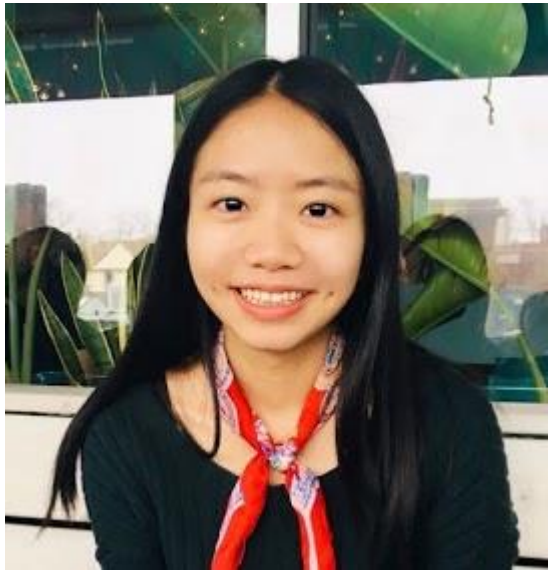
Support Vector Machine II

Jia-Bin Huang
Virginia Tech

Administrative

- HW 1 due tonight
- HW 2 released.

- **Online Scalable Learning Adaptive to Unknown Dynamics and Graphs**
- Yanning Shen, University of Minnesota, Twin Cities
- Tuesday, February 19 9:00am - 10:15am
  1110 Knowledgeworks II



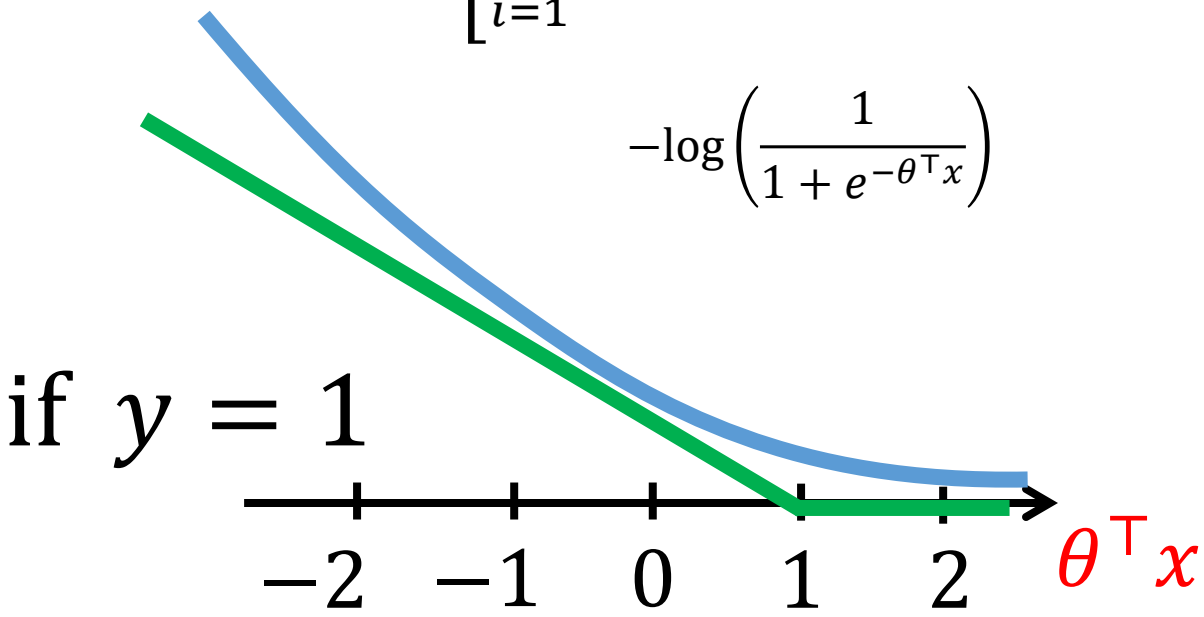
- **Declarative Learning-based Programming for Learning and Reasoning over Spatial Language**
- Parisa Kordjamshidi, Assistant Professor, Tulane University
- Thursday, February 21 9:00am - 10:15am
  110 McBryde

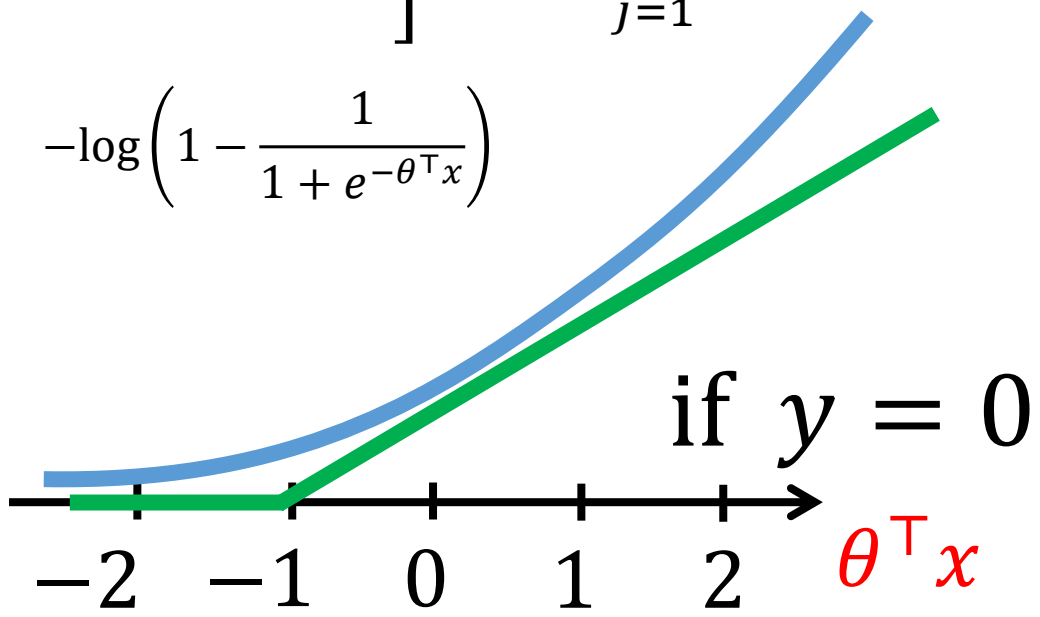
Logistic regression (logistic loss)

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

Support vector machine (hinge loss)

