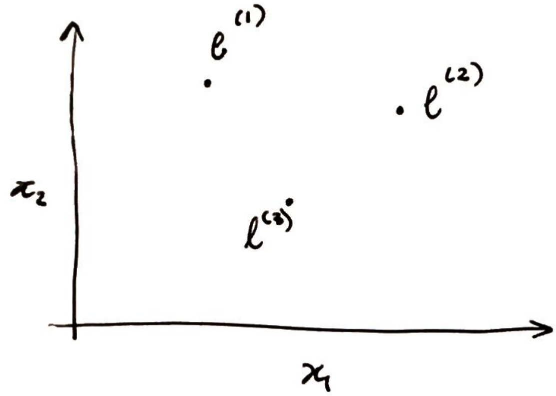
Support Vector Machine (SVM)

Kernel



$\ell^{(1)}, \ell^{(2)}, \ell^{(3)}$: landmarks

→ given x, compute new features based on proximity to these landmarks

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

Gaussian Kernels

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

$$f_3 = \ldots$$

- This similarity function is called **Kernel Function**.
- Note: $\|x - \ell^{(1)}\|^2$ → Euclid distance $= \sum_{j=1}^{n} (x_j - \ell_j^{(1)})^2$ in f_1

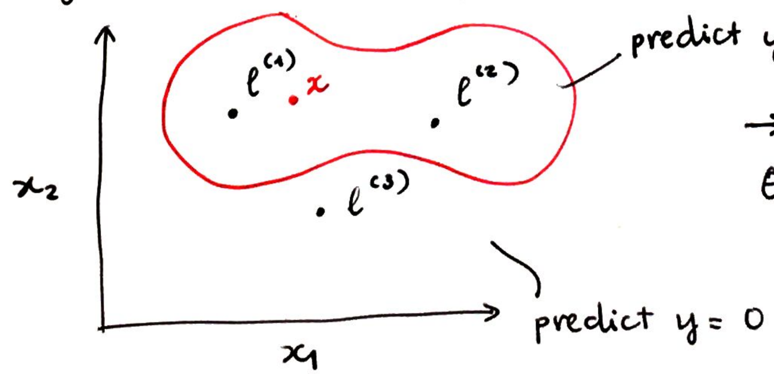
① If $x \approx \ell^{(1)} \Rightarrow$ distance ≈ 0, hence:

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

② If x is far from $\ell^{(1)}$,

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

- Using this, we can get to a non-linear decision bound



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

when, say,

$$\left[\begin{array}{l} \theta_0 = -0.5 \\ \theta_1 = 1 \quad \theta_2 = 1 \quad \theta_3 = 0 \end{array}\right.$$

Ex: x near $\ell^{(1)} \Rightarrow f_1 \approx 1, f_2 \approx 0, f_3 \approx 0 \Rightarrow h(x) = 0.5 \geq 0$

Predict ⇒ $y = 1$

SVM with Kernels

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

Choose $\ell^{(1)} = x^{(1)}, \ell^{(2)} = x^{(2)}, \dots, \ell^{(m)} = x^{(m)}$

Then, given an example x:

$$f_1 = \text{sim}(x, \ell^{(1)})$$
$$f_2 = \text{sim}(x, \ell^{(2)})$$
$$\vdots$$

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

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

$x^{(i)} \rightarrow$

$x^{(i)}$

$$f_1^{(i)} = \text{sim}(x^{(i)}, \ell^{(1)})$$
$$f_2^{(i)} = \text{sim}(x^{(i)}, \ell^{(2)})$$

$f_i^{(i)} = \text{sim}(x^{(i)}, \ell^{(i)}) \longrightarrow$

$$\vdots$$

$= \exp\left(-\frac{0}{2\delta^2}\right)$

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

$f_i^{(i)} = 1$

So $x^{(i)} \in \mathbb{R}^{n+1}$, then $f^{(i)} = \begin{bmatrix} f_0^{(i)} \\ f_1^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix}$ $\qquad f_0^{(i)} = 1 \qquad f \in \mathbb{R}^{m+1}$

- m training examples $\rightarrow$ m landmarks
- Hypothesis: Given x, compute features $f \in \mathbb{R}^{m+1}$

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

  $\hookrightarrow \theta_0 + \theta_1 f_1 + \theta_2 f_2 + \dots + \theta_m f_m$

  we have: $\theta \in \mathbb{R}^{m+1}$

- How to get the set of θ? $\Rightarrow$ using SVM training algo.

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

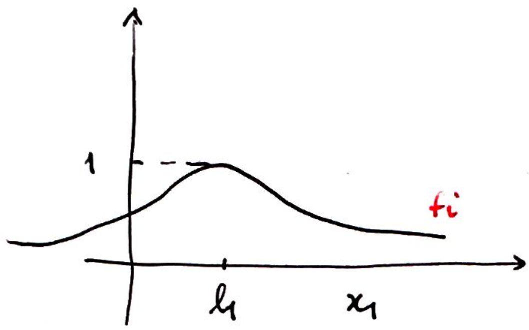
$n = m$

SVM parameters

$C \left(= \frac{1}{\lambda}\right)$
- Large C (ie, small λ) : Lower bias, higher variance
- Small C : high bias, low variance

σ^2

- Large σ^2 ⇒ features f_i vary more smoothly
  ⇒ higher bias, lower variance



- Small σ^2 ⇒ features f_i vary less smoothly
  ⇒ Lower bias, higher variance