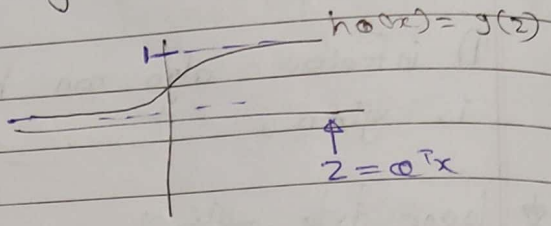
Week 7

Optimization Objective

Alternative view of logistic regn.

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



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

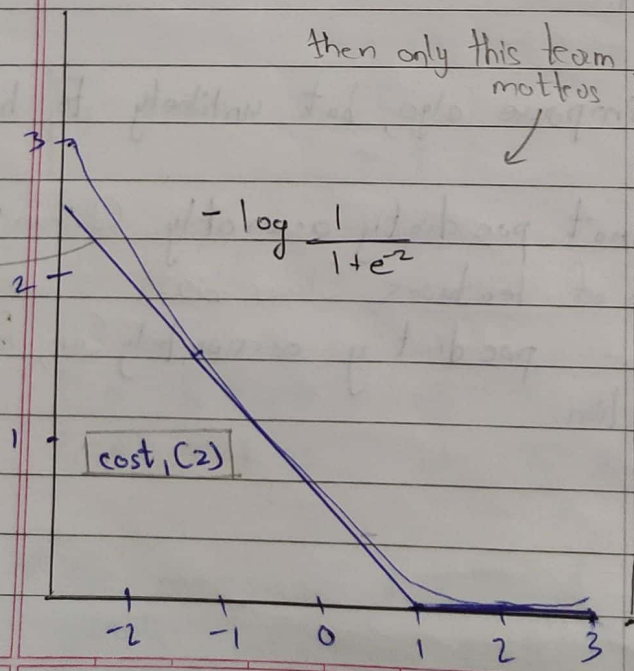
If $y=0$, we want $h_\theta(x) \approx 0$, $\theta^T x << 0$.

Cost of example : $-(y \log h_\theta(x) + (1-y) \log(1 - h_\theta(x)))$

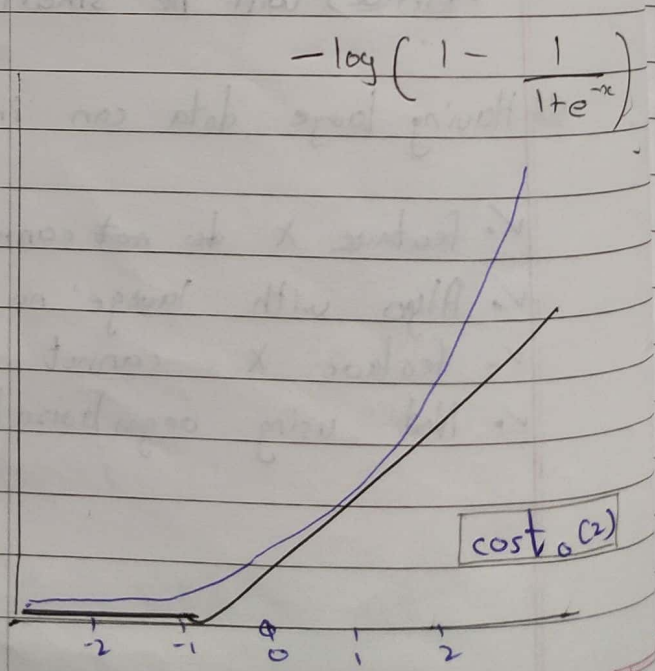
$$= -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y) \log\left(1 - \frac{1}{1+e^{-\theta^T x}}\right)$$

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

$z = \theta^T x$



If $y=0$ (want $\theta^T x << 0$)



★ logistic regn

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

★ Support vector machine.

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

Q) $\min_u \left((u-5)+1\right) \rightarrow u=5 \quad \leftarrow \quad \min_u 10(u-5)^2+10$

→ If constant is multiplicated, it doesn't matter while finding min of a variable.
↳ so we cross out $\frac{1}{m}$

↳ Also, we have crossed out λ & added C
A + λB ← (logistic) λ → controls trade off
CA + B ← (sum) C

think as $C = \frac{1}{\lambda}$

↳ The 2 eqn are not equal, but both will give same value (optimal θ value for the θ).

