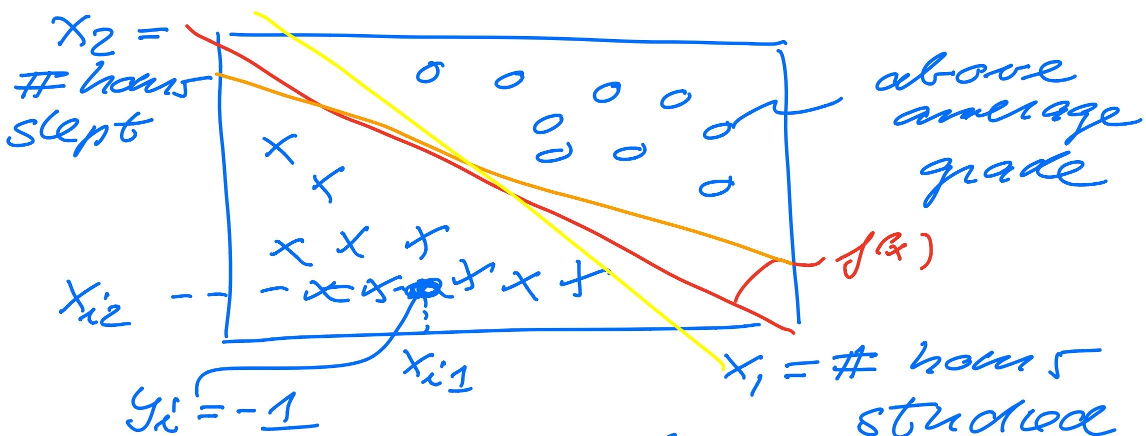
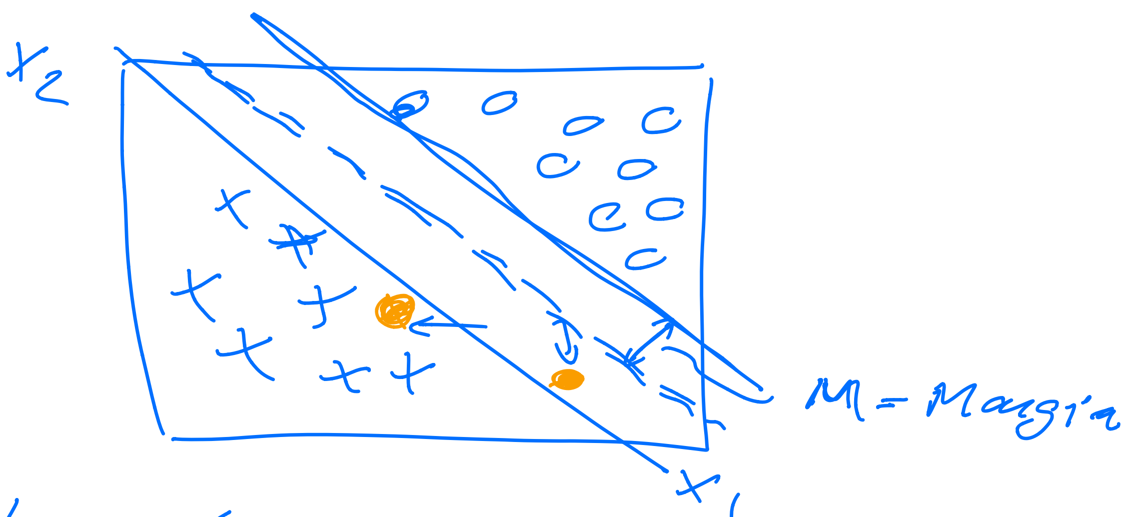
# Lecture November 19

- Linear Classification.




$y_i = \{-1, +1\}$ above average

below average.




## hyperplane

in 2-dim a hyperplane is a straight line

$$b + w_1 x_1 + w_2 x_2 = 0$$

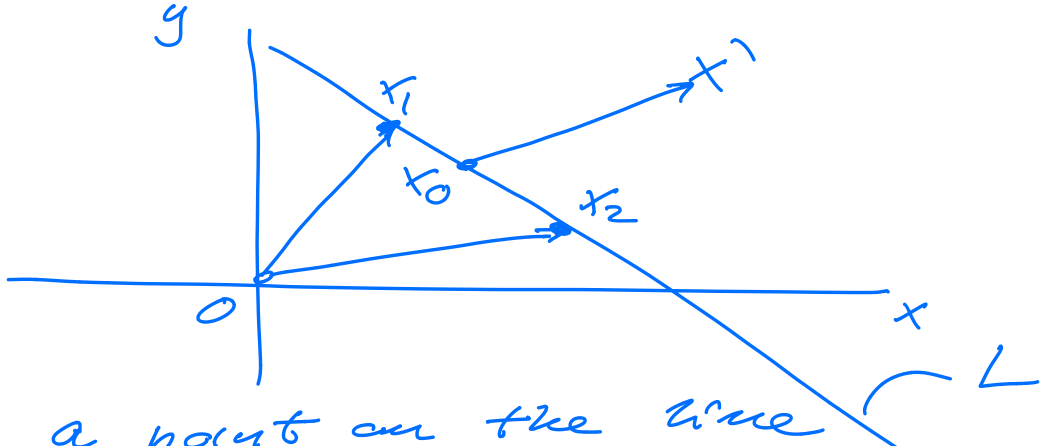
$x^T = [x_1 x_2]$ $\quad$ $w^T = [w_1, w_2]$

$(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$

$f(x) = b + x^T w$

- in 3-dim a hyperplane is a 2-dim plane.
- in p-dim dimension, p-1 hyperplane.

