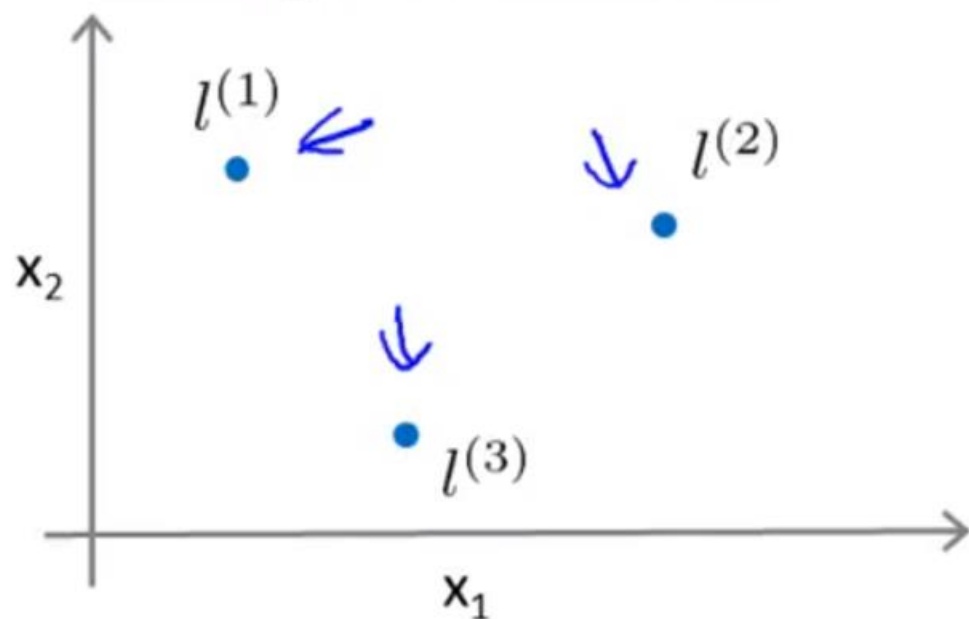


# Kernels-2

*Kernels*

Support Vector Machines

## Choosing the landmarks



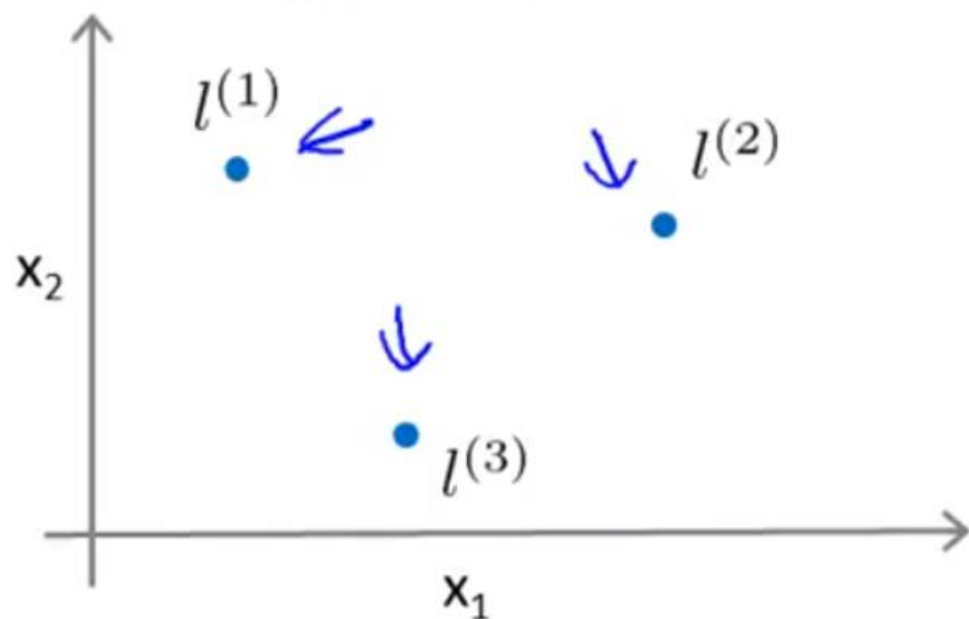
Given  $x$ :

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

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

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

## Choosing the landmarks



Given  $x$ :

