EXAM 19.01.18

Question 1

1). $W_{C1} = 1242 - 5 + 4 + 1 = 1242$ $\quad h_{C1} = 378 \Rightarrow$ After C1: $1242 \times 378 \times 64$

$W_{P1} = \frac{1242-2}{2} + 1 = 621$ $\quad h_{P1} = \frac{378-2}{2} + 1 = 189 \Rightarrow$ After P1: $621 \times 189 \times 64$

$W_{C2} = \frac{621-3}{2} + 1 = 310$ $\quad h_{C2} = \frac{189-3+1}{2} = 94 \Rightarrow$ After C2: $310 \times 94 \times 128$

$W_{P2} = \frac{310-2}{4} + 1 = 78$ $\quad h_{P2} = \frac{94-2}{4} + 1 = 24 \Rightarrow$ After P2: $78 \times 24 \times 128$

2). Layer 1: #params $= 3 \times 5 \times 5 \times 64 + 64 = 4864$

Layer 2: #params $= 64 \times 3 \times 3 \times 128 + 128 = 73856$

Question 2

1). This is a greyscale image of 12×12 pixels so the dimensionality of the data space is $2^{12 \times 12}$. The intrinsic dimensionality is 3: two coordinates for the center of the image and one for the rotation.

2). We want to maximize the data variance after the projection to some direction u_1.

We know that the projected points are $u_1^T x_n$ subject to $u_1^T u_1 = 1$

The variance is: $\frac{1}{6} \sum_{n=1}^{6} (u_1^T x_n - u_1^T \bar{x}_n)^2 = u_1^T S u_1$ with $S = \frac{1}{6} \sum_{n=1}^{6} (x_n - \bar{x}_n)(x_n - \bar{x}_n)^T = X^T X$

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

If we keep the derivative and we impose that is equal to 0.

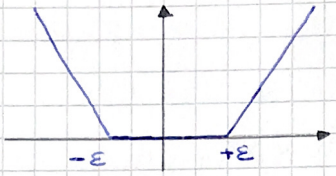
$S u_1 = \lambda_1 u_1 \Rightarrow u_1^T S u_1 = \lambda_1$

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

3). No, usually M is greater than the number of intrinsic dimensions because the principal components of PCA are not latent variables.!

Question 3

1).



The problem of this function is that is not differentiable. Is possible to see that between $-\varepsilon$ and $+\varepsilon$ the value of the function is 0.

2). We introduce some slack variables $\xi_n^+, \xi_n^- \geq 0 \quad \forall n = 1 \dots N$

We know that: $t_n \leq y_n + \varepsilon + \xi_n^+$ and $t_n \geq y_n - \varepsilon - \xi_n^-$

We can derive that: $y_n - \varepsilon \leq t_n \leq y_n + \varepsilon \Rightarrow \xi_n = 0$. The region of the plane between $y_n - \varepsilon$ and $y_n + \varepsilon$ is called ε-tube. The modified error function becomes:

$$J(w) = C \sum_{n=1}^{N} (\xi_n^+ + \xi_n^-) + \frac{1}{2} \|w\|^2$$

Question 4

The linear classification method that i have chosed is least squares. The target function is $f: X \to Y$ with $X \subseteq \mathbb{R}^d$ and the dataset is $\mathcal{D} = \{(x_n, t_n)_{n=1}^N\}$. We want to find $\tilde{W}$ such that $y(x) = \tilde{W}^T \tilde{x}$.

We minimize the error function that is called sum of squares.

$E(\tilde{W}) = \frac{1}{2} Tr\{(T - \tilde{X}\tilde{W})^T (T - \tilde{X}\tilde{W})\}$. Is called sum of squares because the trace operator is simply a sum and because the product between one matrix and the transpose is the square. We want to find $\tilde{W}^* = \arg\min_{\tilde{W}} E(\tilde{W})$

$\tilde{W} = \tilde{X}^\dagger T$ and $y(x) = T^T (\tilde{X}^\dagger)^T \tilde{X}$. This method is not robust to outliers because is simply based on a distance and outliers are sample derived from a different probability so they are very far away from all the other samples and the solution will be affected.

Question 5

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

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

2). Bayes optimal Classifier (Boc) is an optimal classifier, it returns always the optimal solution. Given a target function $f: X \to V$:

$$p(v|x, D) = \sum_{h_i \in H} p(v|x, D, h_i)\, p(h_i|x, D) = \sum_{h_i \in H} p(v|x, h_i)\, p(h_i|D)$$

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

3). Boc is an optimal classifier but can be used if the hypotheses space is not large or if we have analytical solutions, otherwise is not practical and we must use different methods like Naive Bayes Classifier

Question 6

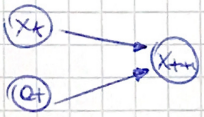
1). The Markov property says that:

Once the current state is known the evolution of a dynamic system does not depend on previous states, actions and observations. The current state contains all the informations needed to predict the future.

Future states are conditionally independent of past states and past observations given the current state.

The knowledge about the current state makes past, present and future observations statistically independent.

2). An MDP can be described as $MDP = \langle X, A, \delta, r \rangle$ with X set of states, A set of actions, δ transition function and r reward function. An HMM can be described as $HMM = \langle X, Z, \pi_0 \rangle$ with X set of states, Z set of observations and π_0 initial distribution. The main difference is that in MDPs we have the property of full observability (states are fully observable), while in HMM no and we need observations.

MDP: 

HMM: