Applied Machine Learning

Lecture 12
Support Vector Machine

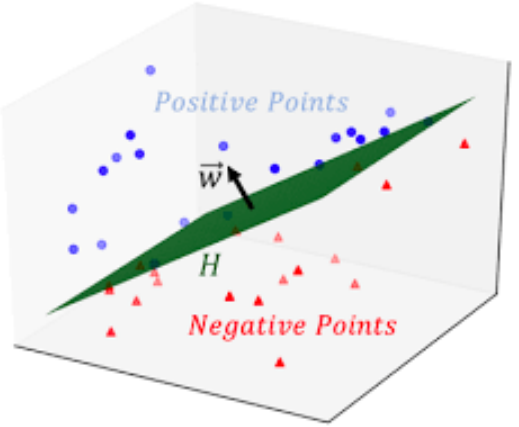
Ekarat Rattagan, Ph.D.

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

Outline

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

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.

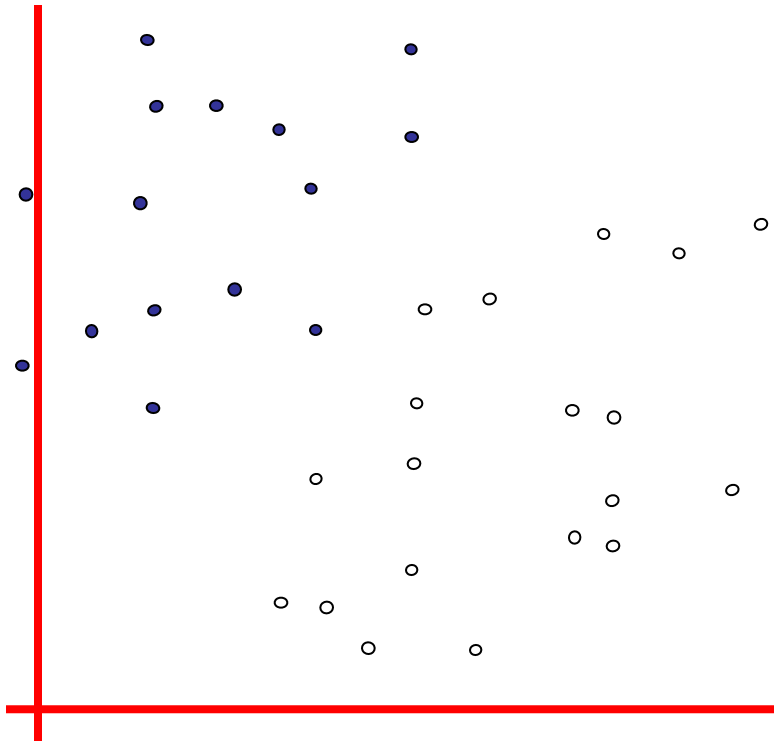
The objective is to find the "maximum-margin hyperplane" that divides the group of points $\vec{x}_i$ for which $y_i = 1$ from the group of points for which $y_i = -1$, so that the distance between the hyperplane and the nearest point $\vec{x}_i$ from either group is maximized.

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

2. Linear Classifiers

denotes +1

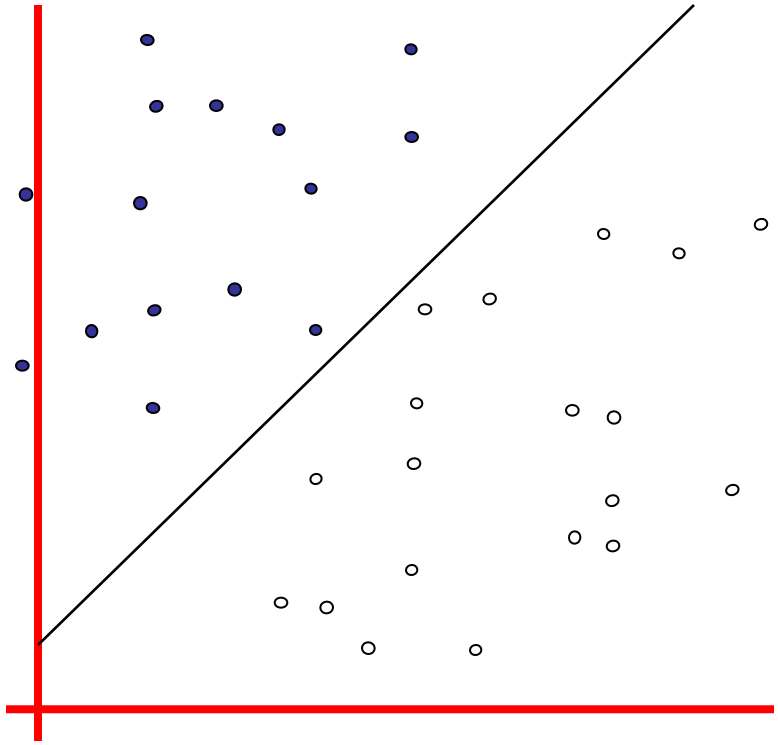
denotes -1



How would you classify this data?

Linear Classifiers

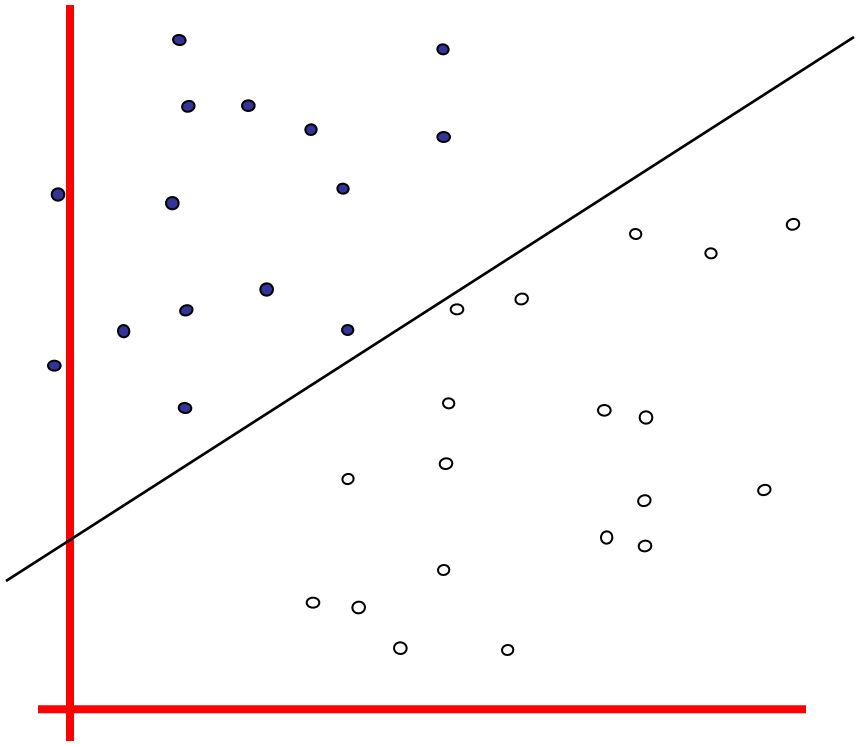
- denotes +1
- denotes -1



How would you classify this data?

Linear Classifiers

- denotes +1
- denotes -1

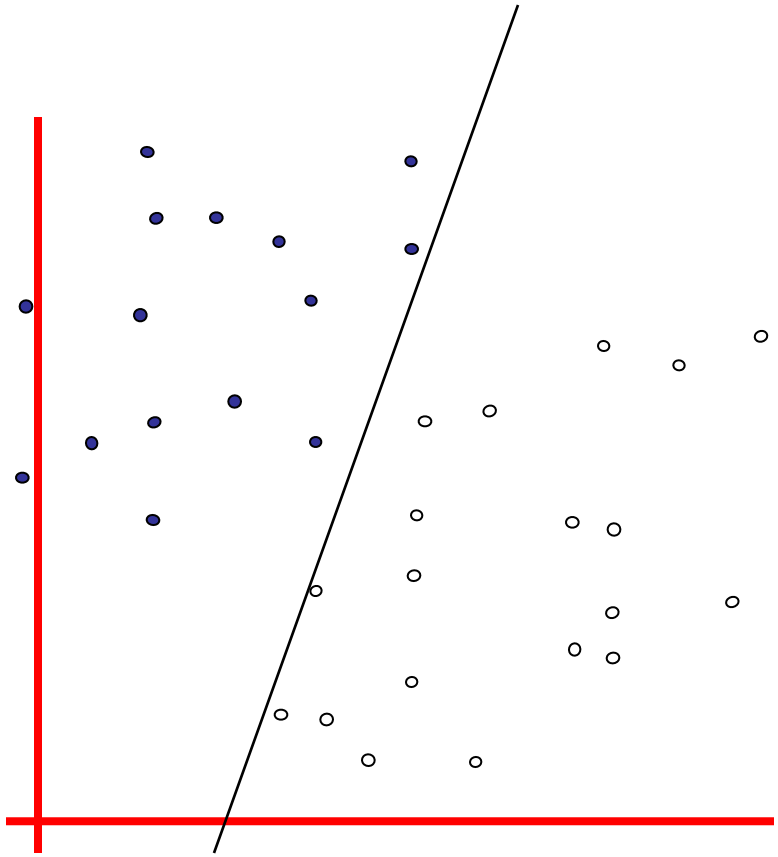


How would you classify this data?