<u>vector algebra</u>




a point on the line

$x : b + w_1 x_1 + w_2 x_2 = 0$

$x = [x_1 x_2]$

$f(x) = x^T w + b = 0$

any two points defined by the vectors $x_1$ & $x_2$ in L

$$\underline{(x_1 - x_2)\, w} = 0$$

vector on $L$

For this to be zero, $w$ must be perpendicular to $L$

Define a vector $w^* = \frac{w}{\|w\|}$

$\|w\| = w^T w$

For any point $x_0$ in $L$

$x_0^T w = -b$

The signed distance of any point $x'$ to $L$ is given

$$(x'^T - \underline{x_0^T})\, w = \frac{1}{\|w\|} \underbrace{(b + x'^T w)}_{f(x)}$$

$x' \Rightarrow x$

$$(x^T - x_0^T)\, w = \frac{f(x)}{\|f'(x)\|}$$

$f(x) = b + x^T w$

$f(x)$ is proportional to the

signed distance from any point $x$ to the hyperplane defined by $f(x) = 0$

$$f(x) = x^T w + b$$

our model. $w_1$ and $w_2$ are unknown parameters.

observations $y_i = \{-1, 1\}$

$$f(x_i) = x_i^T w + b = \{-1, 1\}$$

$$= \begin{bmatrix} x_{i1} & x_{i2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b$$

<u>one possible approach</u>

want to minimize

$$C(b, w) = - \sum_{i \in \text{Mis class}} \underline{\overset{i=1}{y_i} \overset{i=-1}{(x_i^T w + b)}}$$

unknowns $b$ and $w$

$$\frac{\partial C}{\partial w} = - \sum_{i \in \text{Mis class}} y_i x_i$$

$$\frac{\partial C}{\partial b} = -\sum_{\substack{i \in Mis \\ class}} y_i$$

## Problem

- when data are well separated in classes can have many solutions that are independent.

- $$\begin{bmatrix} b \\ w \end{bmatrix} \leftarrow \begin{bmatrix} b \\ w \end{bmatrix} + \underset{\text{learning rate}}{\rho} \begin{bmatrix} -\sum_i y_i \\ -\sum_i y_i x_i \end{bmatrix}$$

- May need many GD iterations. Need to tune $\rho$.

- with no separable data, no convergence.

Instead of $f(x) = 0$, let us try

$$\max_{b,\,W} M \quad \text{subject to}$$

$$y_i(x_i^T W + b) \geq M$$

$$i = 0, 1, \ldots n-1$$

all points are at a signed distance from the decision boundary defined by $b$ and $W$ and $M$

(largest $M$)

$$y_i \frac{(x_i^T w + b)}{\|w\|} \geq M$$

$$y_i(x_i^T w + b) \geq M \underline{\|w\|}$$

we can scale $M = \frac{1}{\underline{\|w\|}}$

$$y_i(x_i^T w + b) \geq \underline{1}$$

- Maximize $M$ or minimize

$\|w\| = w^T w \Rightarrow \frac{1}{2} w^T w$

- subject to (s.t.)

$\underline{y_i(x_i^T w + b) \geq 1}$ for

$\forall i$

# Lagrange formulation

$$\mathcal{L}(w, b, x) = \underline{\frac{1}{2} w^T w}$$
$$- \sum_{i=0}^{n-1} \lambda_i \left( y_i (x_i^T w + b) - 1 \right)$$

Lagrangian multiplier

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=0}^{n-1} \lambda_i y_i x_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_{i=0}^{n-1} \lambda_i y_i = 0$$

$$w = \sum_{i=0}^{n-1} \lambda_i y_i x_i \quad \wedge \quad \sum_{i=0}^{n-1} \lambda_i y_i = 0$$

$$\mathcal{L}(\lambda) = \underline{\sum_{i=0}^{n-1} \lambda_i} - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j$$
$$\times x_i^T x_j \sim [x_j, x_{jk}]$$

subject to

$$\sum \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

in addition we must have

$$\lambda_i [ y_i (x_i^T w + b) - 1 ] = 0$$

$$\forall i$$

min/maximize wrt to $\lambda$ subject to the conditions

$$\min_{\lambda} \frac{1}{2} \lambda^T \begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \cdots \\ y_2 y_1 x_2^T x_1 & x_2^T x_2 y_2 y_2 & \cdots \\ \vdots & & \\ y_n y_1 x_n^T x_1 & \cdots & \end{bmatrix} \times \lambda + (-\mathbb{1}) \lambda$$

