2. no general properties

ML EXAMS 23/03/2018

EXERCISE 1

I. given a dataset D and a hypothesis space $H : h \in H$ overfits training data if exists an hypothesis h' such that

$Error_S(h) < Error_S(h')$ and $error_D(h) > error_D(h')$

II in decision tree is very probable to have overfitting in fact two different solutions for a problem and tree is deeper than other for sure we have overfitting.

for reduce them:

1. stop growing when data split are not ~~signific~~ statically significant
2. grow full tree than post prune

for to do this we need to divide dataset in train and validation and use validation to understand if prune

Question 2

1. Naive bayesian classifier uses conditional independence to approximate the solution

X is conditionally independent of Y given Z

$P(X,Y|Z) = P(X|Y,Z)\,P(Y|Z) = P(X|Z)\,P(Y|Z)$

given a target function $f: X \to V$ where $V = \{v_1 \times v_2 \times \dots \times v_n\}$

knowing that each sample can be expressed as $\langle a_1, \dots, a_N \rangle$

$P(V_i | x, D) = P(V | a_1, \dots, a_M, D) = \dfrac{P(a_1, \dots, a_n | V, D)\,P(V|D)}{P(a_1, \dots, a_n | D)}$

$V_{NB} = \underset{v \in V}{\operatorname{argmax}}\; P(a_1, \dots, a_n | V, D)\,P(V|D)$

now with bayesian assumption $P(a_1, \dots, a_n | V, D) = \prod_{n=1}^{n} P(a_i | v_j, D)$

$$V_{NB} = \underset{v \in V}{\operatorname{argmax}}\; P(v_j|D) \prod_i P(a_i | v_j | D)$$

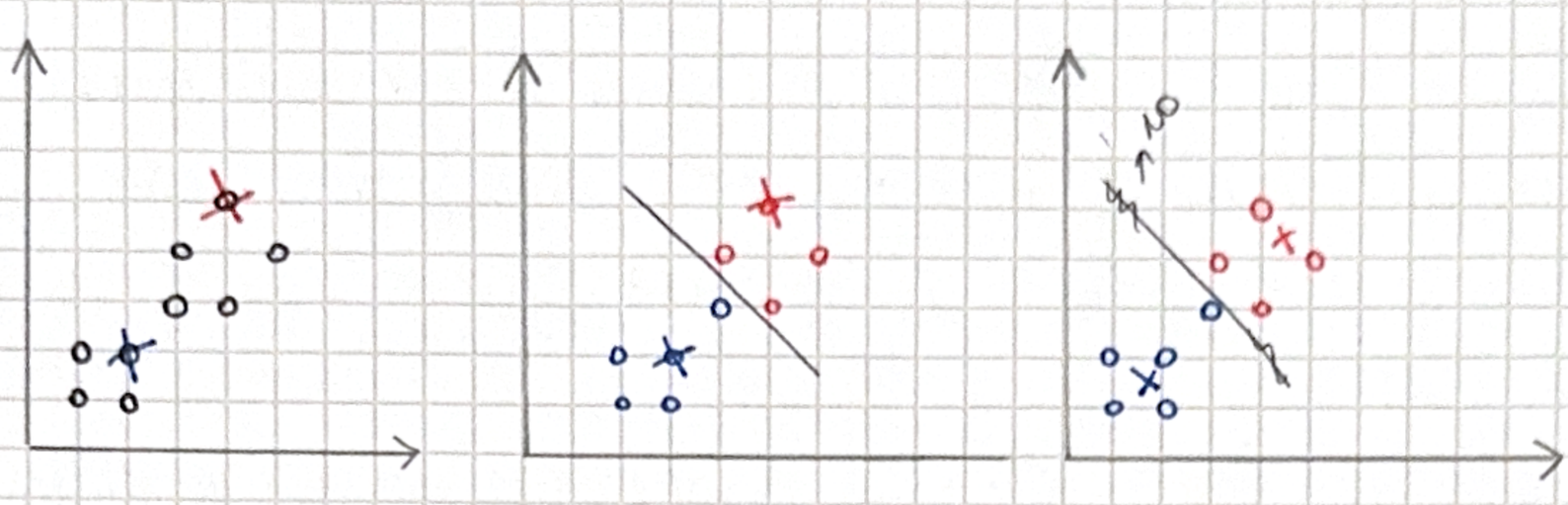
2. the target function is $f: X \to Y$ with $X = \{$ title x abstract x author $\}$ x title $\}$ and $Y = \{ML, KR, RL\}$ the dataset is $D = \{(d_i, y_i)\}_{i=1}^{N}\}$

