$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n) \qquad I(e) = \begin{cases} 1 & e = true \\ 0 & otherwise \end{cases}$$

2. evaluate $\varepsilon_m = \dfrac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \qquad \alpha_m = \ln\left(\dfrac{1-\varepsilon_m}{\varepsilon_m}\right)$

3. update

$$w_n^{(m+1)} \leftarrow w_n^{(m)} e^{\alpha_m I(y(x_n) \neq t_n)}$$

3 output $Y_m = sign\left(\sum_{m=1}^{M} \alpha_m y_m(x)\right)$

ML EXAMS 04/11/2018

EX. 1

1. in a binary classification problem the target function is $f: X \to Y$ where $X = \{FR, NR, NK\}$ and $Y = \{Y, N\}$ the dataset is $D = \{(x_i, y_i)\}_{i=1}^{N}\}$

2. if we use ID3 algorithm for choose which attribute to use we compute the gain

gain = entropy(S) - $\sum_{v \in V} \frac{|S_v|}{|S|}$ entropy(S_v) where with V = values that

$$entropy = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

3.

~~gain~~

$$entropy(F_y) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) =$$

$$entropy(F_N) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.916$$

$$entropy(NR_3) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.916$$

$$entropy(NR_4) = -\frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$entropy(NK_4) = -1\log(1) = 0$$

$$entropy(NKN) = -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$entropy(S) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$

gain (S, F) = entropy(S) - $\frac{2}{5}$ entropy(Fy) - $\frac{3}{5}$ entropy (FN) = 0.0184

gain (S, NR) = entropy (S) - $\frac{3}{5}$ entropy(N3) - $\frac{2}{5}$ entropy (N4) = 0.4214

gain (S, NC) = entropy (S) - $\frac{1}{5}$ entropy(NCy) - $\frac{4}{5}$ entropy (NCN) = 0.171

NR highest gain

EX. 2

1. $h_{MAP} = \arg\max P(D|h)P(h)$ / $P(D)$ if we assume $P(h_i) = P(h_j)$

$h_{ML} = \arg\max P(D|h)$

2. this is false if we introduce 3 hypotheis:

$h_1(x) = +$ $h_2(x) = -$ $h_3(x) = -$ $P(D|h_1) = 0.4$ $P(D|h_2) = 0.3$

$P(D|h_3) = 0.3$

$h_{ME} = \arg\max \{0.4, 0.3, 0.3\} = 0.4$ → class +

now I introduce Bayesian Optimal classifier

$V_{BOC} = \arg\max_{v \in V} \sum_{h_i \in H} P(v|x, h_i) P(h_i|D)$

is an optimal classifier that return the optimal solution

$P(+|x, h_1) = 1$ $P(-|x, h_1) = 0$

$P(+|x, h_2) = 0$ $P(-|x, h_2) = 1$

$P(+|x, h_3) = 0$ $P(-|x, h_3) = 1$

$V_{BOC} = \arg\max \{1 \cdot 0.4 + 0 + 0, 0 + 1 \cdot 0.3 + 1 \cdot 0.3\} = -$

EX. 3

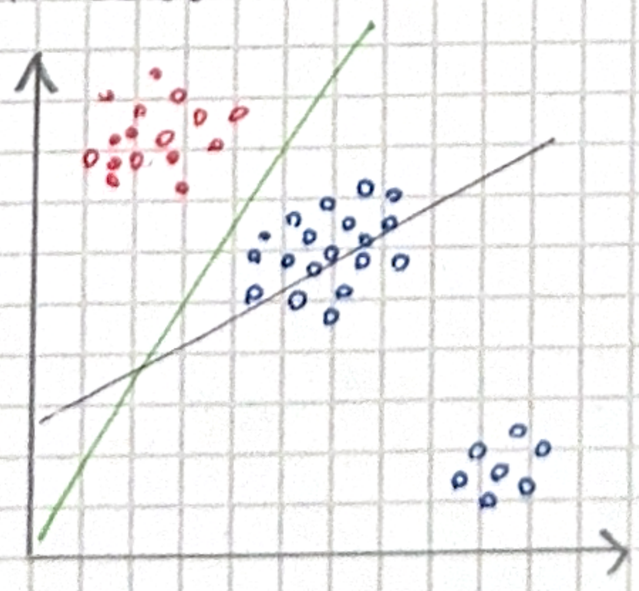
least squares is a linear model for classification. in this method we want minimize an error function that is called sum of squares

$$J(\tilde{w}) = \frac{1}{2} tr((T - \tilde{X}\tilde{w})^T (T - \tilde{X}\tilde{w}))$$

we call this sum of squares because the trace is a sum of element and the product of one matrix for his transpose is the square

$\tilde{w}^* = \tilde{X}^t T$ and $y = \tilde{w}^T \tilde{x} = T^T (X^T)^T \tilde{x}$

is not robust to outliers because is based on distances



EX.4

1.

the gram matrix is defined as $K = XX^T$. if we have a kernell function $k(x, x') = x^T x'$ and a model $y(x, w^*) = \sum_{n=1}^{N} w_n x_n^T x$

the gram matrix is $K = \begin{bmatrix} x_1^T x_1 & \cdots & x_1^T x_M \\ \vdots & \ddots & \vdots \\ x_N^T x_1 & \cdots & x_N^T x_M \end{bmatrix}$

if we have a generic kernell function $k(x, x') \Longrightarrow y(x, w^*) = \sum_{n=1}^{N} w_n k(x_n, x)$

and gram matrix will be $K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_M) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_M) \end{bmatrix}$.