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



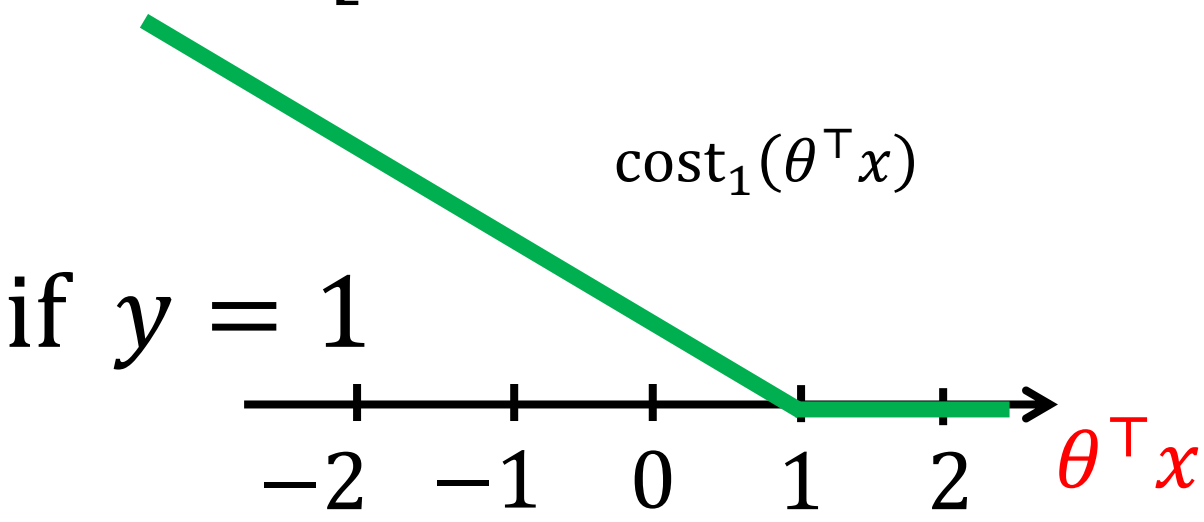


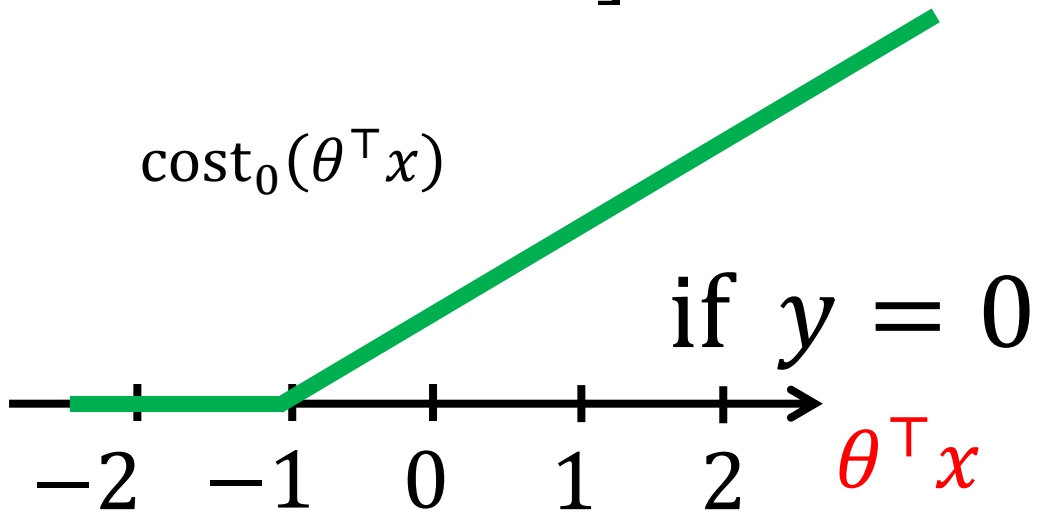
Support Vector Machine

- Cost function
- **Large margin classification**
- Kernels
- Using an SVM

Support vector machine

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





If "y = 1", we want $\theta^{\top}x \geq 1$ (not just ≥ 0)

If "y = 0", we want $\theta^{\top}x \leq -1$ (not just < 0)

Slide credit: Andrew Ng

SVM decision boundary

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

- Let's say we have a very large C …

- Whenever $y^{(i)} = 1$:

  $\theta^\top x^{(i)} \geq 1$

- Whenever $y^{(i)} = 0$:

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

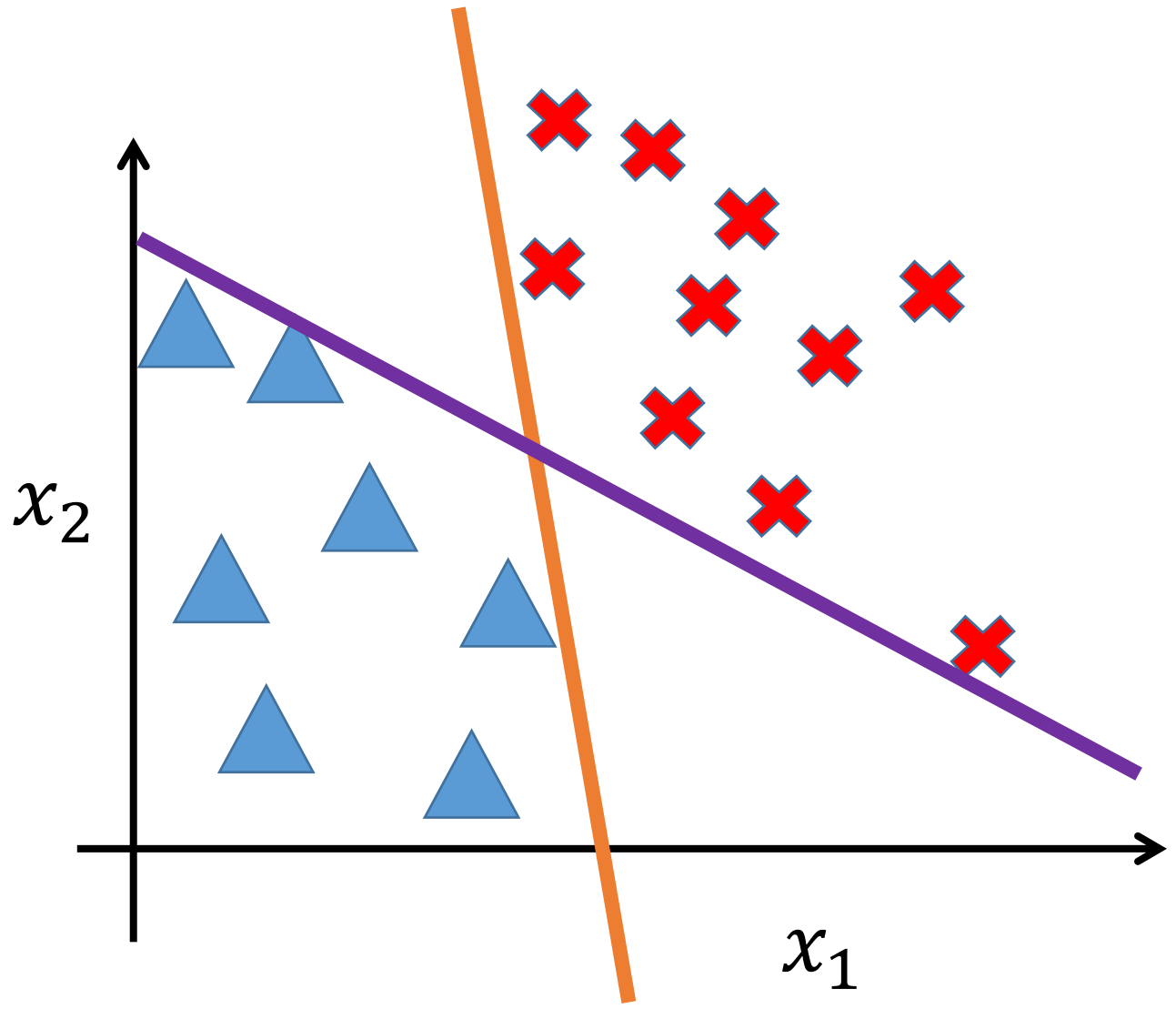
$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

