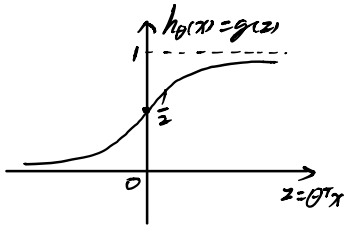
Large Margin classification

alternative view of logistic regression

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ if $y=1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$

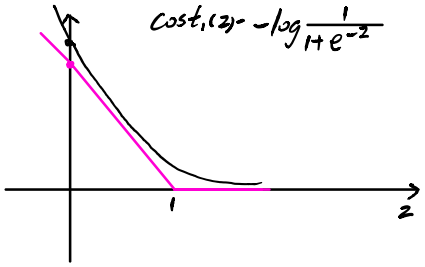
if $y=0$, we want $h_\theta(x) \approx 0$, $\theta^T x \ll 0$



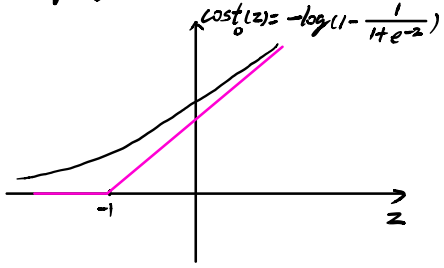
single example cost $= -(y \log h_\theta(x) + (1-y)\log(1-h_\theta(x)))$

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

if $y=1$, want $z = \theta^T x \gg 0$



if $y=0$, want $z = \theta^T x \ll 0$



we replace the black line with two magenta line:

$cost_0(z)$ $cost_1(z)$

logistic regression:

$$\min_\theta \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}(-\log h_\theta(x^{(i)})) + (1-y^{(i)})(-\log(1-h_\theta(x^{(i)})))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

support vector machine:

get rid of m, optima θ will remain unchange, replace $-\log_\theta h_\theta(x^{(i)})$ with $cost_1(\theta^T x^{(i)})$, $1 - \log(1-h_\theta(x^{(i)}))$ with $cost_0(\theta^T x^{(i)})$

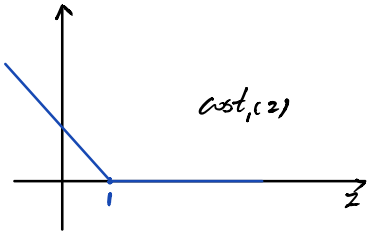
get rid of λ and add C, if $C = \frac{1}{\lambda}$, optima θ remain unchange.

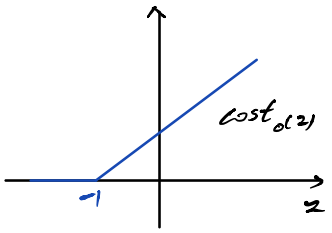
SVM hypothesis:

$$\min_\theta C\sum_{i=1}^{m}\left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1-y^{(i)}) cost_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

(this form just for convention)

Large Margin intuition:





$$\min_\theta C\sum_{i=1}^{m}\left[y^{(i)}\,cost_1(\theta^T x^{(i)}) + (1-y^{(i)})\,cost_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

if $y=1$ we want $\theta^T x \geq 1$ not just ≥ 0

if $y=0$ we want $\theta^T x \leq -1$ not just ≤ 0

if let C be large, $\sum_{i=1}^{m}\left[y^{(i)}\,cost_1(\theta^T x^{(i)}) + (1-y^{(i)})\,cost_0(\theta^T x^{(i)})\right] \approx 0$

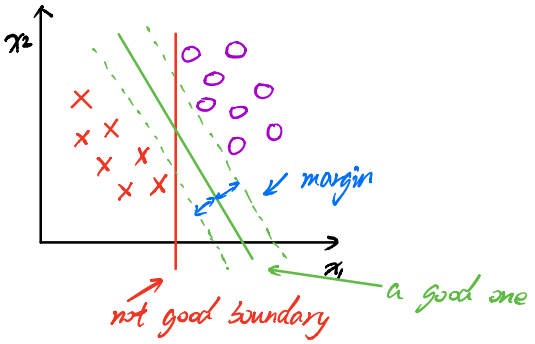
this will be $C\times 0 + \frac{1}{2}\sum_{i=1}^{n}\theta_i^2$:

$$\min_\theta \frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

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

SVM decision boundary example:

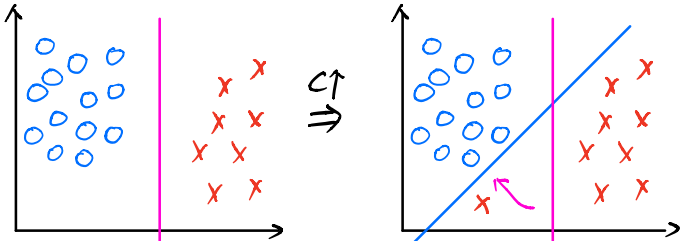
linear separable case:



in presence of outliers

C sensitive

$C\uparrow$ ≒ $\uparrow$ $\lambda\downarrow$ variance$\uparrow$

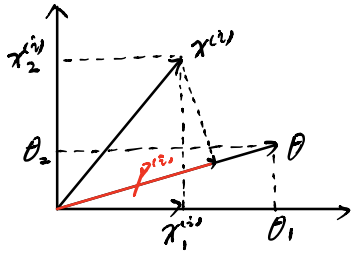


SVM:

$$\min_\theta \frac{1}{2}\sum_{i=1}^{n}\theta_i^2 \qquad \text{s.t. } \theta^T x^{(i)} \geq 1 \text{ if } y=1,\quad \theta^T x^{(i)} \leq -1 \text{ if } y=0$$

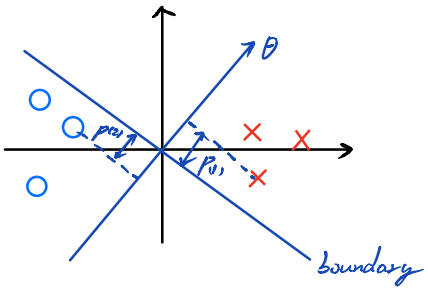
simplified: $n=2$, and $\theta_0=0$

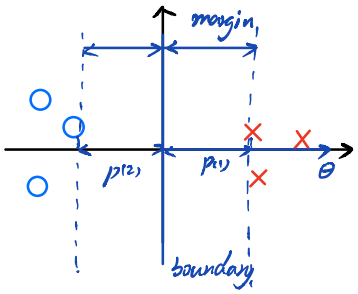
$$\min_\theta \frac{1}{2}(\theta_1^2+\theta_2^2) = \frac{1}{2}(\sqrt{\theta_1^2+\theta_2^2})^2 = \frac{1}{2}\|\theta\|^2$$



$$\theta^T x^{(i)} = p^{(i)}\|\theta\| = \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$$

$$\min_\theta \frac{1}{2}\|\theta\|^2 \qquad \text{s.t. } \begin{cases} p^{(i)}\|\theta\| \geq 1 & \text{if } y=1 \\ p^{(i)}\|\theta\| \leq -1 & \text{if } y=0 \\ \theta_0 = 0 \end{cases} \quad (C \text{ very large})$$





$p^{(1)}\|\theta\| \geq 1 \quad p^{(2)}\|\theta\| \leq -1$

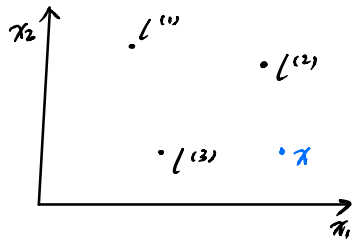
$p^{(1)} > 0$, $p^{(2)} < 0$ and $p^{(1)}, |p^{(2)}|$ quite small so $\|\theta\|$ need to be large, but $\min \frac{1}{2}\|\theta\|^2$ need $\|\theta\|$ to be small.