2. we use a modified error function $J(w) = \sum_{n=1}^{N} \alpha_n x_n^T x$ now we reformulate this function solution

$y(x, w^*) = \sum_{n=1}^{N} \alpha_n k(x_n, x)$ $\Rightarrow$ computationally expensive

EX.5

1. I will use linear regression model for solve this task

$y(x, w) = \sum_{n=1}^{N} w_n \phi(x_n)$ that is linear in w_n the goal is minimize

the error function:

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2$$ this can

we use a sequential algorithm to find a solution

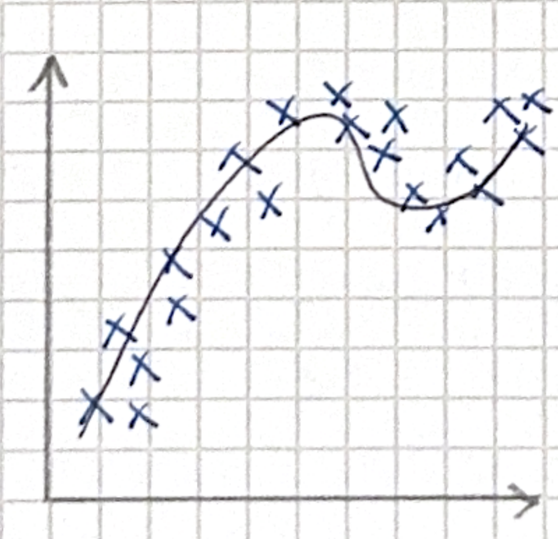
$$w \leftarrow w + \eta (t_n - w^T \phi(x_n)) \phi(x_n)$$

2. the overfitting can be reduced by applying λ regularization factor

$w^* = \arg\min E(w) + \lambda E_w(w)$ with

$$E_w(w) = \frac{1}{2} w^T w$$

3



EX. 6

1. given a dataset $D = \{(x_n, t_n)\}_{n=1}^{N}$ and a target function $f: X \to Y$ kNN is an instance based model.

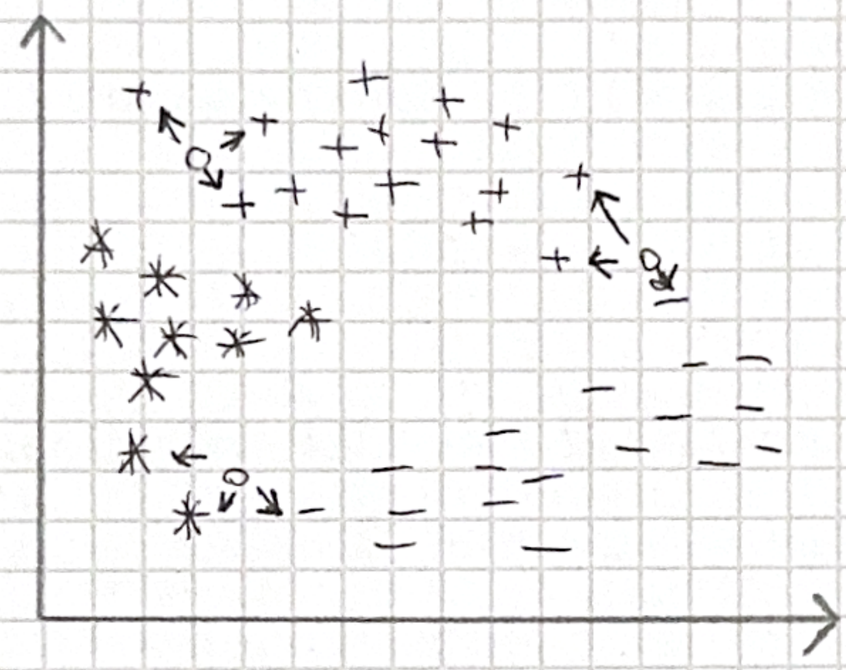
   1. find k nearest neighbor of new instance x
   2. assign to x the most common label among the majority of neighbor

   likelihood of class of new instance x

$$P(c \mid x, D, k) = \frac{1}{k} \sum_{x_n \in N_k(x, D)} \mathbb{I}(t_n = c) \quad \text{where}$$

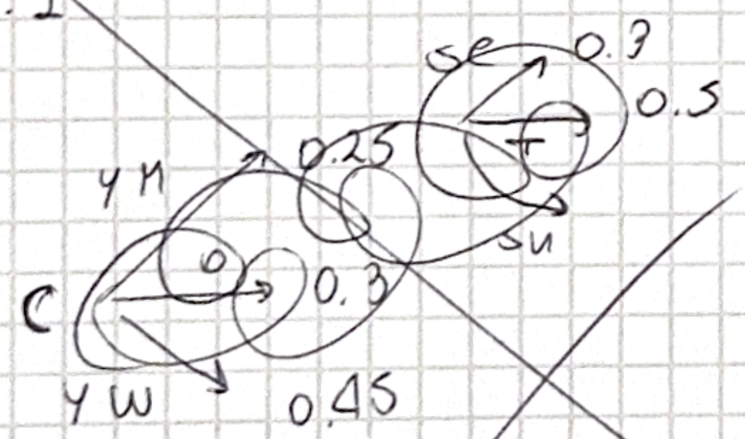
$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

2.



ML EXAMS 20/01/2020

EX. 1



$P(YM) = 0.25$ $P(YW) = 0.45$ $P(O) = 0.3$

$P(YM|S) = 0.3$ $P(YM|T) = 0.5$ $P(YM|R) = 0.2$

$P(YW|S) = 0.5$ $P(YW|T) = 0.3$ $P(YW|R) = 0.2$

$P(O|S) = 0.3$ $P(O|T) = 0.3$ $P(O|R) = 0.4$

$V^A = \text{argmax}$