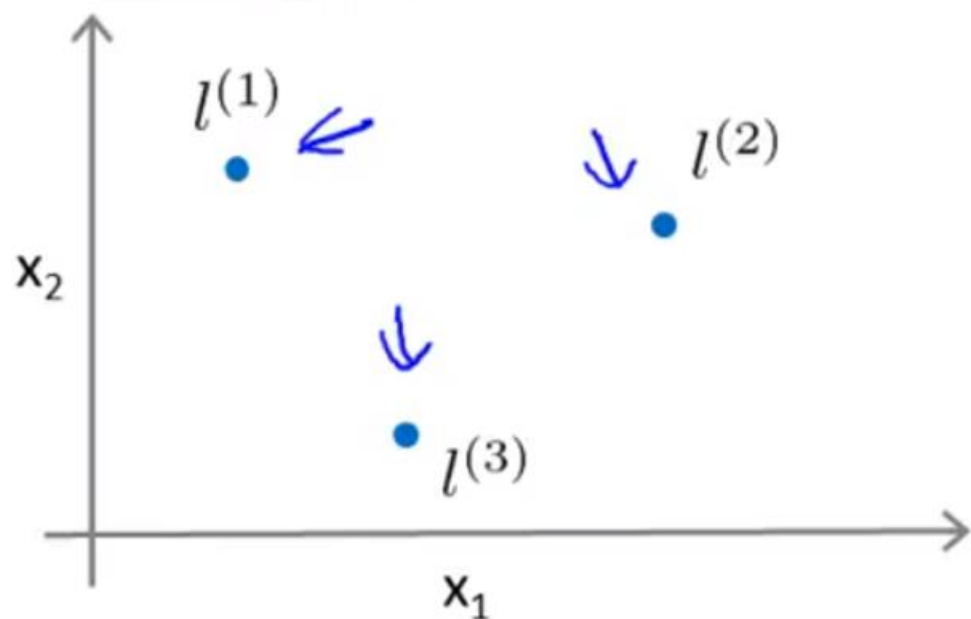
$$\begin{aligned} \rightarrow f_i &= \text{similarity}(x, l^{(i)}) \\ &= \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right) \end{aligned}$$

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

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



## Choosing the landmarks

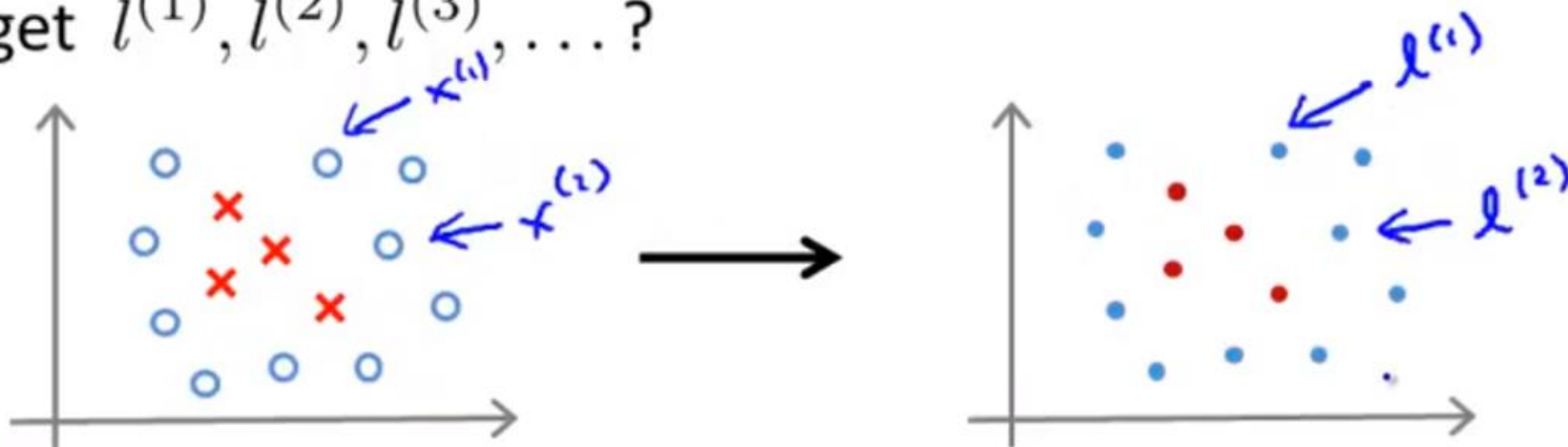


Given  $x$ :