★ SVM hypothesis

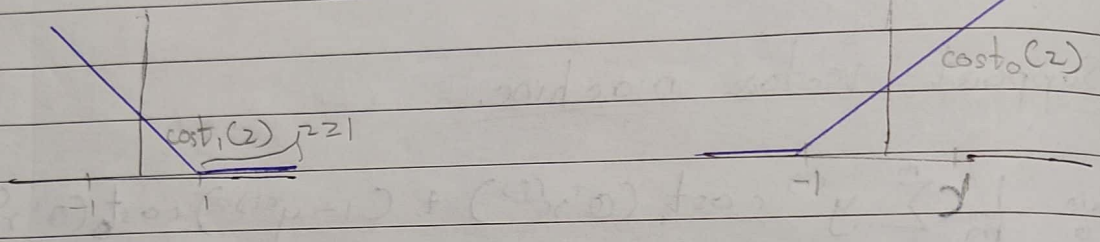
$$\min_{\theta} C\sum_{i=1}^{m}\left[y^{(i)}\,\text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\,\text{cost}_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{i=1}^{n}\theta_j^2$$

Hypothesis →

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

SVM

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



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)

What it take to make cost fun^n 0.

$\Theta^T x \geq 0$ 1 • For tve e.g. i.e. $y=1$, $\text{cost}_1(z)=0$ only when $z \geq 1$

$\Theta^T x \leq 0$ -1 • $y=0$, $\text{cost}_0(z)=0$, • $z \leq 1$

→ This builds extra safty factor in SVM.

☆ SVM decision boundary.

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

If we set C ↑↑, we have to make this term 0 (min)

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

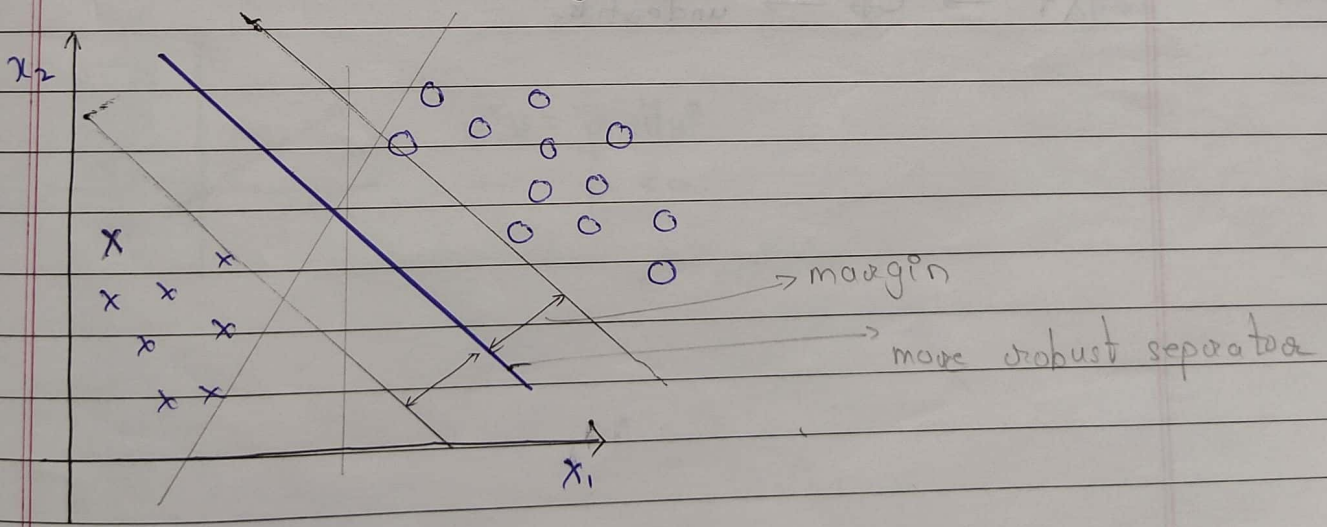
$$\min_{\theta} \; C \times 0 + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

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

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

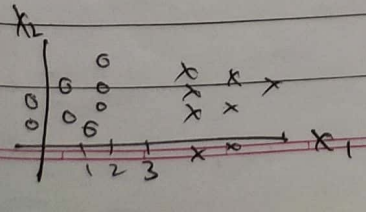
optimization

☆ SVM decision boundary: Linearly separable case.



