EXAM 18.01.19A

Question 1

1). If C = c_1 AND B = b_1 THEN NO ①
If C = c_1 AND B = b_2 THEN YES ②
If C = c_2 AND A = a_1 THEN YES ③
If C = c_2 AND A = a_2 AND B = b_1 THEN YES ④
If C = c_2 AND A = a_2 AND B = b_2 THEN NO ⑤
If C = c_2 AND A = a_3 THEN NO ⑥
If C = c_3 THEN NO ⑦

2). T is consistent with S_1 because of ①.
T " " " S_2 " " ④.
T is not consistent with S_3 because of ⑦.
" " " " " S_4 " " ⑤.

Question 2

1). The maximum a posteriori hypothesis is $h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \frac{p(D|h)p(h)}{p(D)} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)p(h)$

If we assume that $p(h_i) = p(h_j)$ we simplify and we obtain $h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)$.

2). BOC is an optimal classifier, it returns always the most probable prediction.

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

3). The problem of BOC is that we can use it only if we have analytical solutions or if the hypotheses space is not large, otherwise is not practical.

Question 3

1). In linear regression we have a target function $f: X \to Y$ with $X \subseteq \mathbb{R}^d$ and $Y = \mathbb{R}$. The model is $y(x, w) = \sum_{j=1}^{M} w_j \phi_j(x)$ that is still linear in the parameters w. We want to find $w^* = \underset{w}{\operatorname{argmin}}\, E_D(w)$ with $E_D(w)$ error function but we have to discover $E_D(w)$.
We have a dataset $D = \{(x_n, t_n)_{n=1}^{N}\}$ and we know that: $t = y(x, w) + \varepsilon$ with ε additive Gaussian noise. $p(\varepsilon|\beta) = N(\varepsilon|0, \beta^{-1}) \Rightarrow p(t|x_1 \dots x_N, w, \beta) = N(t|y(x, w), \beta^{-1})$
If we use the iid hypothesis: $p(\{t_1 \dots t_N\}|x_1 \dots x_N, w, \beta) = \prod_{n=1}^{N} p(t_n|w^T\phi(x_n), \beta^{-1}) =$

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

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

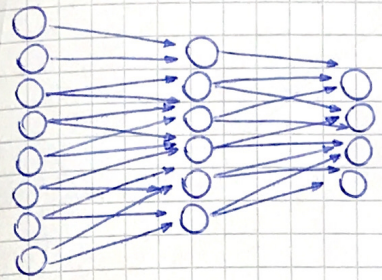
2). We can find w^* in batch or in a sequential way:

if we use least squares $E_S(w) = \frac{1}{2}(t-\phi w)^T(t-\phi w)$ and $w^* = \phi^+ t$.

if we use sequential learning: $\hat{w} \leftarrow \hat{w} + \eta[t_n - w^T\phi(x_n)]\phi(x_n)$

Question 4

1).



$\#params = (8+1)\cdot 6 + (6+1)\cdot 4 = 82$

2). Backpropagation is an algorithm that is used for compute the gradient and the gradient will be propagated thrown all the network. So BackProp is not affected by overfitting or by local minima.

Question 5

1). The dataset is linearly separable because exists a separation surface that splits our instance space into two regions such that different classified instances are separated.

2). I can use a kernel function of the polynomial type $K(x,x') = (\beta x^T x' + \gamma)^d$

Using for example $d=3$.

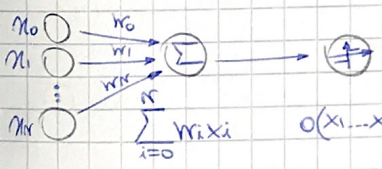
3). $y(x,\alpha) = \text{Sign}\left(w_0 + \sum_{n=1}^{N} a_n x_n^T x\right)$ i apply the kernel trick

$y(x,\alpha) = \text{Sign}\left(w_0 + \sum_{n=1}^{N} a_n K(x_n, x)\right)$

$w_0 = \frac{1}{|SV|}\sum_{x_i \in SV}\left(t_i - \sum_{x_j \in SV} a_j t_j K(x_i, x_j)\right)$

Question 6

1). Perceptron is based on the following structure:



$\sum_{i=0}^{N} w_i x_i \qquad o(x_1 \dots x_d)$

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

For the moment we will consider $o(x) = w^T x$

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

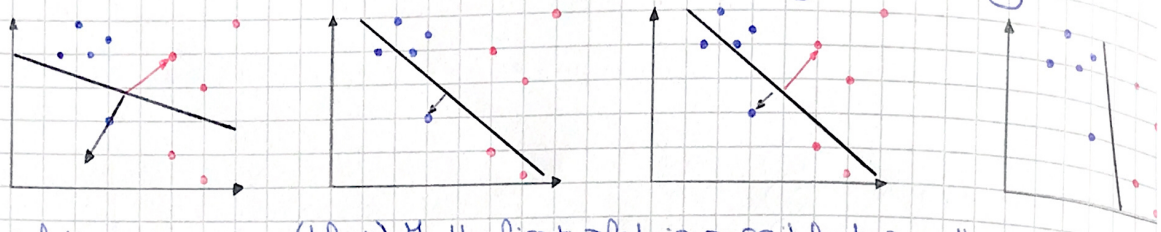
$\frac{\partial E(w)}{\partial w_i} = \sum_{n=1}^{N}(t_n - w^T x_n)(-x_{i,n})$. We find w^* in a sequential way.

$\hat{w} \leftarrow \hat{w} + \Delta w_i$ with $\Delta w_i = -\eta \sum_{n=1}^{N} (t_n - w^T x_n)(x_{i,n})$.

Now we consider another time the sign function: $\Delta w_i = -\eta \sum_{n=1}^{N} (t_n - \text{Sign}(w^T x_n))(x_{i,n})$

η is a very important hyperparameter called learning rate, if η is too small we can take a lot of time to reach the solution, if η is too big we can diverge.

2).



w is plotted as a vector (black). In the first plot is possible to see that we have an error, so we add the vector w to the feature vector of the missclassified point. In the plot 3 is possible to see that we have other missclassified points so we have to sum another time the vector w to the feature vector of the misclassified point.