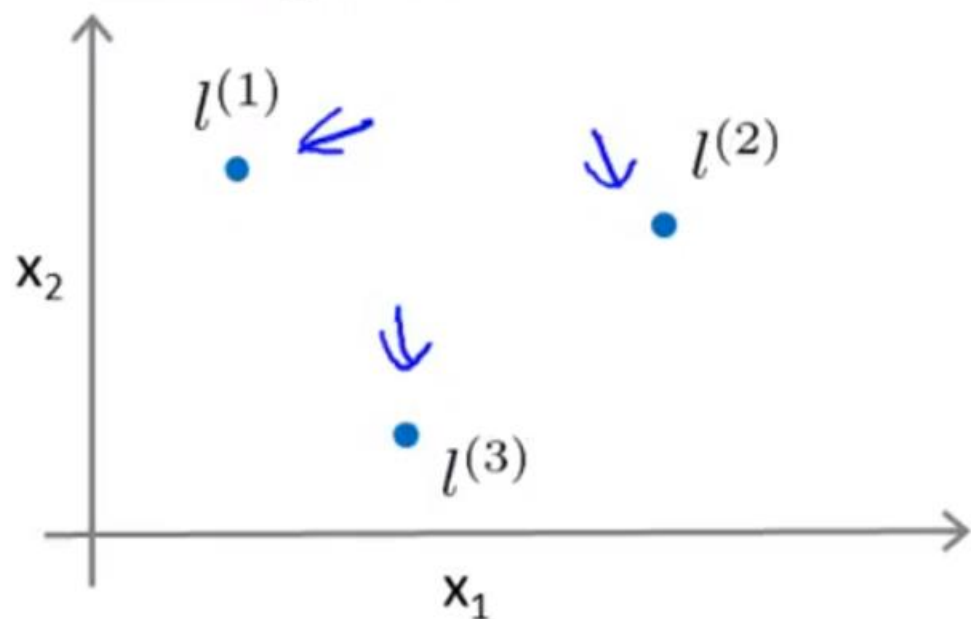
$$\begin{aligned} \rightarrow f_i &= \text{similarity}(x, l^{(i)}) \\ &= \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right) \end{aligned}$$

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

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



## Choosing the landmarks



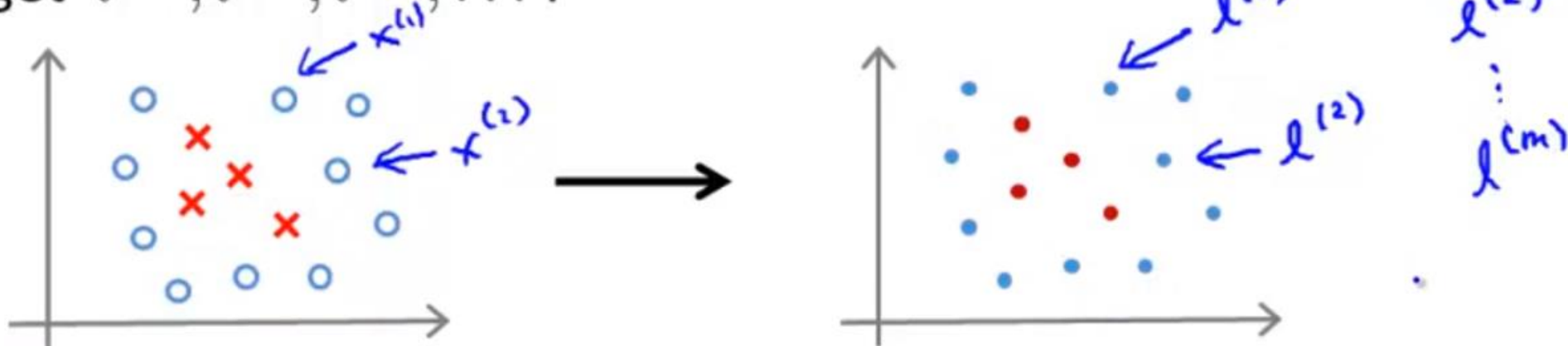
Given  $x$ :

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

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

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

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



## SVM with Kernels

- Given  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,  
choose  $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$ .

## SVM with Kernels

- Given  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,
- choose  $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$ .

Given example  $x$ :

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

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

...

## SVM with Kernels

- Given  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,
- choose  $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$ .