Linear Classifiers

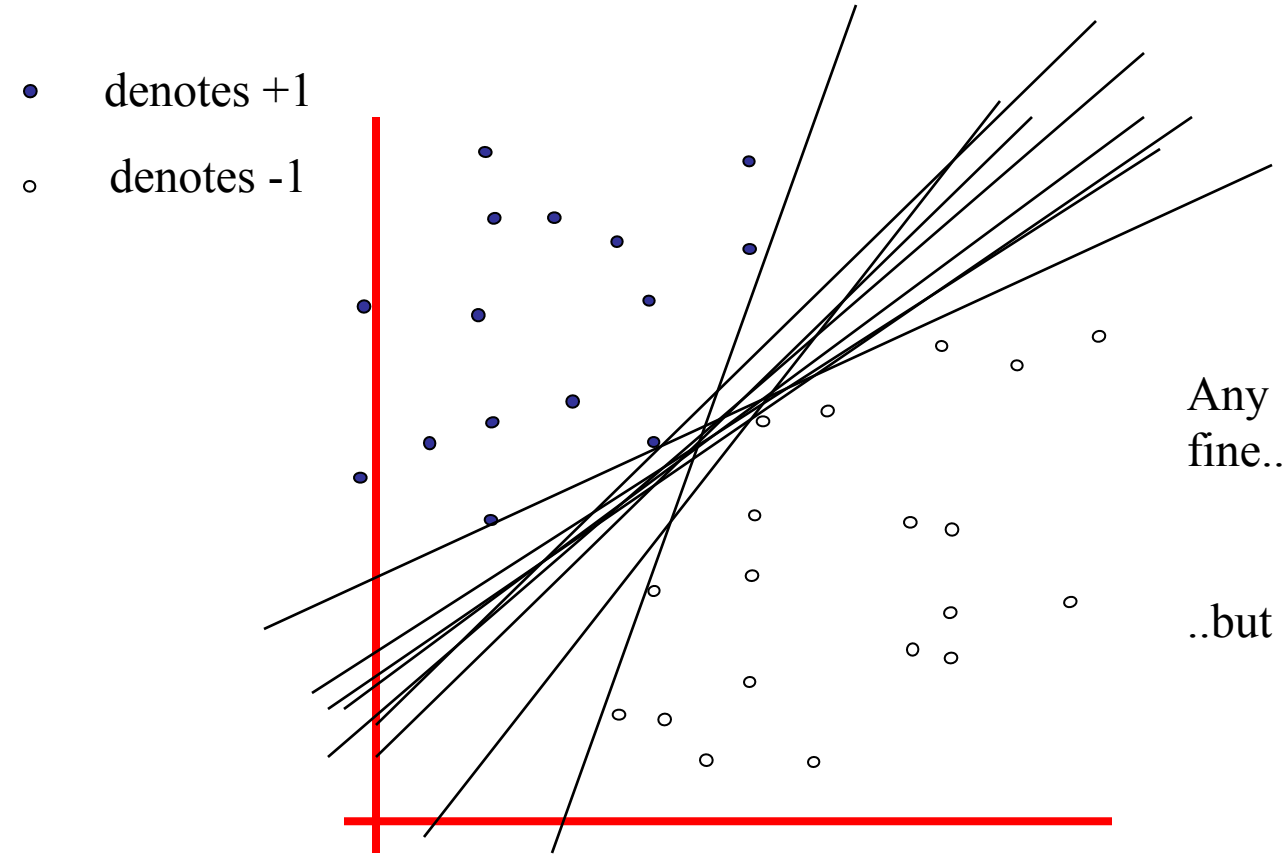
denotes +1

denotes -1



How would you classify this data?