X

$p^{(1)}, |p^{(2)}|$ the biggest, and $\|\theta\|$ can be quite small

✓

Kernel I



given x, compute new feature depending on proximity to landmarks $l^{(1)}$ $l^{(2)}$ $l^{(3)}$

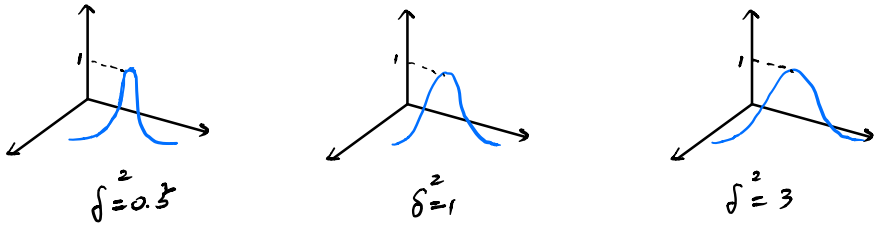
Gauss kernel:

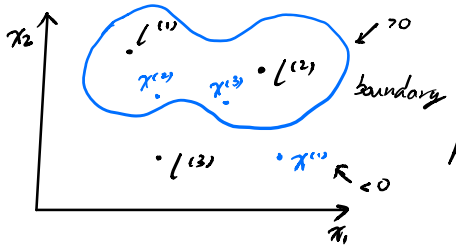
$$f_i = \text{similarity}(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|}{2\sigma^2}\right)$$

if $x \approx l^{(i)}$, $f_i \approx 1$

x far from $l^{(i)}$, $f_i \approx 0$

$\Rightarrow f_i \in (0, 1]$

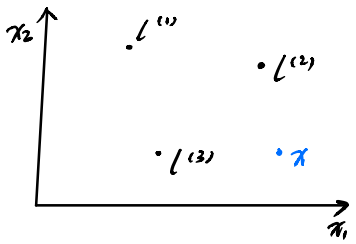




predict 1 when $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

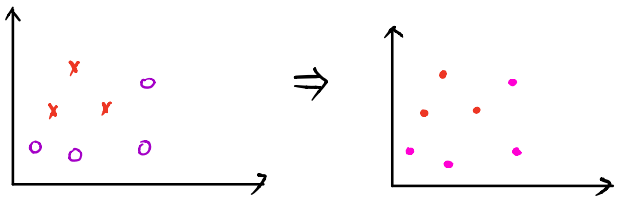
Kernel II:

how to choose landmark? where to get $l^{(1)}, l^{(2)}, l^{(3)} \ldots$?



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

choose $l^{(1)} = x^{(1)}$, $l^{(2)} = x^{(2)}$, $l^{(3)} = x^{(3)}$, $\cdots$, $l^{(m)} = x^{(m)}$



use $f_i = \text{similarity}(x, l^{(i)})$: ← add f_0

$$x^{(i)} \xrightarrow{f} \begin{bmatrix} f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix}$$

$x^{(i)} \in \mathbb{R}^{n+1}$

$f^{(i)} \in \mathbb{R}^{m+1}$

predict 1 if $\theta^T f \geq 0$

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

from θ_1

$C \gg 0$, SVM $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \theta_j^2$

$$\sum_{i=1}^{m} \theta_i^2 = \theta^T \theta = \|\theta\|^2$$ (no kernel / linear kernel)

$\theta^T M \theta$

SVM parameters:

$C \ (= \frac{1}{\lambda})$

$C \uparrow$ bias ↓ variance ↑

$C \downarrow$ bias ↑ variance ↓

σ^2

$\sigma^2 \uparrow$ f_i more smoothly, bias ↑ variance ↓

$\sigma^2 \downarrow$ f_i less smoothly, bias ↓ variance ↑

Use SVM:

- choose C, σ^2, kernal
- do perform feature scaling before using gauss kernal.

multi-class classification

one-vs-all method, train k SVM, distinguish $y=i$ from the others

$i=1,2,\cdots k$, get $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(k)}$, pick class i with largest $(\theta^{(i)})^T x$

LR (logistic regression) vs SVM

n feature $x \in \mathbb{R}^{n+1}$, m = training examples

if $n \gg m$: LR, SVM without kernal (linear kernal)

n is small, m is intermediate : SVM with Gaussian kernal

n is small, m is large : create more features, use LR or SVM without a kernal

neural network is likely to work well for most these setting but slower to train.