ

$\phi(x) = \begin{bmatrix} 0 \\ 8 \\ 0 \\ \vdots \\ 2 \end{bmatrix}$ 20^4 high dimensional

$\phi(x)^T\phi(z) = K(x,z)$