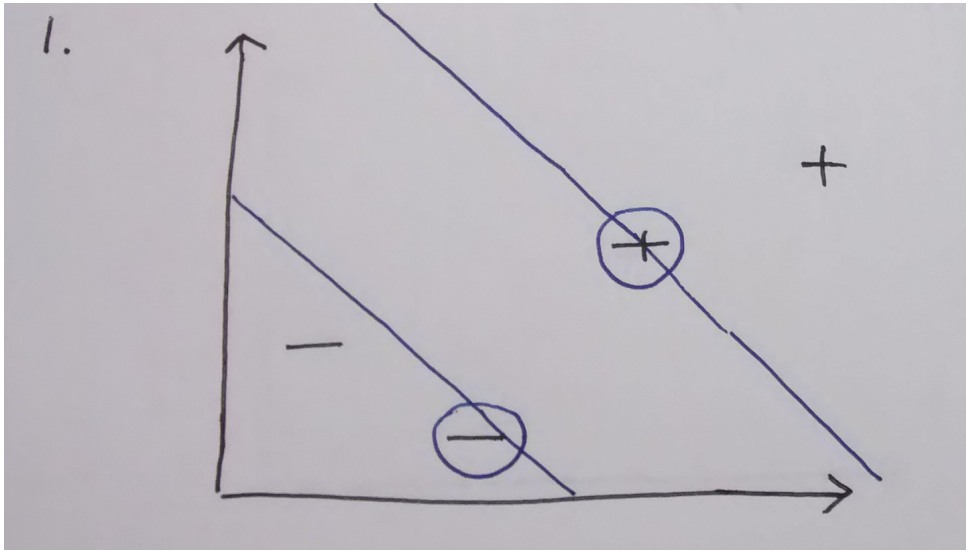
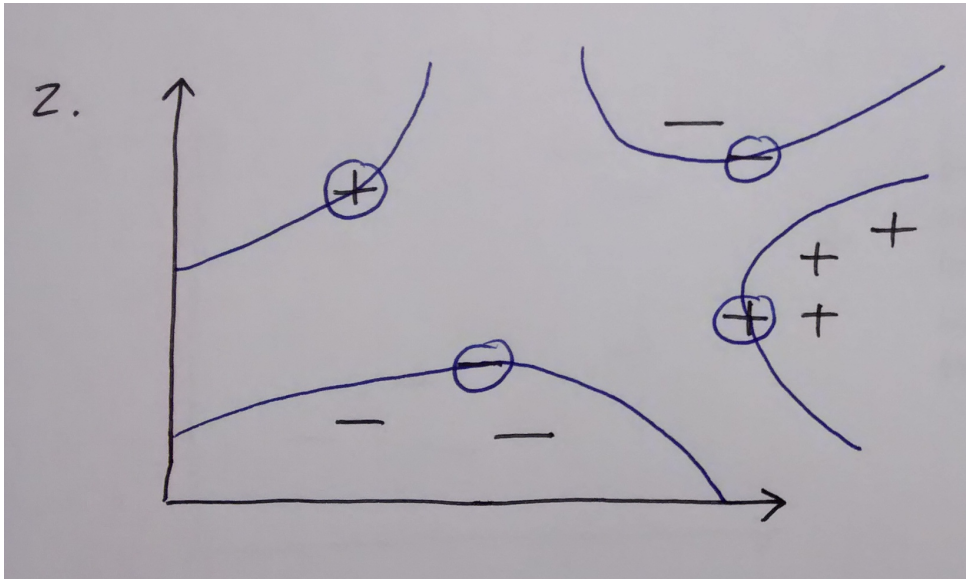
Support Vector Machines

COMP 640

Ryan Spring, Sarah Kim

Quiz Example Solutions



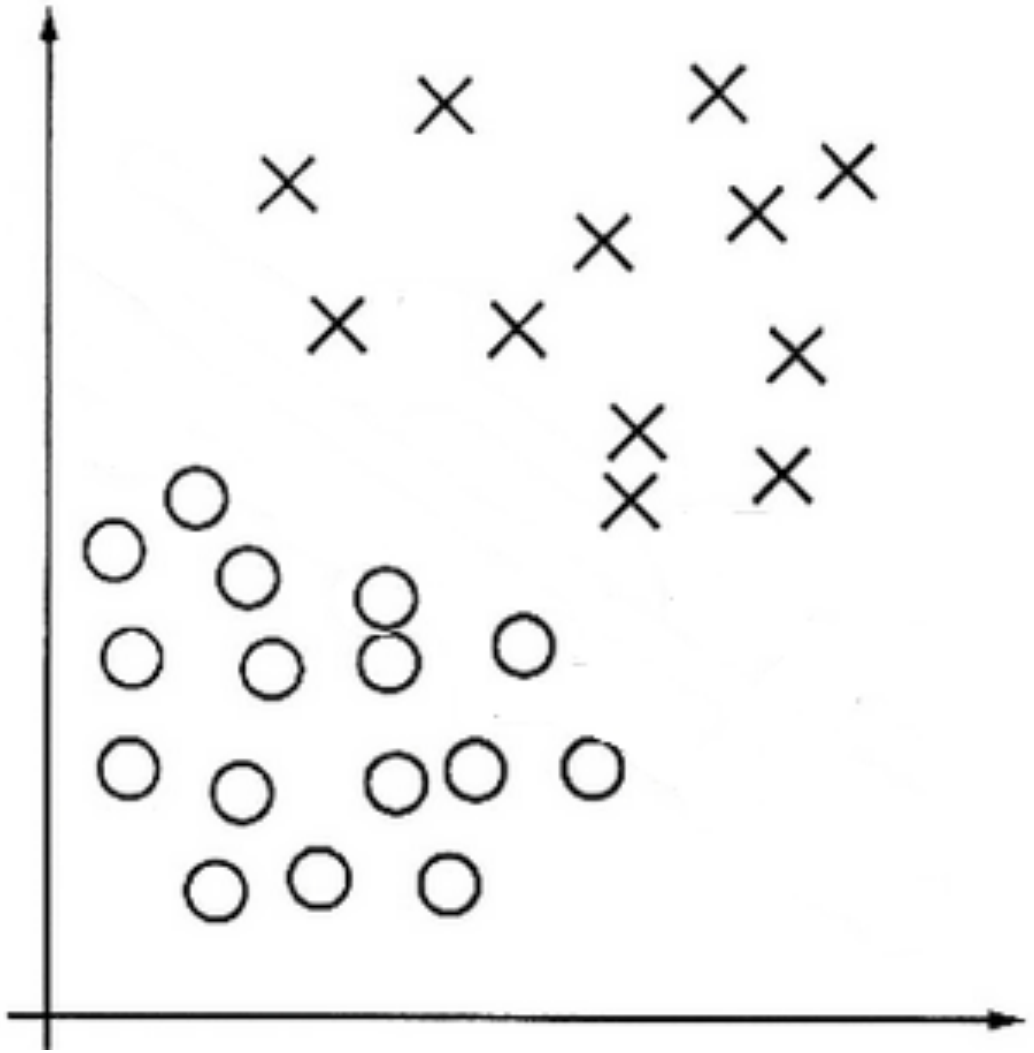


What is classification?