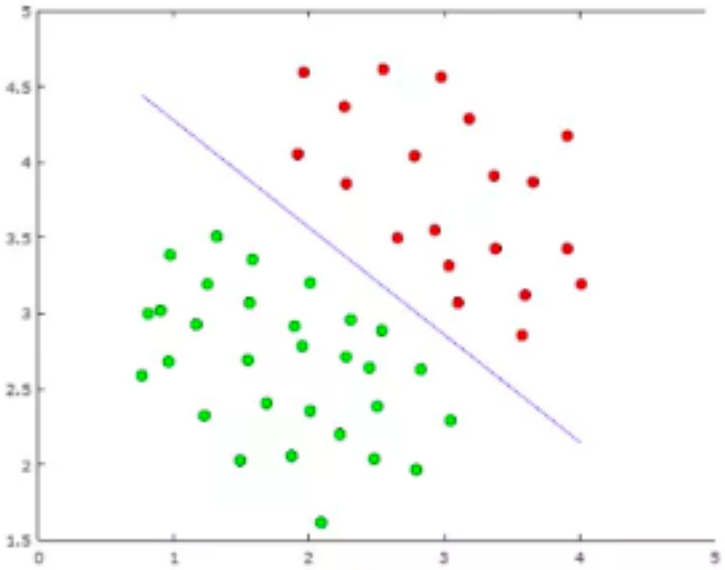
Linear Classifiers



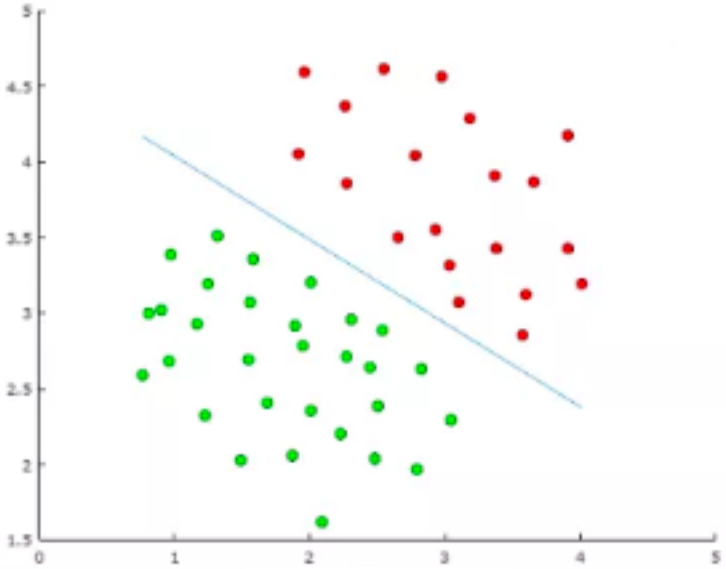
Any of these would be fine..

..but which is best?

Linear Classifiers (SVM VS Logistic Regression)

