- **Review Sunday 3:00, Wu & Chen**
- **Midterm Wednesday Oct 9**
  - One two-sided “cheat sheet”
  - Covers through today (Oct 2)
  - Covers HW 4
    - No late turnin for HW4.
- **Finish the quizzes before the midterm**
- **No class Friday Oct 11 (Fall break)**
  - No office hours that Thus/Fri/Sat/Sun

Support Vector Machines (SVMs)

Lyle Ungar

Learning objectives
Review decision boundaries
Hinge loss
Margin
SVM – primal and dual

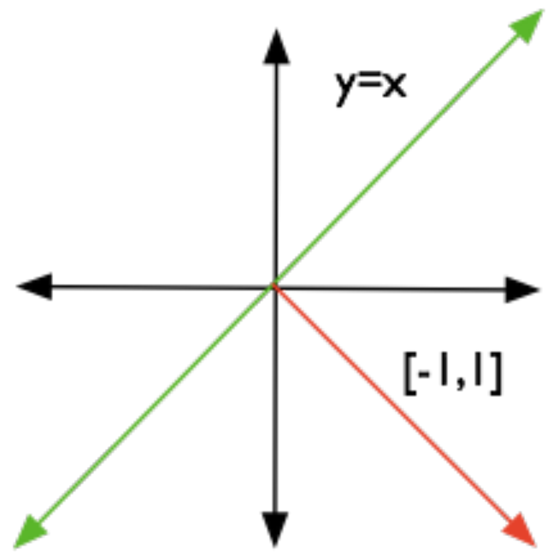
Representing Lines

- How do we represent a line?

$$y = x$$
$$0 = x - y$$
$$0 = [1, -1] \begin{bmatrix} x \\ y \end{bmatrix}$$

- In general, a hyperplane is defined by

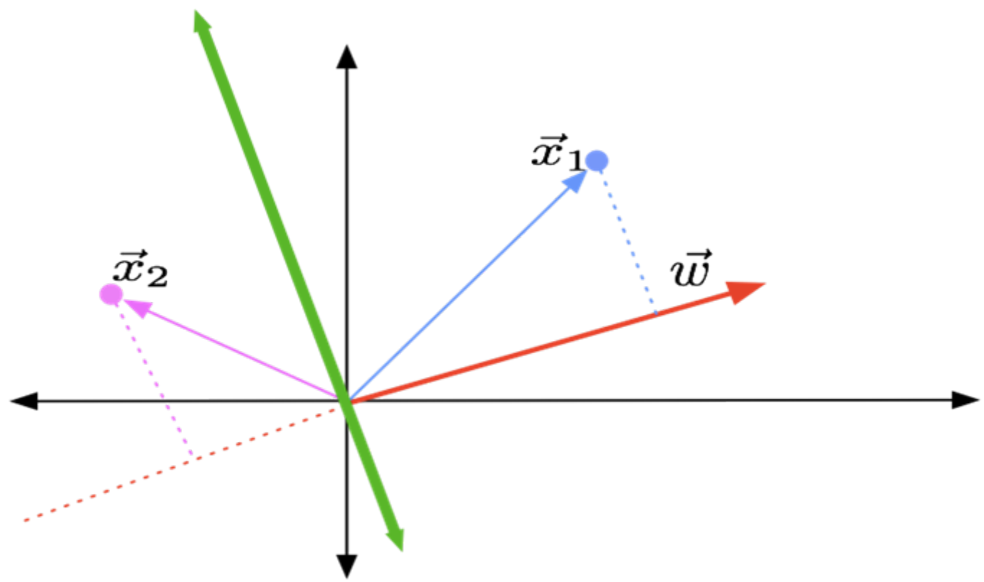
$$0 = \vec{w} \cdot \vec{x}$$



The red vector (w) defines the green plane that is orthogonal to it.

Why bother with this weird representation?

Projections



$(\vec{w} \cdot \vec{x})\vec{w}$ is the projection of $\vec{x}$ onto $\vec{w}$

alternate intuition: recall the dot product of two vectors is simply the product of their lengths and the cosine of the angle between them

