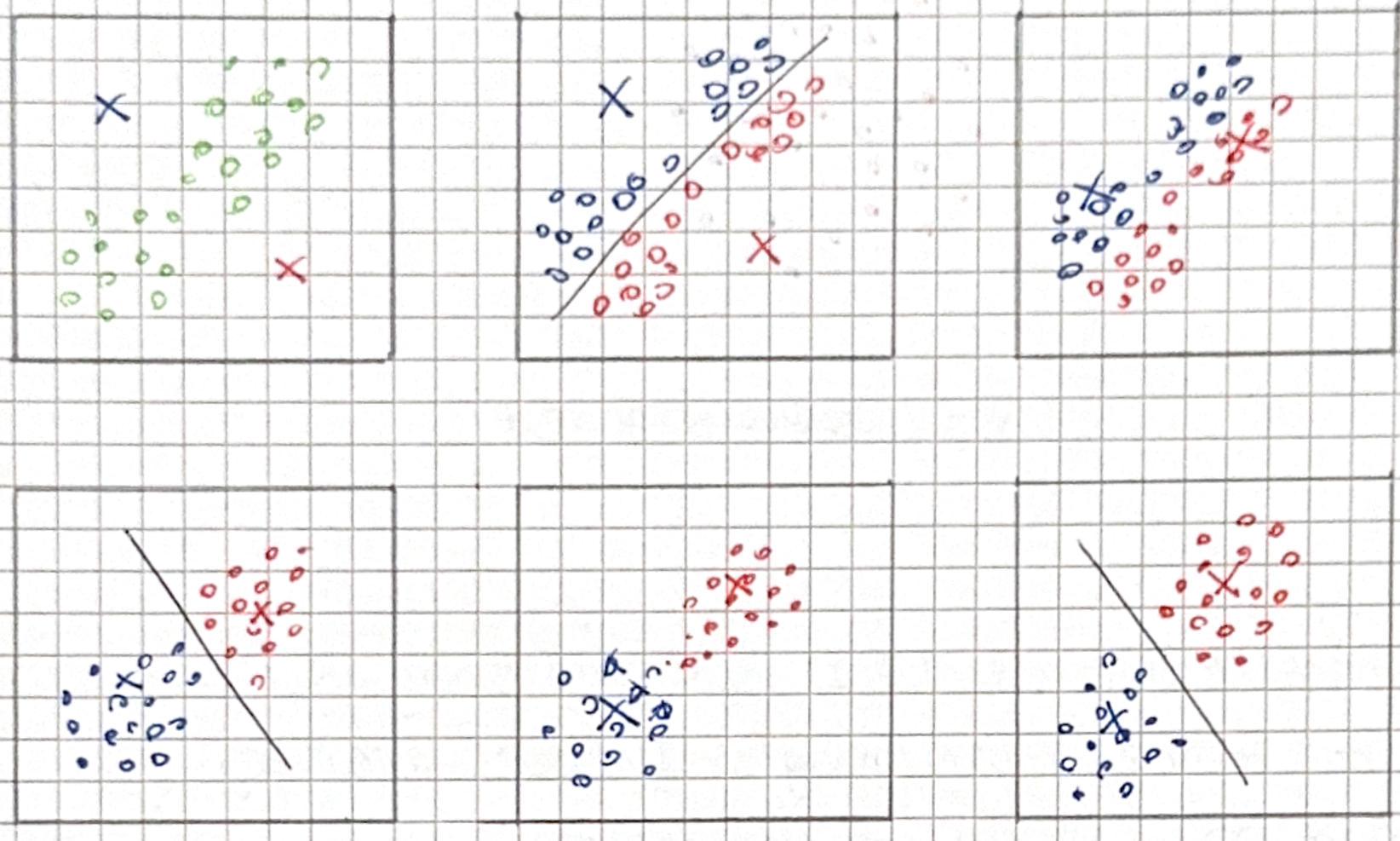
the nearest centroid neither and recompute the centroide

