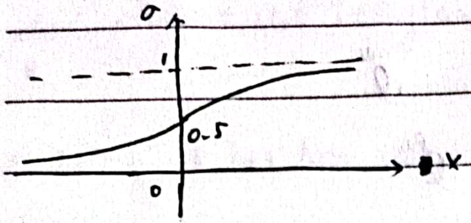
Date. 概率学习方向

Support Vector Machine



$\sigma(x) = \frac{1}{1+e^{-\theta^T x}}$

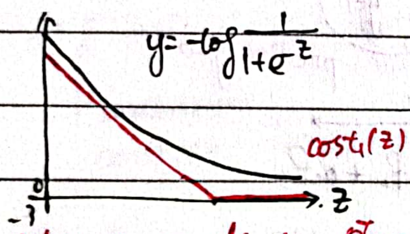
If $y=1$, we want $h_\theta(x) \approx 1$, $\theta^T x >> 0$

单独的(x,y)对Cost Function的增量:

$-y \log h_\theta(x) - (1-y)\log(1-h_\theta(x))$

$= -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y)\log\left(1-\frac{1}{1+e^{-\theta^T x}}\right)$

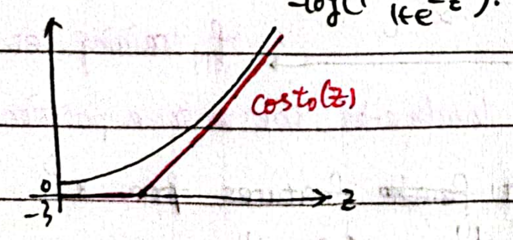
if $y=1$ (want $\theta^T x >> 0$):



When y equals 1, $\theta^T x >> 0$ will make the cost function smaller.

If $y=0$:

$-\log\left(1-\frac{1}{1+e^{-z}}\right)$



support vector machine

$$\min_\theta \underbrace{\sum_{i=1}^{m} y^{(i)} \mathrm{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \mathrm{cost}_0(\theta^T x^{(i)})}_{A} + \frac{\lambda}{2m} \underbrace{\sum_{j=0}^{n} \theta_j^2}_{B}$$

$C \doteq \frac{1}{\lambda}$

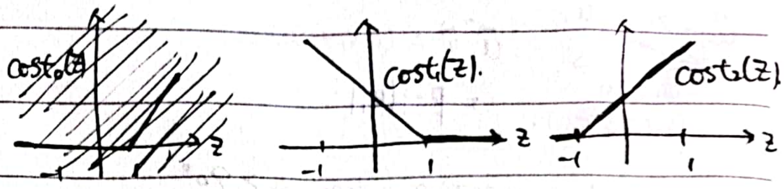
$A + \lambda B \longrightarrow C \cdot A + B$

⇒ Real SVM

$$\min_\theta C \sum_{i=1}^{m} \left[y^{(i)} \mathrm{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \mathrm{cost}_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{i=0}^{n} \theta_j^2$$

$\xRightarrow{\text{get } \theta \text{ that}}$ $h_\theta(x) = \begin{cases} 1, & \text{if } \theta^T x \geq 0 \\ 0, & \text{otherwise} \end{cases}$

The Understanding of Large Margin



If $y=1$, we want $\theta^T x \geq 1$.
If $y=0$, we want $\theta^T x \leq -1$.
} 安全间距

SVM Decision Boundary

$$\min_\theta C \cdot \sum_{i=1}^{m} \left[y^{(i)} \mathrm{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \mathrm{cost}_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{j=1}^{n} \theta_j^2$$

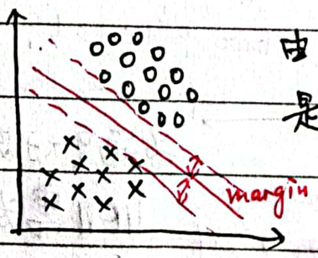
if C is very large, we expect that:

whenever $y^{(i)}=1$: $\theta^T x^{(i)} \geq 1$

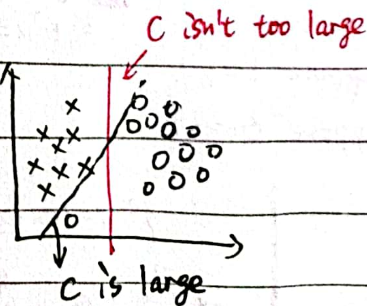
whenever $y^{(i)}=0$: $\theta^T x^{(i)} \leq -1$.

$\Rightarrow \min C \times 0 + \frac{1}{2}\sum_{j=1}^{n} \theta_j^2$

s.t. $\begin{cases} \theta^T x^{(i)} \geq 1, & \text{if } y^{(i)}=1 \\ \theta^T x^{(i)} \leq -1, & \text{if } y^{(i)}=0 \end{cases}$ LP问题 (线性规划)



由SVM产生的decision boundary是一个比较稳健，不擦也的世界.

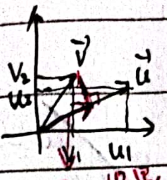


当C很大时，其对异常点十分敏感，会努力地将异常点纳入同类的范围

Date.