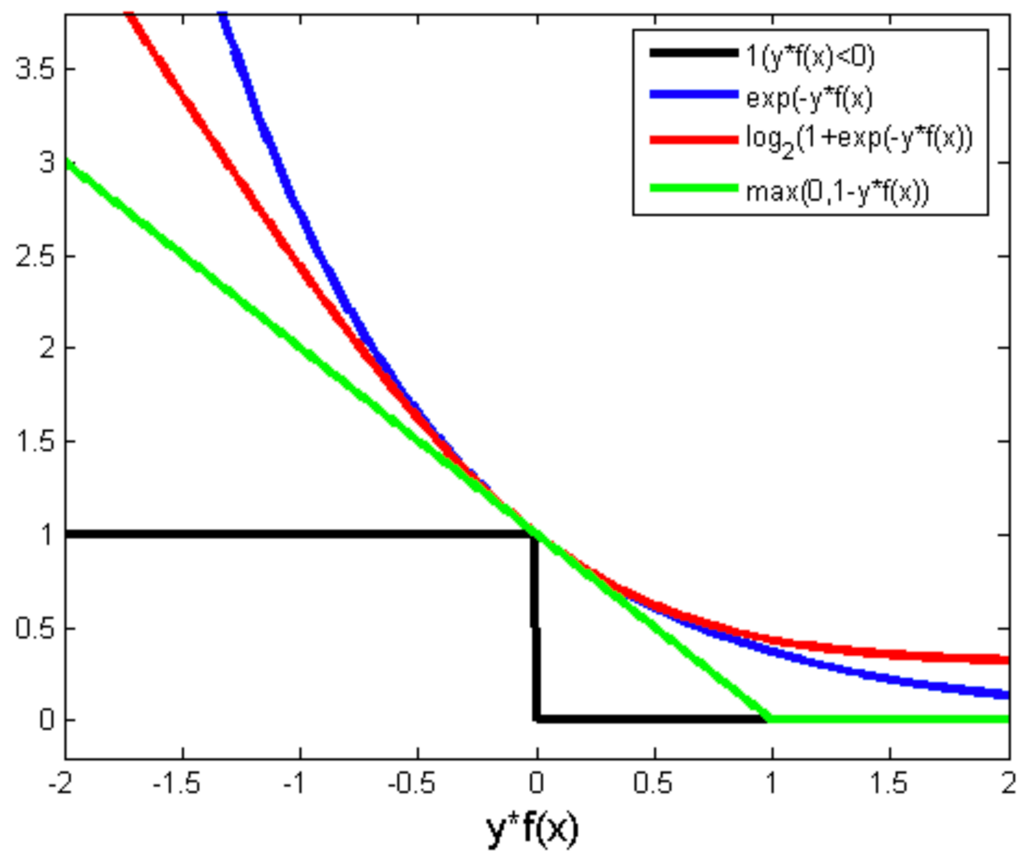
Now classification is easy!

- **Input: x**
- **Model: w**
- **Score:** $\mathbf{w}^T\mathbf{x}$
- **Prediction:** $\mathrm{sgn}(\mathbf{w}^T\mathbf{x})$
- But how do we learn ***w***?

Support Vector Machines (SVMs)

