EXAM 22.07.21

Question 1

1). Given an hypothesis h in the hypotheses space H, h overfits the training data if exists another hypothesis $h' \in H$ such that $error_S(h) < error_S(h')$ and $error_D(h) > error(h')$

2). We can have overfitting in decision trees. If, for example, one tree is much deeper than another tree for the same task for sure we have overfitting. Is possible to have overfitting also in neural networks, for example if we use a wrong number of iterations to train our model.

3). One possible solutions for decision tree is reduced error pruning: we split our training set into training set and validation set. We evaluate the impact on training set of pruning each possible node. We greedily remove the node that most improves validation accuracy.

Question 2

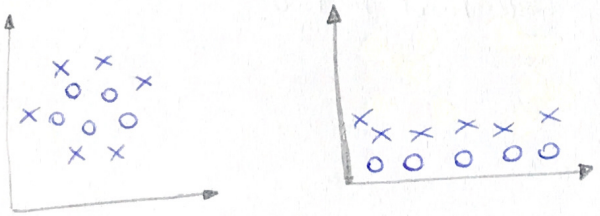
1). This is a regression task. We can use a linear regression model. We want to find $w^* = \arg\min_w E_D(w)$ knowing that the target function is $g: X \to Y$ with $X = \{x_1\}$ and $Y = \mathbb{R}$ and the dataset is $D = \{(x_n, f_n)\}_{n=1}^3$.
We know also that $E_D(w) = \frac{1}{2}\sum_{n=1}^{N} (f_n - w^T\phi(x_n))^2$ and $y(x,w) = \sum_{j=1}^{N} w_j^T \phi_j(x)$

2). A non-optimal solution is obtained using sequential learning:
$\hat{w} \leftarrow \hat{w} + \eta[f_n - w^T\phi(x_n)]\phi(x_n)$
An optimal solution is obtained using least squares:
$E_D(w) = \frac{1}{2}(f - \phi w)^T(f - \phi w) \Rightarrow w^* = \phi^+ f.$

Question 3

1). When the input vector appears only in the form of an inner product x^Tx' we can replace x^Tx' with a kernel function $k(x,x')$.

2).



This solution is obtained using a polynomial kernel $k(x,x') = (\beta x^T x' + \gamma)^d$

Question 4

1). In a convolutional stage is performed the convolution operation:

$$(I * K)(i,j) = \sum_{m \in S_1} \sum_{n \in S_2} I(m,n) \cdot K(i-m, j-n)$$

The property that is used is sparse connettivity but very often we can also find parameter sharing.

It is possible to find also padding in this stage, in this mode we can choose the correct steps for our convolution operation.

2). In parameter sharing we force some parameters to share the same value. In this mode we are reducing the number of parameters that we have to learn.

In sparse connettivity we are saying that outputs depend only on a few inputs. We are not connecting each node to all the other nodes of the successive layers but only to the closest.

Question 5

1). K-means is an unsupervised learning technique. The input is $D = \{x_n\}$ the dataset and the output is $\mu_1, \mu_2 \dots \mu_k$.

- We select the value of k = #clusters.
- We have to split our samples: we can do this randomly or sistematically We take the first k samples and we assign each of them to a 1 elem nt cluster. For the remaining $N-k$ samples we assign each sample to the cluster with the nearest centroid and we recompute the centroid.
- For all the samples we compute the distance from all the clusters and if the sample is not correctly classified we move it to the correct cluster and we recompute the centroid of the modified clusters.
- We repeat the previous step until convergence is achieved.

2). The algorithm will fail if:

- There are ∞ partitions of training samples into k clusters
- For each switch in step 2, the sum of distances from each training sample to that training sample's group centroid increases.

3).