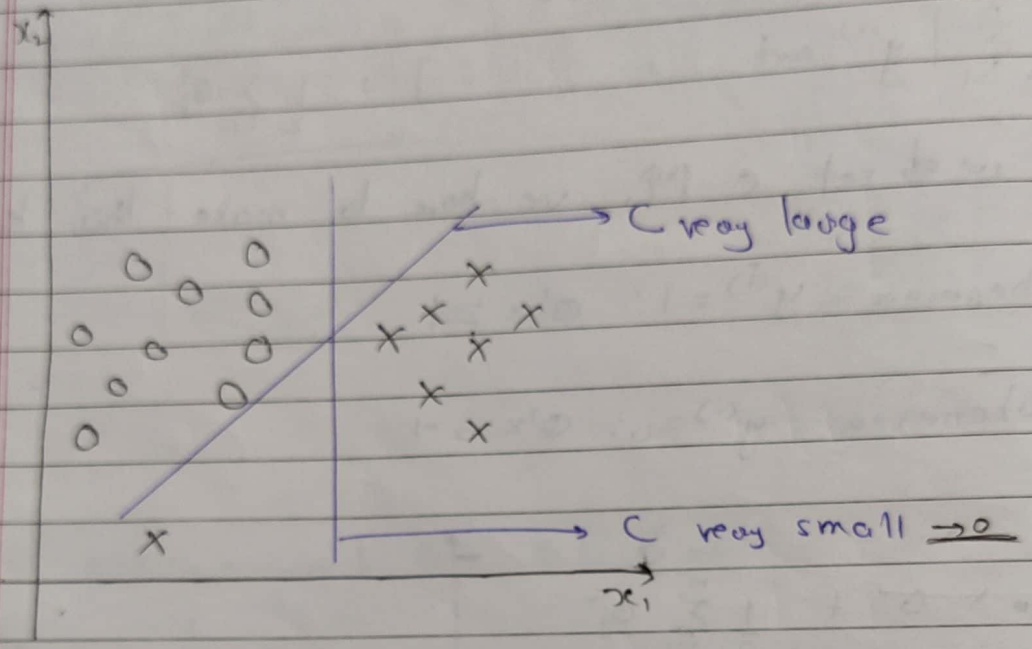
Large margin classifier:- Consequence of optimization problem

(good consequence)



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

* large margin classifier in presence of outliers.



C as if $\frac{1}{\lambda}$:

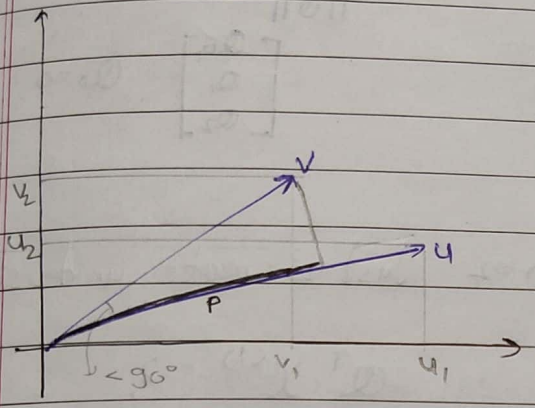
$\lambda \downarrow \rightarrow C \uparrow \longrightarrow$ overfits.
$\lambda \uparrow \rightarrow C \downarrow \longrightarrow$ underfits.

Here we are learning, how doing optimization we got margin classifier.

Page No: 108
Date: 17 / 8 / 19

@ Mathematics behind large Margin Classification (optional)

★ Vector inner product.



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

$$u^T v = v^T u = [u_1, u_2] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = [v_1, v_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$\|u\|$ = length of vector u.

$= \sqrt{u_1^2 + u_2^2} \in \mathbb{R}$

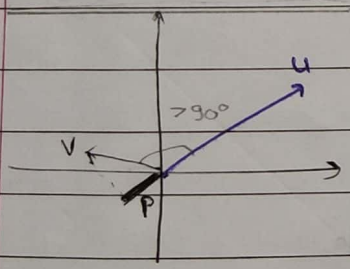
p = length of projection of v onto u.

signed

$u^T v = p \|u\| = v^T u = u_1 v_1 + u_2 v_2$

$p \in \mathbb{R}$

$p > 0$.



$u^T v = p \|u\|$

$p < 0$

✯ SVM Decision Boundary.

