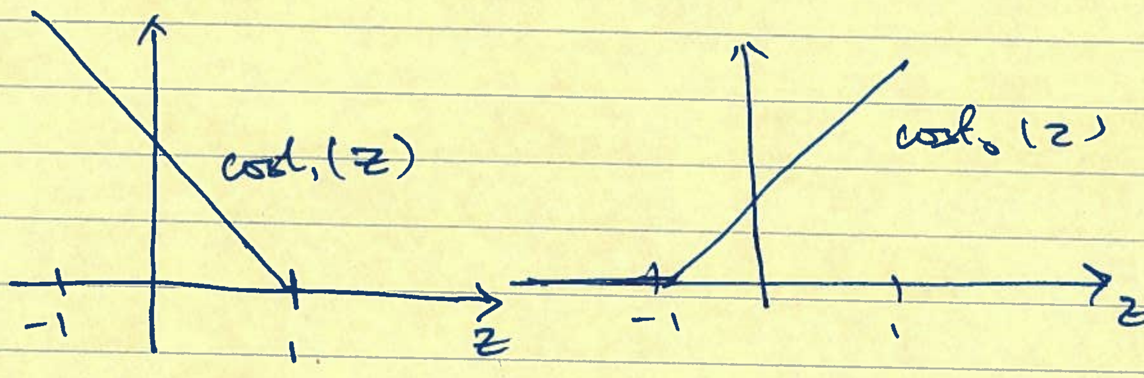
Large Margin Intuition:

SVM

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




If $y=1$, we want $\theta^T x \geq 1$ (not just $\geq 0$)
If $y=0$, we want $\theta^T x \leq -1$ (not just $< 0$)

wherever $y^{(i)} = 1$

$\theta^T x^{(i)} \geq 1$ $\qquad$ $\min C0 + \frac{1}{2} \sum_{1}^{n} \theta_j^2$

wherever $y^{(i)} = 0$:

$\theta^T x^{(i)} \leq -1$

## Kernels - I

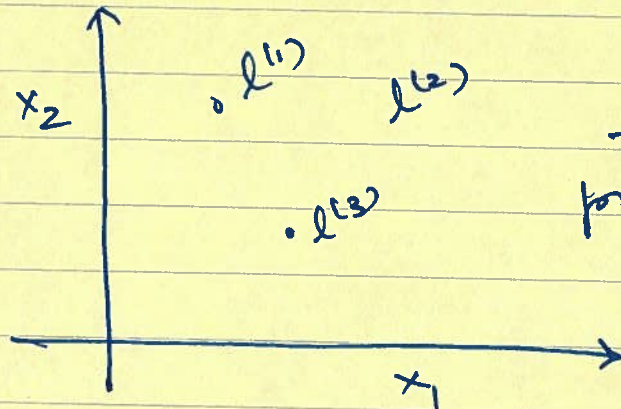
Non-linear Decision Boundary
Predict $y = 1$ if

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \dots \geq$$

$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta_0 + \theta_1 x_1 + \dots \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \dots$$

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




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

Given $x$:

$$f_1 = \text{similarity}(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

$K(x, l^{(i)})$

$$f_2 = \text{similarity}(x, l^{(11)})$$

Kernels (Gaussian kernels) $= \exp\left(-\frac{\|x - l^{(11)}\|^2}{2\sigma^2}\right)$

$$f_3 = \text{similarity}(x, l^{(3)}) = \exp\left(-\frac{\|x - l^{(3)}\|^2}{2\sigma^2}\right)$$

Kernels & similarity

$$f_1 = \text{similarity}(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{\sum_{j=1}^{n}(x_j - l_j^{(1)})^2}{2\sigma^2}\right)$$

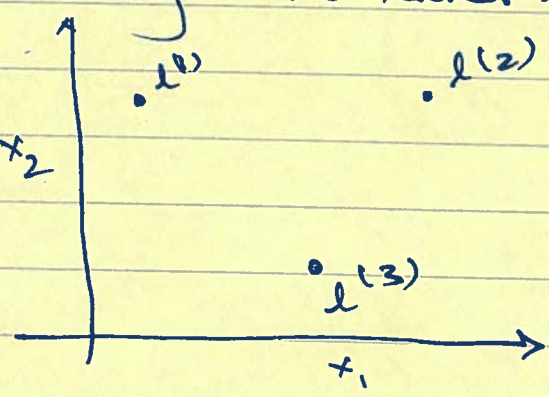
if $x \approx l^{(1)}$

$$f_1 \approx \exp\left(-\frac{0^2}{2\sigma^2}\right) \approx 1$$

if $x$ if far from $l^{(1)}$:

$$f_1 = \exp\left(-\frac{(\text{large number})^2}{2\sigma^2}\right) \approx 0$$

Choosing the landmark.