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

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

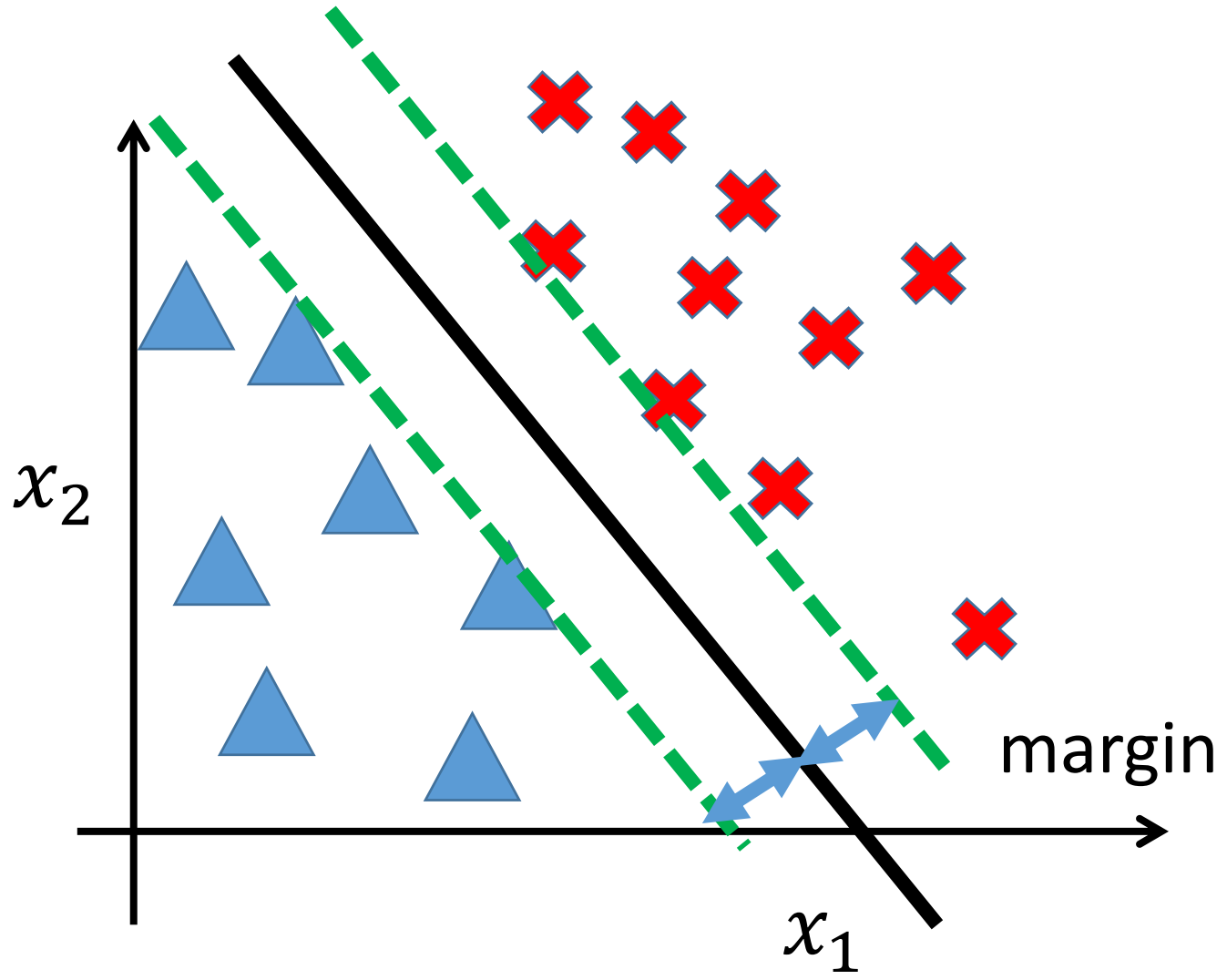
Slide credit: Andrew Ng

SVM decision boundary: Linearly separable case



Slide credit: Andrew Ng

SVM decision boundary: Linearly separable case

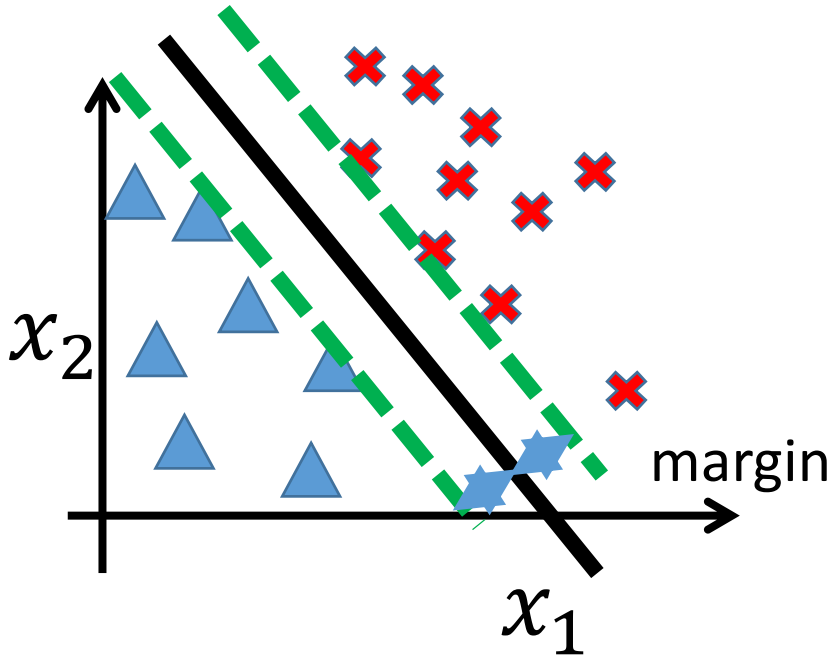


Slide credit: Andrew Ng

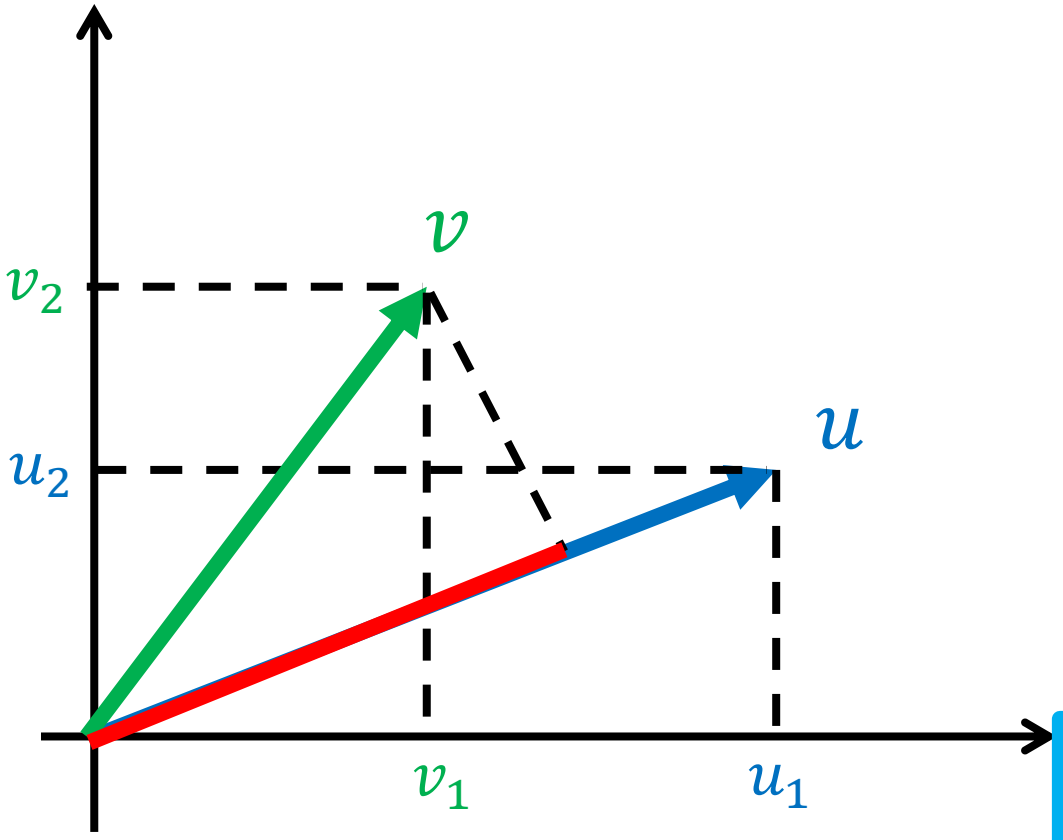
Why large margin classifiers?

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

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$



Vector inner product



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

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

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

p = length of projection of v onto u

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

Slide credit: Andrew Ng

SVM decision boundary

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

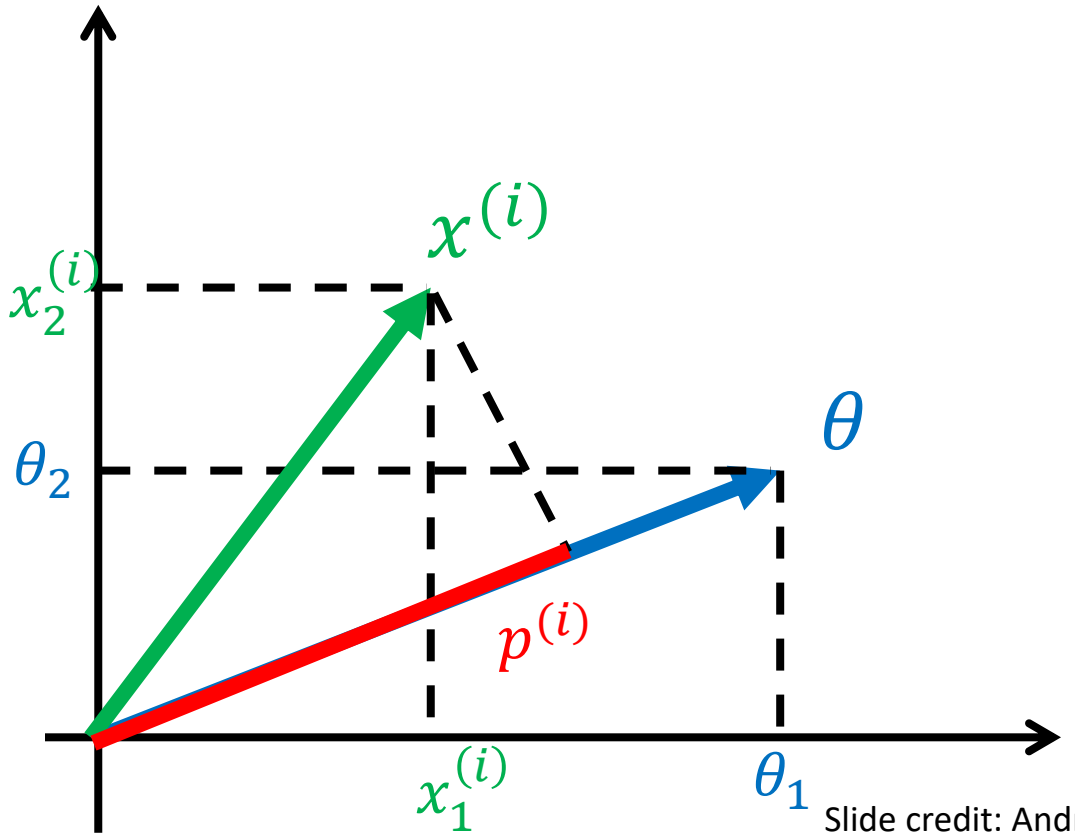
$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

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

Simplication: $\theta_0 = 0, n = 2$

What's $\theta^\top x^{(i)}$?