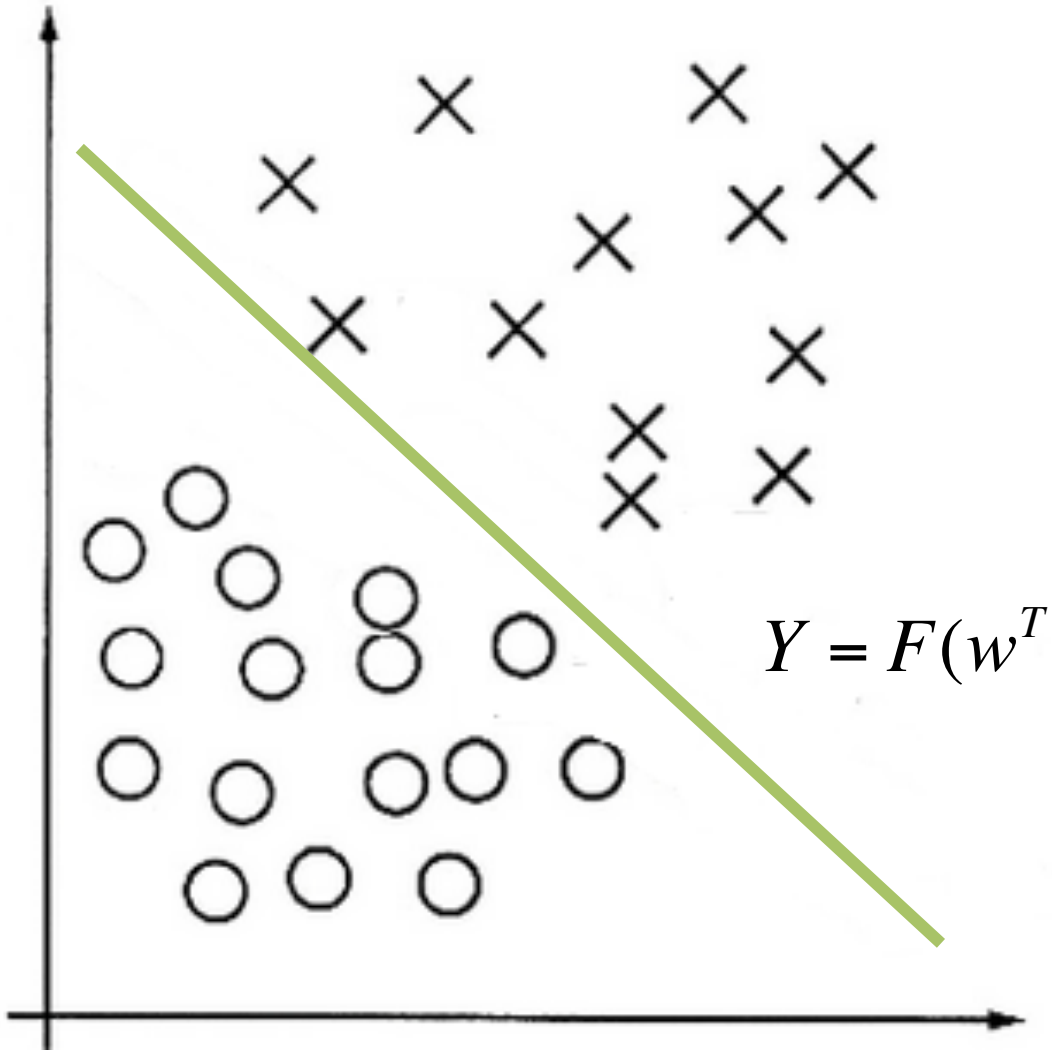
F(x) = -1 Not Spam

F(x) = +1 Spam

How should I divide the data?