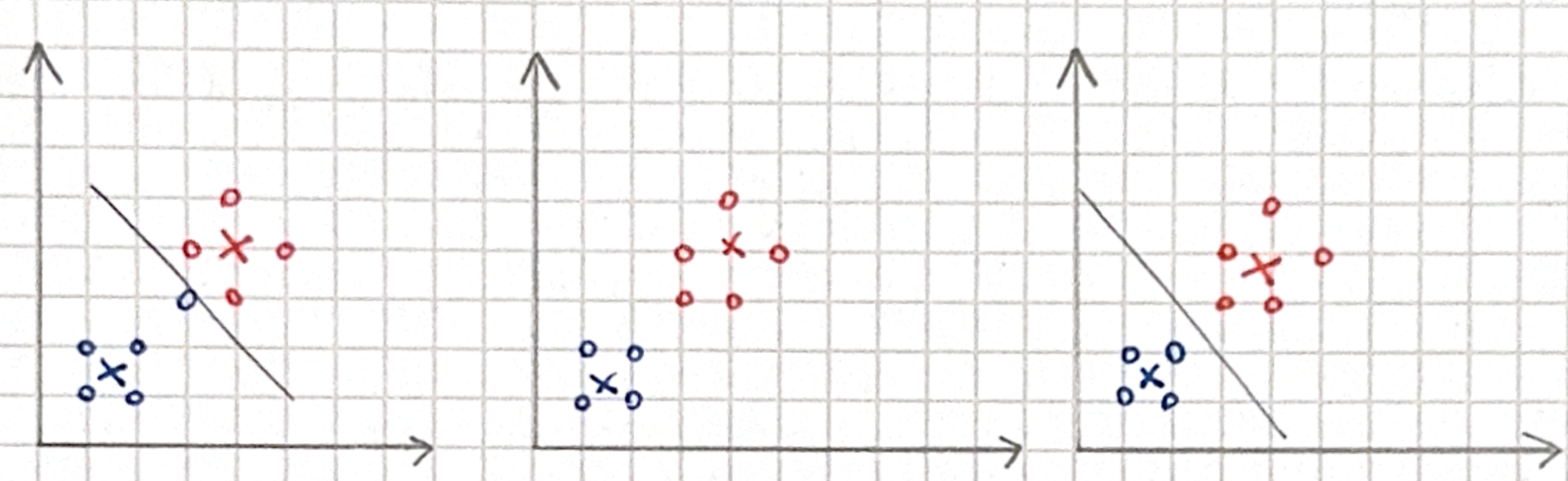
$$V_{NB} = \underset{v \in V}{\operatorname{argmax}}\; P(V|D) \prod_{i=1}^{N} P(d_i | V, D) \qquad \hat{P}(d_i | V, D) = \frac{TF_{ij} + 1}{TF_j + 2}$$

Question 3

1. in a unsupervised learning task we have a target function $f: X \to Y$ and a dataset $D = \{(x_n)_{n=1}^{N}\}$. Is possible to use we don't have labels the datas is compared only by input values

2.





I use K-MEANS to solve this problem

3. we can see an initialy a point is clasified as blue but some iterations is clasifies as RED

Question 4.

1. clasification problem $f: X \to C$ with dataset $D = \{(x_n, t_n)_{n=1}^{N}\}$

   ① find k nearest neighboors of new instance x

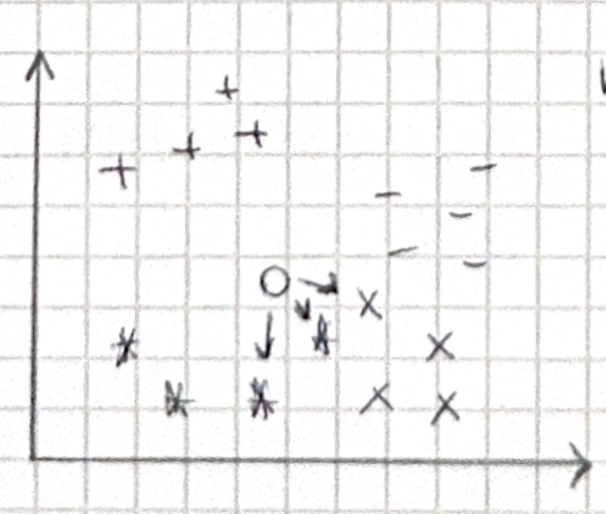
   ② assign to x the most common label among the majority of neighboors

   likelihood of class c of new instance x is:

$$P(c|x,D) = \frac{1}{K} \sum_{x_n \in N_k(x_n, D)} \mathbb{I}(t_n = c)$$

   where $\mathbb{I}(e) = \begin{cases} 1 & \text{if } e = \text{true} \\ 0 & \text{otherwise} \end{cases}$

2.