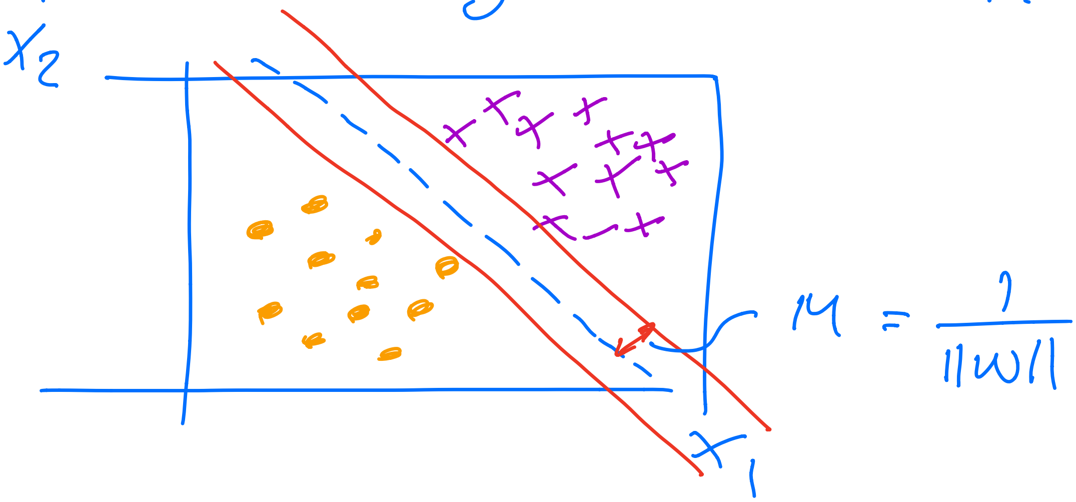
subject to $y^T \lambda = 0$

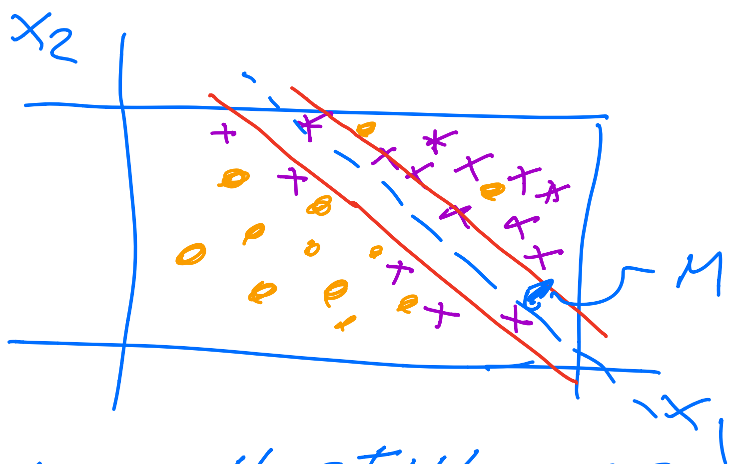
and $0 \leq \lambda < \infty$

The points defined by $x_i$ are called the support vectors.

Convex optimization problem. Python $\boxed{\text{CVXopt}}$

Hard margin case $M$




We will still maximize $M$, but will allow for some points which are on the wrong side of the margin

Introduce new hyperparameter

$\xi^T = [\xi \qquad \xi]$

$\ldots \xi_0, \ldots, \xi_{n-1}]$

$$\frac{1}{\|w\|} y_i (x_i^T w + b) \geq M(1 - \xi_i)$$

$$\forall i \quad \xi_i \geq 0 \quad \sum_{i=0}^{n-1} \xi_i \leq \text{constant}$$

$$M = \frac{1}{\|w\|}$$

$$y_i (x_i^T w + b) \geq 1 - \xi_i$$

Slack variable

Need to minimize

$$\frac{1}{2} w^T w + C \sum_{i=0}^{n-1} \xi_i$$

s.t.

$$y_i (x_i^T w + b) \geq 1 - \xi_i \quad \forall i$$

Lagrange formulation

$$\mathcal{L}(w, b, \underline{\lambda}, \gamma, \xi, x) = \frac{1}{2} w^T w + C \underline{\sum_{i=0}^{n-1} \xi_i} - \sum_{i=0}^{n-1} \left( \lambda_i \left[ y_i (x_i^T w + b) - (1 - \underline{\xi_i}) \right] \right.$$

$n-1$

$$- \sum_{i=0} \gamma_i \underline{\xi_i}$$

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \lambda_i y_i x_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_i \lambda_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \lambda_i - \gamma_i = 0 \Rightarrow$$

$$\underline{\lambda_i} = C - \underline{\gamma_i}$$

$C$ = hyperparameter.