Applied Machine Learning

Lecture 12
Support Vector Machine

Ekarat Rattagan, Ph.D.

Slides adapted from Andrew's tutorials: http://www.cs.cmu.edu/~awm/tutorials .

Classification

x, y

Outline

Classical ML

logistic R

Decision Tree

K-NN

NN

SVM

1. Definition
2. Linear classifiers
3. How SVM works?
4. Cost function / loss function
5. Optimization

binary classification

yes/no

one vs rest

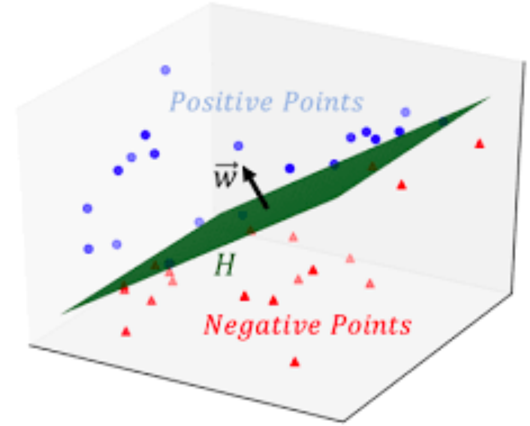
Constraint based

IoT

- better
- Mem
- Com

$\vec{x}_1$ y
$\vec{x}_2$ y
$\vec{x}_3$ y
$\vec{x}_m$ y

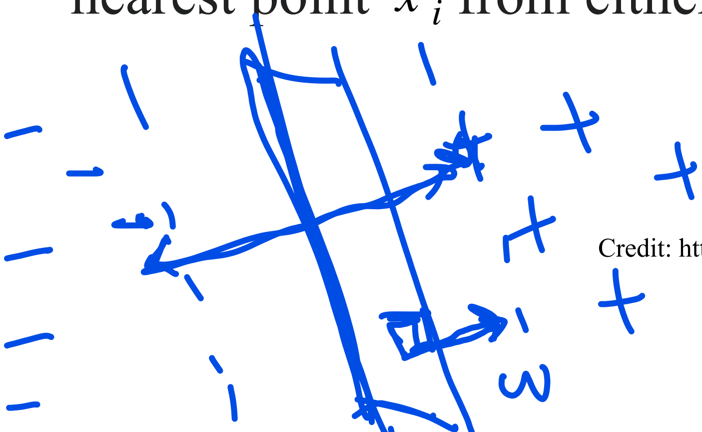
1. Definition



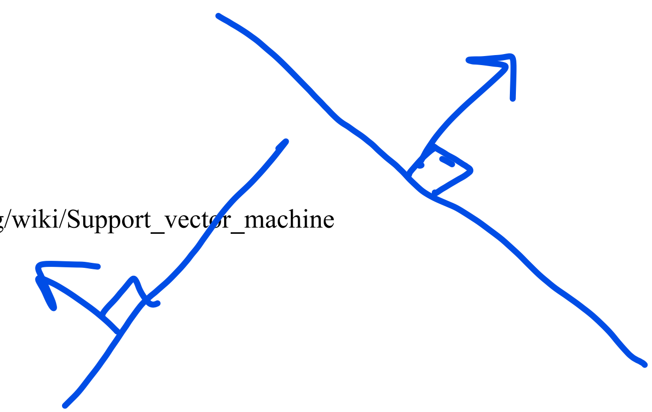
https://waterprogramming.wordpress.com/2019/01/29/intro-to-machine-learning-part-4-support-vector-machines/

Given a training dataset of points, $(\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m)$,where y_i are either +1 or −1, each indicating the class to which the point $\vec{x}_i$ belongs.

The objective is to find the "maximum-margin hyperplane" that divides the group of points $\vec{x}_i$ for which $y_i = 1$ from the group of points for which $y_i = -1$, so that the distance between the hyperplane and the nearest point $\vec{x}_i$ from either group is maximized.