$$\theta^\top x^{(i)} = p^{(i)}\|\theta\|^2$$



Slide credit: Andrew Ng

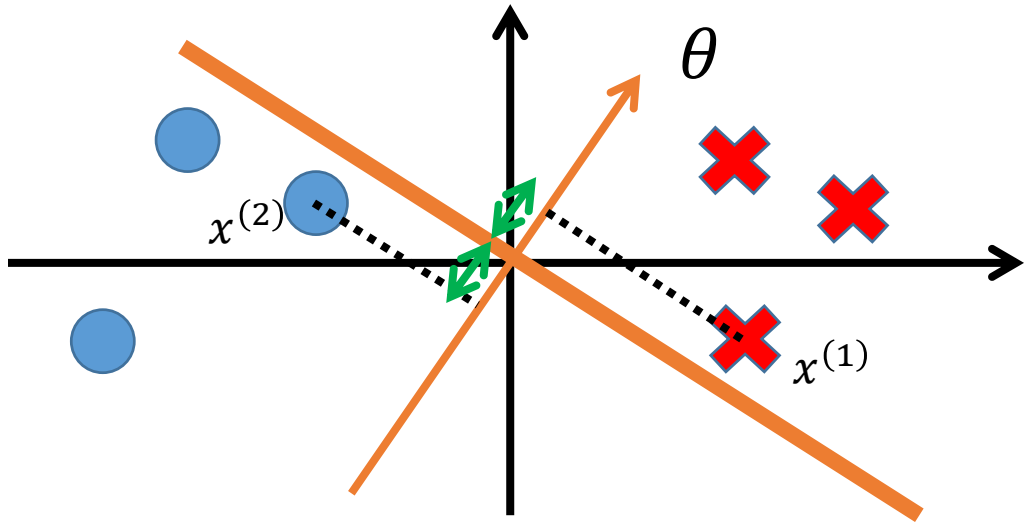
SVM decision boundary

$$\min_{\theta} \frac{1}{2}\|\theta\|^2$$

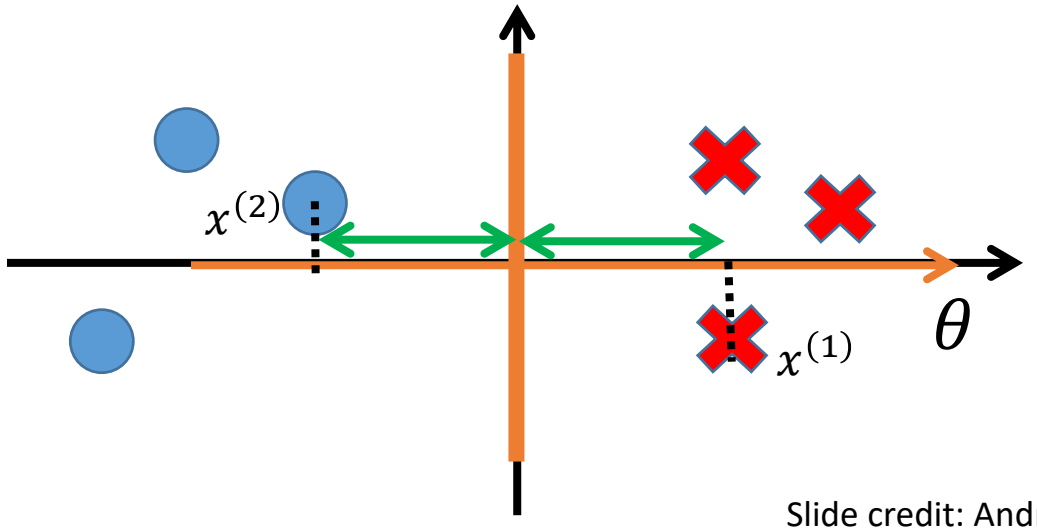
$$\text{s.t.} \quad p^{(i)}\|\theta\|^2 \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$p^{(i)}\|\theta\|^2 \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

Simplication: $\theta_0 = 0, n = 2$

$p^{(1)}, p^{(2)}$ small $\rightarrow \|\theta\|^2$ large

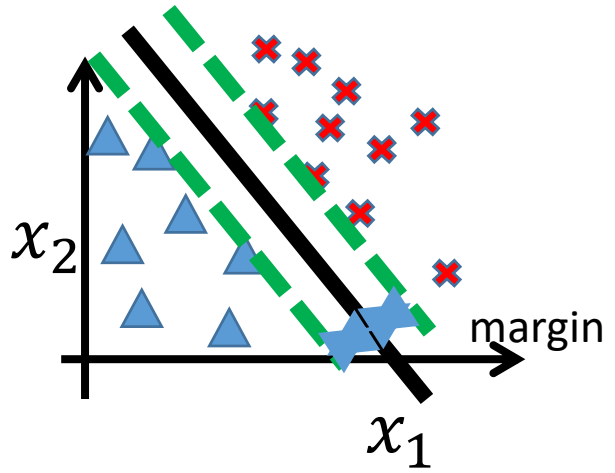


$p^{(1)}, p^{(2)}$ large $\rightarrow \|\theta\|^2$ can be small



Slide credit: Andrew Ng

Rewrite the formulation



Let $\theta_j = \theta'_j/\gamma$, with $\|\theta'\|^2 = 1$

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

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

$$\max_{\theta',\gamma} \quad \gamma$$

$$\text{s.t.} \quad \|\theta'\|^2 = 1$$
$$\theta'^\top x^{(i)} \geq \gamma \quad \text{if} \quad y^{(i)} = 1$$
$$\theta'^\top x^{(i)} \leq -\gamma \quad \text{if} \quad y^{(i)} = 0$$

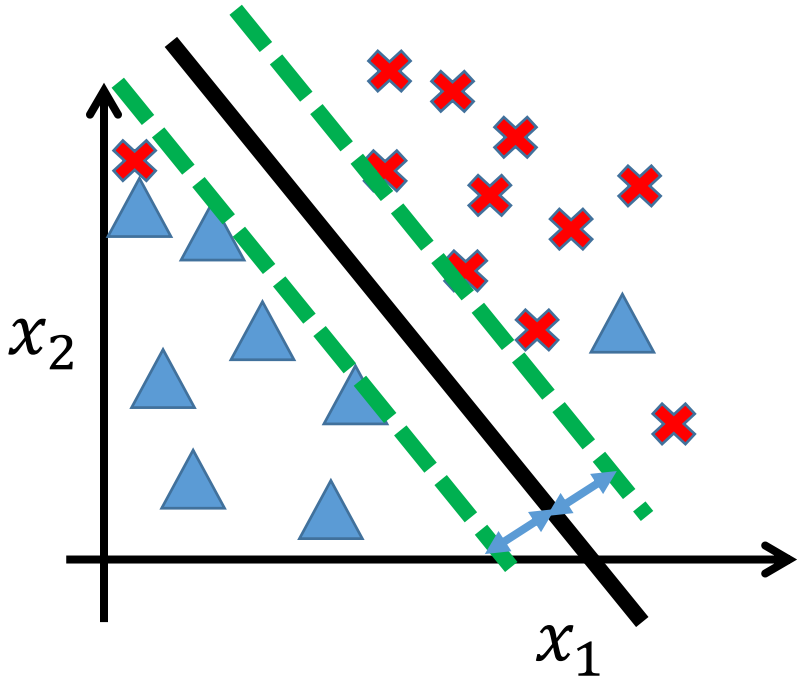
$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 = \min_{\theta',\gamma} \frac{1}{2}\sum_{j=1}^{n} \frac{\theta_j'^2}{\gamma^2} = \max_{\gamma} \gamma$$

$$\theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1 \quad \rightarrow \quad \theta'^\top x^{(i)} \geq \gamma \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0 \quad \rightarrow \quad \theta'^\top x^{(i)} \leq -\gamma \quad \text{if} \quad y^{(i)} = 0$$

Data not linearly separable?

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

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$



$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 + C(\#\text{misclassification})$$

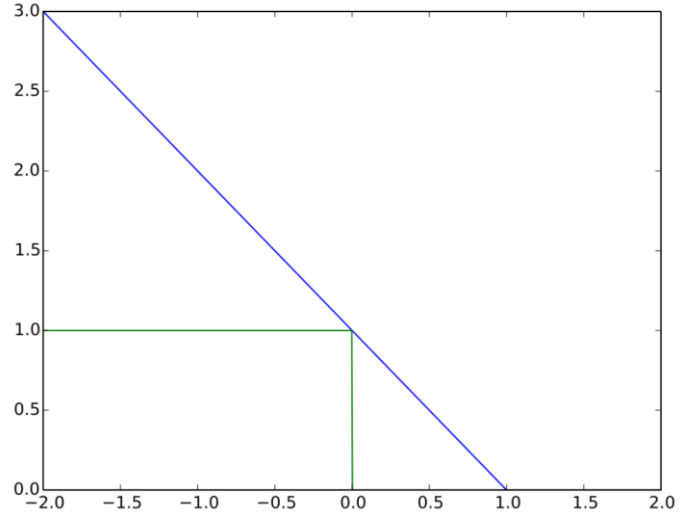
$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

← NP-hard ☹

Convex relaxation

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 + C(\#\text{misclassification})$$