Given example  $\underline{x}$ :

- $f_1 = \text{similarity}(x, l^{(1)})$
  - $f_2 = \text{similarity}(x, l^{(2)})$
  - $\dots$
- $\swarrow x^{(i)}$

The vector  $f$  calculates the proximity of the point  $x$  with  $m$  different landmarks:

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



## SVM with Kernels

- Given  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,
- choose  $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$ .

Given example  $\underline{x}$ :

$$\begin{aligned} \rightarrow f_1 &= \text{similarity}(x, l^{(1)}) \\ \rightarrow f_2 &= \text{similarity}(x, l^{(2)}) \\ &\vdots \end{aligned}$$

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

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

$$\begin{aligned} \underline{x}^{(i)} \rightarrow & \begin{aligned} f_1^{(i)} &= \text{sim}(x^{(i)}, l^{(1)}) \\ f_2^{(i)} &= \text{sim}(x^{(i)}, l^{(2)}) \\ &\vdots \\ f_i^{(i)} &= \text{sim}(x^{(i)}, l^{(i)}) = \exp(-\frac{0}{2\sigma^2}) = 1 \\ &\vdots \\ f_m^{(i)} &= \text{sim}(x^{(i)}, l^{(m)}) \end{aligned} \end{aligned}$$

## SVM with Kernels

- Given  $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ ,
- choose  $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$ .

Given example  $\underline{x}$ :

$$\begin{aligned} \rightarrow f_1 &= \text{similarity}(x, l^{(1)}) \\ \rightarrow f_2 &= \text{similarity}(x, l^{(2)}) \\ &\vdots \end{aligned}$$

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

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

$$\underline{x^{(i)}} \rightarrow \begin{bmatrix} f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix} \quad \begin{aligned} f_1^{(i)} &= \text{sim}(x^{(i)}, l^{(1)}) \\ f_2^{(i)} &= \text{sim}(x^{(i)}, l^{(2)}) \\ &\vdots \\ f_i^{(i)} &= \text{sim}(x^{(i)}, l^{(i)}) = \exp\left(-\frac{0}{2\sigma^2}\right) = 1 \\ &\vdots \\ f_m^{(i)} &= \text{sim}(x^{(i)}, l^{(m)}) \end{aligned}$$

$$\underline{x^{(i)}} \in \mathbb{R}^{n+1} \quad (\text{or } \mathbb{R}^n)$$
$$f^{(i)} = \begin{bmatrix} f_0^{(i)} \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix} \quad f_0^{(i)} = 1$$

## SVM with Kernels

Hypothesis: Given  $x$ , compute features  $f \in \mathbb{R}^{m+1}$   
Predict “y=1” if  $\theta^T f \geq 0$

## SVM with Kernels

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

$\rightarrow$  Predict "y=1" if  $\underline{\theta}^T \underline{f} \geq 0$   $\theta_0 f_0 + \theta_1 f_1 + \dots + \theta_m f_m$

## SVM with Kernels

Hypothesis: Given  $\underline{x}$ , compute features  $\underline{f} \in \mathbb{R}^{m+1}$   $\Theta \in \mathbb{R}^{n+1}$   
→ Predict “y=1” if  $\underline{\theta}^T \underline{f} \geq 0$   
 $\theta_0 f_0 + \theta_1 f_1 + \dots + \theta_m f_m$

Training:

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

## SVM with Kernels

Hypothesis: Given  $\underline{x}$ , compute features  $\underline{f} \in \mathbb{R}^{m+1}$   $\Theta \in \mathbb{R}^{n+1}$

→ Predict "y=1" if  $\theta^T f \geq 0$   $\theta_0 f_0 + \theta_1 f_1 + \dots + \theta_m f_m$

Training:

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

~~$\theta^T x^{(i)}$~~   $\theta^T f^{(i)}$

WE used to make predictions by looking at an  $x$ .

Now, we calculate  $f^i$ 's for each  $x$  and then predict using the  $f^i$ 's.



