EXAM 20.01.20

Question 1

1). $p(YH) = 0.25$ $p(YW) = 0.45$ $p(O) = 0.3$

$p(S|YH) = 0.3$ $p(T|YH) = 0.5$ $p(R|YH) = 0.2$

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

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

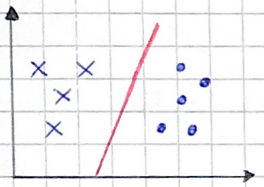
$$v^* = \underset{c \in C}{\text{argmax}} \{p(T|c)p(c)\} = \underset{c \in C}{\text{argmax}} \{0.5 \cdot 0.25, 0.3 \cdot 0.45, 0.3 \cdot 0.3\} =$$

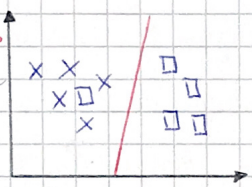
$$= \text{argmax}\{0.125, 0.135, 0.09\} = YW$$

2). $\alpha = \sum_{c \in C} p(T|c)p(c) = 0.125 + 0.135 + 0.09 = 0.35$ (Likelihood)

Question 2

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

2). 

3). 

4). The first dataset is linearly separable. Its possible to see that this solution is computed with SVM because SVM aims at maximum margin with the better accuracy. Let x_n be the closest point to the separation surface $\bar{h} = \bar{w}_0 + \bar{w}^T x = 0$. We define margin $\frac{|y(x_n)|}{\|w\|}$. To find the margin we have to find $\min_{n=1\dots N} \frac{t_n(w^T x_n + w_0)}{\|w\|}$. But we want to find the maximum margin: $w_0^*, w^* = \underset{w_0, w}{\text{argmax}} \frac{1}{\|w\|} \min_{n=1\dots N} t_n[w^T x_n + w_0]$

Scaling we are not affecting our solution: $t_n y(x_n) = 1$ and $t_n y(x_n) \geq 1 \ \forall n = 1 \dots N$

In this mode $w^*, w_0^* = \underset{w_0, w}{\text{argmin}} \frac{1}{2}\|w\|^2$. Now we move to a Lagrangian domain.

$w^* = \sum_{n=1}^{N} \alpha_n^* t_n x_n$. We know that, using the KKT hypothesis, if $t_n y(x_n) \geq 1 \Rightarrow \alpha_n^* = 0$

We obtain that $w^* = w_0^* + \sum_{x_j \in SV} \alpha_j^* t_j x^T x_j = 0$ with $SV = \{x_n \in D | t_n y(x_n) = 1\}$

Also the second solution is computed with SVM. In fact, we can use SVM if the dataset is not perfectly linearly separable introducing slack variables $\xi_n \geq 0$, $t_n y(x_n) \geq 1 - \xi_n \quad \forall n = 1 \dots N$ and the modified optimal value w^*:

$$w^* = \arg\min_w \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} \xi_n$$

Question 3

1). In PCA we want to maximize the data variance after the projection to some direction u_1. The projected points are $u_1^T x_n$ subject to $u_1^T u_1 = 1$.

The data variance is: $\frac{1}{N}\sum_{n=1}^{N}(u_1^T x_n - u_1^T \bar{x}_n)^2 = u_1^T S u_1$ with $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

and $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x}_n)(x_n - \bar{x}_n)^T = X^T X$

We want to find $\max_{u_1} u_1^T S u_1$ S.t. $u_1^T u_1 = 1 \iff \max_{u_1} u_1^T S u_1 + \lambda_1(1 - u_1^T u_1)$

If we set the derivative equal to $0 \Rightarrow Su_1 = \lambda u_1 \Rightarrow u_1^T S u_1 = \lambda_1$

S is the eigenvalue associated to the eigenvalue λ_1. S is called first principal component.

2). Yes, because the intrinsic dimensionality of the dataset is 3.

Question 4

Layer 1: $|N_1| = 5\times5\times3\times16 + 16$

Layer 2: $|N_2| = 3\times3\times16\times32 + 32$

Layer 3: $|N_3| = 5\times3\times32\times64 + 64$

FC Layer: $|N_{FC}| = 200 \times 10$

A suitable loss function is categorical cross-entropy:

$J(N) = E[-\ln p(t|x, N)]$ If we assume additive gaussian noise:

$p(t, x, N) = N(t|f(x, N), \beta I) \Rightarrow J(N) = E\left[\frac{1}{2}\|t - f(x, N)\|^2\right]$

Question 5

1). The model can be defined as: $y(x, w) = \sum_{n=1}^{N} w_n \phi_n(x) = w^T \phi(x)$. We know that the target value is $t = y(x, w) + \varepsilon$ with ε additive gaussian noise. If the noise is gaussian $p(\varepsilon|\beta) = N(\varepsilon|0, \beta^{-1}) \Rightarrow p(t|x_1 \dots x_N, w, \beta) = N(t|y(x, w), \beta^{-1})$.

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

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

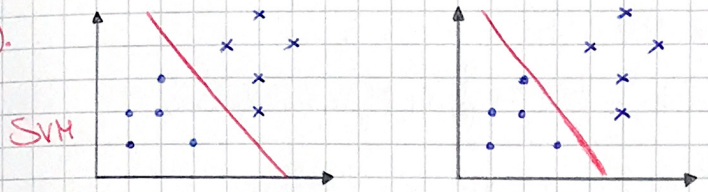
2). We consider the error function $E_D(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2$.

We can use a sequential learning algorithm to find $w^* = \underset{w}{\operatorname{argmin}}\, E_D(w)$.

$\hat{w} \leftarrow \hat{w} + \eta [t_n - w^T \phi(x_n)] \phi(x_n)$

Question 6

1).



2). The main difference is that SVM aims at maximum margin with the better accuracy, while perceptron no. Perceptron stop the iterations when finds a possible separation surface simply summing the vector related to the separation surface with the feature vector of the misclassified point. For this reason very often we find a separation surface that is very close to the points. SVM try to maximize this margin and for this reason we have the plotted separation surface.

3). I prefer SVM because, using the principle of maximum margin, we have a very low probability of misclassification, while in perceptron no. Another advantage of SVM is that we can use it also if the dataset is not perfectly linearly separable.