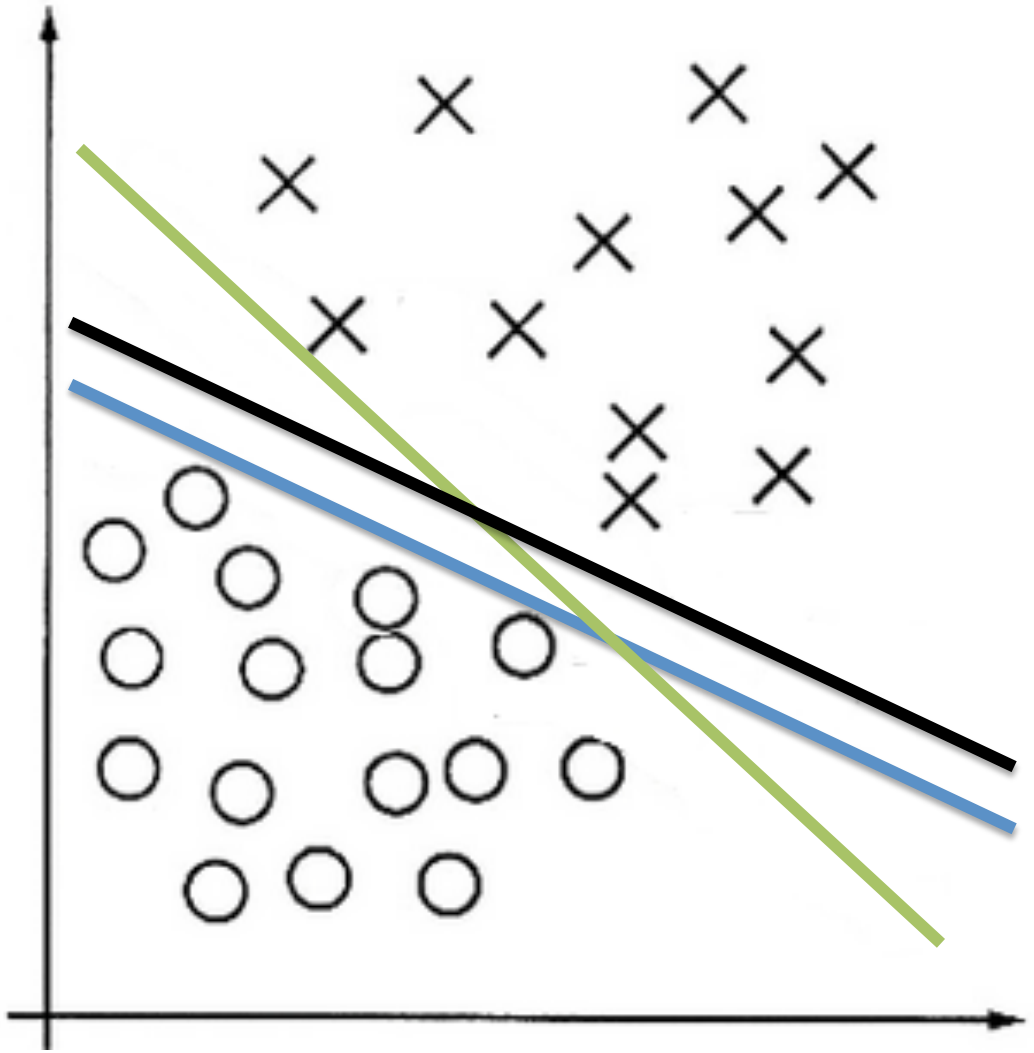


Linear Classifier

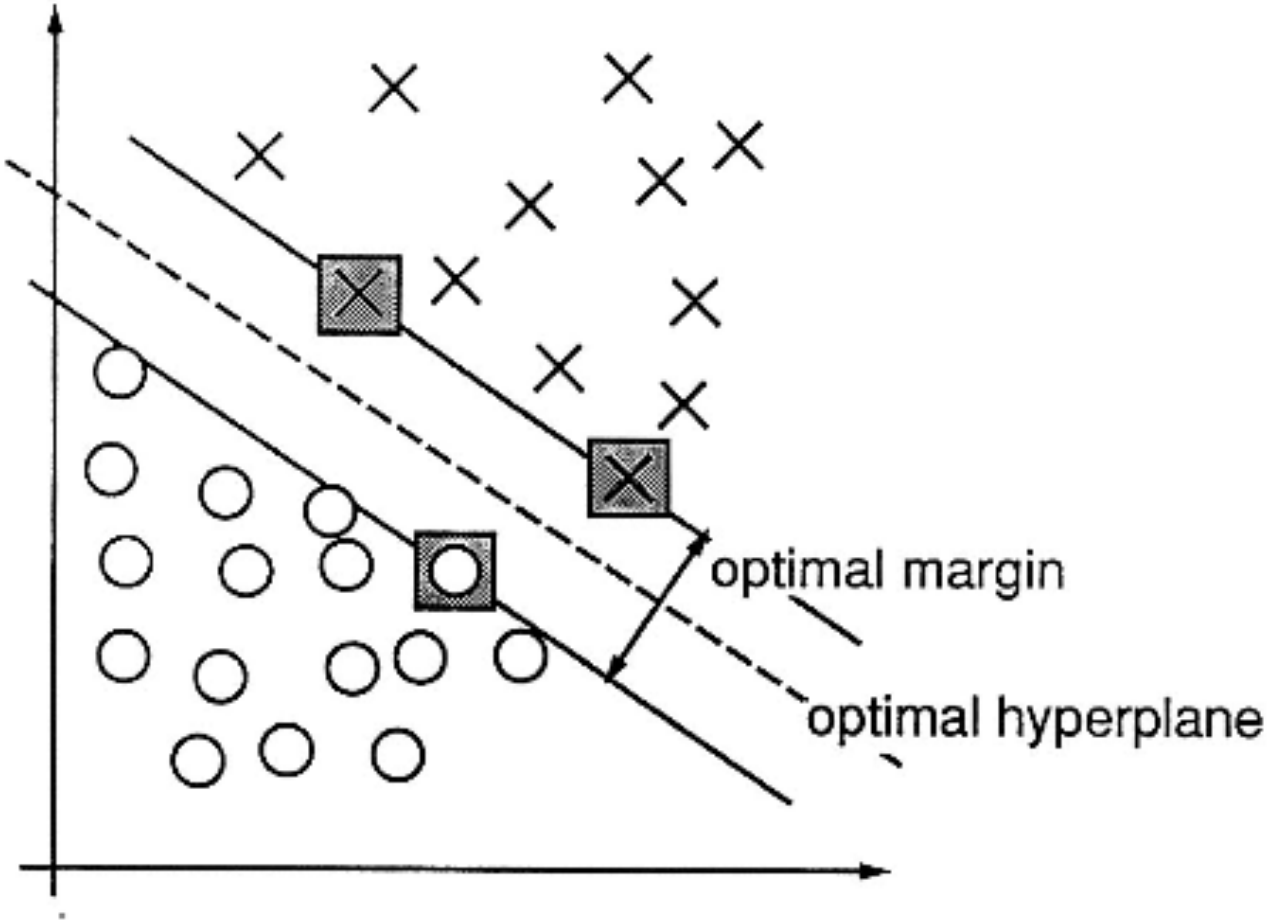


$$Y = F(w^T x) = \sum w_i x_j$$

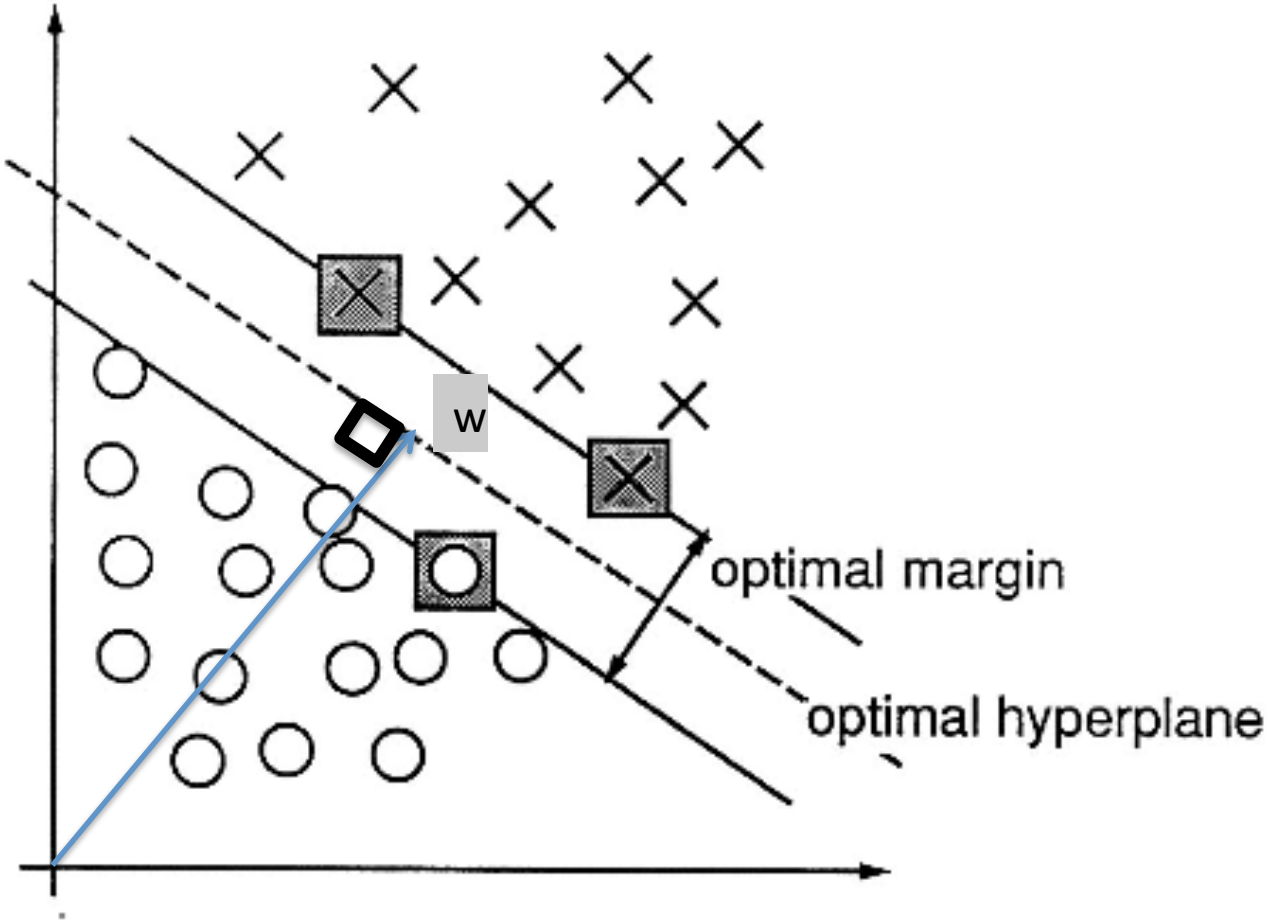