4. repeate step 3 until convergence

terminazion: ~~condizione per~~ the num of distance from each training sample is decrease (from centroid)

there are only finetly partizions of training example into k cluster

6.2



ML EXAMS 22/06/2022

EX 1

1.1

we have overfitting when given a datase D exist an hypothesis h' that $error_S(h) < error(h')$ and $error_D(h) > error_D(h')$

where $S \subseteq D$ and $h, h' \in$ Hypothesis spaa

1.2

we are sure that we have overfitting with decision tree when for same solution we have two different trees one more deeper than other

~~to consider two decision tree T and T' obtain with ID3 algorithm~~

there are two methods for reduce overfitting in decision tree

- REDUCE ERROR PRUNING we must split training data in validation and we evaluate impact on validation set of pruning each possible node. greedly remove the one that most improves validation set accuracy

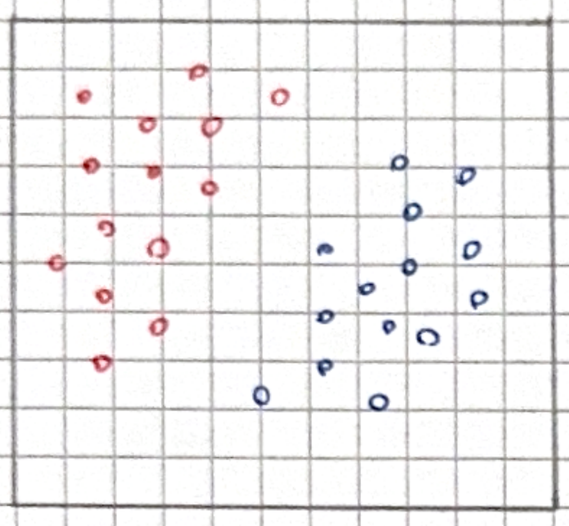
- RULE POST PRUNING (C4.5): if accuracy after pruning is greater respect before we have no overfitting

EX. 2

2.1

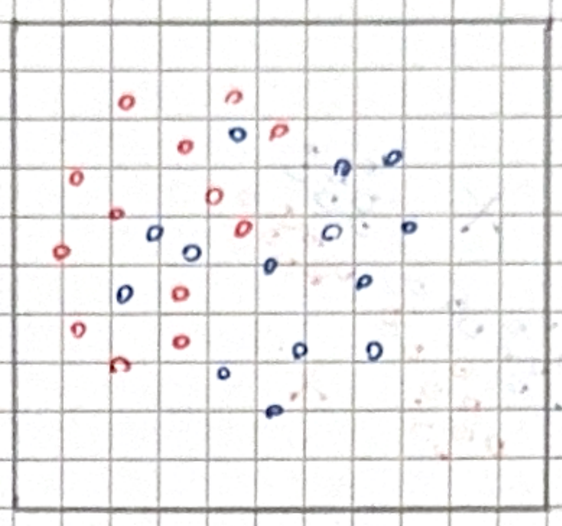
data is linearly separable if exist a surface that split data in two regions such that differently classified instances are separated

2.2



LINEARLY SEPARABLE

2.3



NOT LINEARLY SEPARABLE

2.4

in case of linearly separable dataset (2.2) SVM can easily solve this problem because aims to maximum margin for better accuracy a margin is defined as $\frac{|y(x_n)|}{\|w\|}$ the goal is

$$\arg\max_w \left(\min \frac{|y(x_n)|}{\|w\|} \right) = \arg\max_w \frac{1}{\|w\|} \min |y(x_n)| =$$

$$= \arg\max_w \frac{1}{\|w\|} \min [t_n * [w^T x_n + w_0]]$$

to an optimization problem can be solved using lagrangians multipliers and if also we evaluating for all $SV = \{ x_k \in D \mid t_k (y_k(x_k)) = 1 \}$ we obtain

$$w_0^* = \frac{1}{|SV|} \sum_{x_k \in SV} t_k \left(- \sum_{x_j \in S} a_j^* t_j x_k^T x_j \right)$$

$$y(x, w) = \text{sign} \left(\sum_{x_k \in SV} a_k^* t_k x'^T x_k - w_0^* \right)$$

for the second can see the fact the data is not linearly separable we need to use soft margins constraint in detail