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

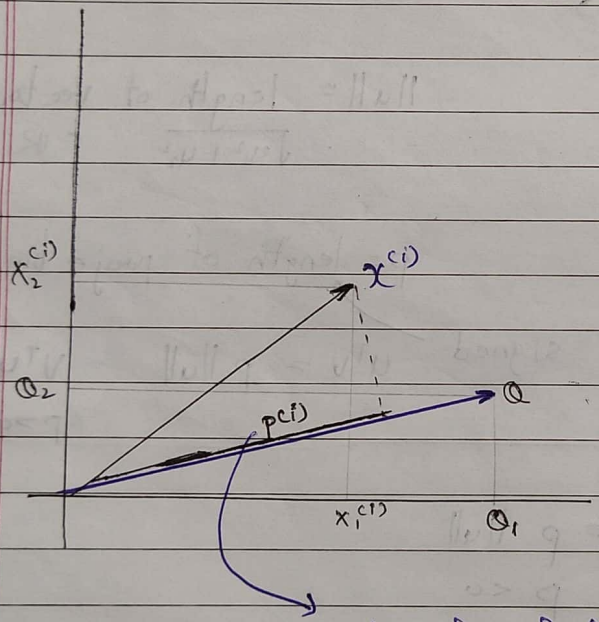
$\|\theta\|$

$$\begin{bmatrix}\theta_0\\ \theta_1\\ \theta_2\end{bmatrix} \quad \theta_0 = 0$$

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

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

simplification: • $\theta_0 = 0$; $n = 2$ Just to make understand better.



$\theta^T x^{(i)} = 2$.

$\uparrow \quad \uparrow$

$u^T \quad v$

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

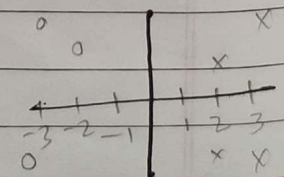
$= \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$

value of p of i^{th} example.

Q1 $\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$, s.t $\|\theta\| p^{(i)} \geq 1$ if $y^{(i)} = 1$

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

At optimal value of θ, value of $\|\theta\| = ?$



$\frac{1}{2} \times 2 = 1$

$\frac{1}{2}$

Ans

SVM Decision Boundary

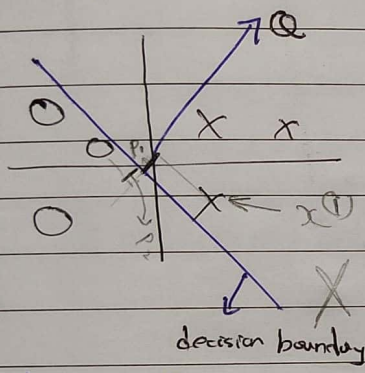
$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 = \frac{1}{2}\|\theta\|$$

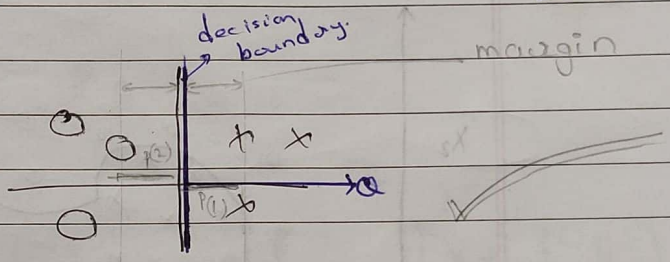
s.t. $p^{(i)} \cdot \|\theta\| \geq 1$ if $y^{(i)}=1$
$p^{(i)} \cdot \|\theta\| \leq -1$ if $y^{(i)}=-1$
} C very large

where $p^{(i)}$ is the projection of $x^{(i)}$ onto the vector θ.

simplification : $\theta_0 = 0$

19/8/19





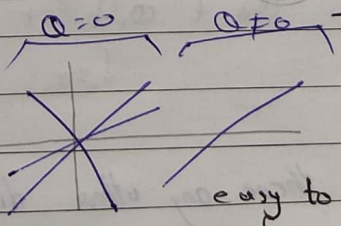
$p^{(1)}$, $p^{(2)}$ are bigger, so $\|\theta\|$ can be small now.
$p^{(i)} \|\theta\| \geq 1$

- let say this is decision boundary.
- Not a good decision boundary because examples are ~~not~~ very close to it.
- SVM will not choose such DB.

SVM Hypothesis

Magnitude of margin is $p^{(1)}$, $p^{(2)}$, ..
So, in SVM we want large $p^{(i)}$
∴ large margin

$p_0^{(1)} > 0$, $p_0^{(2)} < 0$



for both cases optimization holds true, it was easy to prove for $\theta_0 = 0$.