- **Minimize hinge loss**
  - With regularization
- **”Large margin” methods**

Loss functions for classification



Hinge loss

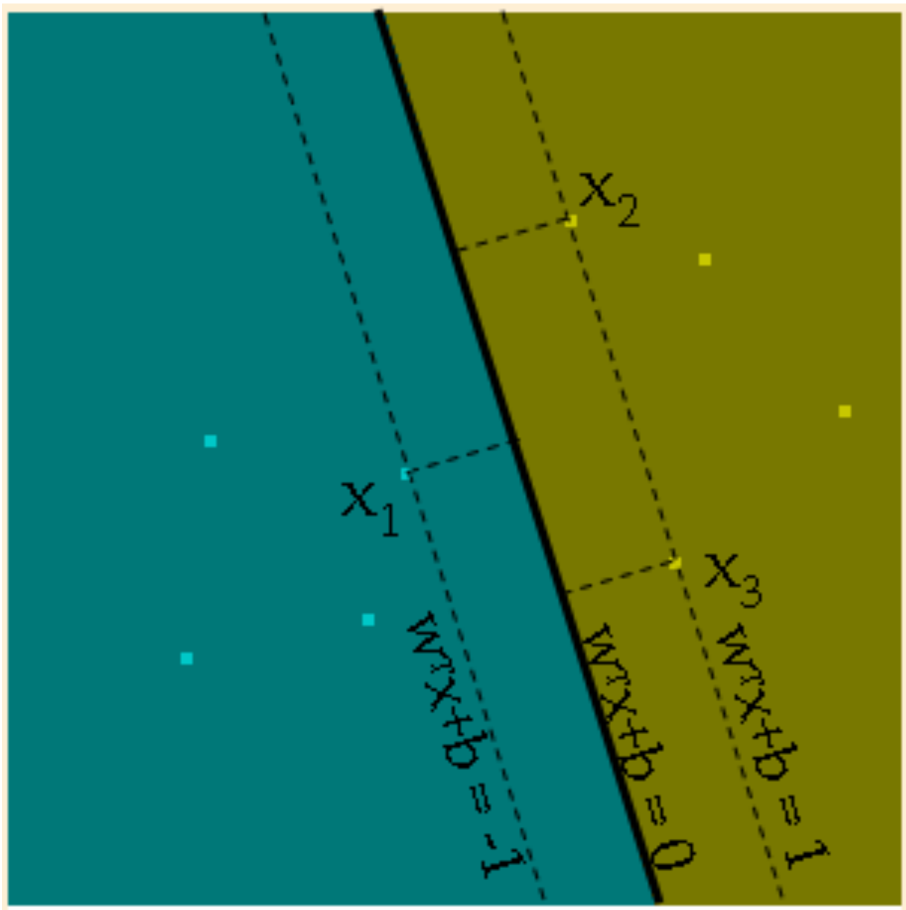
SVM: Hinge loss, ridge penalty

$h(x) = sign(w^T x + b)$

$$\min_{\mathbf{w},b,\xi\geq 0} \quad \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_i \xi_i$$

$$\xi_i = \max(0, 1 - y_i(\mathbf{w}^\top\mathbf{x}_i + b))$$

0 if score is right by 1 or more
(hinge loss)



$$h(\mathbf{x}) = sign(\mathbf{w}^\top \mathbf{x} + b)$$

Arbitrarily normalize

$$\mathbf{w}^\top \mathbf{x} + b = \pm 1$$

Compute margin

$$\mathbf{w}^\top \mathbf{x}_1 + b = -1 \quad \text{and} \quad \mathbf{w}^\top \mathbf{x}_2 + b = 1$$

Maximize margin

$$\mathbf{w}^\top(\mathbf{x}_2 - \mathbf{x}_1) = 2 \quad \rightarrow \quad \frac{\mathbf{w}^\top}{2||\mathbf{w}||_2}(\mathbf{x}_2 - \mathbf{x}_1) = \frac{1}{||\mathbf{w}||_2}$$

Max margin interp. of SVM

Separable SVM primal:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, n$$

Use Lagrange Multiplier magic

Separable SVM dual:

$$\max_{\alpha \geq 0} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0$$

$$\mathbf{w} = \textstyle\sum_i \alpha_i y_i \mathbf{x}_i$$

Most constraints are non-binding so most α_i are zero.

The x_i with nonzero α_i are support vectors.

Kernelized separable dual:

$$\max_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0$$

$$\mathbf{w}^\top \mathbf{x} + b = \textstyle\sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$

Max margin interp. of SVM

Separable SVM primal:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n$$

The non-separable case

Hinge primal:

$$\min_{\mathbf{w},b,\xi\geq 0} \quad \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i = \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

Generalize it!

Hinge primal:

$$\min_{\mathbf{w},b,\xi\geq 0} \quad \frac{1}{2}\|\mathbf{w}\|_p^p + C\|\xi\|_q^q$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

The non-separable dual

Hinge dual:

$$\max_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \leq C, \quad i = 1, \ldots, n$$

Support Vector Machines

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials . Comments and corrections gratefully received.