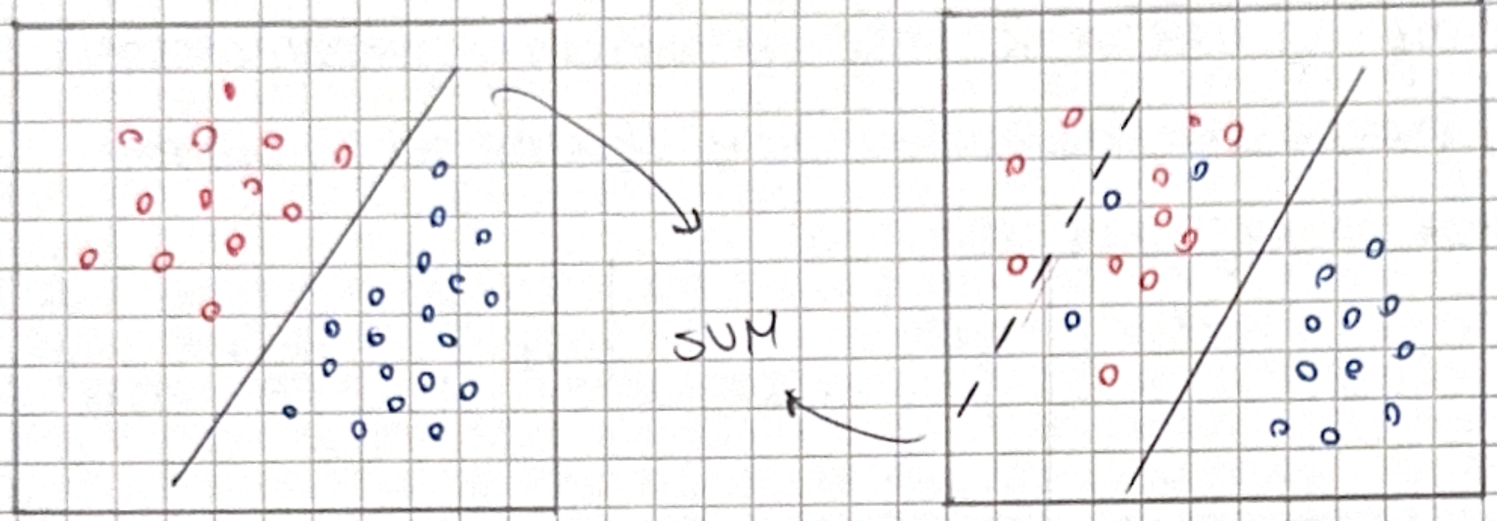
$\xi = 0$ if points is ON or inside the correct margin boundary

$0 < \xi_n \leq 1$ if points is inside the margin but in correct boundaries

$\xi_n > 1$ if points in on wrong side of boundary

in this case the problem can be solved:

$$w^*, w_0^* = \arg\min \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi_n$$



↳ SVM with soft margins constraints

EX.3

3.1

given an dataset $D = \{(x_n, t_n)\}_{n=1}^{N}$ we can define a linear model for solve regression problem in this way:

$$y(x, w) = \sum_{i=1}^{M} w_i^* \phi_i(x_i) \quad \text{still linear in } w$$

the number of trainable parameters is $M+1$ (1 is bias term)

3.2

the ~~real~~ ~~definition regression~~ target value ~~was~~ affected by additive noise ε can be written

$$\forall\, t = y(x, w) + \varepsilon$$

if we assume gaussian noise $P(\varepsilon|\beta) = \mathcal{N}(\varepsilon|0, \beta^{-1})$ so we have

$P(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1})$

if we assume iid observation we have

$$P(\{t_1, \dots, t_n\} | x_1, \dots x_n, w, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_n | w^T \phi(x_n), \beta^{-1}) =$$

$$= -\beta \cdot \frac{1}{2} \sum_{n=1}^{N} [t_n - w^T \phi(x_n)]^2 - \frac{N}{2} \ln(2\pi\beta^{-1})$$

maximum likelihood

$$\arg\max_w P(\{t_1, \dots, t_n\} | x_1, \dots, x_n, \beta)$$

correspond least squares

$$\arg\min_w E_D(w) = \arg\min_w \frac{1}{2} \sum_{n=1}^{N} [t_n - w^T \phi(x_n)]^2$$