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

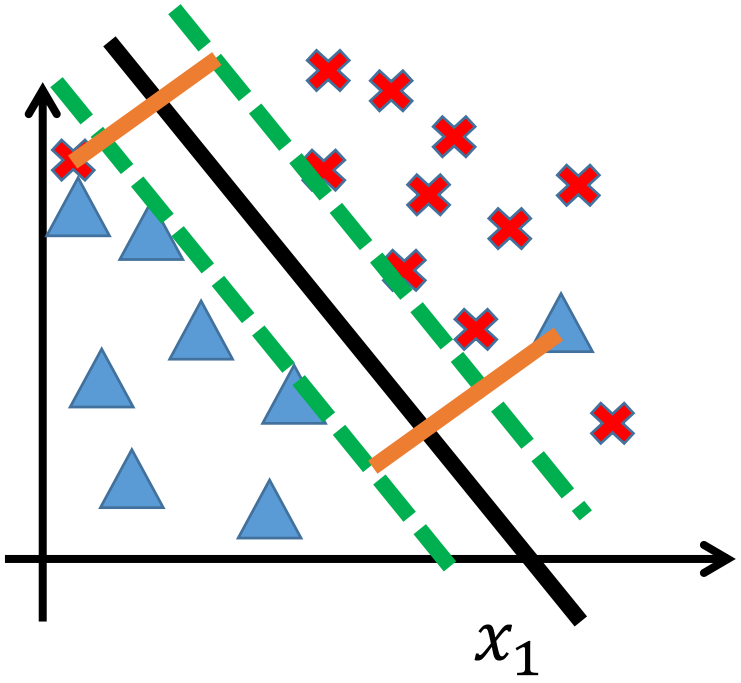


NP-hard ☹

$\xi^{(i)}$: slack variables

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 + C\sum_i \xi^{(i)}$$

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 - \xi^{(i)} \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 + \xi^{(i)} \quad \text{if} \quad y^{(i)} = 0$$
$$\xi^{(i)} \geq 0 \quad \forall\, i$$



Hinge loss

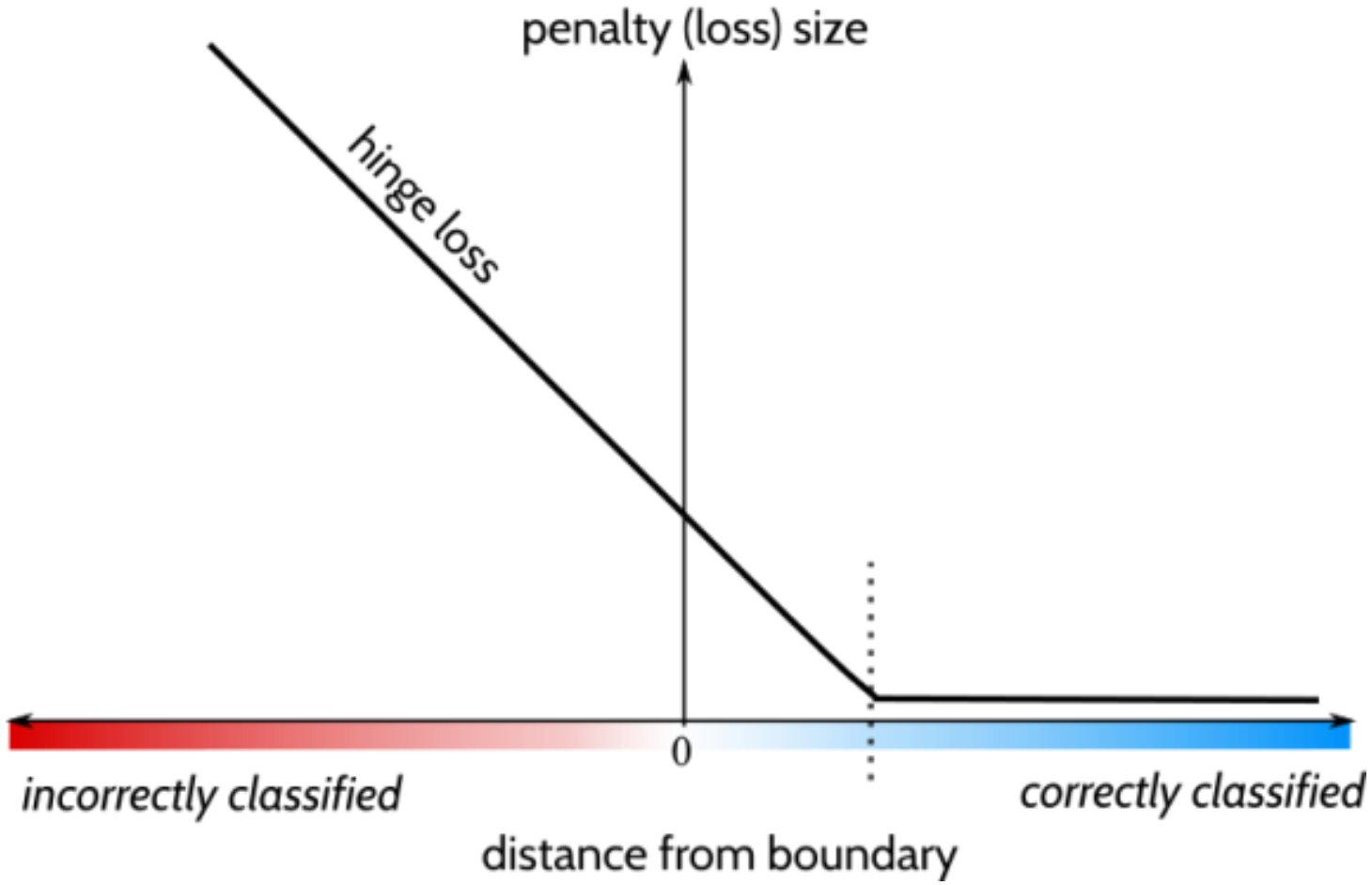


Image credit: https://math.stackexchange.com/questions/782586/how-do-you-minimize-hinge-loss

Hard-margin SVM formulation

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

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 \quad \text{if} \quad y^{(i)} = 0$$

Soft-margin SVM formulation

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 + C\sum_i \xi^{(i)}$$

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 - \xi^{(i)} \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 + \xi^{(i)} \quad \text{if} \quad y^{(i)} = 0$$
$$\xi^{(i)} \geq 0 \quad \forall\, i$$

