SVM Decision Boundary

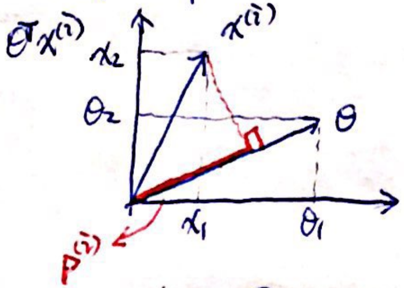
cost: $\min_\theta \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 = \frac{1}{2}\left(\sqrt{\theta_1^2+\theta_2^2}\right)^2 = \frac{1}{2}\|\theta\|^2$

$\sqrt{\theta_1^2+\theta_2^2} \to \|\theta\|$

s.t. $\theta^T x^{(i)} \ge 1$, if $y^{(i)}=1$

$\theta^T (x^{(i)}) \le -1$, if $y^{(i)}=0$

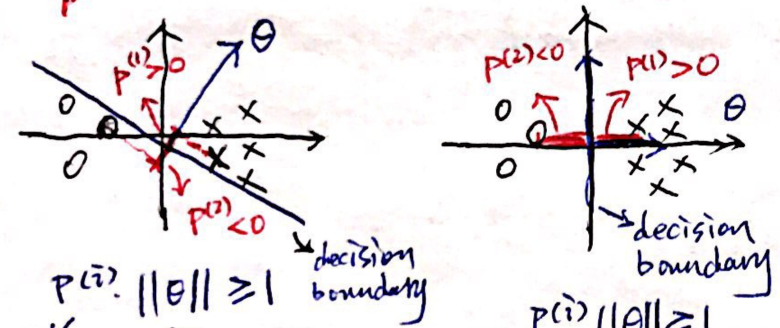
Simplify example: $\theta_0 = 0$, $n=2$



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

$p^{(i)} \cdot \|\theta\| \ge 1$, if $y^{(i)}=1$

$p^{(i)} \cdot \|\theta\| \le -1$, if $y^{(i)}=0$



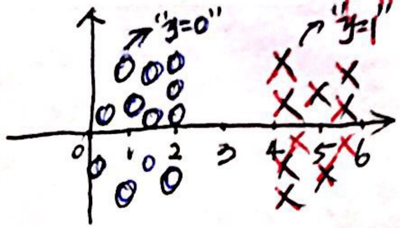
$p^{(i)} \cdot \|\theta\| \ge 1$
↙ small ↘ large.

small margin ⇒ bad decision boundary

$p^{(i)} \|\theta\| \ge 1$
↙ large ↘ small

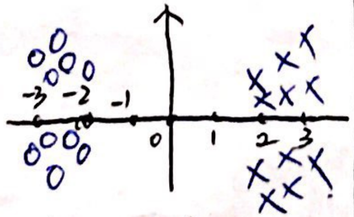
large margin ⇒ good

Question: suppose $\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2$. What values might the SVM given for θ?



$\theta_0 = -3$, $\theta_1 = 1$, $\theta_2 = 0$

Question: Given the training data, what is $\|\theta\|$?



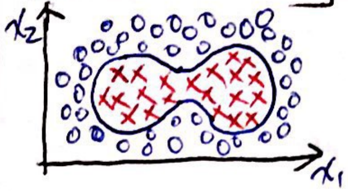
$P = +2 / -2$

$\|\theta\| = \frac{1}{2}$

NOV 1, 2018

Kernel Method

Two ways to achieve non-linear decision boundary:



(1) Polinomil: Predict "$y=1$" if

$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \ge 0$

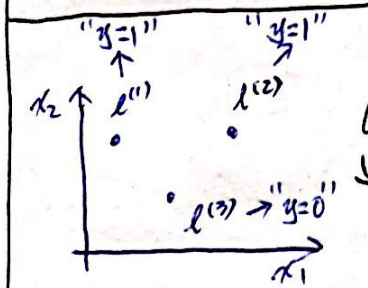
$h_\theta(x) = \begin{cases} 1 & \text{if } \theta_0 + \theta_1 x_1 + \cdots + \ge 0 \\ 0 & \text{if } \theta_0 + \theta_1 x_1 + \cdots + < 0 \end{cases}$

Let $f_1 = x_1$, $f_2 = x_2$, $f_3 = x_1 x_2$, $f_4 = x_1^2$, $f_5 = x_2^2$

then we check $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \cdots \ge$ or < 0

(2) kernels: f could be a function instead of polinomil

Kernel and Similarity



Given x, compute new features depending on proximity to landmarks $l^{(1)}, l^{(2)}, l^{(3)}$.

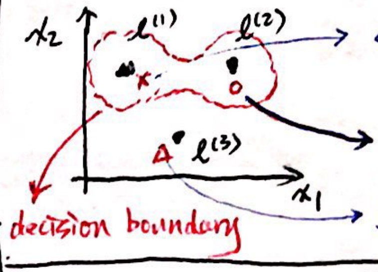
$f_1 = \text{Similarity}(x, l^{(1)}) = \cdots$

$f_2 = \text{Similarity}(x, l^{(2)}) = \exp\left(-\frac{\|x - l^{(2)}\|^2}{2\sigma^2}\right)$ → Gaussian Kernels.

$f_3 = \text{Similarity}(x, l^{(3)})$

if x is close to $l^{(i)}$, $f_i \to 1$

if x is far away from $l^{(i)}$, $f_i \to 0$ $\Leftarrow \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{j=1}^{n}(x_j - l_j^{(1)})^2}{2\sigma^2}\right)$



Suppose we learned $\theta_0 = -0.5$, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 0$

$f_1 \approx 1$, $f_2 \approx 0$, $f_3 \approx 0 \Rightarrow \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \ge 0$ Predict "$y=1$"

$f_1 \approx 0$, $f_2 \approx 1$, $f_3 \approx 0 \Rightarrow$ predict "$y=1$"

$f_1 \approx 0$, $f_2 \approx 0$, $f_3 \approx 1 \Rightarrow$ predict "$y=0$"

SVM with Kernels

Given $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})$

choose $l^{(1)} = x^{(1)}$, $l^{(2)} = x^{(2)}$, $\cdots$, $l^{(m)} = x^{(m)}$

Given example x:

$\left.\begin{aligned} f_1 &= \text{Similarity}(x, l^{(1)}) \\ f_2 &= \text{Similarity}(x, l^{(2)}) \\ &\vdots \end{aligned}\right\} m$

$f_0^{(i)} = 1$

We get new features

$\mathbb{R}^{n+1} \leftarrow x^{(i)} \rightarrow \begin{bmatrix} f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix}$ $f^{(i)} \in \mathbb{R}^{m+1}$

Hypothesis: Given x, comput f

predict "$y=1$" if $\theta^T f \ge 0$

Large C: Low bias; High Variance

Small C: High bias, Low variance

Large σ^2: H b, L V ($f \to$ smooth)

Small σ^2: L b, H V ($f \to$ not smooth)