Given $x$:

$$f_i = \text{similarity}(x, l^{(i)})$$

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

Predict $y = 1$ if $\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 \geq 0$

Where to get $l^{(1)}, l^{(2)}, l^{(3)}, \ldots$ ?

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

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

Given example $x$:

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

$f_2 = \quad " \quad (x, l^{(2)})$

$f = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \quad f_0 = 1$

For training example $(x^{(i)}, y^{(i)})$

$$\boxed{\begin{aligned} f_1^{(i)} &= \text{sim}(x^{(i)}, l^{(1)}) \\ f_2^{(i)} &= \text{sim}(x^{(i)}, l^{(2)}) \\ &\vdots \\ f_m^{(i)} &= \text{sim}(x^{(i)}, l^{(m)}) \end{aligned}}$$

$\Leftarrow f_i^{(i)} = \text{sim}(x^{(i)}, l^{(i)}) = \exp\left(-\frac{0}{2\sigma^2}\right) =$

$f^{(i)} = \begin{bmatrix} f_0^{(i)} \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix}$

$f_0^{(i)} = 1$

Hypothesis

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

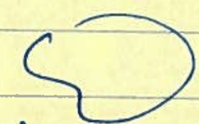
Predict "$y = 1$" if $\theta^T f \geq 0$

Training

$$\sum_j \theta_j = \theta^T\theta \leftarrow \|\theta\|^2 \qquad \theta = \begin{bmatrix}\theta_1 \\ \vdots \\ \theta_m\end{bmatrix} \text{ ignoring } \theta_0$$

$$\theta^T M \theta$$

inc $C$
dec $\sigma^2$

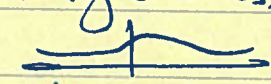


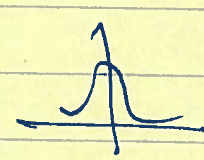
dec $C$ inc $\sigma^2$

$C = \frac{1}{\lambda}$ Large $C$ : Lower bias, high var
Small $C$ : Higher bias, low var.

large $C$ means small $\lambda$
small $C$ " large $\lambda$.

Large $\sigma^2$ : features $f_i$ vary more smoothly. High bias, low var.



Small $\sigma^2$, features $f_i$ vary less smoothly
lower bias, higher var, 

If
Overfit: Dec $C$, Inc $\sigma^2$

Need to specify parameter $c$:
Choice of $c$.

Eg No Kernel ("linear kernel")
Predict "$y=1$" if $\theta^T x \geq 0$.

Gaussian Kernel
$x \in \mathbb{R}^n$ $n$ small

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

where $l^{(i)} = x^{(i)}$

Need to choose $\sigma^2$.

$\|x - l\|^2$ $\quad v = x - l$

$$\|v\|^2 = v_1^2 + v_2^2 + \dots + v_n^2$$
$$= (x_1 - l_1)^2 + (x_2 - l_2)^2 + (x_n - l_n)^2.$$

Off shelf kernels.
Polynomial Kernels $k(x, l) = (x^T l)^2$
$(x^T l)^3$, $(x^T l + 1)^3$
$(x^T l + 5)^4$

String kernel, chi-squared kernel, histogram intersection kernel.

Multi class Classification

one vs all
(Train K SVMs)

Logistic Reg vs SVM

If n is large (relative m)

n = num of features
m = num of train ex
Eg $n \geq m$ $n = 10{,}000$

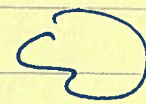
($m = 10 \ldots 1000$)
Use logistic reg or SVM without a kernel ("linear kernel")

If n is small, m is intermediate
Use SVM with Gaussian.
$n = 1-1000$, $m = 10 - 10{,}000$)
if n is small, m is large
($n = 1-1000$, $m = 50{,}000$ or $>$)
create/add more features
then use logistic reg
or SVM without a kernel.

Overfit : Dec c, inc $a^2$

inc c

dec $a^2$



dec c inc $a^2$.

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

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

Try using NN with large hid

Use SVM with Gauss Ker

create/add new poly fea

Use SVM with a lin ker, withou new

$x^{(i)} \in \mathbb{R}^2$, dec bound is st. line.

It is imp to perform feature

norm before using Gau ker

The max value of Gauss ker

sim$(x, l^{(i)})$ is 1