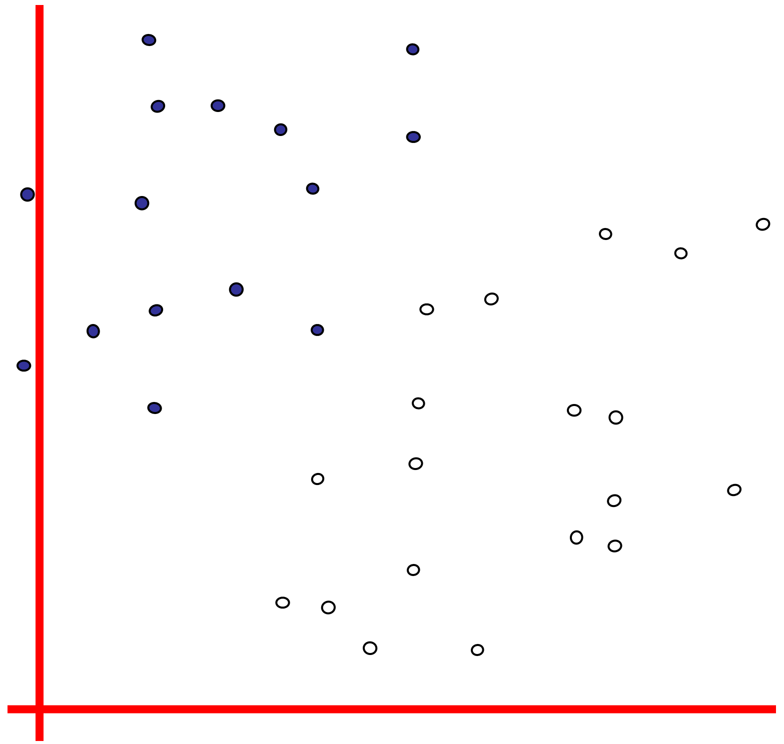


Credit: https://en.wikipedia.org/wiki/Support_vector_machine



2. Linear Classifiers

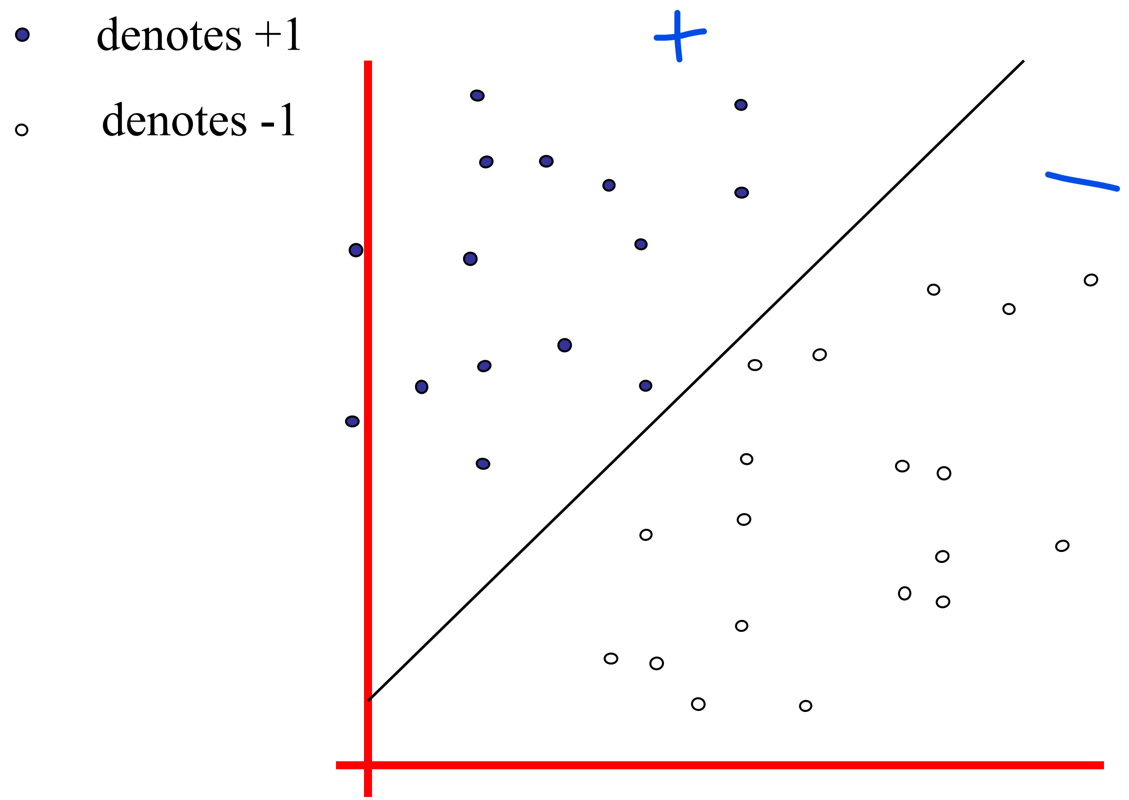
- denotes +1
- denotes -1



How would you classify this data?