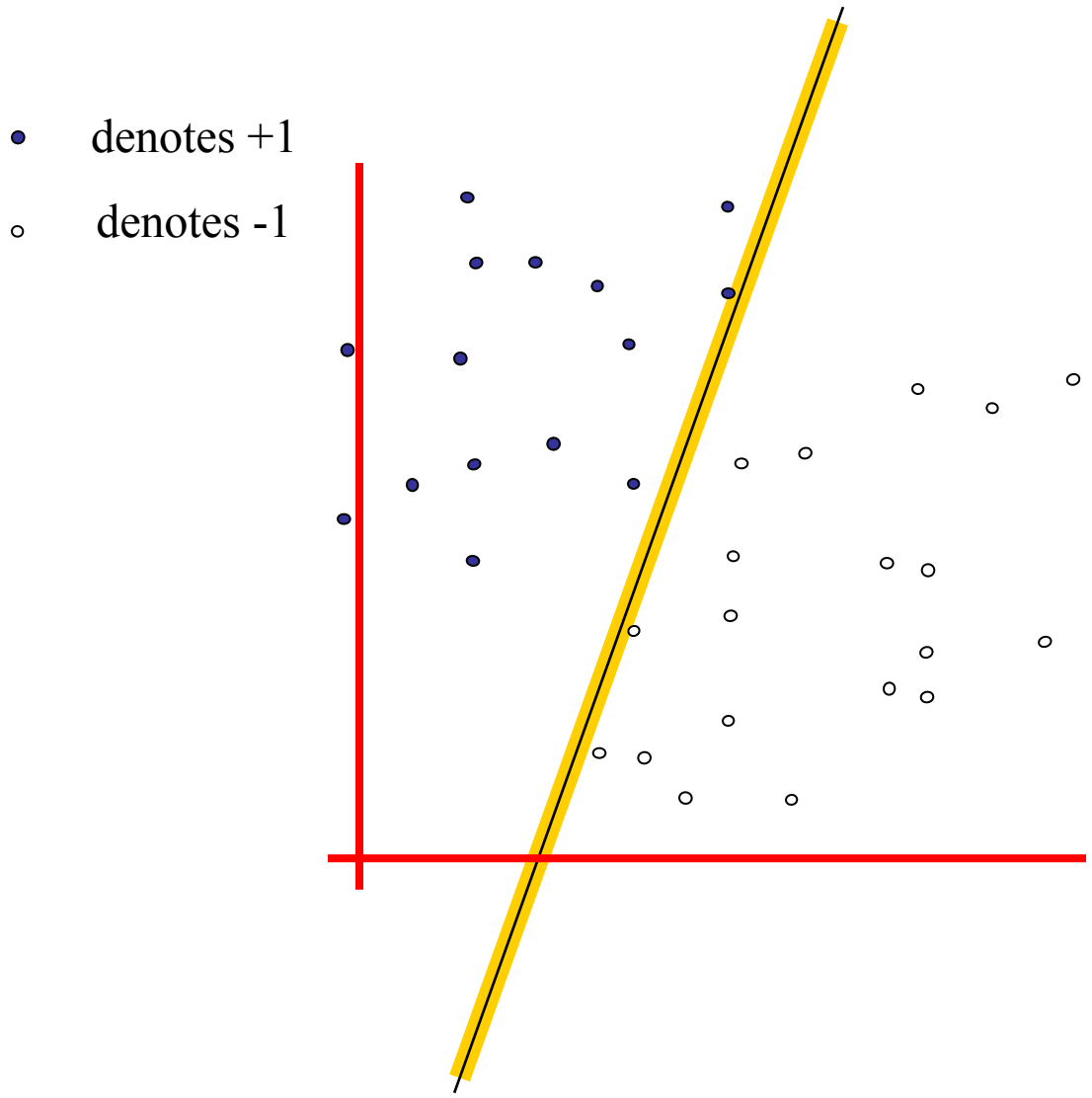


SVM



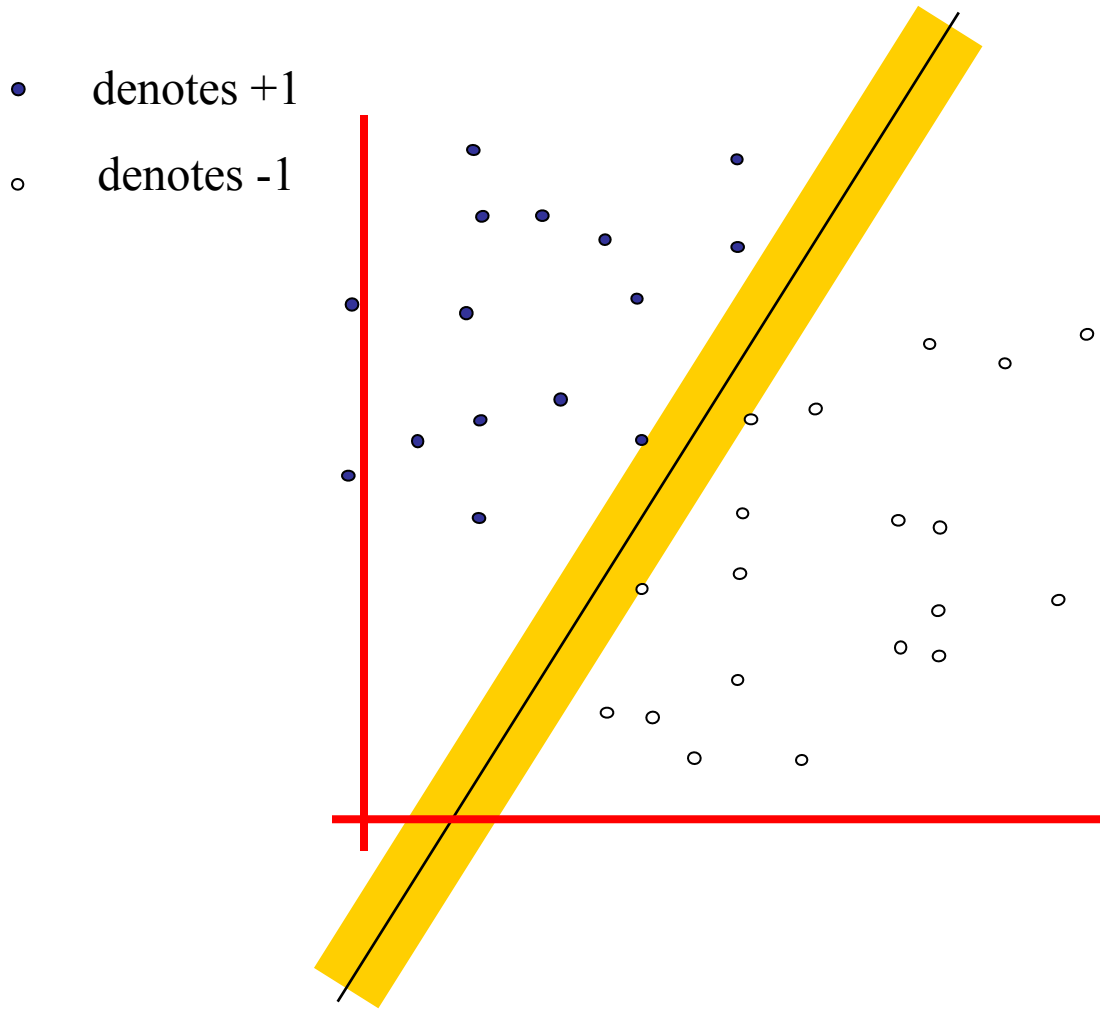
Logistic Regression

Classifier Margin



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

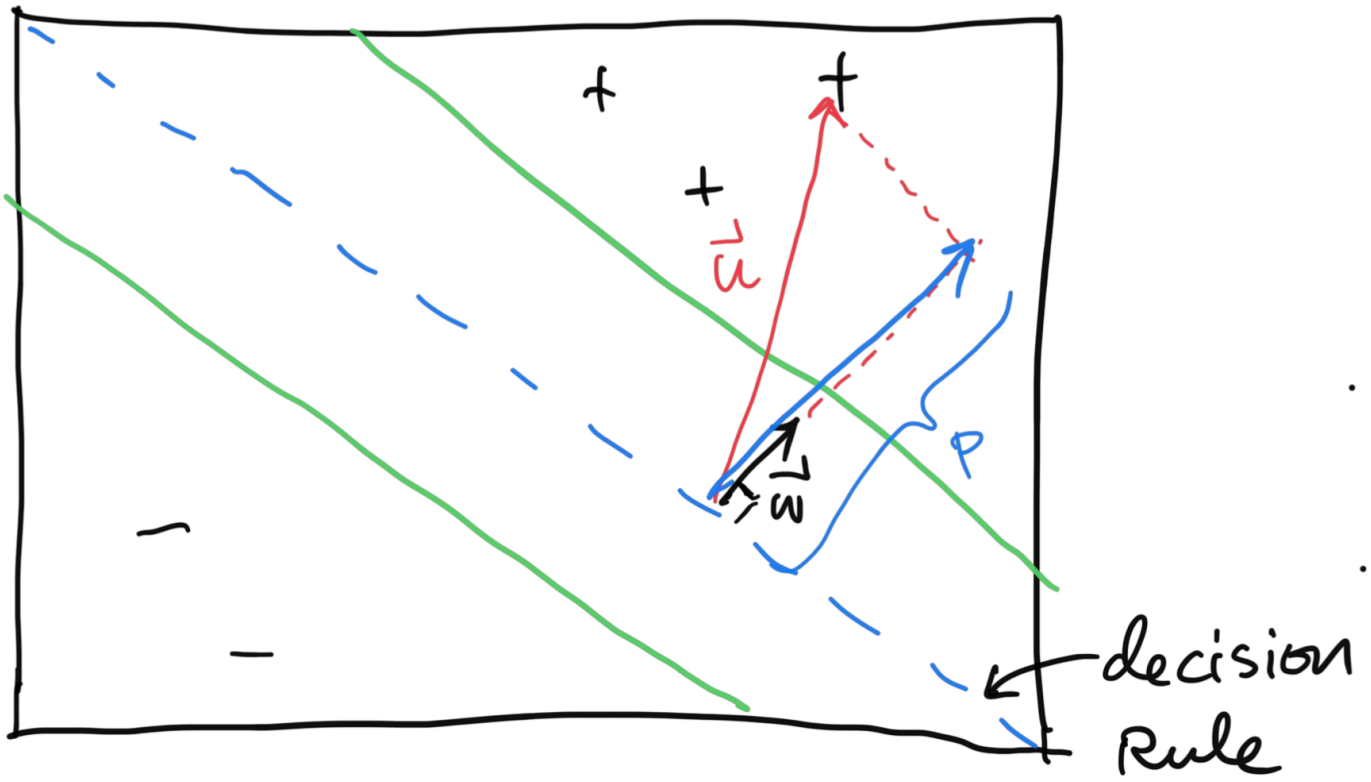