Andrew W. Moore

School of Computer Science

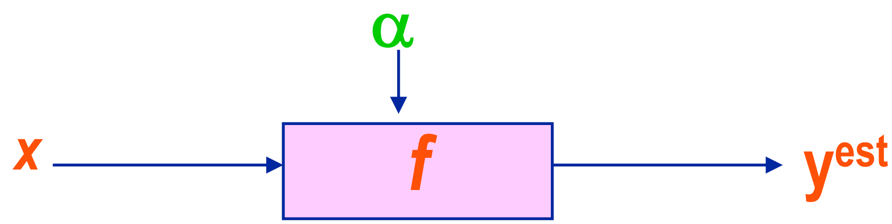
Carnegie Mellon University

www.cs.cmu.edu/~awm

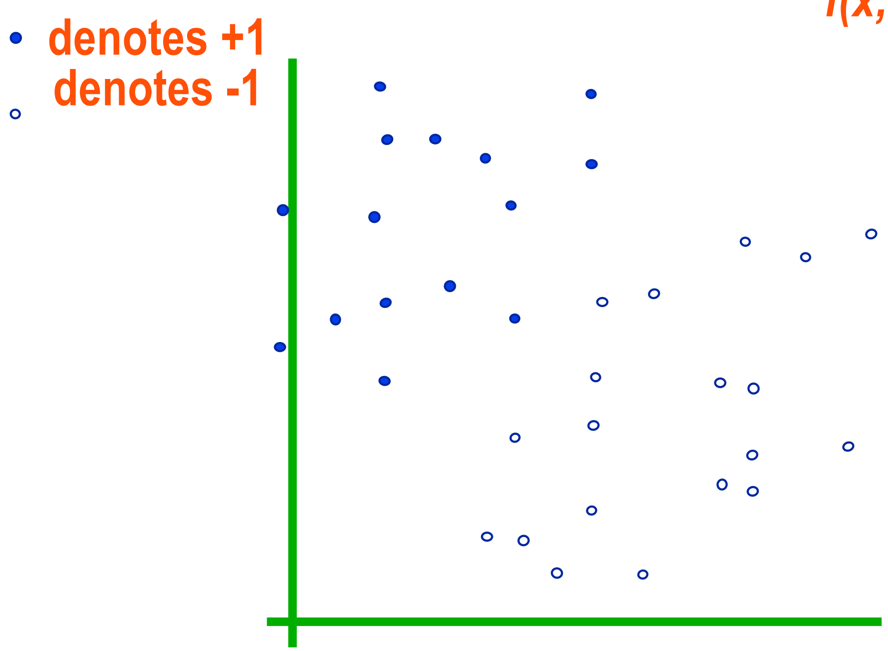
awm@cs.cmu.edu

With minor modifications by Lyle Ungar

Linear Classifiers

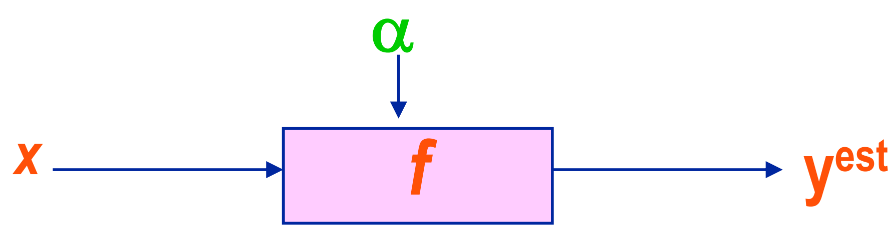


$f(x,w,b) = sign(w. x - b)$

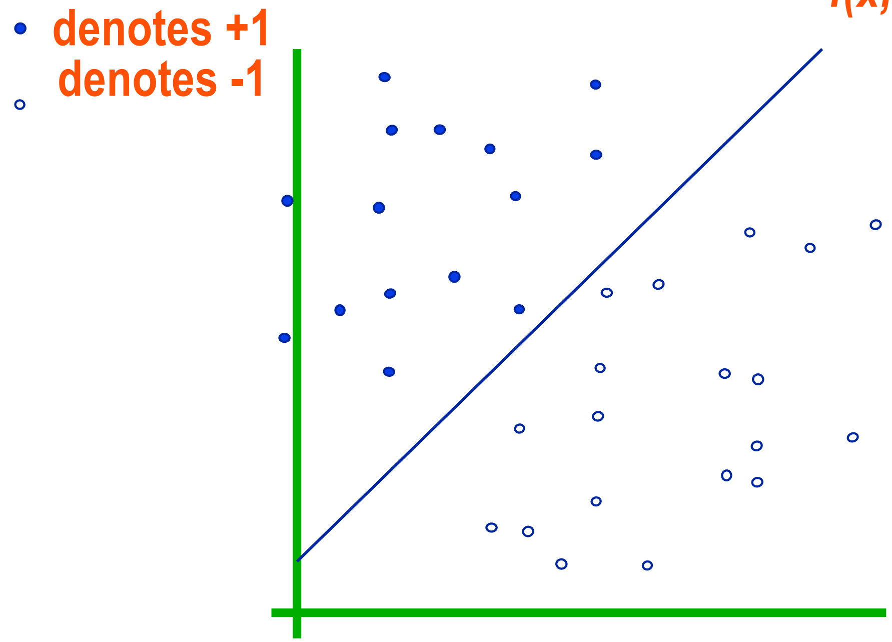


How would you classify this data?