Linear Classifiers

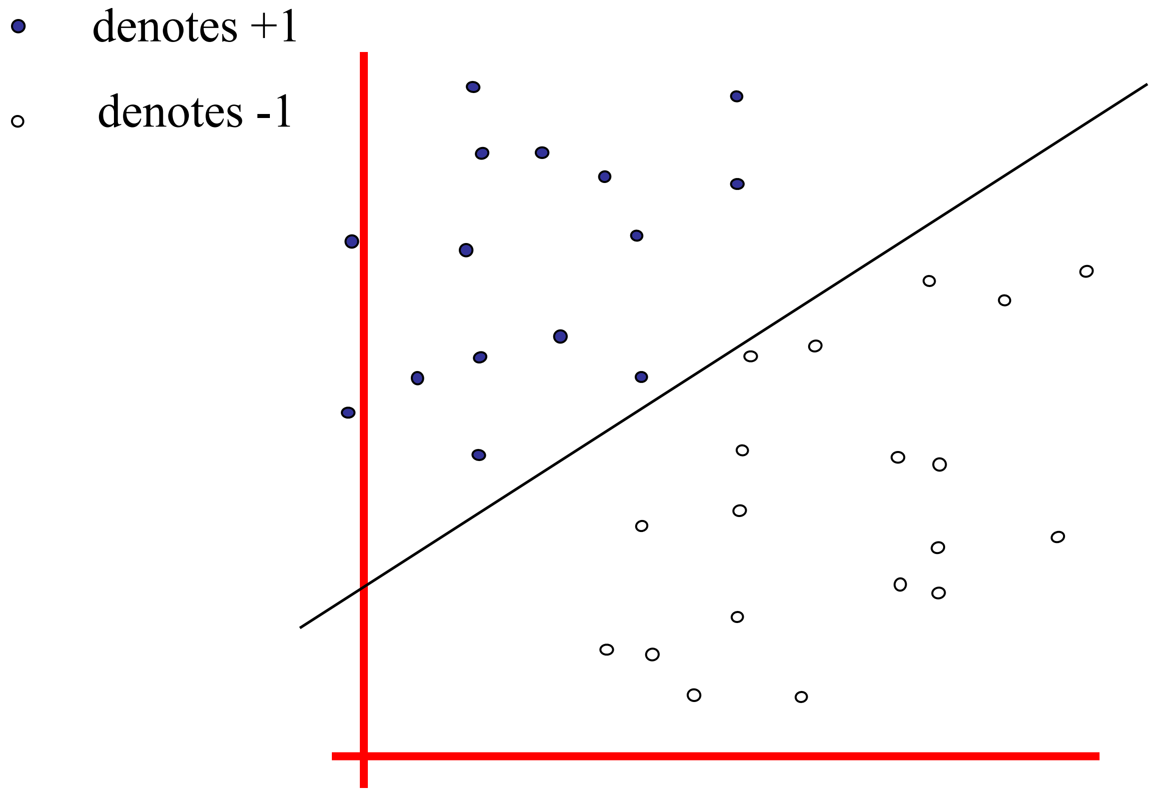
100%



How would you classify this data?