Maximum Margin



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

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

3. How SVM works

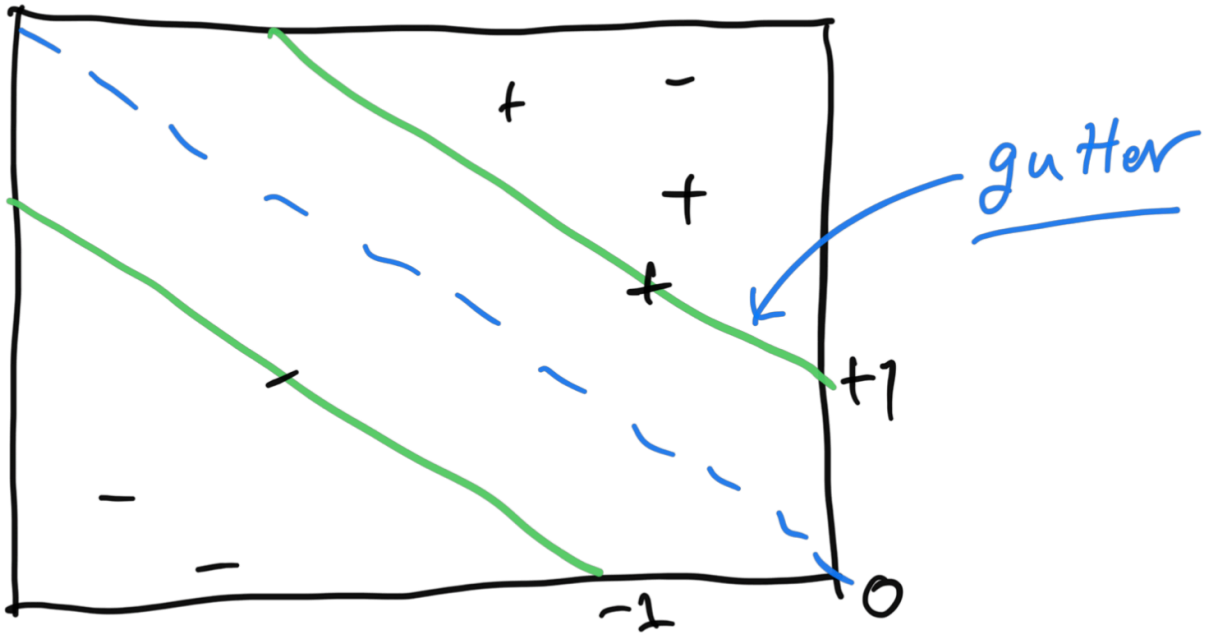


$$\vec{p}_{rojection} = \vec{u} \cdot \vec{w} \quad \text{if} \quad \|w\| = 1$$

If $\vec{u} \cdot \vec{w} \geq C$ the example is classified as +.

$eq(1)$

Let C = -b, then Decision rule = $\vec{w} \cdot \vec{u} + b \geq 0$, the example is classified as +.