Linear Classifiers

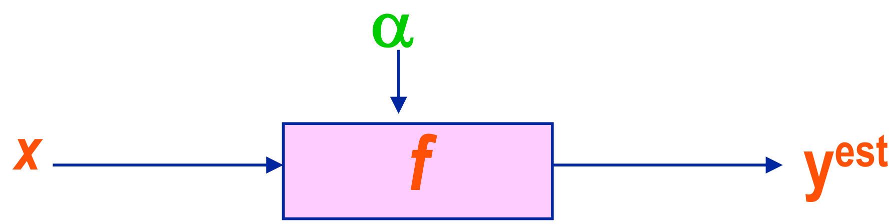


$f(x,w,b) = sign(w.\ x - b)$

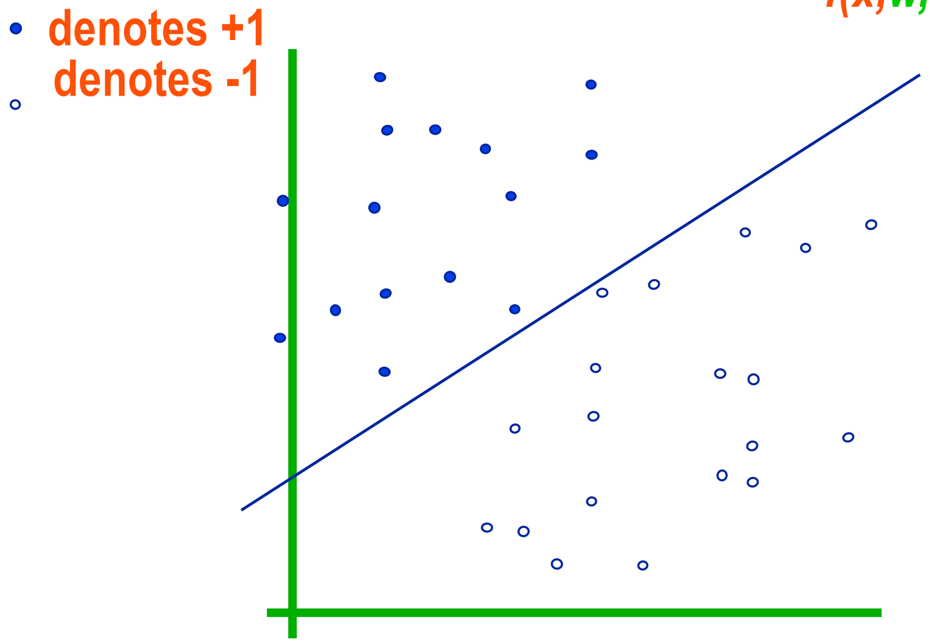


How would you classify this data?

Linear Classifiers

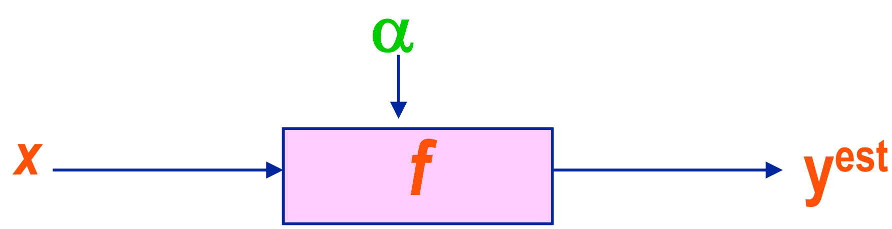


$f(x,w,b) = sign(w.\ x - b)$

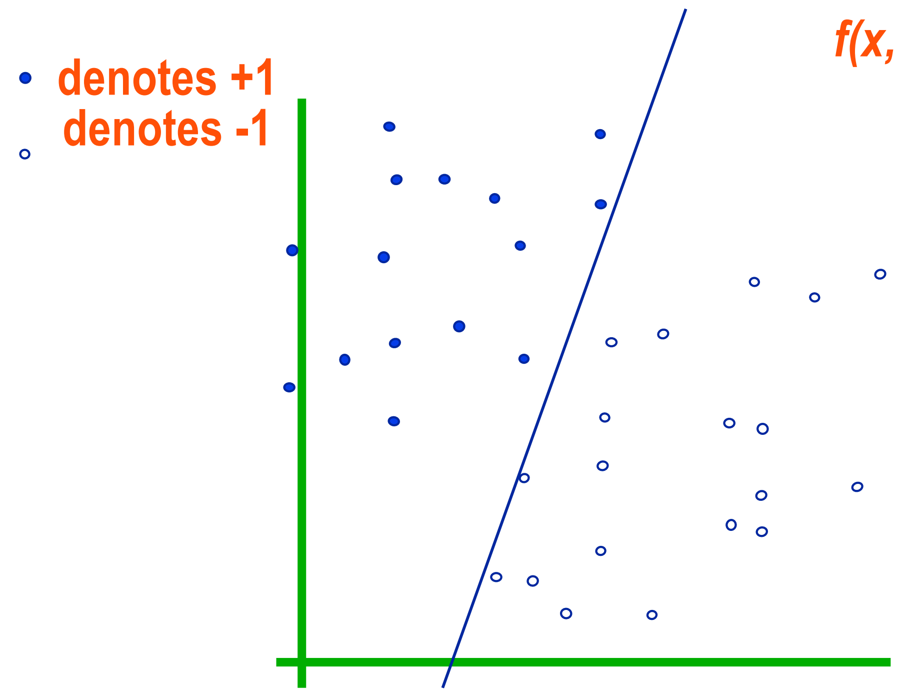


How would you classify this data?

