EXAM TEST 20.12.18 (K-MEANS)

## Question 1

1). The gram matrix is $K = XX^T$ that represents the inner product between the input vector and the transpose of the input vector. If we use a kernel function $k(x, x') = x^T x' \longrightarrow y(x) = \sum_{n=1}^{N} a_n x_n^T x$ and $K = \begin{pmatrix} x_1^T x_1 & \cdots & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & \cdots & x_N^T x_N \end{pmatrix}$

If we have a generical kernel function $k(x, x') \longrightarrow y(x) = \sum_{n=1}^{N} a_n k(x_n, x)$

and $K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$

2). Is possible to obtain a kernelized version of regression. The error function becomes $J(w) = \sum_{n=1}^{N} E(y_n, t_n) + \lambda \|w\|^2$ with $E(y_n, t_n) = (y_n - t_n)^2$

We can use the equations defined precedently and we obtain: $y(x, w^*) = \sum_{n=1}^{N} a_n x_n^T x$

We apply the kernel trick $\longrightarrow y(x, w^*) = \sum_{n=1}^{N} a_n k(x_n, x)$

This is really computationally expensive, in fact requires $|D|^2$ matrices multiplications.

## Question 2

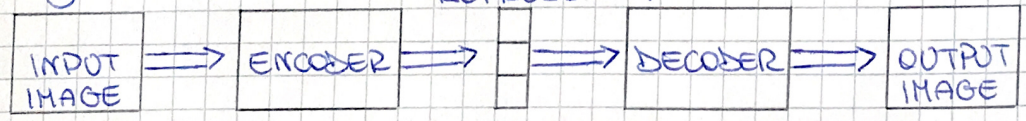
1). $\dim(W_1) = 128 \times 10 = 1280$ $\quad\quad$ $\dim(W_2) = 10 \times 50 = 500$

2). $h = g(W^T x + c)$ with $g(a) = \max(0, a)$ Relu function

$y(x, \theta) = w^T \max(0, W^T x + c) + b$

## Question 3

1). The goal of an autoencoder is dimensionality reduction. An autoencoder is composed by 2 neural networks, an encoder and a decoder. The structure is the following:



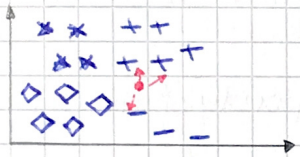

In autoencoder usually we have hidden layers with reduced size that are also called bottlenecks. The training is based on recostruction loss.

Given a dataset $\{x_n\}$, autoencoders are trained with the sample $x_n$ in input and in output.

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

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

$N_K(x_n, D)$ K nearest neighbors of $x_n$.

2). 

K=3 The sample will be classified as +.

## Question 6

1). We can use boosting, a simple ensemble method. We split our dataset D into M bootstraped datasets $D_1 \dots D_M$ and we train each model $y_i(x)$ using the dataset $D_i$ $\forall i = 1 \dots M$

$$y_{\text{~~ensemble~~}}(x) = \frac{1}{M} \sum_{k=1}^{M} y_k(x)$$

2). No, i think that there are no properties that have to be satisfied.