Multiple Possible Solutions



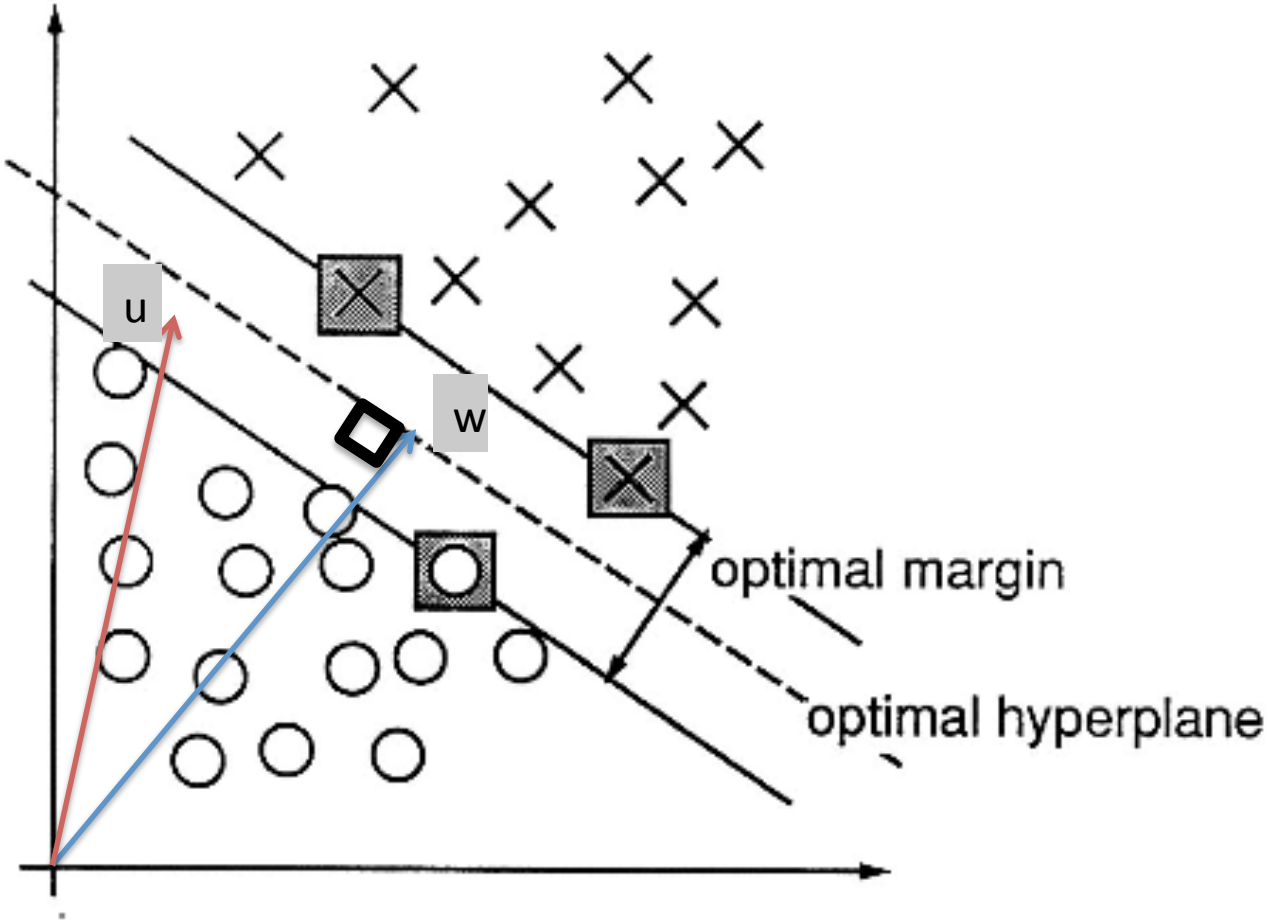
Defining Features of SVM



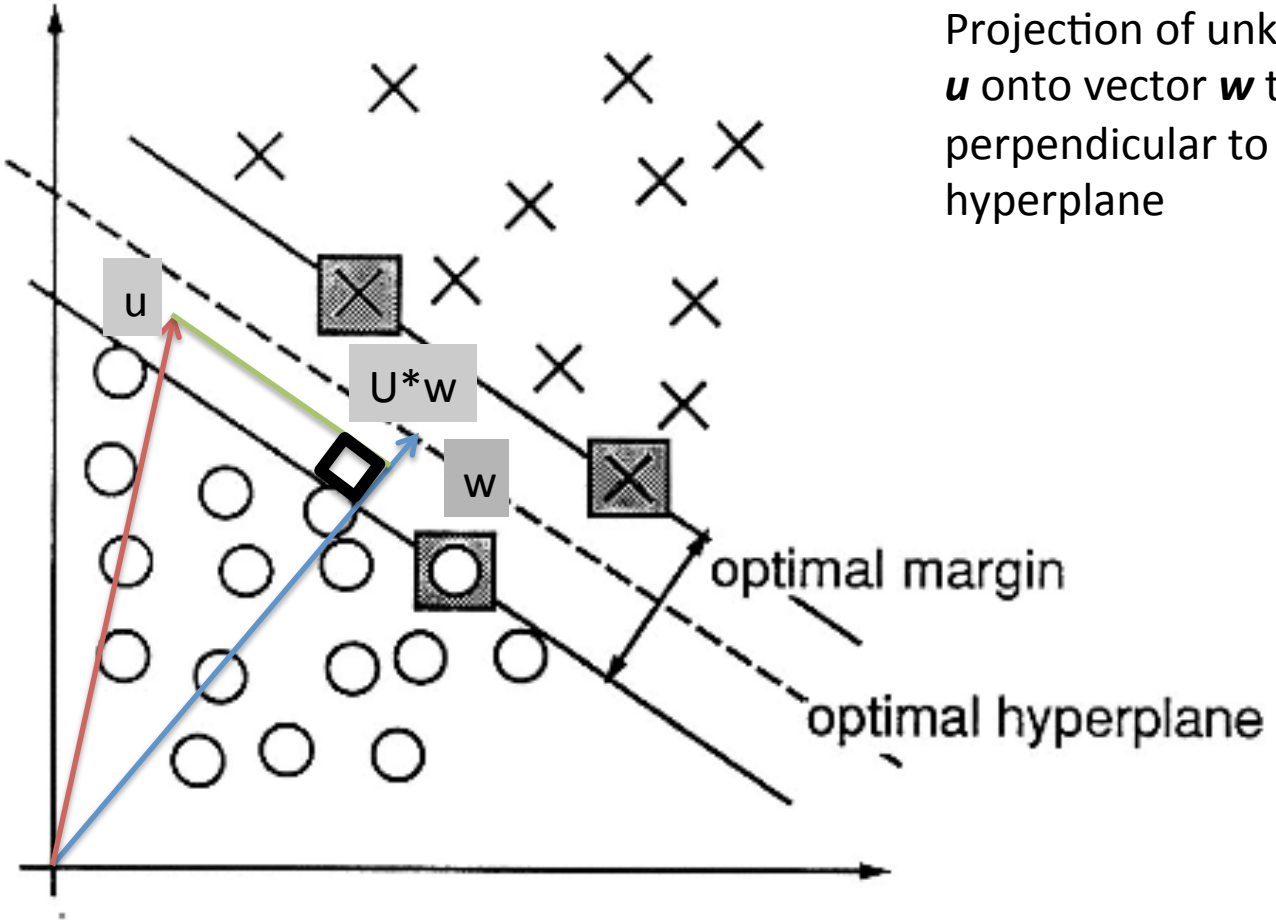
How SVM works



Unknown Data (1)

