EXAM TEST 1 2021/22

Question 1

1). This is a binary classification task. The target function is $f: X \to \{Y, N\}$ with $X = \{F \times NR \times NK\}$ and the dataset is $D = \{(x_i, y_i)_{i=1}^{N}\}$.

2). Since we are using the ID3 algorithm we select the attributes that maximize the information gain. $IG(S,a) = ent(S) - \sum_{v \in values\{A\}} \frac{|S_v|}{|S|} ent(S_v)$. The information gain is based on the concept of entropy: $ent(S) = \sum_{i=1}^{C} -p_i \log_2 p_i$. Entropy measures the impurity of our information.

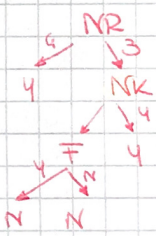
3). $ent(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

$IG(S,F) = ent(S) - \frac{2}{5}ent(F_Y) - \frac{3}{5}ent(F_N) = 0.971 - (0.916 \cdot 0.6) - 0.4 = 0.019$

$ent(F_Y) = 1 \quad ent(F_N) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.916$

$IG(S,NR) = ent(S) - \frac{3}{5}ent(NR_3) - \frac{2}{5}ent(NR_4) = 0.971 - (0.6 \cdot 0.916) = 0.4214$

$ent(NR_4) = 0 \quad ent(NR_3) = 0.916$



$IG(S,NK) = -\frac{1}{5}ent(NK_Y) - \frac{4}{5}ent(NK_N) = ent(S) - (0.8 \cdot 1) = 0.171$

$ent(NK_Y) = 0 \quad ent(NK_N) = 1$

$IG(S,F) = -\frac{2}{3}ent(F_N) - \frac{1}{3}ent(F_Y) + ent(S) = 0.971 - (1 \cdot 0.3) \cdot 2 = 0.304$

$ent(F_N) = 1 \quad ent(F_Y) = 0$

$ent(NK_Y) = 0 \quad ent(NK_N) = 0$

$IG(S,NK) = ent(S) - \frac{2}{3}ent(NK_N) - \frac{1}{3}ent(NK_Y) = 0.971$

Question 2

1). Given a dataset D and an hypotheses space H we are interested in $p(H|D)$. We can not compute this probability but, knowing $h \in H$, we can compute $p(h|D) = \frac{p(D|h)p(h)}{p(D)}$.

Now we introduce the maximum a posteriori hypothesis: $h_{MAP} = \underset{h \in H}{\text{argmax}} \frac{p(D|h)p(h)}{p(D)} =$

$= \underset{h \in H}{\text{argmax}}\, p(D|h)p(h)$; we can in fact ignore the denominator in the argmax.

If i assume that $p(h_i) = p(h_j)$ i can simplify and i obtain the maximum likelihood hypothesis: $h_{ML} = \underset{h \in H}{\text{argmax}}\, p(D|h)$

2). This statement is false. We do an example:

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

$h_{ML} = \underset{h \in H}{\text{argmax}}\, p(D|h) = \text{argmax}\{0.4, 0.3, 0.3\} = +$

Now we introduce Bayes Optimal Classifier (BOC). BOC is an optimal learner, no other methods with the same hypotheses space and the same prior knowledge can outperform this method. $V_{MB} = \underset{v \in V}{\text{argmax}} \sum_{h_i \in H} p(v|x, h_i)\, p(h_i|D)$

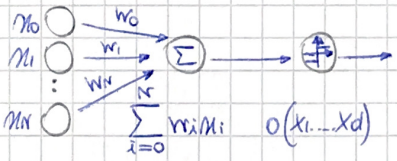
$p(+|x, h_1) = 1$ $p(+|x, h_2) =$ $p(+|x, h_3) = 0$

$p(-|x, h_1) = 0$ $p(-|x, h_2) = 1$ $p(-|x, h_3) = 1$

$V_{MB} = \underset{v \in V}{\text{argmax}}\{((1 \cdot 0.4) + 0 + 0, (0 + (1 \cdot 0.3) + (1 \cdot 0.3)\} = -$

Question 3

Perceptron is based on the following structure:



$o(x_1 \dots x_d) = \begin{cases} +1 & \text{if } w_0 + w_1 x_1 + \dots + w_d x_d > 0 \\ -1 & \text{otherwise} \end{cases}$

$\sum_{i=0}^{N} w_i x_i$ $o(x_1 \dots x_d)$ For the moment we consider $o(x) = w^T x$.

