EXAM 12.04.19

Question 1

1). Given a dataset D and an hypotheses space H: $h \in H$ overfits training data if exists another hypothesis $h' \in H$ such that $error_S(h) < error_S(h')$ and $error_S(h) > error_S(h$

2). In decision trees is very common to have problems of overfitting. If we have two solutions for a problem and one tree is deeper than the other for sure we have overfitting. We have 2 possible techniques to avoid overfitting:

- Reduced error pruning: we split our training set into training set and validation set. We evolute the impact on validation set of pruning each possible node. We prune the node that most improves validation accuracy.
- Rule post pruning: we generate all the tree and we move to a logical representation of the tree. We compute the accuracy and we start pruning. If the accuracy after pruning is greater than the accuracy before pruning we have overfitting and we can prune.

Question 2

1). Naive Bayes Classifier is based on the property of conditional indipendence. X is conditionally indipendent from Y given Z if $p(X,Y|Z) = p(X|Z)p(Y|Z)$.

Given a target function $f: X \to V$ with $V = \{v_1, v_2 \dots v_k\}$ knowing that each instance x can be described as $\langle a_1, a_2 \dots a_N \rangle$

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

Now we introduce the conditional indipendence hypothesis:

$p(a_1 \dots a_N|v,D) = \prod_{i=1}^{N} p(a_i|v,D)$ Each sample is mutually conditionally indipendent from another.

$$v_{NB} = \underset{v \in V}{\operatorname{argmax}}\; p(v|D) \prod_{i=1}^{N} p(a_i|v,D)$$

Is possible to see that this solution, differently from the BoC is not optimal but could used also if the hypotheses space is large or if we don't have analytical solutions.

2). This is a classification task. The target function is $f: X \to Y$ with $X = \{$title x author x abs x site$\}$ and $Y = \{HL, HR, PL\}$. The dataset is $D = \{(d_i, y_i)_{i=1}^{N}\}$.

$$v_{NB} = \underset{v \in V}{\operatorname{argmax}}\; p(v|D) \prod_{i=1}^{N} p(a_i|v,D)$$

We know that $p(d_i|v,\Delta) = \prod_{i=1}^{L} p(d_i = w_u|v,\Delta)$ with $L = \text{length}(d_i)$ and $p(d_i = w_u|v,\Delta)$ probability that the word u occurs in the document i.

$$\hat{P}(d_i|v,\Delta) = \frac{TF_{ij}+2}{TF_j+1}$$

Question 3

1). $p(\varphi|\tilde{w}) = \prod_{n=1}^{N} y_n^{p_n}(1-y_n)^{(1-p_n)}$ with $y_n = \sigma(\tilde{w}^T\tilde{x}_n)$

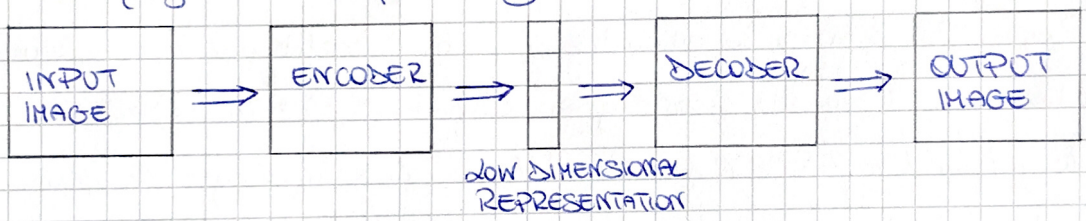
We want to minimize the cross-entropy error function $E(w) = -\sum_{n=1}^{N}[p_n \log(y_n) + (1-p_n)\log(1-y$

The parameters of the model that have to be learned are the numbers of hours e_i that the student i has attended a course and the number of hours s_i that the student has studied for the exam.

Question 4

1). An autoencoder is a combination of two NNs, a decoder and a decoder. The training is based on reconstruction loss and the hidden layers of the nets have a reduced size (bottlenecks). An autoencoder takes as input and as output the same sample x_n from the dataset $\{x_n\}$ and uses it for training.

2).



Question 5

1). In a dynamic system we have the property of fully observability if we can see the state resulting from the execution of one action. Also if we have non deterministic actions we can not predict the state resulting from the execution but we can observe it.

2). We can define an MDP as: $MDP = \langle X, A, \delta, r\rangle$. X is the set of states, A is the set of actions, $\delta: X \times A \to X$ is the transition function and $r: X \times A \to \mathbb{R}$ is the reward function. We can define a POMDP as: $POMDP = \langle X, A, Z, \delta, r, o\rangle$. The differences with the previous model are: Z is the set of observations, δ is the probability of the transition function and o is the probability of the observations and $p(x_0)$ is the probability of the initial state.

2).