Linear Classifiers

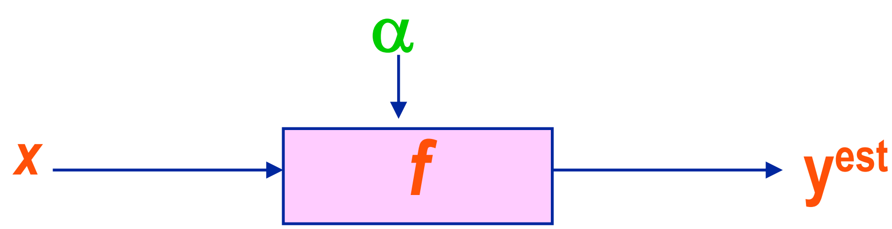


$f(x,w,b) = sign(w.\ x - b)$

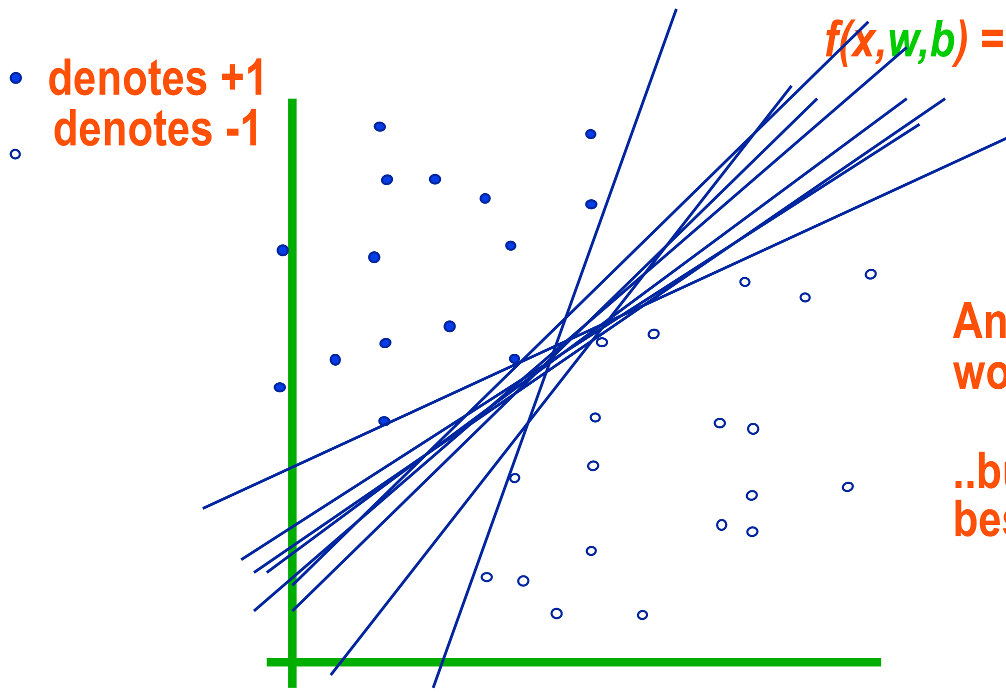


How would you classify this data?