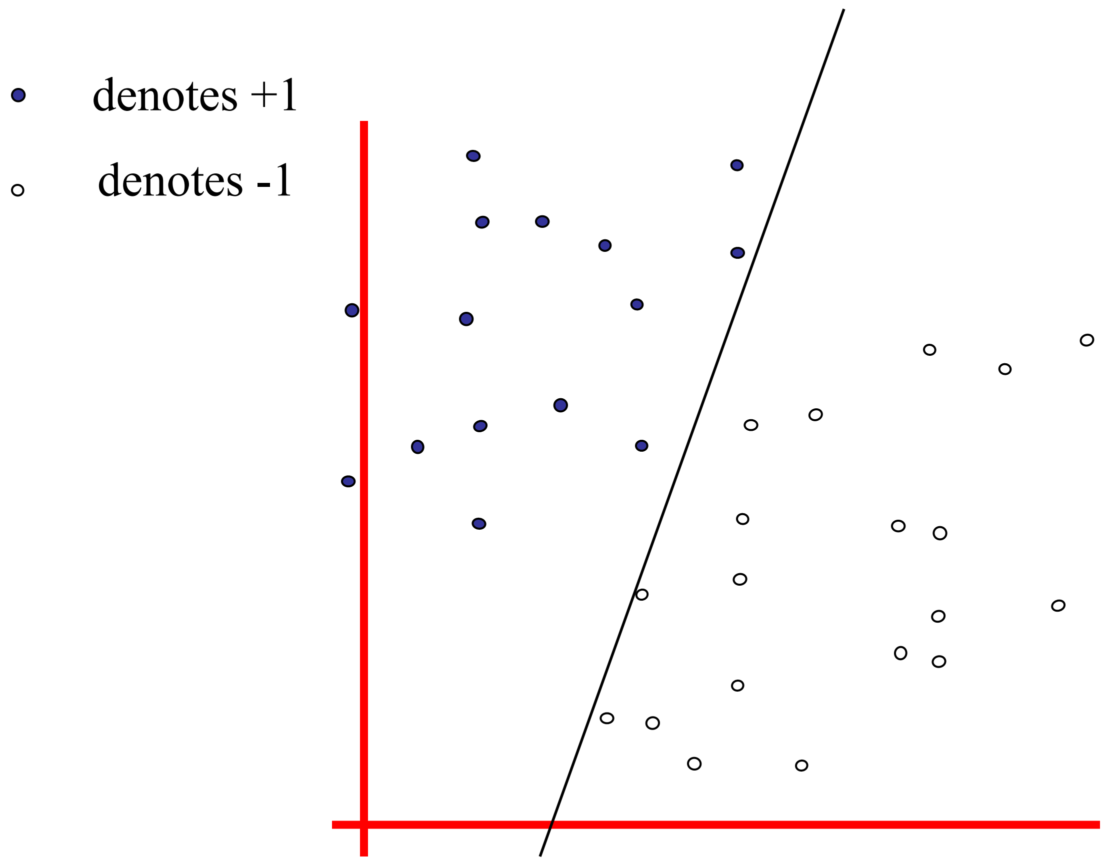
Linear Classifiers

100%
Accuracy



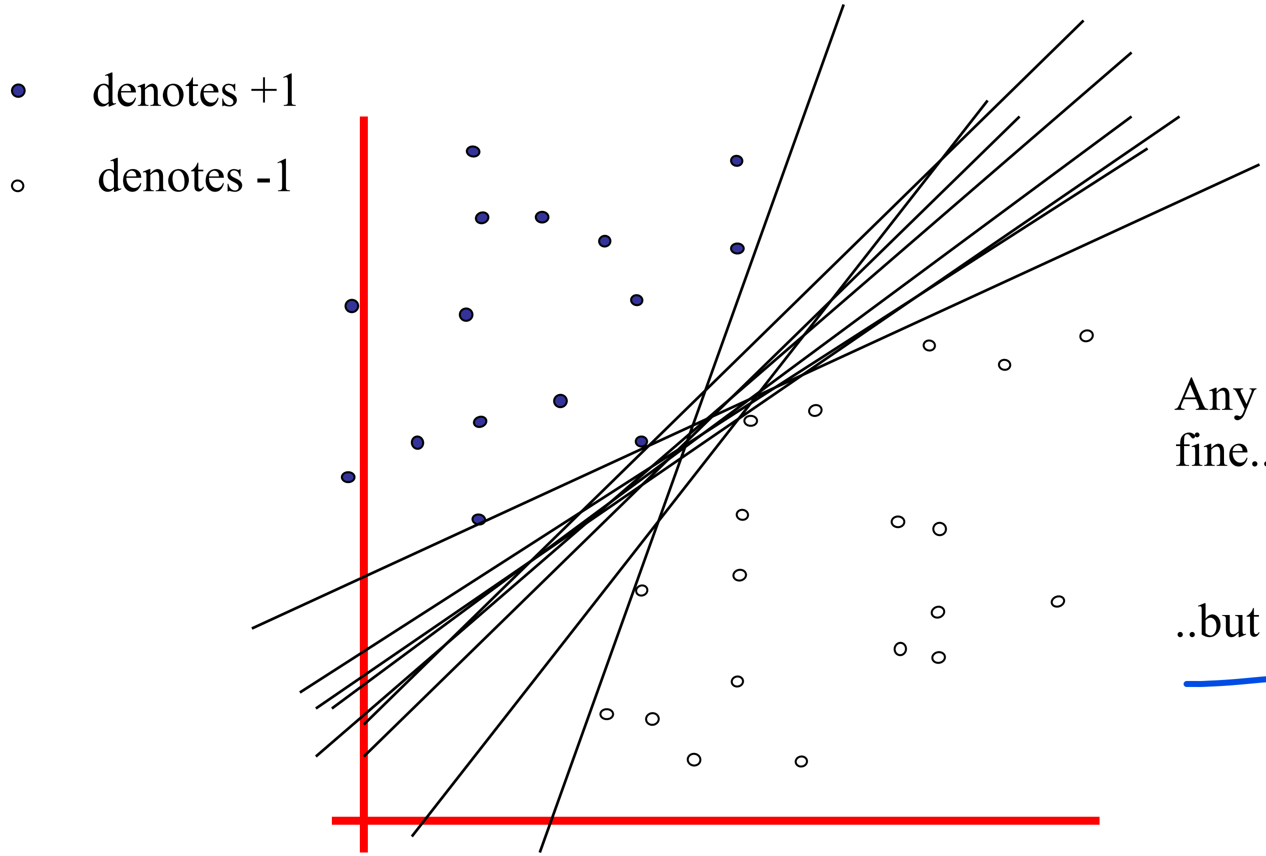
How would you classify this data?

Linear Classifiers



How would you classify this data?

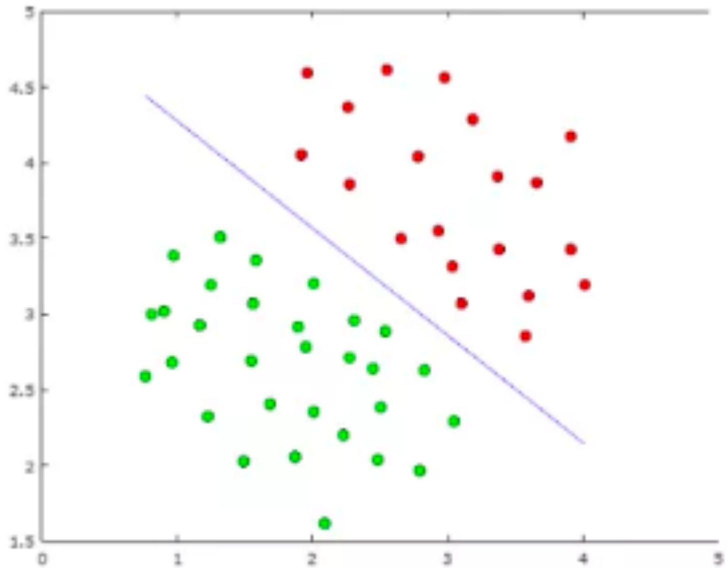
Linear Classifiers



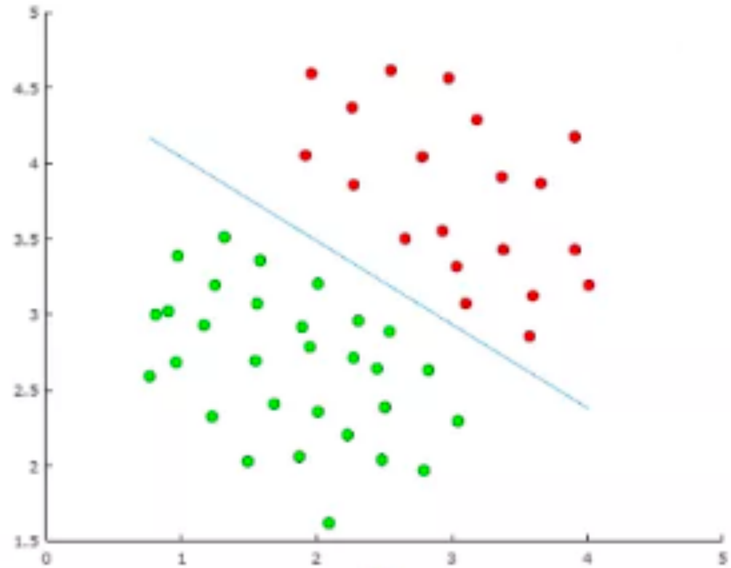
Any of these would be fine..

..but which is best?