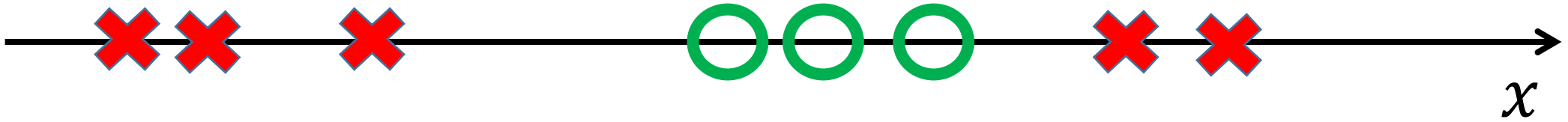
Support Vector Machine

- Cost function
- Large margin classification
- **Kernels**
- Using an SVM

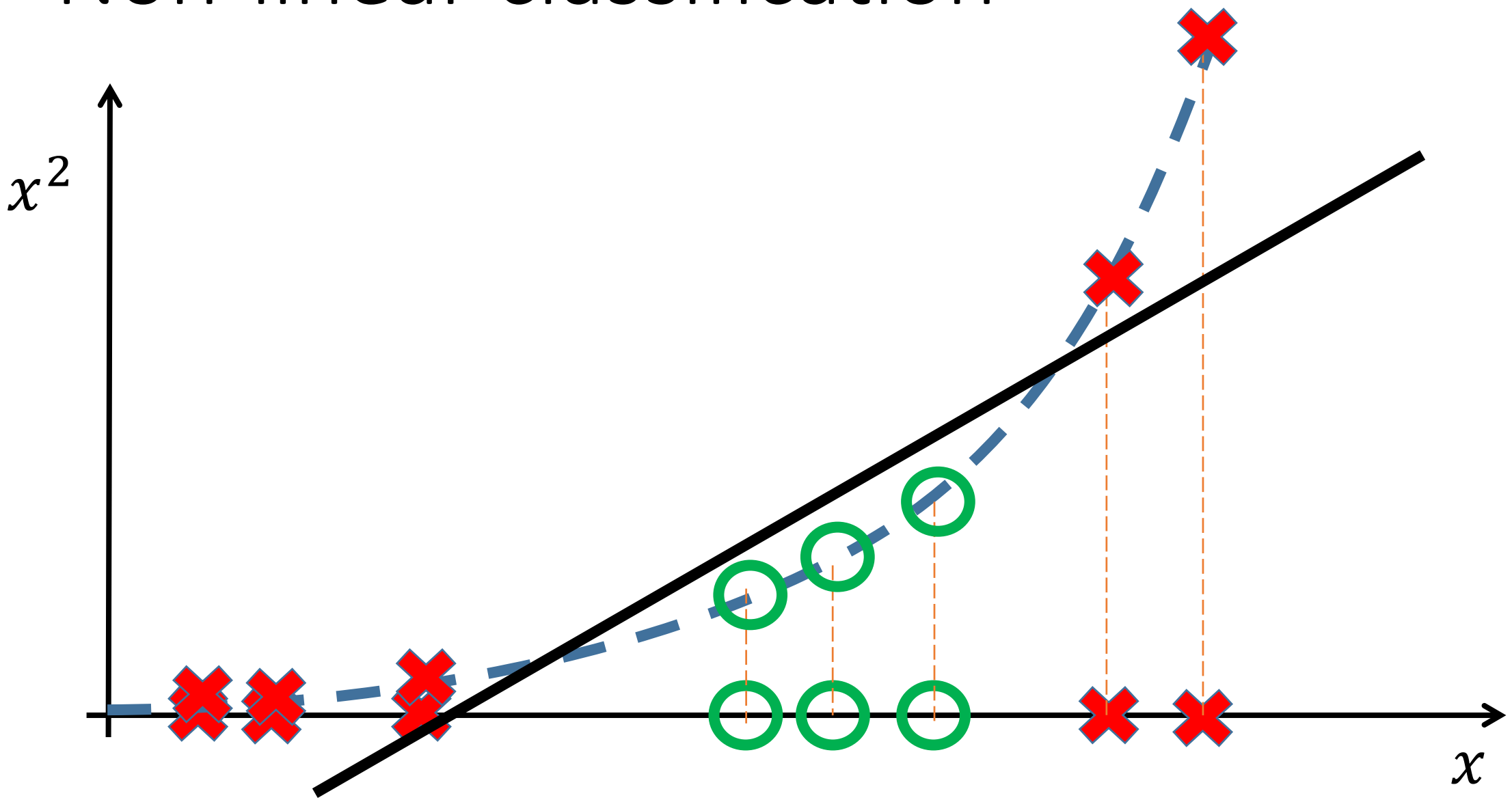
Non-linear classification

- How do we separate the two classes using a hyperplane?

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



Non-linear classification



Kernel

- $K(\cdot,\cdot)$ a legal definition of inner product:

$\exists \phi: X \to R^N$
s.t. $K(x,z) = \phi(x)^\top \phi(z)$

Why Kernels matter?

- Many algorithms interact with data only via **dot-products**
- Replace $x^\top z$ with $K(x, z) = \phi(x)^\top \phi(z)$
- Act *implicitly* as if data was in the higher-dimensional ϕ-space

Example

$K(x,z) = (x^{\mathsf{T}}z)^2$ corresponds to

$(x_1, x_2) \rightarrow \phi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$

$$\phi(x)^{\mathsf{T}}\phi(z) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)\left(z_1^2, z_2^2, \sqrt{2}z_1z_2\right)^{\mathsf{T}}$$

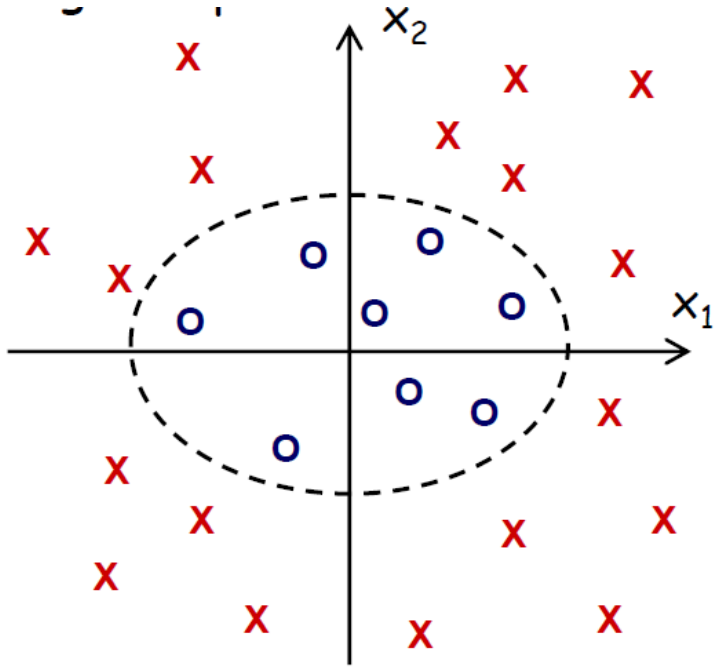
$$= x_1^2z_1^2 + x_2^2z_2^2 + 2\mathrm{x}_1\mathrm{x}_2\mathrm{z}_1\mathrm{z}_2 = (\mathrm{x}_1\mathrm{z}_1 + \mathrm{x}_2\mathrm{z}_2)^2$$

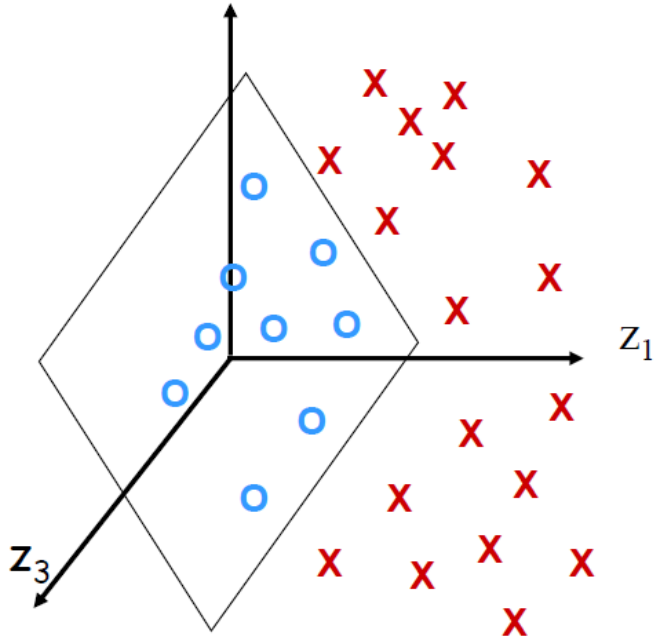
$$= (x^{\mathsf{T}}z)^2 = K(x,z)$$

Example

$K(x, z) = (x^\top z)^2$ corresponds to

$(x_1, x_2) \rightarrow \phi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$