For optimization objective we want $p^{(i)}\|\theta\| \geq 1$
$p^{(i)}$ very small
so $\|\theta\|$ must be very large

$p^{(2)}\|\theta\| \leq -1$
small must be large → invalid

But in optimization objective, we have to make θ small.
To min we are doing optimization objective

Sadguru

Page No: 111

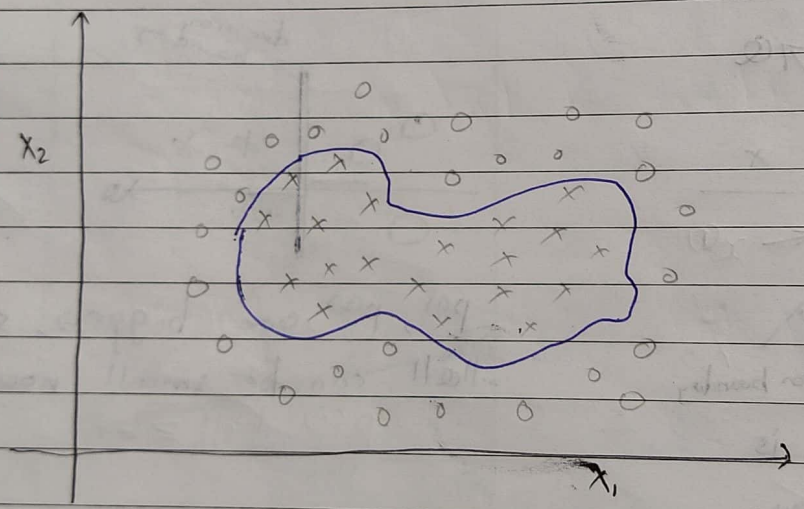
Date: / /

⑥ ☆ Kernels I

☆ Hence, We are starting at adapting SVM, in order to develop complex non-linear classifier.

→ For this we use technique called kernels.

☆ Non-linear Decision Boundary.



Predict $y=1$ if $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 \geq$

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

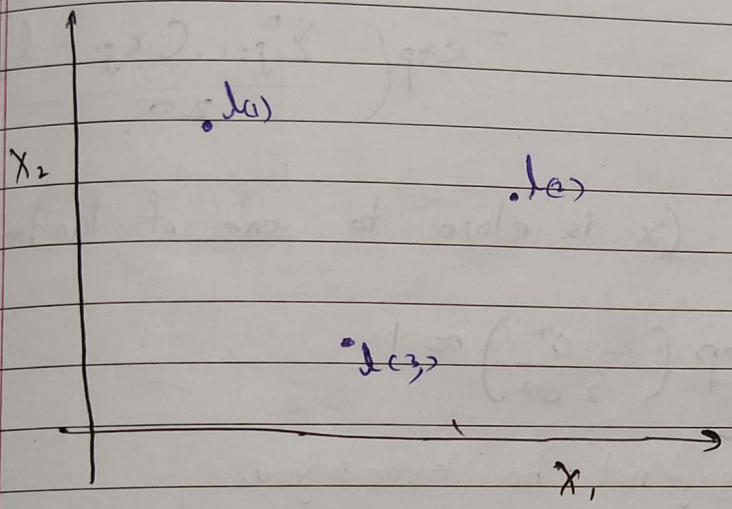
$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \dots$$

$f_1 = x_1 \quad f_2 = x_2 \quad f_3 = x_1 x_2 \quad f_4 = x_1^2 \quad f_5 = x_2^2$

Is there any other different / better choice of features f_1, f_2, f_3 because, considering higher degree term is expensive.

So, now we are going to find replacement of f_1, f_2, f_3 ...

Here we are going to define only 3 new features, unlike real world problem. lineup 1, lineup 2, lineup 3.



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

We are defining new features as →

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

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

$$f_2 = \text{similarity}(x, l^{(2)}) = \exp\left(-\frac{\|x - l^{(2)}\|^2}{2\sigma^2}\right)$$

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

Gaussian Kernel → $k(x, l^{(i)})$

What kernals actually do

PTO

★ Kernels & similarity

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

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

If $x \approx l^{(1)}$: (x is close to one of landmark)

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

↑ distance may not be exactly 0.

If x is far far $l^{(1)}$:

$$f_1 = \exp\left(-\frac{(\text{large no})^2}{2\sigma^2}\right) \approx 0$$

This features measures how similar x is from one of landmark

f = 0 , x far from landmark
f = 1 , x close to landmark

each land mark defines a new feature →

$l^{(1)} \rightarrow f_1$
$l^{(2)} \rightarrow f_2$
$l^{(3)} \rightarrow f_3$
↑
x

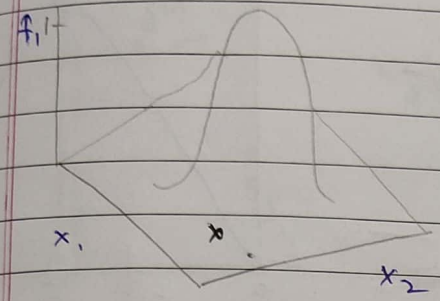
Given one example x, we have to compute 3 features f_1, f_2, f_3 by similarity b/w $l^{(i)}$ & $x^{(j)}$

first landmark

eg. $l^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$, $f_1 = \exp\left(\frac{-\|x - l^{(1)}\|^2}{2\sigma^2}\right)$

$\sigma^2 = 1$

Plotting these features.