EX. 4

4.1

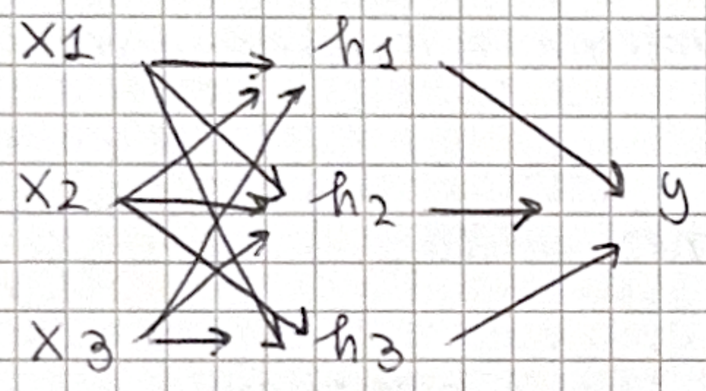
feed forward network → informations flow from input to output without any loop and network is composition of elementary functions in an acyclic graph.

in our case considering $f: \mathbb{R}^3 \to \mathbb{R}$ the parametric model can be written ~~$f(x,\theta) = f(x,\theta_1,\theta_2,\theta_3)$ where $f(x,\theta)$ approximation of $f(x)$ and θ are parameters of network~~

$$f(x,\theta) = f^{(L)}(f^{(L-1)}(\dots f^{(2)}(x_1, x_2, x_3, \theta^{(1)}); \dots; \theta^{(L-1)}); \theta^{(L)})$$

where L are the layers of network, θ are parameter and x_i are the input

considering an example a two layers of network we will have



~~wrong~~

where h_i are hidden units

$h = g(w^T x + c)$ with g the activation function

output $y = w^T h + b$

4.2

a suitable loss depends on task we are try to solve in this case due the fact $f: \mathbb{R}^3 \to \mathbb{R}$ I suppose in a regression task so I will use ~~MSE~~ loss function $\frac{1}{N}\sum_{n=1}^{N}(t_n - x_n)^2$

4.3

I will use stochastic gradient descent:

Require learning rate $\varepsilon > 0$

Require initial values of $\theta^{(1)}$

$k \leftarrow 1$

while stopping criterion not met DO

sample a subset (minibatch) $\{x^{(1)}, \dots, x^{(m)}\}$ from dataset

compute the gradient estimate $g = \frac{1}{m}\nabla_\theta \sum_i L(f(x^{(i)}, \theta^{(k)}), t^{(i)})$

update $\theta^{(k+1)} \leftarrow \theta^{(k)} - \varepsilon g$

$k \leftarrow k+1$

end while

$\nabla_\theta L(f(x^{(i)}, \theta^{(k)}), t^{(i)})$ is compute with BACK PROPAGATION

EX. 5

5.1

MDP = $\langle X, A, \delta, r \rangle$

X = room

A = clean and move

δ: $X \times A \to$ transition function

r: reward function

5.2

considering N = 3 we can derive the optimal policy

N=1	N=2	N=3

$\tilde{\pi}$ → policy have to maximize the cumulative reward in particular

if $r(s,a) = +1$ clean room

$r(s,a) = -2$ move to room

$r(s,a) = 0$ otherwise

$\tilde{\pi} = \{((N=2, \text{clean}), \text{move to } N_2), ((N1, \text{dirty}), \text{clean}), ((N2, \text{dirty}), \text{clean}), ((N=2, \text{clean}), \text{move to } N3\}, ((N=3, \text{dirty}), \text{clean}), ((N=3, \text{clean}), \text{stop}))$

initial state = N=1

EX. 6

6.1

$$P(x) = \sum_{n=1}^{3} \pi_n N(x, \mu_N, \Sigma_N)$$

x: input data μ_N = mean Σ: covariance matrix

weight of component π_j $k=3$ but $\sum_k \pi_k = 1 \to k-1=2$

μ_n is a vector in two dimension so # params of $\mu_n = 6$

Σ_k is a symmetric covariance matrix $d=2$ so # params $d \times \frac{(d+1)}{2} = 3$