EXAM 18.01.14 5

Question 1

1). This dataset is linearly separable because exists a separation surface that splits our instance space into 3 regions such that different classified instances are separated.

2). I can use a kernel function $K(x,x') = (\beta x^T x' + \gamma)^d$, for example with $d=3$. This is a polynomial kernel function.

3). We have to use also in this case the kernel trick:

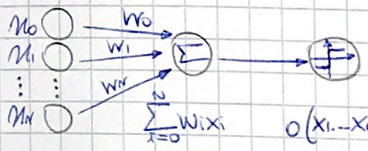
$y(x,\alpha) = \text{sign}\left(w_0 + \sum_{n=1}^{N} \alpha_n x_n^T x\right)$ is possible to see an inner product.

$y(x,\alpha) = \text{sign}\left(w_0 + \sum_{n=1}^{N} \alpha_n K(x_n, x)\right)$ We have applied the kernel trick.

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

Question 2

1). The perceptron model is based on the following structure:



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

For the moment for simplicity we consider only $o(x) = w^T x$.

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

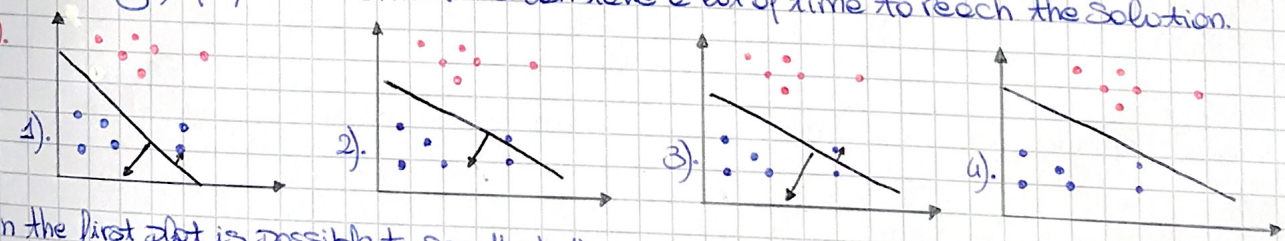
We compute the gradient: $\frac{\partial E(w)}{\partial w_i} = \sum_{n=1}^{N}(t_n - w^T x_n)(-x_{i,n})$

In perceptron we try to found the optimal solution using a sequential method:

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

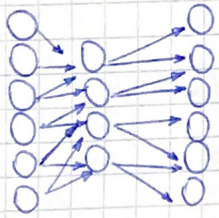
η is a very importantant hyperparameter called learning rate. If η is too large we can diverge, if η is too small we can take a lot of time to reach the solution.

2).



In the first plot is possible to see that there are misclassified point. We sum the vector related to the separation surface with the feature vector of the misclassified point and we obtain the plot number 2. In the plot number 3 we do the same and we obtain the correct separation surface in the plot 4.

Question 3



$\#\text{parameters} = (7\times4)+(5\times6) = 58 = (6\times4+4)+(4\times6+6)$

Backpropagation is not affected by local minima because is not a learning algorithm but is an algorithm used to compute the gradient and to propagate the gradient throw all the network. For this same reason is not affected by overfitting.

Question 4

1). In linear regression we have a target function $f: X \to Y$ with $X \subseteq \mathbb{R}^d$ and $Y = \mathbb{R}$ and a dataset $\mathcal{D} = \{(x_n, t_n)_{n=1}^N\}$. The goal is to find $w^* = \arg\min_w E_{\mathcal{D}}(w)$. We have a model $y(x,w) = \sum_{j=1}^{M} w_j \phi$ and we know that the target value is $t = y(x,w) + \varepsilon$ with ε additive gaussian noise. Using the information about the gaussian $p(\varepsilon|\beta) = \mathcal{N}(\varepsilon|0,\beta^{-1}) \Rightarrow p(t|x_1 \dots x_N, w, \beta) = \mathcal{N}(t|y(x,w),\beta$

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

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

We have found the error function $E_{\mathcal{D}}(w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2$

2). For the batch mode we can use least squares: $E_{\mathcal{D}}(w) = \frac{1}{2}(t-\phi w)^T(t-\phi w)$ and $w^* = \phi^+ t$.

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

Question 5

1). $h_{MAP} = \arg\max_{h\in H} p(\mathcal{D}|h)\frac{p(h)}{p(\mathcal{D})} = \arg\max_{h\in H} p(h|\mathcal{D}) = \arg\max_{h\in H} p(\mathcal{D}|h)p(h)$

If i assume that $p(h_i) = p(h_j)$ i can simplify and i obtain the ML hypothesis: $h_{ML} = \arg\max_{h\in H} p(\mathcal{D}|h)$

2). given a target function $f: X \to V$ knowing that each instance can be described as $\langle a_1 \dots a_N \rangle$:

$$p(v|x,\mathcal{D}) = p(v|a_1 \dots a_N, \mathcal{D}) = \frac{p(a_1 \dots a_N|v,\mathcal{D})p(v|\mathcal{D})}{p(a_1 \dots a_N|\mathcal{D})} = p(a_1 \dots a_N|v,\mathcal{D})p(v|\mathcal{D})$$

$v_{NB} = \arg\max_{v\in V} p(a_1 \dots a_N|v,\mathcal{D})p(v|\mathcal{D})$. Now we have to introduce a simplification. Each sample is mutually conditionally independent from the other $\to p(a_1 \dots a_N|v,\mathcal{D}) = \prod_{i=1}^{N} p(a_i|v,\mathcal{D})$

We obtain that: $v_{NB} = \arg\max_{v\in V} p(v|\mathcal{D})\prod_{n=1}^{N} p(a_i|v,\mathcal{D})$

3). NBC is not an optimal classifier differently from BOC because is based on the conditional independence assumption, but can be used in all the cases.