## SVM with Kernels

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

→ Predict "y=1" if  $\underline{\theta}^T \underline{f} \geq 0$

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

$$\underline{\theta} \in \mathbb{R}^{n+1}$$

Training:

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

Annotations:

- $\underline{\theta}^T \underline{f}^{(i)}$  (with  $\underline{\theta}^T \underline{x}^{(i)}$  crossed out)
- $\underline{\theta}^T \underline{f}^{(i)}$  (with  $\underline{\theta}^T \underline{x}^{(i)}$  crossed out)
- Green box around  $\frac{1}{2} \sum_{j=1}^m \theta_j^2$  with  $n=m$  and  $\rightarrow \theta_0$

## SVM with Kernels

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

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

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

$$\theta \in \mathbb{R}^{n+1}$$

Training:

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

Annotations:   
 -  $\theta^T f^{(i)}$  is underlined in red in the cost functions.   
 - A green box highlights the regularization term  $\frac{1}{2} \sum_{j=1}^m \theta_j^2$ .   
 - A blue circle with a red 'X' and "= m" is above the sum, with an arrow pointing to  $\theta_0$ .   
 - A blue circle with a red 'X' and "(i)" is above the  $\theta^T f^{(i)}$  term, with an arrow pointing to  $\theta$ .   
 -  $n = m$  is written above the green box.

$$- \sum_j \theta_j^2 = \theta^T \theta \quad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_m \end{bmatrix} \quad (\text{ignore } \theta_0)$$

Annotations:   
 -  $\theta^T M \theta$  is underlined in blue.   
 -  $\|\theta\|^2$  is written in blue with an arrow pointing to  $\theta^T \theta$ .



## SVM parameters:

$C \left( = \frac{1}{\lambda} \right)$ .  $\rightarrow$  Large C: Lower bias, high variance.  
 $\rightarrow$  Small C: Higher bias, low variance.

(small  $\lambda$ )

(large  $\lambda$ )

## SVM parameters:

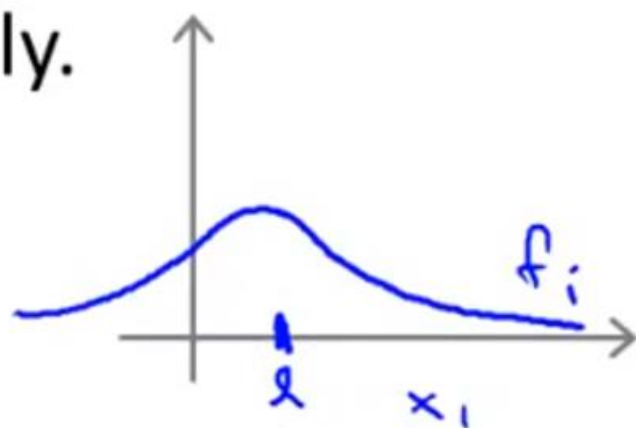
$C \left( = \frac{1}{\lambda} \right)$ .  $\rightarrow$  Large  $C$ : Lower bias, high variance.  
 $\rightarrow$  Small  $C$ : Higher bias, low variance.

(small  $\lambda$ )

(large  $\lambda$ )

$\sigma^2$  Large  $\sigma^2$ : Features  $f_i$  vary more smoothly.  
 $\rightarrow$  Higher bias, lower variance.

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



## SVM parameters:

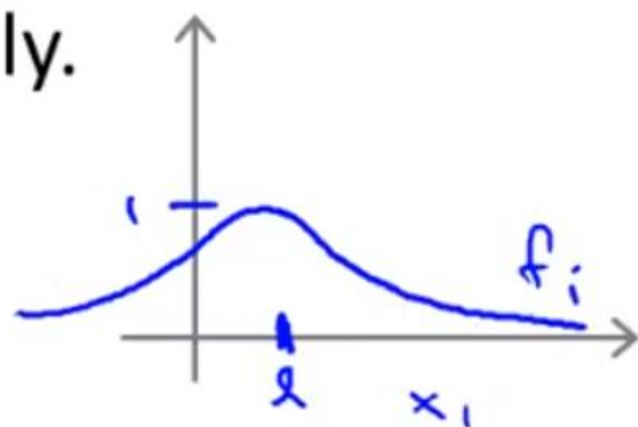
$C \left( = \frac{1}{\lambda} \right)$ .  $\rightarrow$  Large  $C$ : Lower bias, high variance.  
 $\rightarrow$  Small  $C$ : Higher bias, low variance.

(small  $\lambda$ )

(large  $\lambda$ )

$\sigma^2$  Large  $\sigma^2$ : Features  $f_i$  vary more smoothly.  
 $\rightarrow$  Higher bias, lower variance.

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



Small  $\sigma^2$ : Features  $f_i$  vary less smoothly.  
Lower bias, higher variance.