Linear Classifiers (SVM VS Logistic Regression)

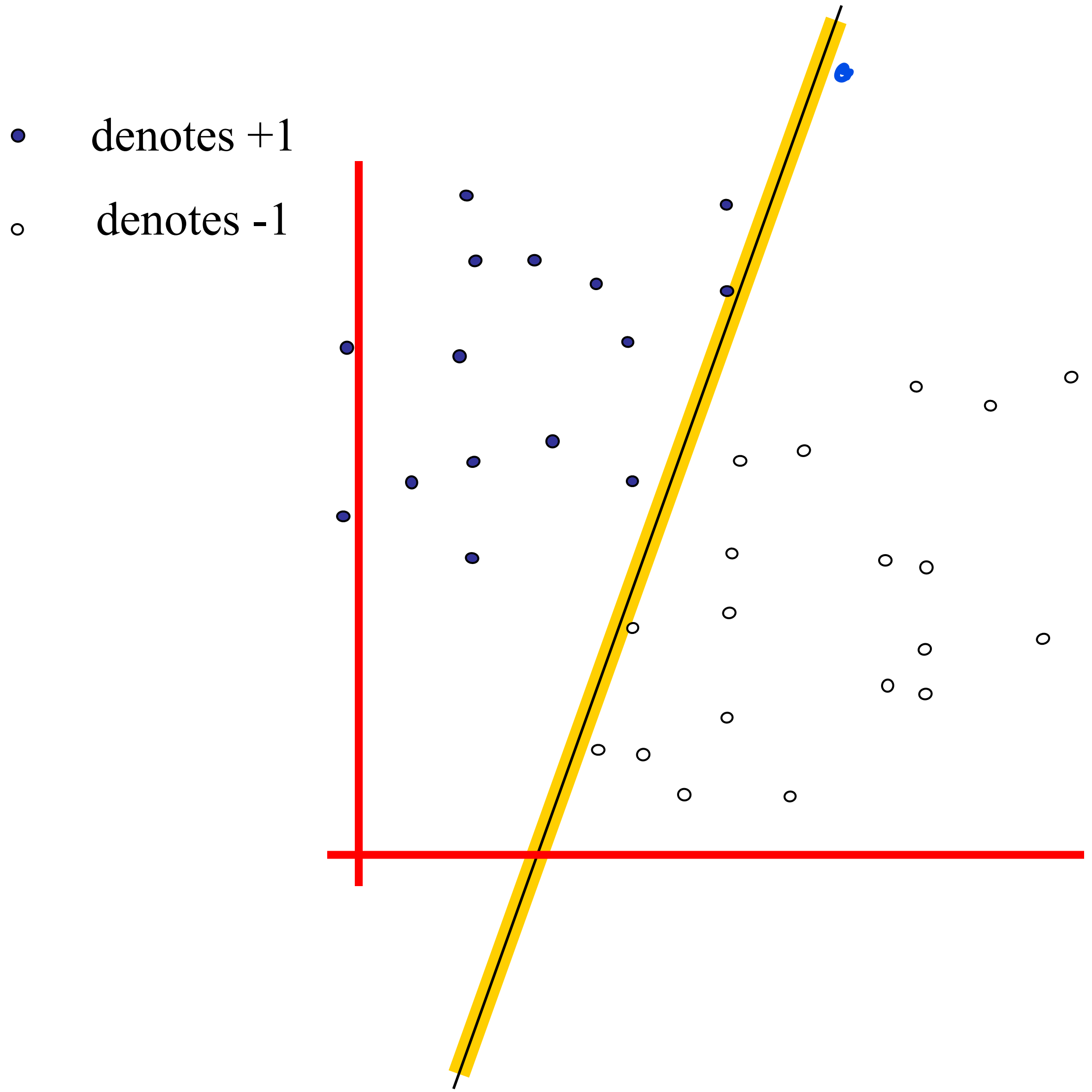


SVM



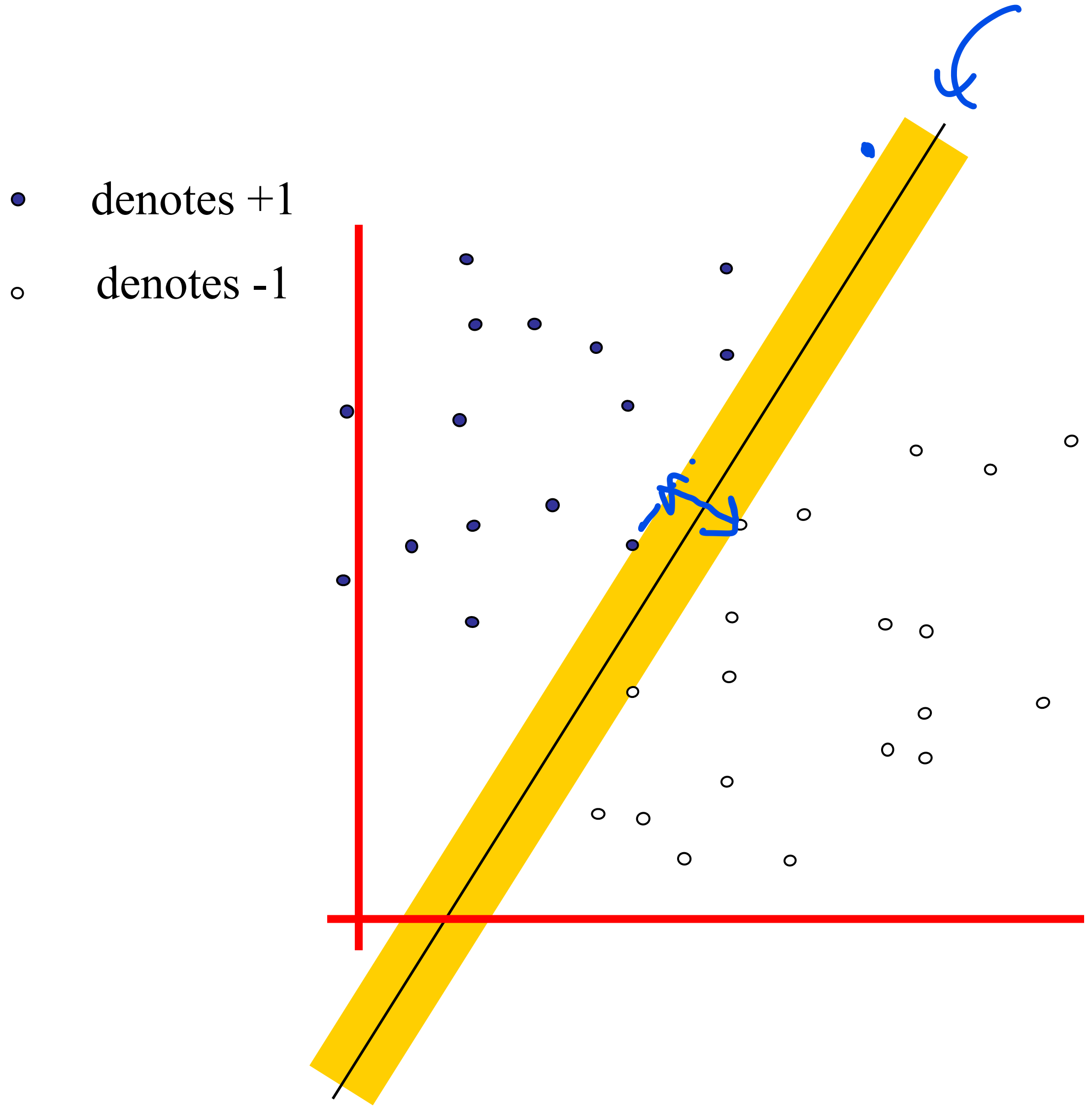
Logistic Regression

Classifier Margin



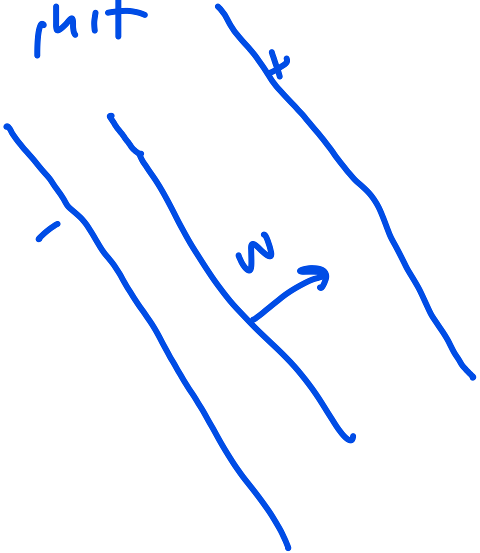
Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

Maximum Margin



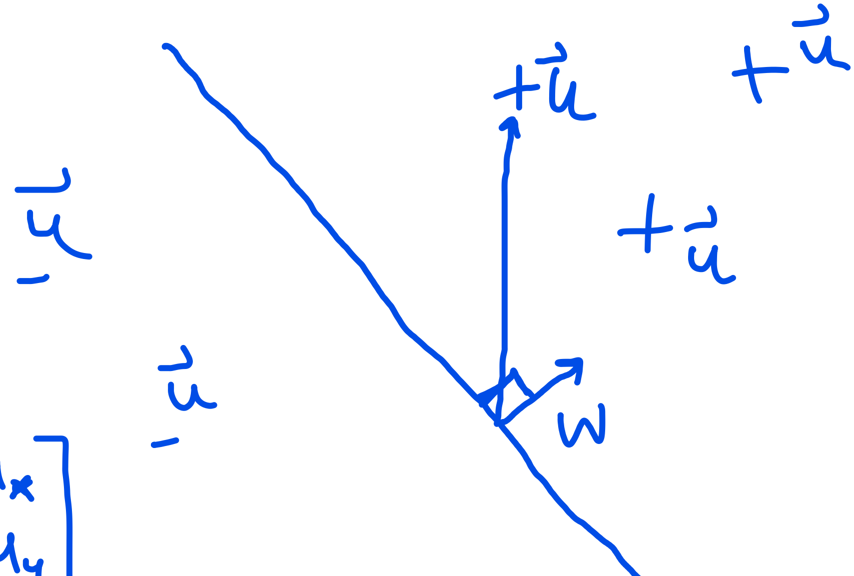
The maximum margin linear classifier is the linear classifier with the maximum margin.