Math in SVM

Vector Inner Product



$\vec{u}=\begin{bmatrix}u_1\\u_2\end{bmatrix}$ $\vec{v}=\begin{bmatrix}v_1\\v_2\end{bmatrix}$

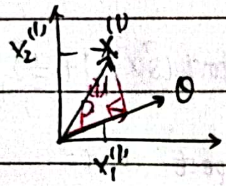
$\vec{u}^T\vec{v}=u_1v_1+u_2v_2$

$=p\cdot\|u\|$

p (投影).

p may be negative when $\theta>90°$.

SVM Decision Boundary

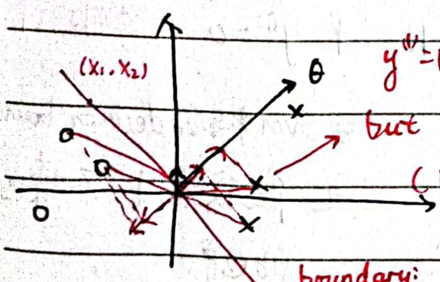


$\theta^T x^{(1)}=p^{(1)}\|\theta\|$

Thus，对于规划问题

$\min\limits_\theta \frac{1}{2}\sum_{j=1}^n \theta_j^2$

s.t. $\begin{cases} p^{(i)}\|\theta\|\geq 1, & \text{if } y^{(i)}=1 \\ p^{(i)}\|\theta\|\leq -1, & \text{if } y^{(i)}=0 \end{cases}$



$y^{(i)}=1$

but $p^{(i)}\|\theta\|\geq 1$?

($p^{(i)}$ is too small!)

为满足大于等于关系，

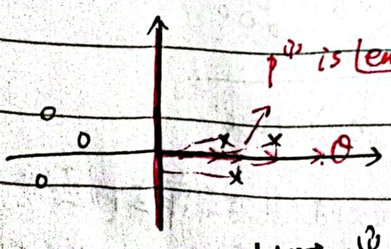
$\|\theta\|$需为较大值

boundary: $\theta^T x+b=0$.

$\Rightarrow \theta^T x=0$.

~~(Suppose b=0).~~

(Suppose $\theta_0=0$)



$p^{(i)}$ is enough large!

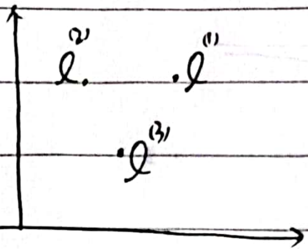
约束条件

可以被满足

此时 $p^{(i)}$ (绝对值) 较大.

$\|\theta\|$不用很大，也可满足约束条件 $|p^{(i)}\|\theta\||\geq 1$

Kernels (核函数).



Build Kernels:

$f_1=e^{-\frac{\|x-l^{(1)}\|^2}{2\sigma^2}}$

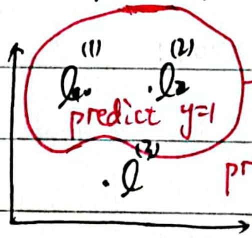
$f_2=e^{-\frac{\|x-l^{(2)}\|^2}{2\sigma^2}}$

$f_3=e^{-\frac{\|x-l^{(3)}\|^2}{2\sigma^2}}$ ~ $\frac{\sum_{i=1}^n (x_i-l_i^{(3)})^2}{2\sigma^2}$ 向量模

σ可调. σ↑时，f较缓；σ↓时，f较陡斜

$h_\theta(x)=\theta_0+\theta_1 f_1+\theta_2 f_2+\theta_3 f_3$.

eg:

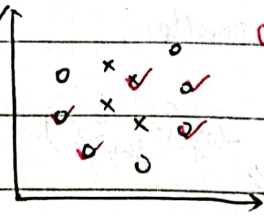


$h_\theta(x)=0$

predict y=0

$h_\theta(x)=-0.5+f_1+f_2+0\cdot f_3$

How to get Kernels?



① get landmarks: choose some dates of training set as landmarks [The same position]

② get ~~featu~~ features from $x^{(i)}$

$x^{(i)}\begin{cases} f_1^{(i)}=\text{sim}(x^{(i)},l^{(1)}) \\ f_2^{(i)}=\text{sim}(x^{(i)},l^{(2)}) \\ \vdots \\ f_m^{(i)}=\text{sim}(x^{(i)},l^{(m)}) \end{cases}$ $\Rightarrow$ 这些作为输入的特征

No.

Date.

③ get θ^T from training.

$$\min_\theta C \sum_{i=1}^{m} \left[y^{(i)} \text{cost}_1(\theta^T f^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T f^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

④ If θ^T has been trained, we predict dates like this:

Given x, compute feature $f \in \mathbb{R}^{m+1}$

Predict "$y=1$" if $\theta^T f \geq 0$.

i.e. $\theta_0 f_0 + \theta_1 f_1 + \cdots + \theta_n f_m \geq 0$

SVM Parameters:

C { large C: Lower bias, high variance — overfitting
Small C: Higher variance, lower bias — underfitting

σ^2 { large σ^2: Features f_i vary more Smoothly — underfitting
~~contrast~~
contrary: less smoothly — overfitting.

扫描全能王 创建