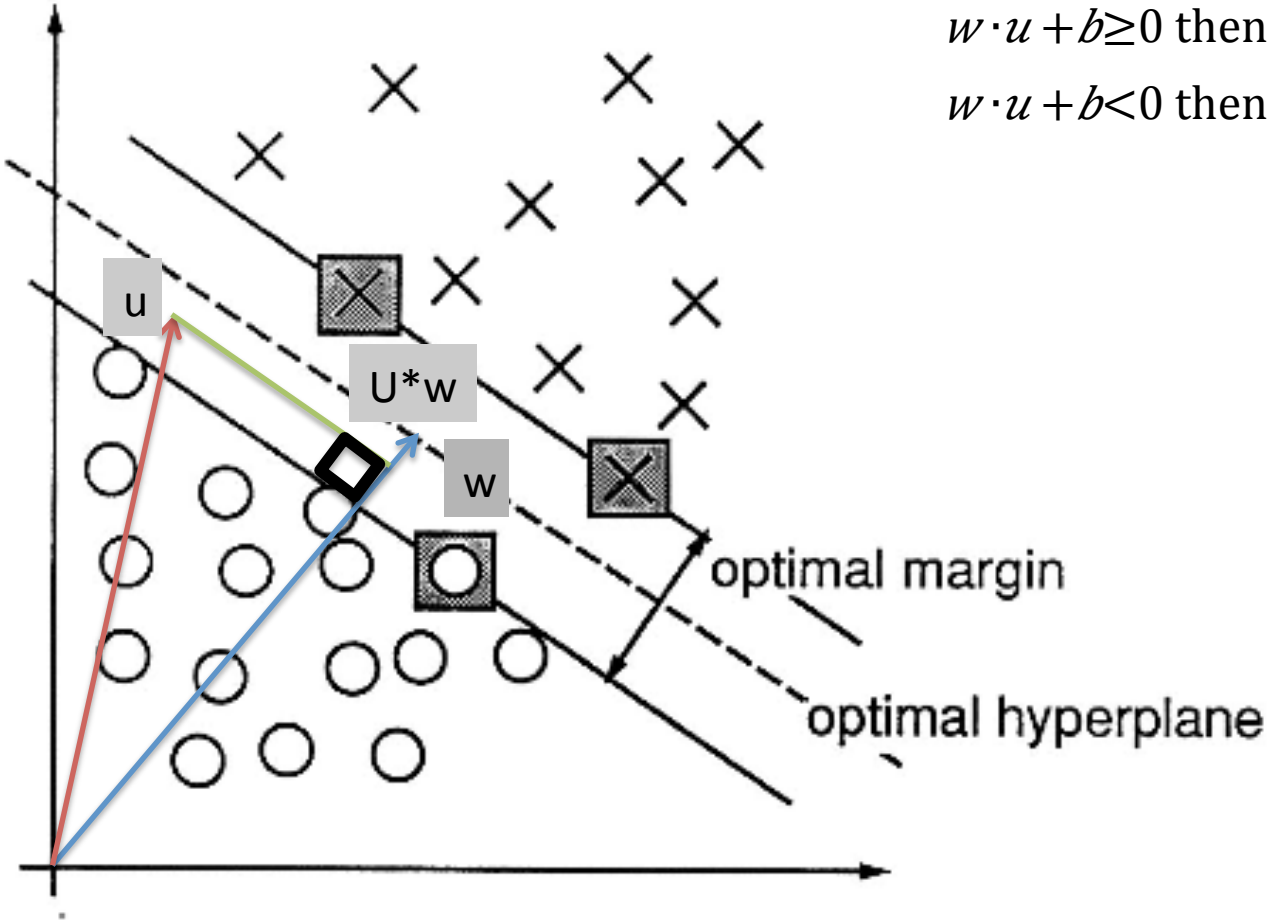


Unknown Data (2)



Projection of unknown item ***u*** onto vector ***w*** that is perpendicular to the hyperplane

SVM Decision Rule

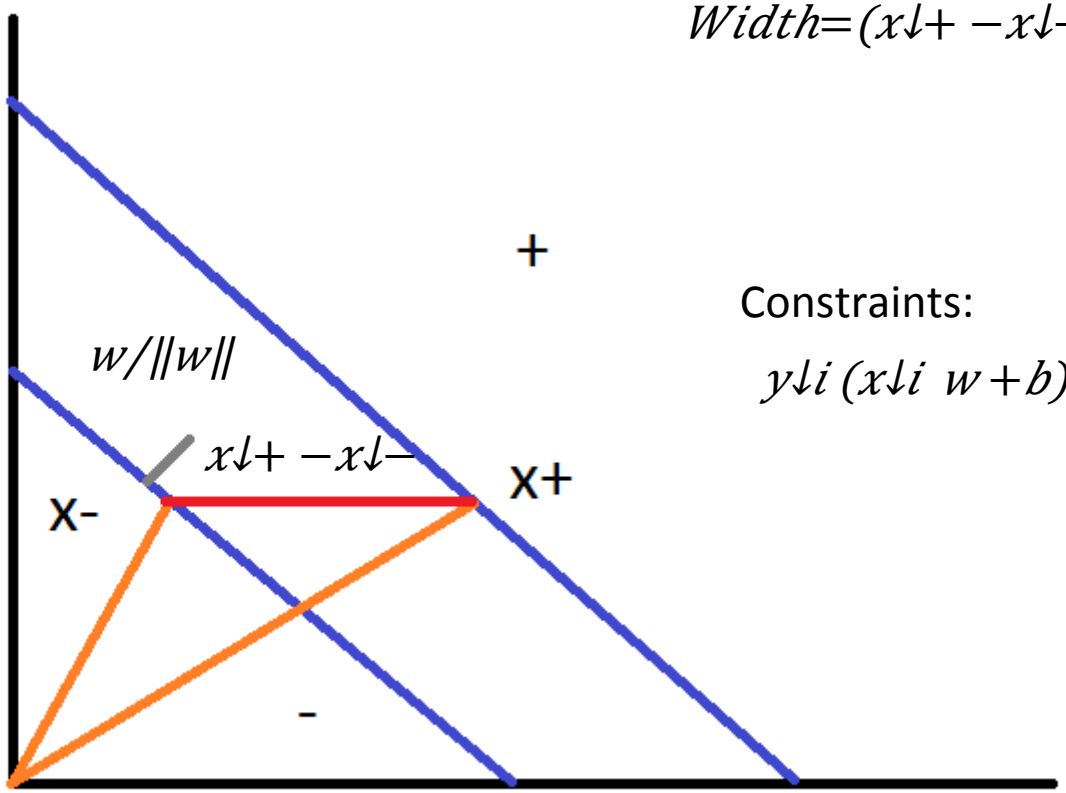


$w \cdot u + b \geq 0$ then X↓+

$w \cdot u + b < 0$ then X↓−

Learning SVM-Minimizing w

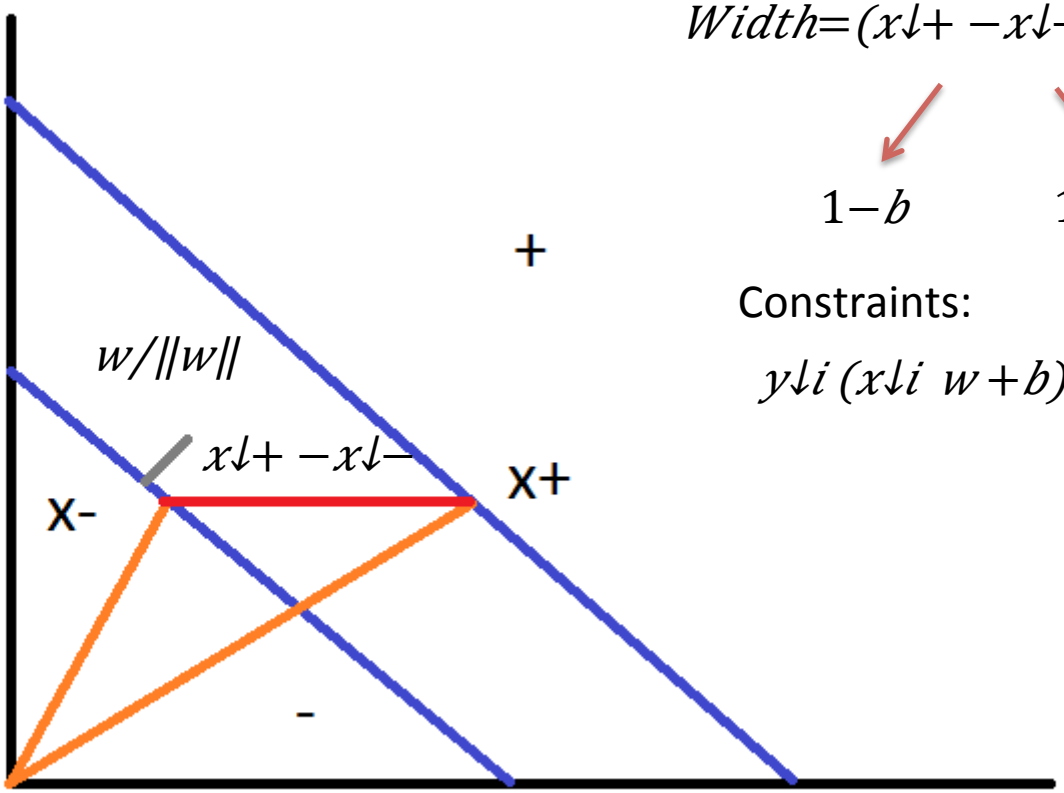
$$Width=(x\downarrow+ - x\downarrow-)\cdot w/\|w\|$$