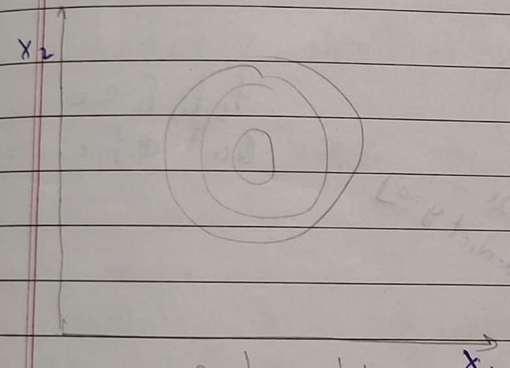


when $x = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$ exactely,

the $f_1 = 1$ (max)

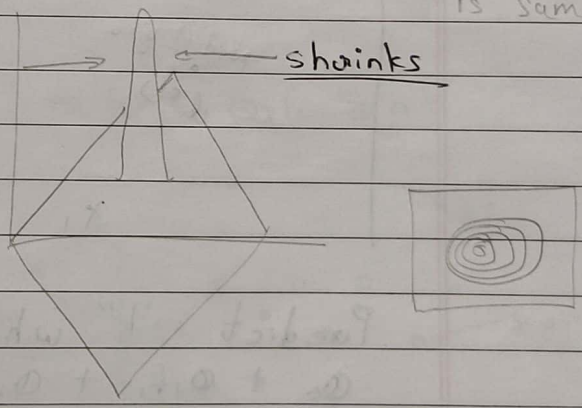
When x moves farther from $\begin{bmatrix} 3 \\ 5 \end{bmatrix}$ f_1 moves closer to 0.

So, f_1 is measure b/w



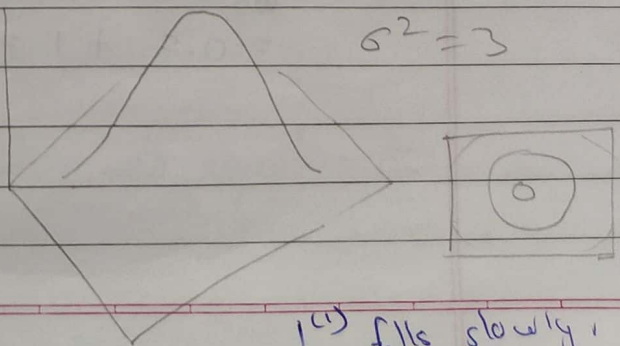
contour plot.

If $\sigma^2 = 0.5$, everything other is same.



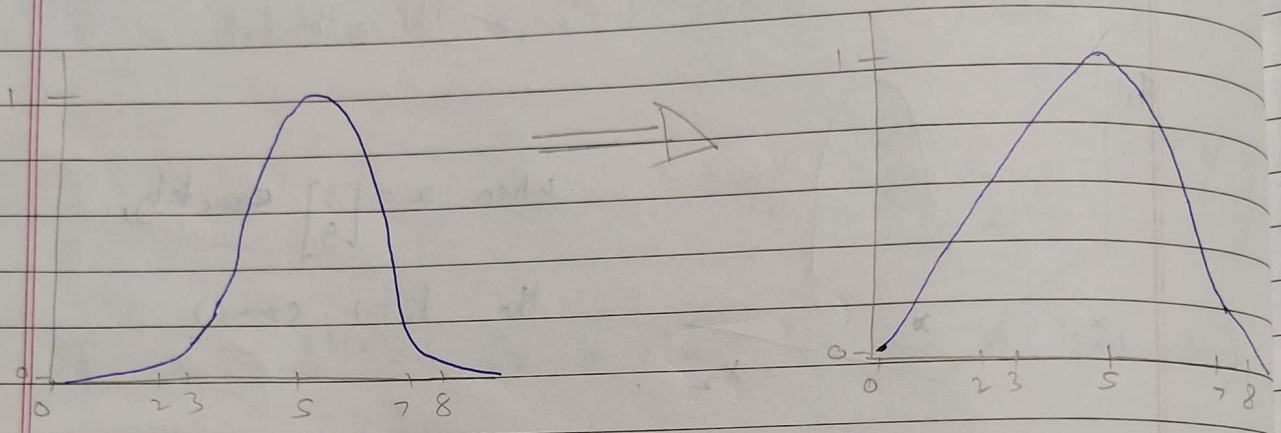
Moving from $x = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$, f_1 falls close to 0, more rapidly.