This is the simplest kind of SVM (Called an Linear SVM) init



How SVM works

Find w that maximizes margin



dot product

$$\vec{w} \cdot \vec{u} = \begin{bmatrix} w_x \\ w_y \end{bmatrix} \cdot \begin{bmatrix} u_x \\ u_y \end{bmatrix}$$

$$= w_x \times u_x + w_y \times u_y$$

Ex. $\vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\vec{u} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

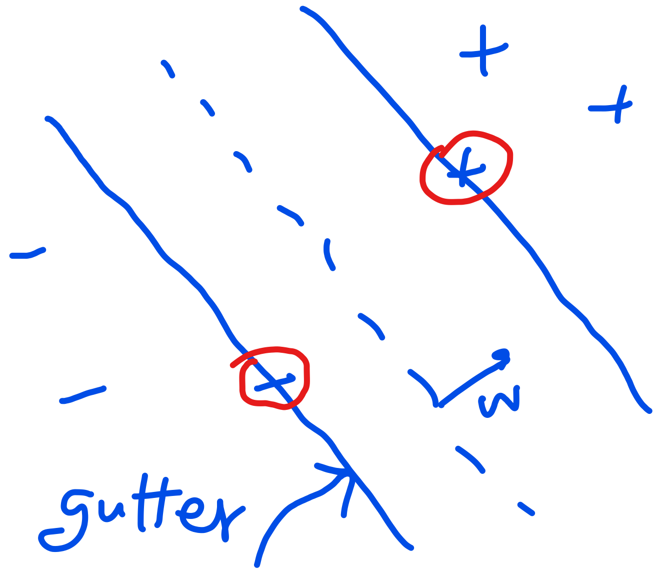
$\vec{w} \cdot \vec{u} = 2 + 2 = 4$

Define $\vec{u} \cdot \vec{w} \geq C$, $\vec{u}$ is classied as $+$

Let $C = -b$, so that $\vec{u} \cdot \vec{w} \geq -b$,

$\vec{u} \cdot \vec{w} + b \geq 0$

Eq. 1



Let $\vec{w} \cdot \vec{u}_{\oplus} + b \geq +1$

$\vec{w} \cdot \vec{u}_{\ominus} + b \leq -1$

$y_i = +1$ for all positive samples

$y_i = -1$ for all negative samples

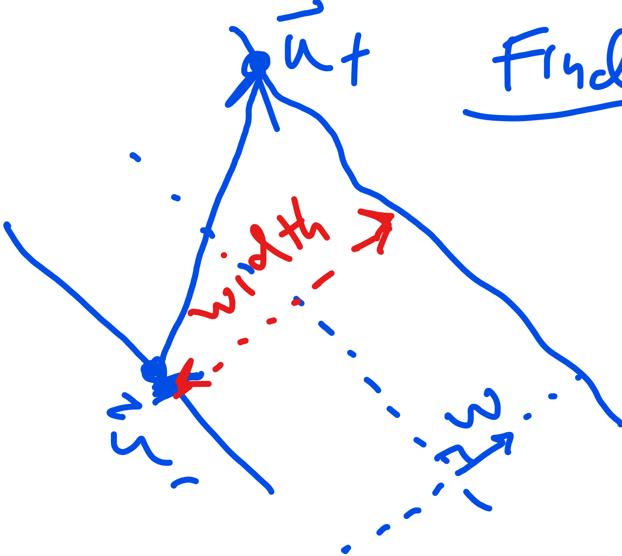
$$y_i(\vec{w} \cdot \vec{u}_i + b) - 1 \geq 0$$

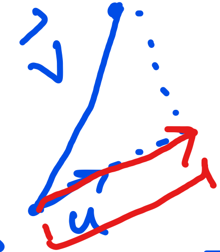
$$y_i(\vec{w} \cdot \underline{\vec{u}}_i + b) - 1 = 0 \rightarrow \text{gutter}$$

E.g. 2

$$-1(w\underline{u} + b) - 1 = -b - 1$$

Find margin width





$\vec{v}$ dot $\vec{u}$, a unit vector

$$\text{width} = (\vec{u}_+ - \vec{u}_-) \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

$$\text{width} = \frac{\vec{w}\cdot\vec{u}_+ - \vec{w}\cdot\vec{u}_-}{\|\vec{w}\|} = \frac{(1-b)-(-b-1)}{\|\vec{w}\|}$$