We want to minimize the error function $E(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - o_n)^2 = \frac{1}{2}\sum_{n=1}^{N}(t_n - w^T x_n)^2$.

We compute the gradient: $\frac{\partial E(w)}{\partial w_i} = \sum_{n=1}^{N}(t_n - w^T x_n)(-x_{i,n})$. We want to find w^* and we do this in a sequential way: $\hat{w} \leftarrow \hat{w} + \eta \sum_{n=1}^{N}(t_n - w^T x_n) x_{i,n} \Rightarrow \hat{w} \leftarrow \hat{w} + \eta \sum_{n=1}^{N}(t_n - \text{sign}(w^T x_n)) x_{i,n}$

This method is robust to outliers, where outliers are samples derived from a different probability distribution.

Question 4

1). We want to minimize the modified error function: $E(w) = \sum_{n=1}^{N} E(y_n, t_n) + \lambda\|w\|^2$

knowing that we have a dataset $\mathcal{D} = \{(x_n, t_n)_{n=1}^{N}\}$ and $E(y_n, t_n) = (y_n - t_n)^2$

We can find $y(x, w^*) = \sum_{n=1}^{N} a_n x_n^T x$ and now we apply the kernel trick:

$y(x, w^*) = \sum_{n=1}^{N} a_n k(x_n, x)$. We can do this knowing that $a = (XX^T + \lambda I_n)^{-1}$, but is really computationally expensive because requires $|\mathcal{D}|^2$ matrices multiplications.

2). The gram matrix is $K = XX^T$ and in this case is $K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_N) \\ k(x_N, x_1) & k(x_N, x_N) \end{pmatrix}$

Question 5

1). We can use a linear regression model. We simply rename the dataset for simplicity: $\mathcal{D} = \{(x_i, y_i)_{i=1}^{N}\}$. We know that $y(x, w) = \sum_{j=1}^{N} w_j x$ and the target value is $t = y(x, w) + \varepsilon$ with ε additive gaussian noise. If we assume that: $p(\varepsilon|\beta) = \mathcal{N}(\varepsilon|0, \beta^{-1})$ i obtain that $p(t|x_1 \dots x_N, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1})$. Now i consider the iid hypothesis.

$p(\{t_1 \dots t_n\}|x_1 \dots x_N, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T\phi(x_n), \beta^{-1}) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|w^T\phi(x_n), \beta^{-1}) =$

$= -\beta \cdot \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T\phi(x_n))^2 - \frac{N}{2} \ln(2\pi\beta^{-1})$. The error function becomes:

$E_{\mathcal{D}}(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T\phi(x_n))^2$ and we want to find $w^* = \operatorname{argmin}_w E_{\mathcal{D}}(w)$

We can do this in a sequential way: $\hat{w} \leftarrow \hat{w} + \eta[t_n - w^T\phi(x_n)]\phi(x_n)$

2). We can avoid overfitting using regularization:

$\operatorname{argmin}_w E_{\mathcal{D}}(w) + \lambda E_W(w)$ with $\lambda > 0$ regularization factor and $E_W(w) = \frac{1}{2} w^T w$

Question 6

1). We compute the k nearest neighbors of a new instance $x \notin \mathcal{D}$.

We assign to x the most common label among the majority of neighbors.

$p(c|x, \mathcal{D}, k) = \frac{1}{k} \sum_{x_n \in S} \mathbb{I}(c = t_n)$ with $\mathbb{I}(e) = \begin{cases} 1 \text{ if } e \text{ is true} \\ 0 \text{ otherwise} \end{cases}$

and $S = r_k(x_n, \mathcal{D})$ k nearest neighbors of x_n.