Slide credit: Maria-Florina Balcan

Example kernels

- Linear kernel

$$K(x, z) = x^\top z$$

- Gaussian (Radial basis function) kernel

$$K(x, z) = \exp\left(-\frac{1}{2}(x - z)^\top \Sigma^{-1}(x - z)\right)$$

- Sigmoid kernel

$$K(x, z) = \tanh(a \cdot x^\top z + b)$$

Constructing new kernels

- Positive scaling

$$K(x, z) = cK_1(x, z), c > 0$$

- Exponentiation

$$K(x, z) = \exp\big(K_1(x, z)\big)$$

- Addition

$$K(x, z) = K_1(x, z) + K_2(x, z)$$

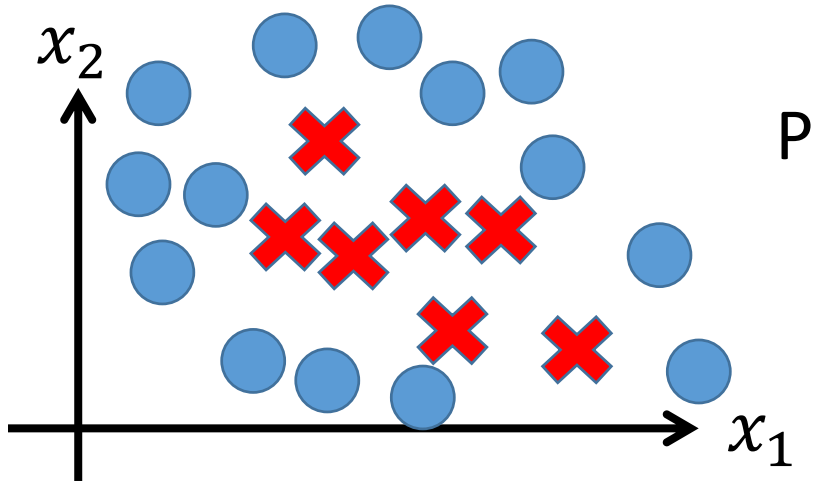
- Multiplication with function

$$K(x, z) = f(x)K_1(x, z)f(z)$$

- Multiplication

$$K(x, z) = K_1(x, z)K_2(x, z)$$

Non-linear decision boundary



Predict $y = 1$ if

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \cdots \geq 0$$

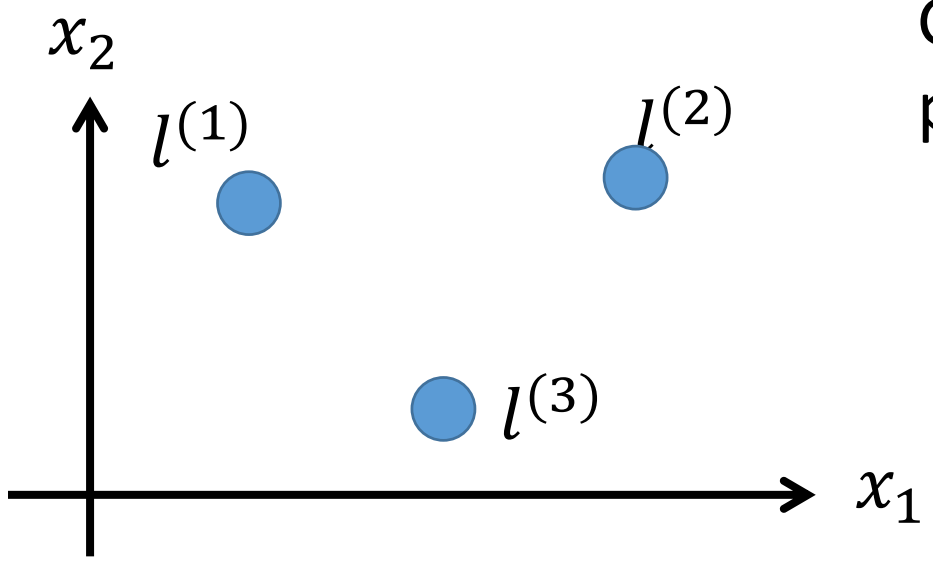
$$\theta_0 + \theta_1 f_1 + \theta_2 f_2 + \theta_3 f_3 + \cdots$$

$$f_1 = x_1, f_2 = x_2, f_3 = x_1 x_2, \cdots$$

Is there a different/better choice of the features $f_1, f_2, f_3, \cdots$?

Slide credit: Andrew Ng

Kernel



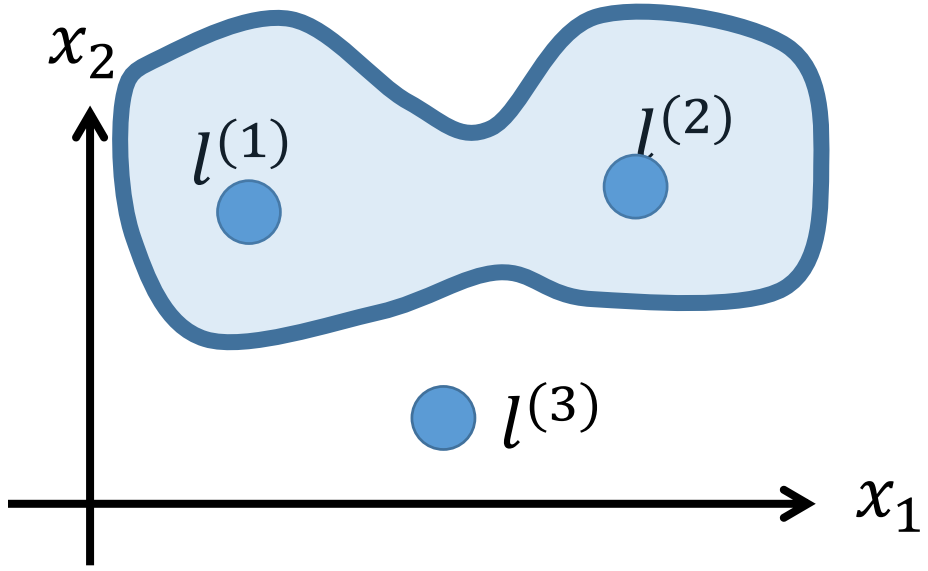
Give x, compute new features depending on proximity to landmarks $l^{(1)}$, $l^{(2)}$, $l^{(3)}$

$$f_1 = \text{similarity}(x, l^{(1)})$$
$$f_2 = \text{similarity}(x, l^{(2)})$$
$$f_3 = \text{similarity}(x, l^{(3)})$$

Gaussian kernel

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

Slide credit: Andrew Ng



Predict $y = 1$ if

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

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

$$f_1 = \text{similarity}(x, l^{(1)})$$
$$f_2 = \text{similarity}(x, l^{(2)})$$
$$f_3 = \text{similarity}(x, l^{(3)})$$

Slide credit: Andrew Ng

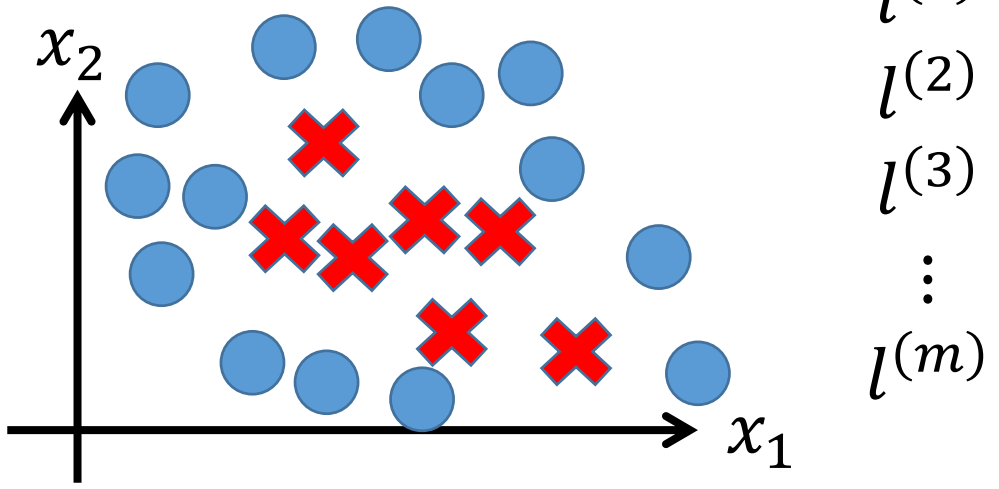
Choosing the landmarks

- Given x

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

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

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



Slide credit: Andrew Ng

SVM with kernels

- Given $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right)$
- Choose $l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, l^{(3)} = x^{(3)}, \cdots, l^{(m)} = x^{(m)}$
- Given example x:
  - $f_1 = \text{similarity}\left(x, l^{(1)}\right)$
  - $f_2 = \text{similarity}\left(x, l^{(2)}\right)$
  - $\cdots$
- For training example $\left(x^{(i)}, y^{(i)}\right)$:
  - $x^{(i)} \rightarrow f^{(i)}$

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

Slide credit: Andrew Ng

SVM with kernels

- Hypothesis: Given x, compute features $f \in \mathbb{R}^{m+1}$
  - Predict $y = 1 \quad \text{if} \quad \theta^\top f \geq 0$

- **Training (original)**

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

- **Training (with kernel)**

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

Support vector machines (Primal/Dual)

- Primal form

$$\min_{\theta,\xi^{(i)}} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 + C\sum_i \xi^{(i)}$$

$$\text{s.t.} \quad \theta^\top x^{(i)} \geq 1 - \xi^{(i)} \quad \text{if} \quad y^{(i)} = 1$$
$$\theta^\top x^{(i)} \leq -1 + \xi^{(i)} \quad \text{if} \quad y^{(i)} = 0$$
$$\xi^{(i)} \geq 0 \quad \forall\, i$$

- Lagrangian dual form

$$\min_{\alpha} \frac{1}{2}\sum_i\sum_j y^{(i)}y^{(j)}\alpha^{(i)}\alpha^{(j)}{x^{(i)}}^\top x^{(j)} - \sum_i \alpha^{(i)}$$

$$\text{s.t.} \quad 0 \leq \alpha^{(i)} \leq C_i$$
$$\sum_i y^{(i)}\alpha^{(i)} = 0$$

SVM (Lagrangian dual)

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j y^{(i)} y^{(j)} \alpha^{(i)} \alpha^{(j)} \boxed{x^{(i)\top} x^{(j)}} - \sum_i \alpha^{(i)}$$

s. t. $0 \leq \alpha^{(i)} \leq C_i$

$$\sum_i y^{(i)} \alpha^{(i)} = 0$$

Replace $x^{(i)\top} x^{(j)}$ with $K(x^{(i)}, x^{(j)})$

Classifier: $\theta = \sum_i \alpha^{(i)} y^{(i)} x^{(i)}$

- The points $x^{(i)}$ for which $\alpha^{(i)} \neq 0 \rightarrow$ Support Vectors