$$= \frac{1-b+b+1}{\|\vec{w}\|}$$

$$= \frac{2}{\|\vec{w}\|_2} \quad \text{— Eq. 3}$$

$\|\vec{w}\|_2 \rightarrow$ L2 norm

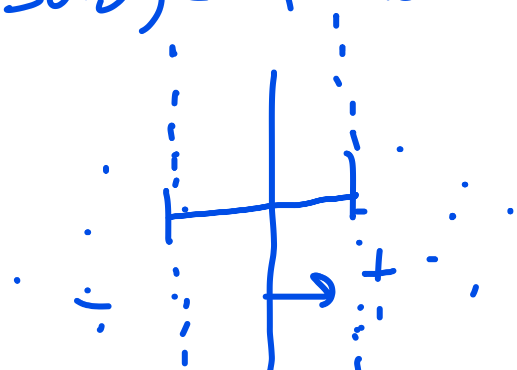
Objective

Lagrangian

maximize $\dfrac{2}{\|w\|_2}$

minimize $\|w\|_2 \rightarrow \frac{1}{2}\|w\|_2^2$

subject to $y_i(\vec{w}\cdot\vec{u}_i + b) - 1 \geq 0, \; i=1,\dots,m$



Hard SVM

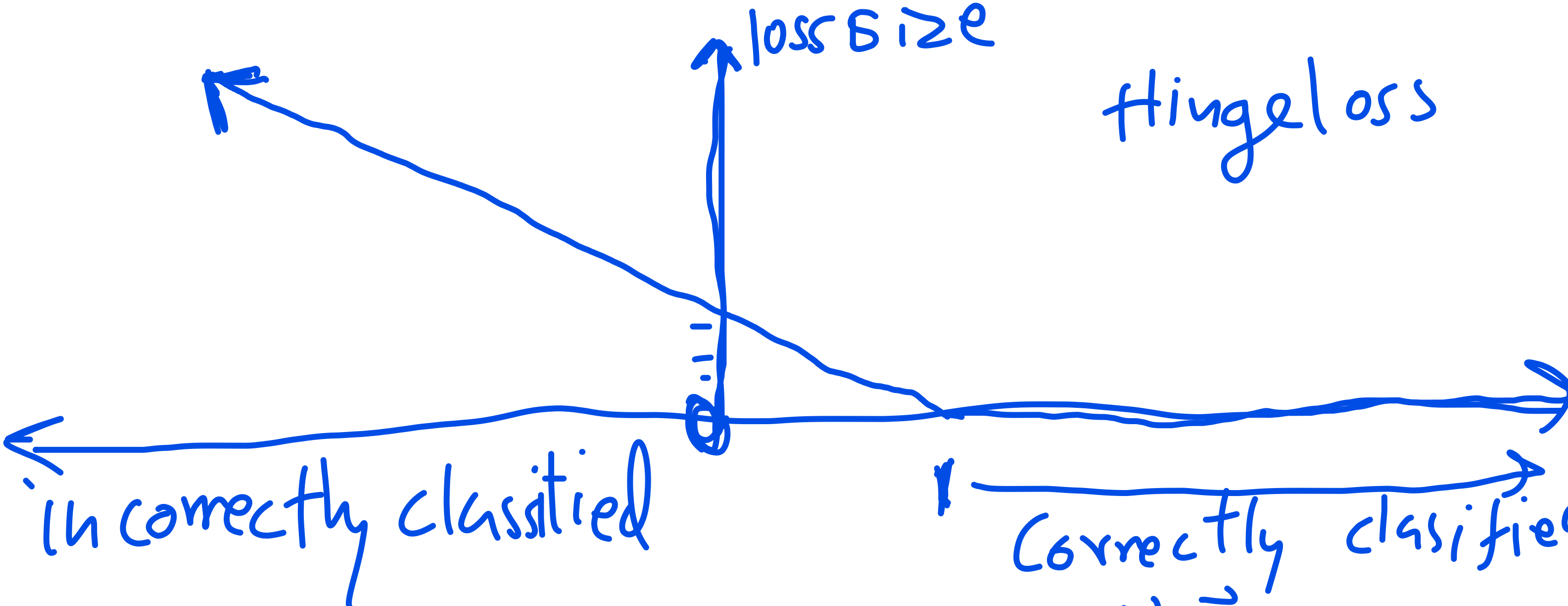
Linear SVM

Penalty parameter

$$\overbrace{J(\vec{w}, b) = \underbrace{\frac{1}{2}\lambda \|\vec{w}\|^2}_{\text{Regularization term}} + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max\left(0, 1 - y_i(\vec{w}\vec{u}_i + b)\right)}_{\text{Hinge loss}}}^{\text{Loss function}}$$

$$\underset{\vec{w}, b}{\operatorname{argmin}} \; J(\vec{w}, b)$$

Hingeloss



$$\text{Hinge Loss}(\vec{w}, b, \vec{x}, y) = \begin{cases} 0, & \text{if } y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \\ 1 - y_i(\vec{w} \cdot \vec{x}_i + b), & \text{else} \end{cases}$$

$$J(\vec{w}, b) = \frac{\lambda}{2} \|\vec{w}\|_2^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y(\vec{w} \cdot \vec{x}_i + b))$$

$$\frac{\partial J}{\partial \vec{w}} = \boxed{\lambda \vec{w}} + \begin{cases} 0 \\ 1 - y_i(\vec{w}\,\vec{x}_i + b) \end{cases}$$

$$\downarrow$$

$$\cancel{1} - y_i \vec{w} \vec{x}_i - \cancel{y_i b}$$

$$\downarrow$$

$$-y_i \vec{x}_i$$

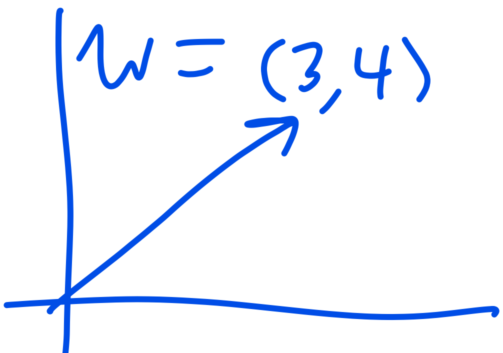
$$\downarrow$$

$$\boxed{-\frac{1}{m} \sum_{i=1}^{m} y_i \vec{x}_i}$$

$$\frac{\partial J}{\partial b} = \boxed{-\frac{1}{m} \sum_{i=1}^{m} y_i}$$

$$\frac{\partial \|\vec{w}\|_2^2}{\partial \vec{w}} = \left(\sqrt{\vec{w}^T\vec{w}}\right)^2 = \vec{w}^2$$

$$= 2\vec{w}$$



$$\|w\|_2^2 = \left(\sqrt{3^2+4^2}\right)^2$$

$$= 25$$

Learning GD

For each iteration: α is learning rate.

$$w' \longleftarrow w - \alpha \frac{\partial J}{\partial w}$$

$$b' \longleftarrow b - \alpha \frac{\partial J}{\partial b}$$

- Stochastic GD $\rightarrow$ one sample