Linear Classifiers



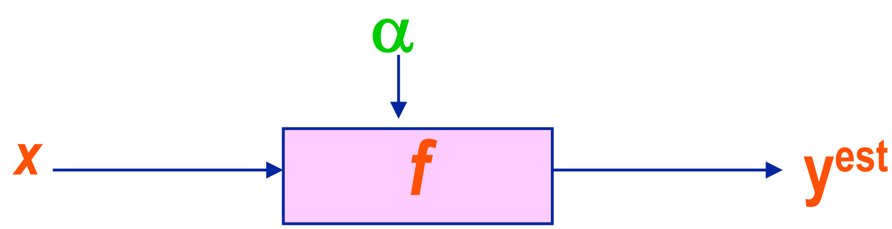
$f(x,w,b) = \mathrm{sign}(w.\ x - b)$



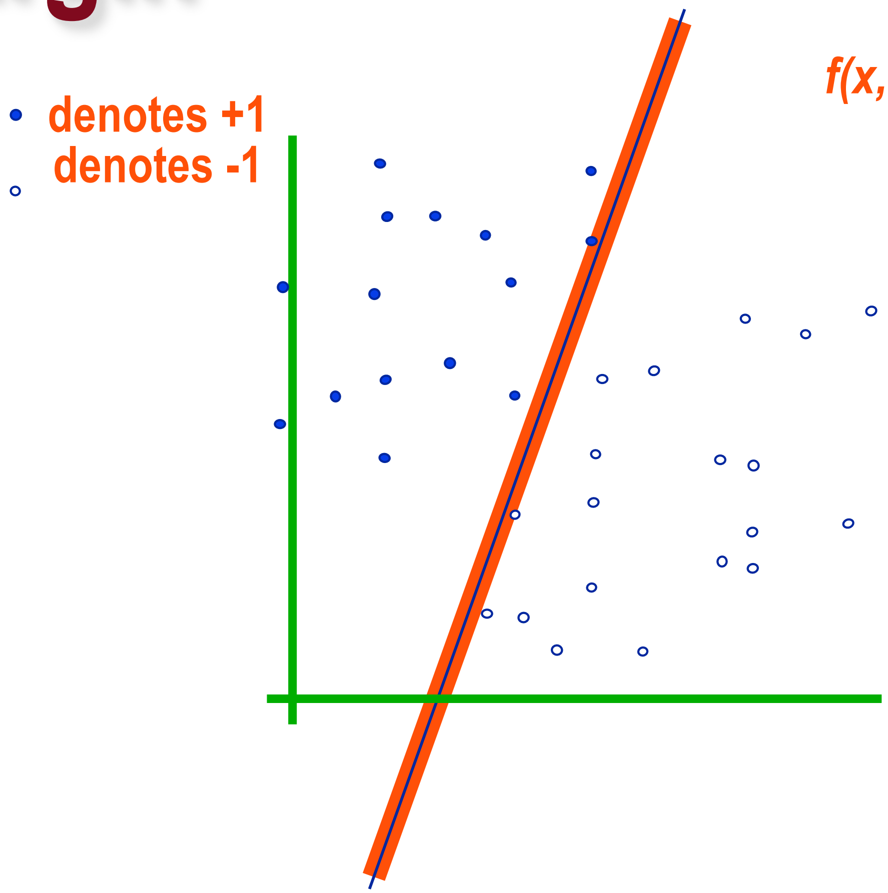
Any of these would be fine..

..but which is best?