Constraints:

$$y\downarrow i\ (x\downarrow i\ \ w+b)-1=0$$

Learning SVM-Minimizing w



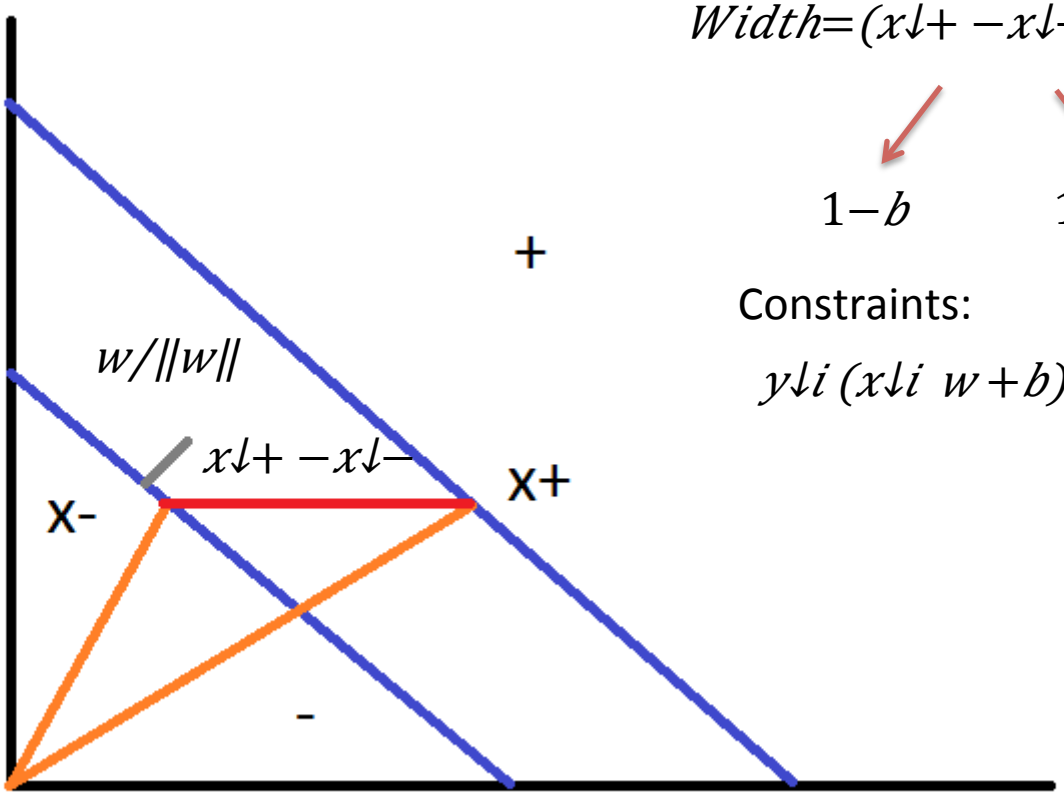
$$Width=(x\downarrow+ - x\downarrow-)\cdot w/\|w\|$$

$1-b$ $1+b$

Constraints:

$$y\downarrow i\ (x\downarrow i\ \ w+b)-1=0$$

Learning SVM-Minimizing w



$$Width=(x\downarrow+ - x\downarrow-)\cdot w/\|w\| = 2/\|w\|$$

$1-b$ $1+b$

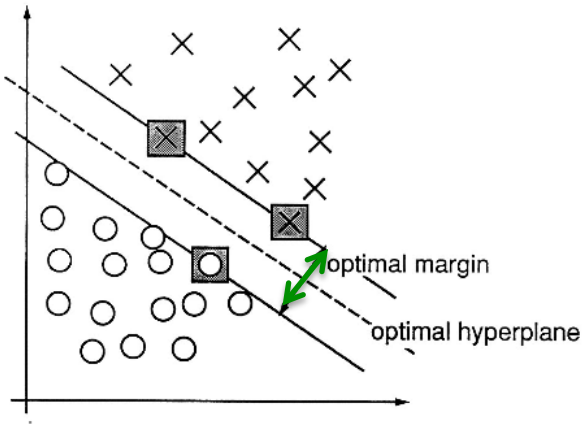
Constraints:

$$y\downarrow i\ (x\downarrow i\ \ w+b)-1=0$$

Learning SVM – Minimizing w

Distance between projections of training data:

$$p(w,b) = \min_{\{x:y=1\}} \frac{x \cdot w}{|w|} - \max_{\{x:y=-1\}} \frac{x \cdot w}{|w|}$$



When maximizing this distance:

$$p(w_0,b_0) = \frac{2}{|w_0|} = \frac{2}{\sqrt{w_0 \cdot w_0}}$$

Minimize this

Learning SVM – Penalizing misclassification

Hinge Loss Function

$$C\sum_{i}^{N}\max(0,1-y_i f(x_i))$$

Primal Form

$$f(x) = w^T x + b \quad \leftarrow \text{Classifier}$$

For $w \in \Re^d$:

$$\min_{w \in \Re^d} \| w \|^2 + C \sum_i^N \max(0, 1 - y_i f(x_i))$$

Minimize w
Maximize margin

Penalizing misclassification
(Hinge Loss)

Challenges

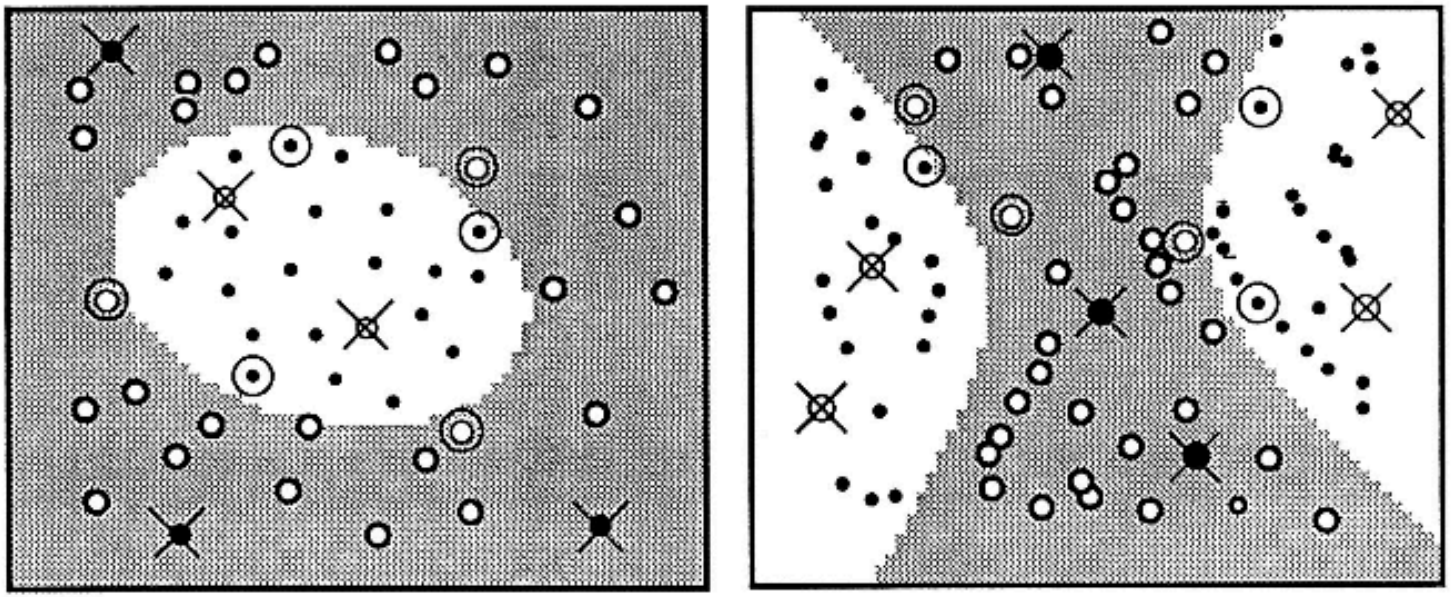


Figure 5. Examples of the dot-product (39) with $d = 2$. Support patterns are indicated with double circles, errors with a cross.