We define

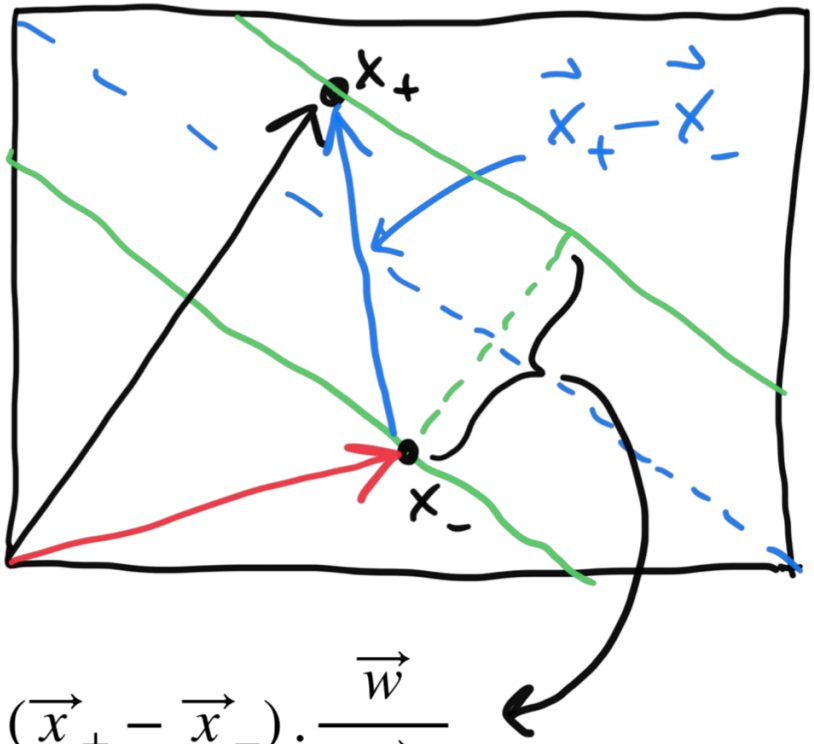
$$\vec{w} \cdot \vec{x}_+ + b \geq +1$$

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

For mathematically convenient, we define y_i such that $y_i = +1$ for + samples and $y_i = -1$ for - samples. Thus, we obtain

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

,and $y_i(\vec{w} \cdot \vec{x}_i + b) - 1 = 0$, for x_i in gutter $eq(2)$



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

From eq(2), we get $((1-b)-(-1-b)) \cdot \frac{\vec{w}}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}$

Minimize $\|\vec{w}\|$ or $\frac{1}{2}\|\vec{w}\|^2$ $eq(3)$

Objective

minimize $\frac{1}{2}\lambda\|\vec{w}\|^2$

subjected to $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$

Cost function

Let $f(x_i) = \vec{w} \cdot \vec{x}_i + b$

$$J(\vec{w}, b) = \frac{1}{2}\lambda \|\vec{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} max(0, 1 - y_i f(x_i))$$

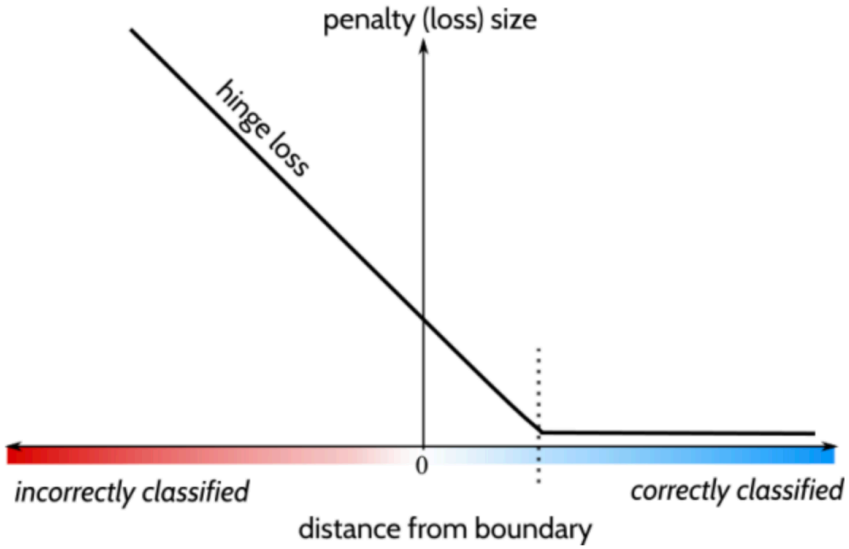
This is called **regularization**; used to prevent overfitting!

hinge loss

4. Cost function

Let $f(x_i) = \vec{w} \,.\, \vec{x}_i + b$

Hinge loss



$$L^{hinge}(f(x_i), y_i) = \begin{cases} 0, & \text{if } y_i * f(x_i) \geq 1 \\ 1 - y_i * f(x_i), & \text{else} \end{cases}$$

Convenient shorthand

$$L^{hinge}(f(x_i), y_i) = (1 - y_i * f(x_i))_+$$

5. Optimization

$$J(\vec{w}, b) = \frac{1}{2}\lambda \|\vec{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} max(0, 1 - y_i f(x_i))$$

$$\frac{\partial \frac{1}{2}\lambda \|w\|^2}{\partial \vec{w}} = \lambda \vec{w}$$

$$\frac{\partial max(0, 1 - y * f(x_i))}{\partial \vec{w}} = \begin{cases} 0, & \text{if } y_i * f(x_i) \geq 1 \\ -\frac{1}{m}\sum_{i=1}^{m} y_i x_i, & \text{else} \end{cases}$$

$$\frac{\partial J(\vec{W}, b)}{\partial b} = -\frac{1}{m}\sum_{i=1}^{m} y_i$$