Classifier Margin

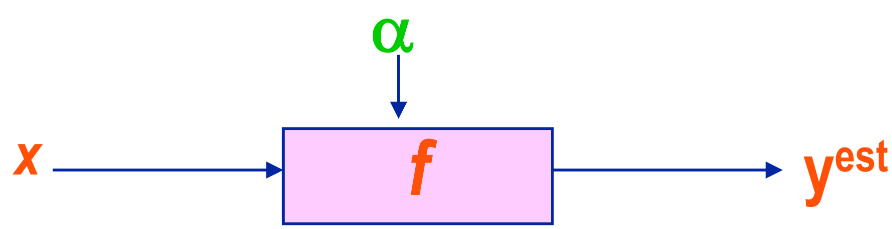


$f(x,w,b) = sign(w.\ x - b)$

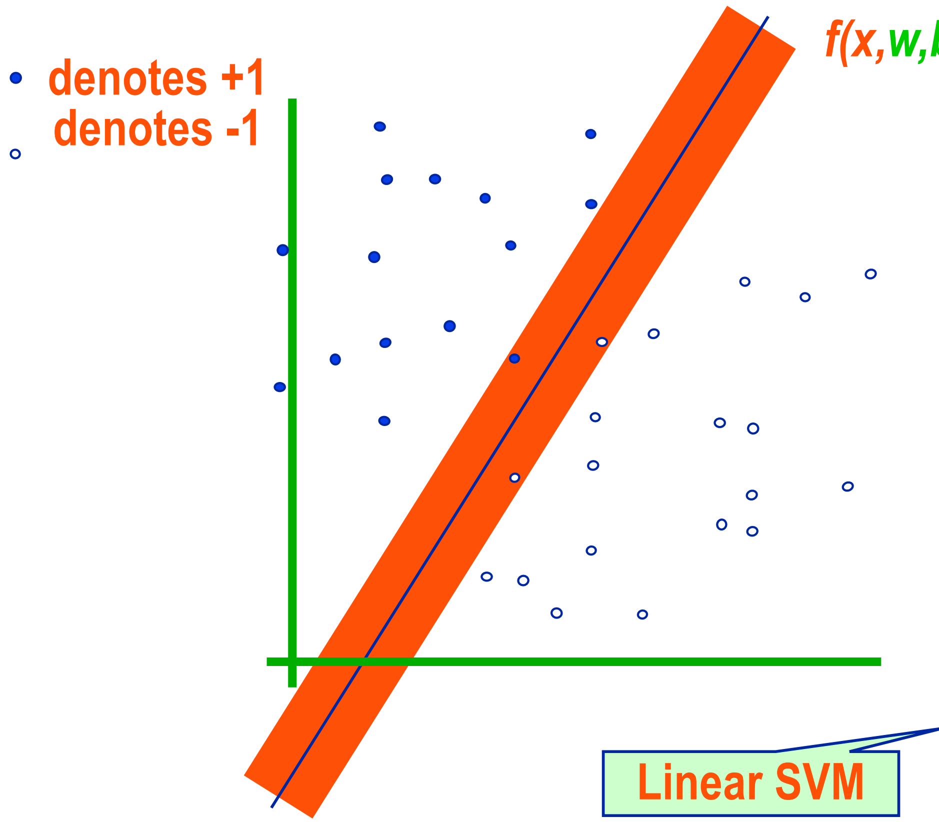


Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Maximum Margin

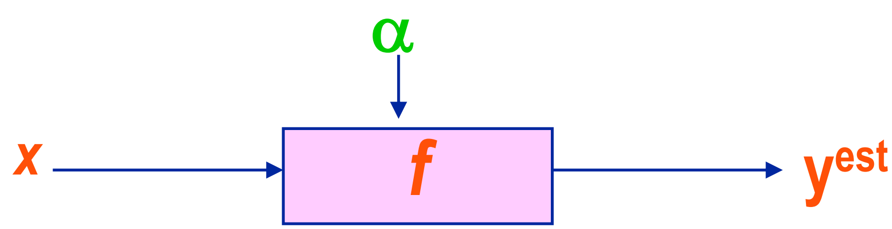


$f(x,w,b) = sign(w. x - b)$



The maximum margin linear classifier is the linear classifier with the, um, maximum margin.
This is the simplest kind of SVM (Called an LSVM)

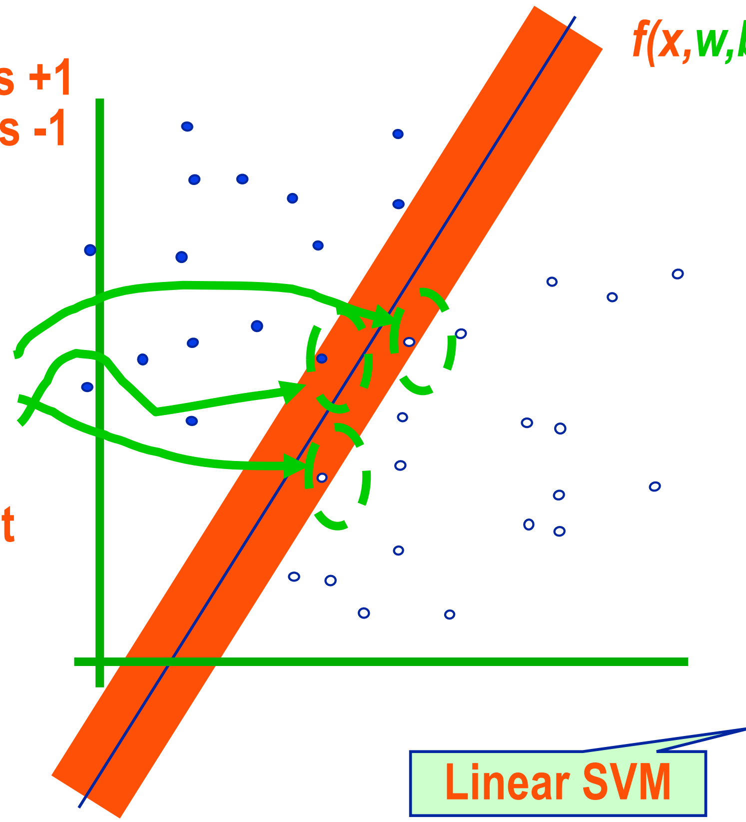
Maximum Margin



$f(x,w,b) = sign(w.\ x - b)$

• denotes +1

∘ denotes -1



Support Vectors are those datapoints that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin.
This is the simplest kind of SVM (Called an LSVM)