1. Handling error (slack vars.)
2. Handling non-linearly separable data (kernels)

1. Handling Error - Slack Variables

$$\xi_i \geq 0$$

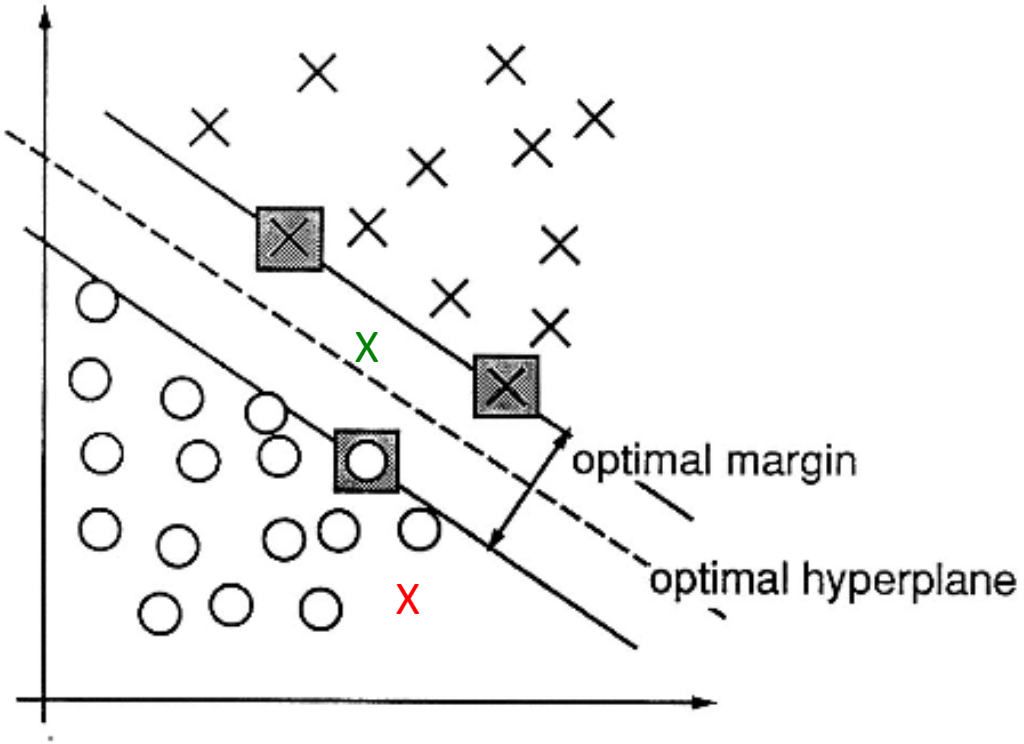
All data points

$$0 < \xi \leq 1$$

Inside the margin

$$\xi > 1$$

Misclassified



Slack Formulation

$$\min_{w \in \Re, \xi_i \in \Re+} \| w \|^2 + C \sum_{i}^{N} \xi_i$$

Subject to

$$y_i (w^T x_i + b) \geq 1 - \xi_i \qquad \text{For i = 1…N}$$

2. Non-Linear Separation - Dual Form

Solution w can be written as linear combo of training data:

$$w = \sum_{j=1}^{N} a_j y_j x_j$$

Substitute w in primal classifier $f(x) = w^T x + b$

$$f(x) = \left(\sum_{j=1}^{N} \alpha_j y_j x_j \right)^T x + b = \sum_{i}^{N} \alpha_i y_i (x_i^T x) + b$$

Dual Form Problem

For $w \in \Re^N$:

$$\max_{\alpha_i \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{jk} \alpha_j \alpha_k y_j y_k (x_j^T x_k)$$

Subject to $0 \leq \alpha \leq C$ for $\forall i$, and $\sum_i \alpha_i y_i = 0$

Kernel Trick

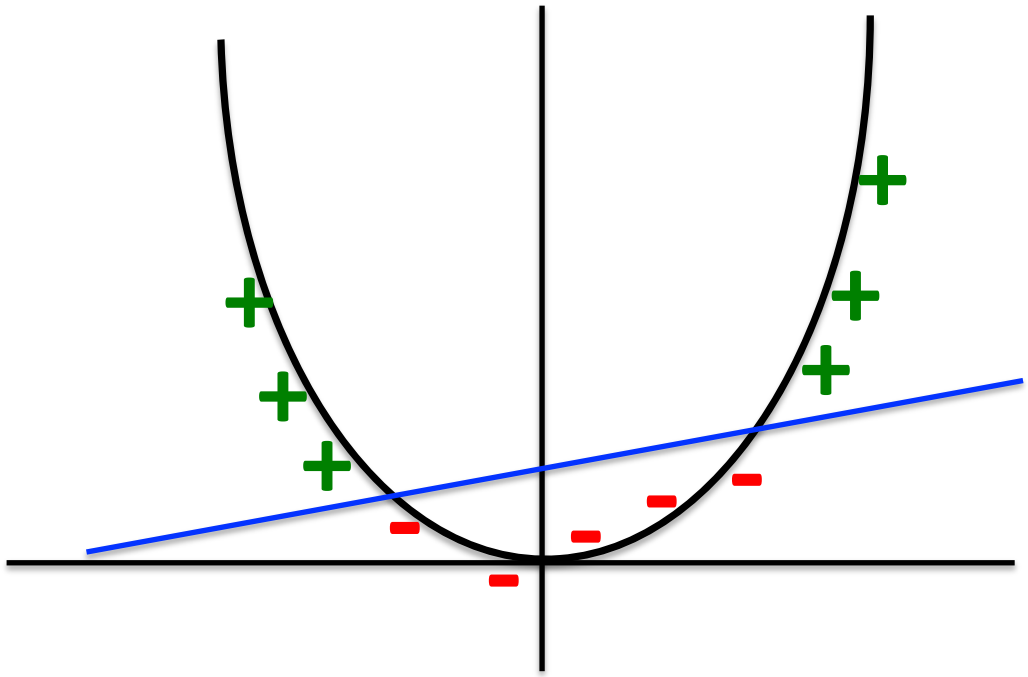
Dual Form Classifier: $f(x) = \sum_i \alpha_i y_i (x_i^T x) + b$

Kernel Classifier: $f(x) = \sum_i \alpha_i y_i k(x_i, x) + b$, where $k(x_i, x) = (x_i^T z)$