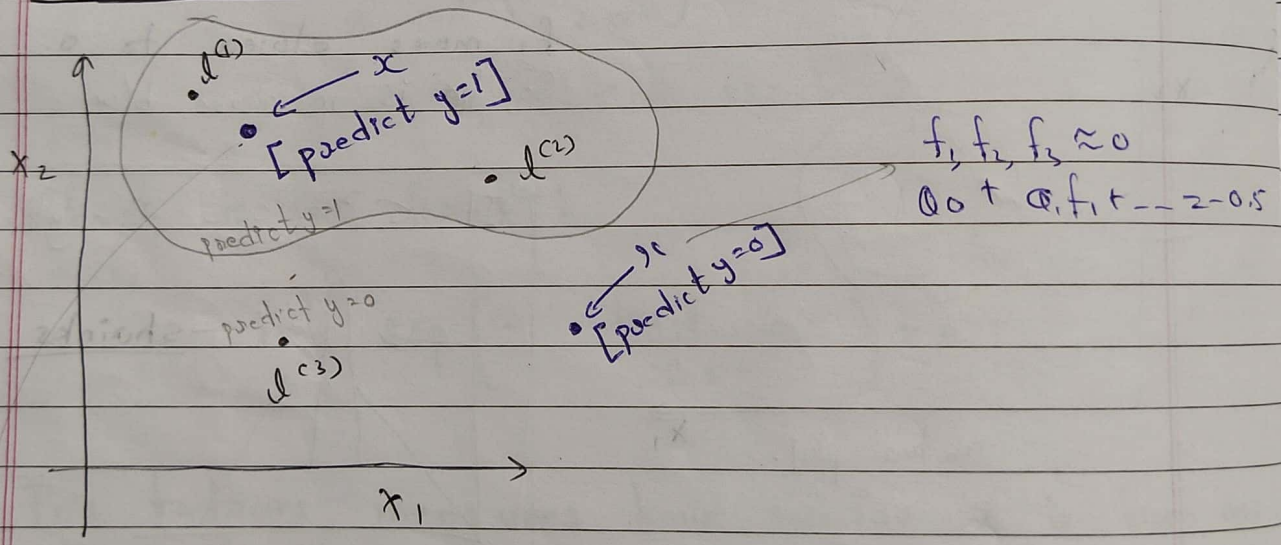
$\sigma^2 = 3$



$l^{(1)}$ falls slowly.

Q) Consider a 1-D example with one feature x_1. Suppose $l^{(1)} = 5$. To the right is plot of $f_1 = \exp\left(\frac{-|x - l^{(1)}|}{2\sigma^2}\right)$ when $\sigma^2 = 1$. Suppose we ~~$\sigma^2 = 1$~~ now change $\sigma^2 = 4$.





Predict "1" when

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

$\theta_0 = -0.5, \quad \theta_1 = 1, \theta_2 = 1, \theta_3 = 0$

$\theta_0 + \theta_1 \times 1 + \theta_2 \times 0 + \theta_3 \times 0$
$= -0.5 + 1 = 0.5 \geq 0$

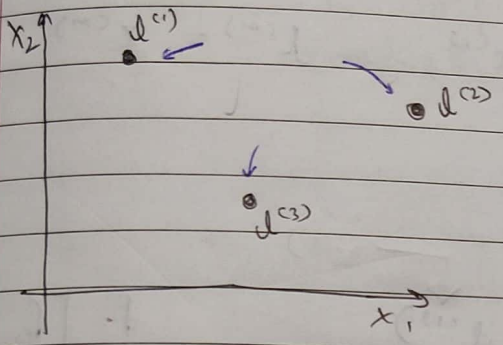
$f_1 \approx 1, f_2 \approx 0, f_3 \approx 0$