Question 5

1.

BACKPROPAGATION: is an algorithm used to compute the gradient is not a learning algorithm and is used in ANN. In this way we propagate the gradient through all the net

FORWARD AND BACKWARD:

in the forward step we compute the value of the parameter of the function (given x, t compute a h and $J = L(t,y)$

in the backward we compute the gradient (given x, t, a, h, y, J

compute $\partial J / \partial w_{ij}^{(k)}$

SGD: is a learning algorithm that is used to find the minimum or a maximum of a function in an iterative method and is a optimization analysis

2.

FORWARD STEP

Require: depth ℓ, weight matrices $W^{(i)}$ bias parameters $b^{(i)}$, x input y target value

$h^{(0)} = x$

for $k = 1, \dots, \ell$ DO

$a^{(k)} = b^{(k)} + W^{(k)} h^{(k-1)}$

$h^{(k)} = f(a^{(k)})$

end for

$y = h^{(\ell)}$

$J = L(t,y)$

BACKWARD STEP

$g \leftarrow \nabla_y J = \nabla_y L(t,y)$

for $k = \ell, \dots, 1$ DO

$g \leftarrow \nabla_{a^{(k)}} J = g \odot f'(a^{(k)})$

$\nabla_{b^{(k)}} J = g$

$\nabla_{W^{(k)}} J = g (h^{(k-1)})^T$

$g \leftarrow \nabla_{h^{(k-1)}} J = (W^{(k)})^T g$

end for

Question 6

1. we have a target function $f: X \to y$ with $X \subseteq \mathbb{R}^D$ and $y = \mathbb{R}$ an model is $y(x, w) = \sum_{n=1}^{N} w_n \phi_n(x)$ that is linear in w and $D = \{(x_n, t_n)\}_{n=1}^{N}$

we know that $t = y(x,w) + \varepsilon$ where ε is additive Gaussian noise $p(\varepsilon | \beta) = \mathcal{N}(\varepsilon | 0, \beta^{-1})$ assuming that we have

~~now we use~~ $p(t | x, w, \beta) = \mathcal{N}(t | y(x,w), \beta^{-1})$

now we use iid hypothesis the error function will be:

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2 \text{ and } w^* = \arg\min(E_D(w))$$

2. we can use least squares $ED(w) = \frac{1}{2}(t-\phi(w))^T(t-\phi(w))$ and
$w^* = \phi^\dagger t$
we can also use sequential algorithm $\hat{w} \leftarrow \hat{w} - \eta[t_n - w^T\phi(w)]\phi(x_n)$

ML EXAMS 18/01/2018 A

Exercise 1

1.
if $C = c_1$ and $B = b_1$ then ~~YES~~ NO ①
if $C = c_1$ and $B = b_2$ then YES ②
if $C = c_2$ and $B = a_1$ then YES ③
if $C = c_2$ and $A = a_2$ and $B = b_1$ ④
if $C = c_2$ and $A = a_2$ and $B = b_1$ ⑤
if $C = c_2$ and $A = a_3$ then NO ⑥
if $C = c_3$ ~~and~~ then NO ⑦

2.
is consistent with S_1 because of ①
is consistent with S_2 because of ④
is NOT consistent with S_3 because of ⑦
is NOT consistent with S_4 because of ⑤

Exercise 2

1. the maximum aposteriori hypothesis is $h_{MAP} = \arg\max_{h \in H} P(D|h)\frac{P(h)}{P(D)}$
if we assume $P(h_j) = P(h_i)$ we simplify and
we obtain $h_{ML} = \arg\max_{h \in H} P(D|h)$

2. Boc is an optimal classifier that returns the most probable ~~classification~~ prediction:
$v_{NB} = \arg\max_{v \in V} \sum_{h_i \in H} P(v|x, D_{h_i}')P(h_i|x, D) = \sum_{h_i \in H} P(v|x, h_i)P(h_i|D)$

3. we can use it if we have analytical solutions or if hypothesis space is not to large otherwise is not practical