Knowledge of inner product is key

Example: Polynomial Kernel

$$k(x, x') = (1 + x^T x')^2$$



Experiments - Classifying Numbers

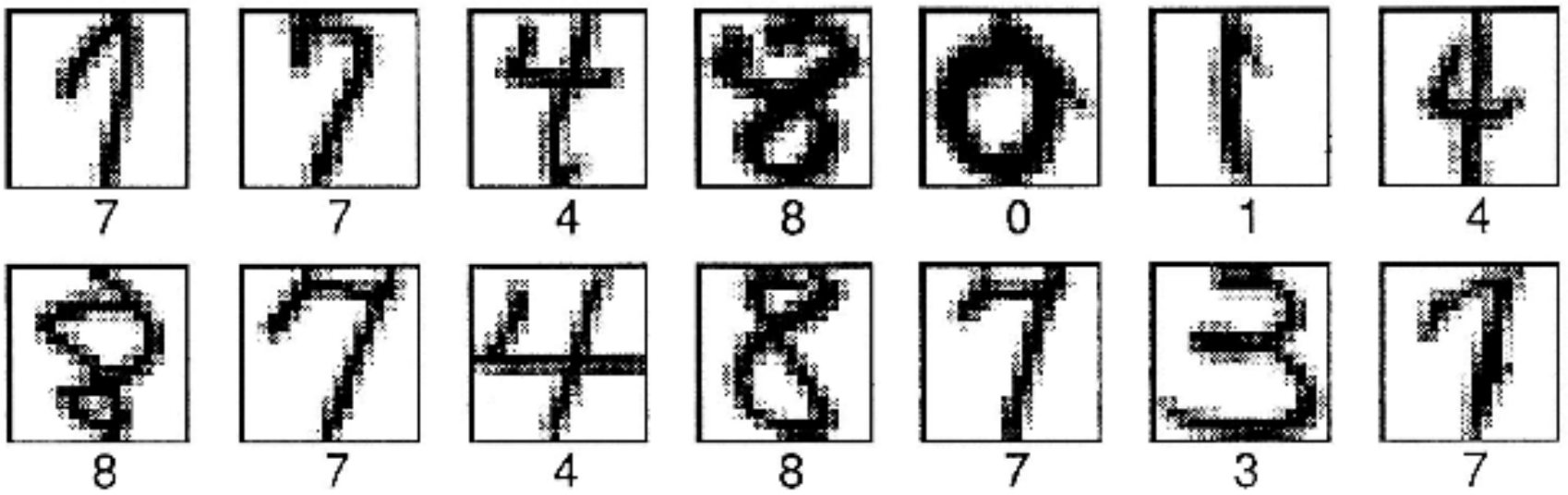


Figure 6. Examples of patterns with labels from the US Postal Service digit database.

- Postal (16x16 pxls): 7,300 training, 2,000 test
- NIST (28x28 pxls): 60,000 training, 10,000 test

Error remains constant with increasing feature space size

Table 2. Results obtained for dot products of polynomials of various degree. The number of "support vectors" is a mean value per classifier.

Degree of polynomial	Raw error, %	Support vectors	Dimensionality of feature space
1	12.0	200	256
2	4.7	127	~33000
3	4.4	148	$\sim 1 \times 10^6$
4	4.3	165	$\sim 1 \times 10^9$
5	4.3	175	$\sim 1 \times 10^{12}$
6	4.2	185	$\sim 1 \times 10^{14}$
7	4.3	190	$\sim 1 \times 10^{16}$

Training time?

Comparison with other classifiers

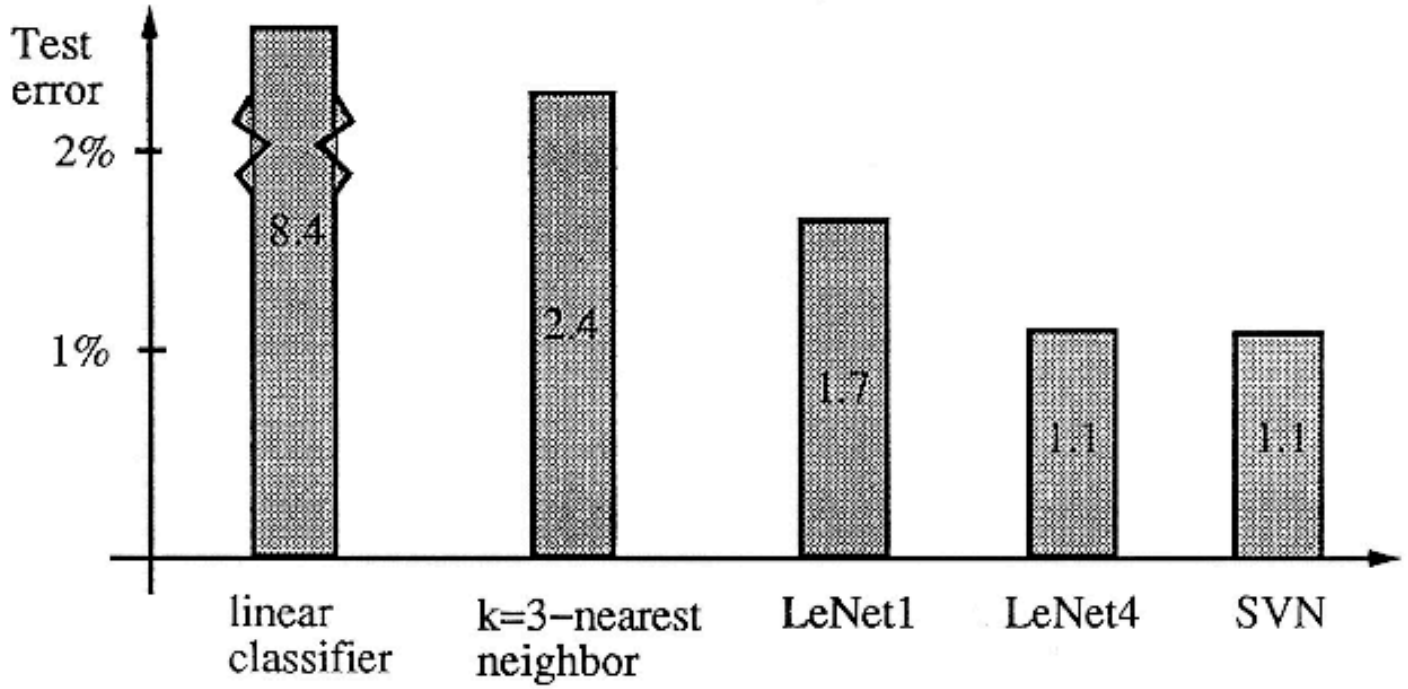


Figure 9. **Results from the benchmark study.**

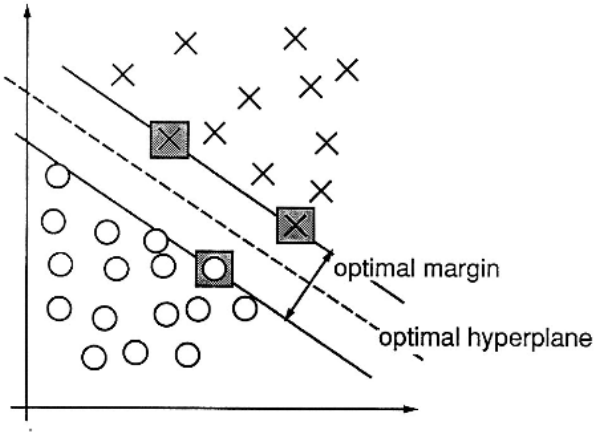
Advantages over Neural Net and kNN

- Neural Net
  - Global optimum not guaranteed
    - Non-convex cost function
  - Several parameters require tuning

- kNN
  - Curse of dimensionality

Conclusions about SVM

- Optimal hyperplane for classification
- Universal learning machine
  - Slack variables (error)
  - Kernels (non-linear separation)



- Knowledge of inner products is key

Other Resources

- Andrew Zisserman's lectures
  - http://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf
  - http://www.robots.ox.ac.uk/~az/lectures/ml/lect3.pdf
- MIT AI Course Video
  - https://www.youtube.com/watch?v=_PwhiWxHK8o