SVM parameters

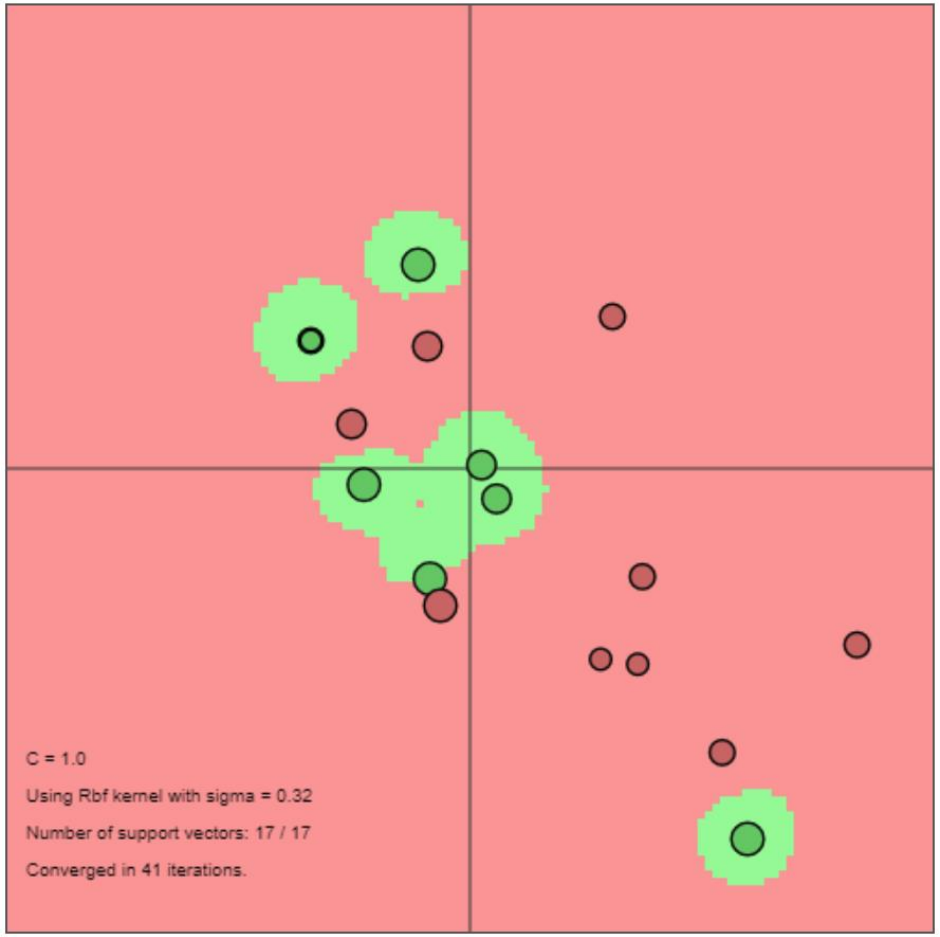
- $C\left(=\frac{1}{\lambda}\right)$
  Large C: Lower bias, high variance.
  Small C: Higher bias, low variance.

- 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

Slide credit: Andrew Ng

SVM Demo

- https://cs.stanford.edu/people/karpathy/svmjs/demo/



SVM song

- https://www.youtube.com/watch?v=g15bqtyidZs

Video source:

Support Vector Machine

- Cost function
- Large margin classification
- Kernels
- **Using an SVM**

Using SVM

- SVM software package (e.g., liblinear, libsvm) to solve for θ

- Need to specify:
  - Choice of parameter C.
  - Choice of kernel (similarity function):

- Linear kernel: Predict $y = 1 \quad \text{if} \quad \theta^\top x \geq 0$
- Gaussian kernel:
  - $f_i = \exp(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2})$, where $l^{(i)} = x^{(i)}$
  - Need to choose σ^2. Need proper feature scaling

Slide credit: Andrew Ng

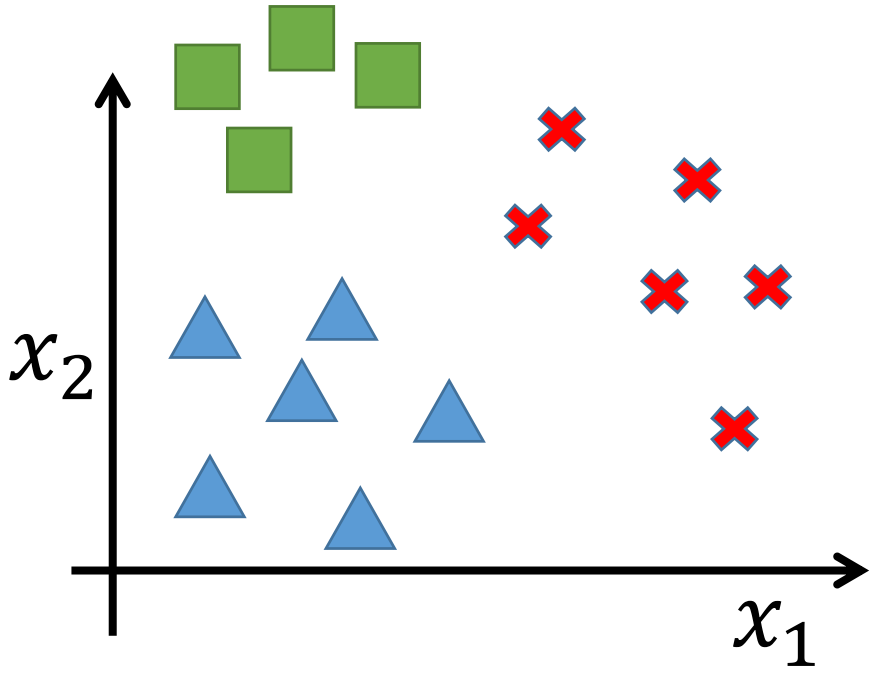
Kernel (similarity) functions

- Note: not all similarity functions make valid kernels.

- Many off-the-shelf kernels available:
  - Polynomial kernel
  - String kernel
  - Chi-square kernel
  - Histogram intersection kernel

Slide credit: Andrew Ng

Multi-class classification



- Use one-vs.-all method. Train K SVMs, one to distinguish $y = i$ from the rest, get $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(K)}$

- Pick class i with the largest ${\theta^{(i)}}^{\top} x$

Slide credit: Andrew Ng

Logistic regression vs. SVMs

- $n =$ number of features ($x \in \mathbb{R}^{n+1}$), $m =$ number of training examples

1. **If n is large (relative to m)**: $(n = 10{,}000, \mathrm{m} = 10 - 1000)$
   → Use logistic regression or SVM without a kernel (“linear kernel”)
2. **If n is small, m is intermediate**: $(n = 1 - 1000, \mathrm{m} = 10 - 10{,}000)$
   → Use SVM with Gaussian kernel
3. **If n is small, m is large**: $(n = 1 - 1000, \mathrm{m} = 50{,}000+)$
   → Create/add more features, then use logistic regression of linear SVM

Neural network likely to work well for most of these case,
but slower to train

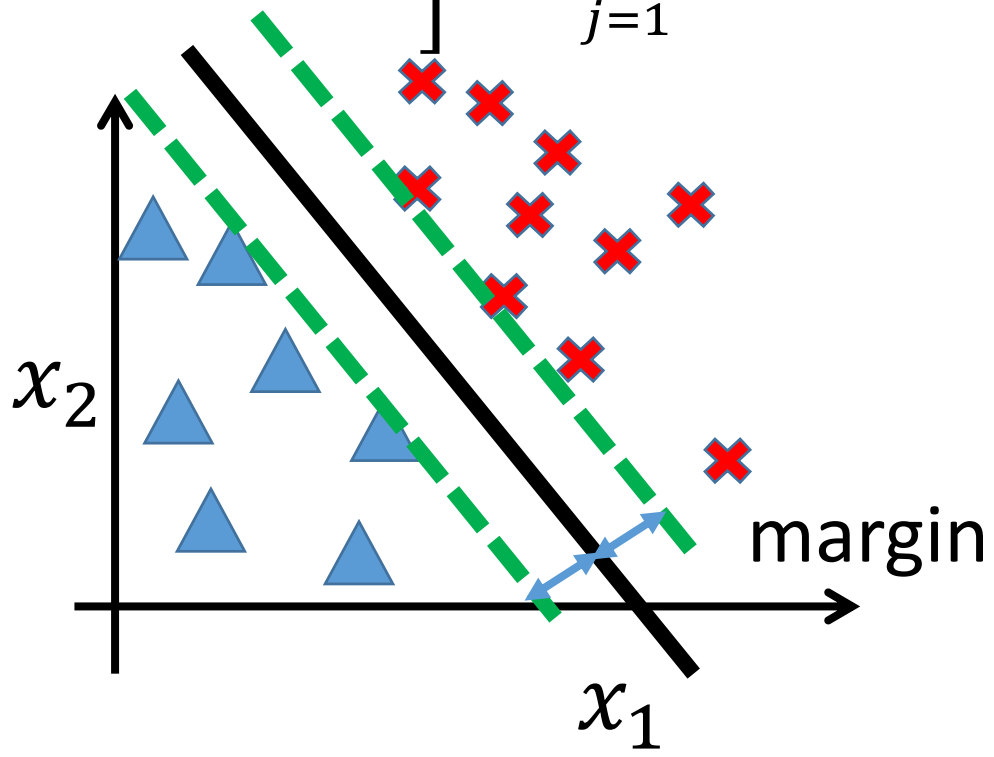
Slide credit: Andrew Ng

Things to remember

- **Cost function**

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

- **Large margin classification**

- **Kernels**

- **Using an SVM**