closer to $l^{(1)}$, far from $l^{(2)}, l^{(3)}$

⑫ Kernels II

How to choose landmarks $l^{(1)}, l^{(2)}, l^{(3)}, \ldots$

☆ Choosing the landmarks



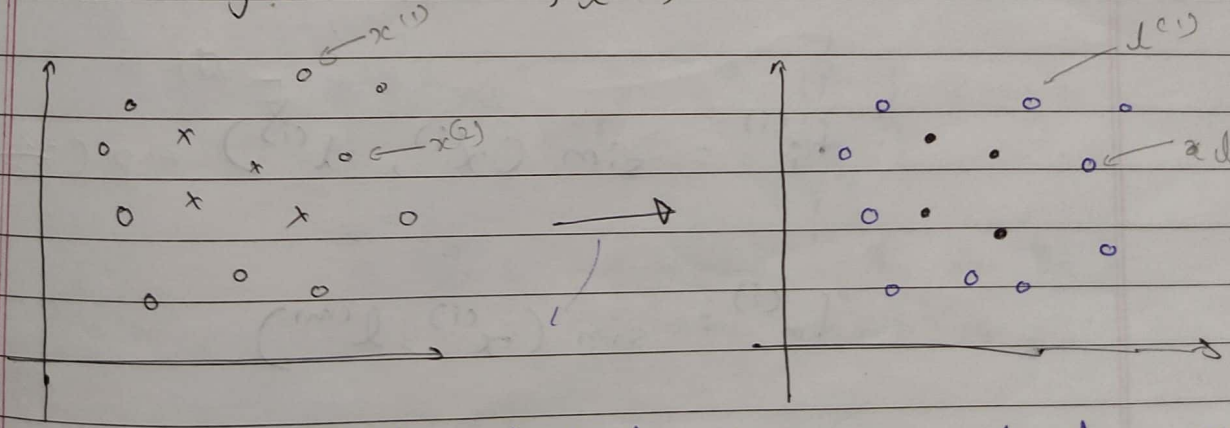
Given x:

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

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

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

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



We are going to put landmarks as same location as training examples, all.

Color of dots on right figure is not significant

✱ SVM with Kernels.

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

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

Given example x:

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

$f_2 = \text{similarity}(x, l^{(2)})$

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

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

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

$x^{(i)} \longrightarrow$ $f_2^{(i)} = \text{sim}(x^{(i)}, l^{(2)})$

$f_i^{(i)} = \text{sim}(x^{(i)}, l^{(i)}) = \exp\left(\frac{-0}{2\sigma^2}\right) = 1$ ($l^{(i)} = x^{(i)}$)

$f_m^{(i)} = \text{sim}(x^{(i)}, l^{(m)})$

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

$$f^{(i)} = \begin{bmatrix} f_0^{(i)} \\ f_1^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix}, \quad f_0^{(i)} = 1$$

This will represent my i^{th} training example with new features $(f_0, f_1, \ldots f_m)$

★ SVM with kernels

- **Hypothesis :**

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

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

$\theta_0 f_0 + \theta_1 f_1 + \dots \theta_m f_m$

→ This is how we do prediction using θ.

How to calculate θ.

- **Training :** (Here we are using new features)

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

here $n=m$ and $(\theta_1, \theta_2, \dots)$

$-\sum_j \theta_j^2 = \theta^T \theta \rightarrow \|\theta\|^2$ $\quad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}$ ignoring θ_0

$\theta^T M \theta$ ... sometimes written like this to optimize computations.

★ SVM parameters

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

prone to overfitting large C: lower bias, high variance. (small λ)
prone to underfitting Small C: Higher bias, low variance. (large λ)

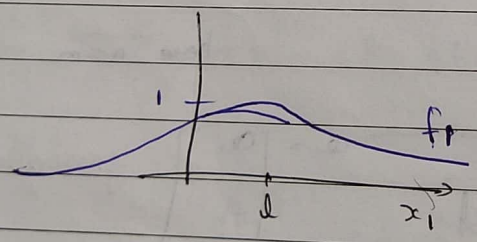
σ^2

appears in gaussian kernel.

- large σ^2:

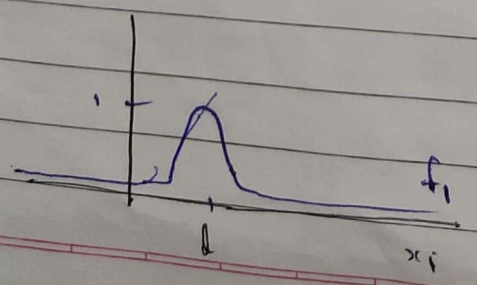
Features f_i vary more smoothly.
Higher bias, lower variance.

$$\exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$$ -- If we have only 1 feature x and landmark l we only



- Small σ^2:

Features f_i vary less smoothly.
Lower bias, higher variance.



Q If SVM overfits what should we do.

✓. Decrease C

✓. Increase σ^2