EXAM TEST 04.11.19

Question 1

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

2). If we use the ID3 algorithm we choose the attribute with the highest information gain: $IG(S,a) = \text{Entropy}(S) - \sum_{v \in V} \frac{|S_v|}{|S|} \text{entropy}(S_v)$ with V = values $\{a\}$

With the entropy = $\sum_{i=1}^{N} -p_i \log_2 p_i$ that measures the impurity of the informat

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

$IG(S,F) = ent(S) - \frac{2}{5}ent(F_N) - \frac{3}{5}ent(F_Y) = 0.0194$

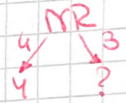
$ent(F_N) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.916 \quad ent(F_Y) = 1$

$IG(S, NR) = ent(S) - \frac{2}{5}ent(NR_4) - \frac{3}{5}ent(NR_3) = 0.4214 \longrightarrow$ HIGHEST IG!

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

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

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



We continue using this technique.

Question 2

1). Given a classification task with a dataset D and an hypotheses space H

we are interested in $p(H|D)$. We know that $p(h|D) = \frac{p(D|h)p(h)}{p(D)}$

We want to find the Maximum A Posteriori (MAP) hypothesis: $h_{MAP} = \underset{h \in H}{\text{argmax}} \frac{p(D|h)p(}{p(D)}$

$= \underset{h \in H}{\text{argmax}}\, p(D|h)p(h)$. If we assume that $p(h_i) = p(h_j)$ we can further simp

fy and we obtain the maximum likelihood (ML) hypothesis $h_{ML} = \underset{h \in H}{\text{argmax}}\, p(D|h)$

2). This is false. We introduce 3 hypotheses:

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

$h_{ML} = \underset{h \in H}{\text{argmax}} \{0.4, 0.3, 0.3\} = +$

Now we introduce Bayes Optimal Classifier:

$v_{BOC} = \underset{v \in V}{\operatorname{argmax}} \sum_{h_i \in H} p(v|x, h_i)\, p(h_i|D)$.

Bayes optimal classifier is an optimal classifier, in fact it returns always the optimal solution.

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

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

$v_{NB} = \operatorname{argmax}\{1 \cdot 0.4 + 0 + 0,\ 0 + 1 \cdot 0.3 + 1 \cdot 0.3\} = -$

This confirms that the hypothesis is false.

Question 3

Least squares is a linear model for classification. In this method we want to minimize an error funcion that is also 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 function sum of squares beca

se the trace is simply the sum of the elements in the main diagonal and The product of one matrix for the transpose is simply the square.

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

Question 4

1). The gram matrix is $K = XX^T$. If we have a kernel 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 $K = \begin{pmatrix} x_1^T x_1 & x_1^T x_n \\ x_n^T x_1 & x_n^T x_n \end{pmatrix}$

If we have a generic kernel function $k(x, x') \rightarrow y(x, w^*) = \sum_{n=1}^{N} w_n k(x_n, x)$ and the gram matrix K is $K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_n) \\ k(x_n, x_1) & k(x_n, x_n) \end{pmatrix}$

2). We use the modified error function $\tilde{J}(w) = \sum_{n=1}^{N} E(y_n, t_n) + \lambda \|w\|^2$

We know that $E(y_n, t_n) = (y_n - t_n)^2$ and $y(x, w^*) = \sum_{n=1}^{N} a_n x_n^T x$. Now we kernelize this solution: $y(x, w^*) = \sum_{n=1}^{N} a_n k(x_n, x)$. This is a solution for our problem but is really computationally expensive, in fact requires $|D|^2$ matrices multiplications.

Question 5

1). We can use a linear regression model. We know that $y(x,w) = \sum_{j=1}^{M} w_j \phi_j(x)$.

The target value is $t = y(x,w) + \varepsilon$ with ε additive gaussian noise.

$p(\varepsilon|\beta) = N(\varepsilon|0, \beta^{-1})$. With this hypothesis we obtain: $p(t|x_1 \dots x_N, w, \beta) = N(t|y(x,w),$

Using the iid hypothesis: $p(\{t_1 \dots t_N\}|x_1 \dots x_N, w, \beta) = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta^{-1}) =$

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

The error function is $E(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T\phi(x_n))^2$. We use sequential learning to find the solution: $\hat{w} \leftarrow \hat{w} + \eta[(t_n - w^T\phi(x_n))\phi(x_n)]$

2). We can reduce overfitting using regularization:

$w^* = \arg\min_w E_D(w) + \lambda E_W(w)$ with $\lambda > 0$ regularization factor.

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

Question 6

1). We compute the k nearest neighbors of a new instance $x \notin D$ and then we assign to x the most common label among the majority of neighbors.

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