Instructions. (16 points)
You have 2 hours to answer this exam.
Answer all questions on this paper, just circle the correct answer. If you make a mistake, mark a cross through your wrong choice and circle your next alternative. This paper is the only one that you have to deliver at the end of the exam.
For the multiple choice questions each correct answer scores the value of question divided by the number of corrects answers of the question (2 or 3).
Wrong answers penalize 1/3 of the value of the question for questions with only one answer and 1/2 of the value of the question for each wrong answer in the multiple choice questions.

(2pts) 1. Assume that you have a concept description language that only allows pure conjunctive formulas and we have a domain with the following attributes:

<table>
<thead>
<tr>
<th>Attr1</th>
<th>Attr2</th>
<th>Attr3</th>
<th>Attr4</th>
<th>Attr5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>A, B, C</td>
<td>1, 2, 3</td>
<td>a, b</td>
<td>1, 2, 3, 4</td>
</tr>
</tbody>
</table>

Given the following examples:

+ (A,B,2,a,3)
- (B,B,1,a,3)
+ (A,C,2,a,1)
- (B,A,1,a,3)
+ (A,B,2,a,4)
- (A,C,3,b,3)
- (A,C,2,a,2)

What concept learns the Version Space algorithm?

(a) (A, *, 2, a, *)
(b) There is not enough examples to converge
(c) The examples are inconsistent
(d) (A, *, 2, *, *)

(2pts) 2. We want to build a system able to control a process with four states \{A, B, C, D\} were we can perform the actions a and b. The following figure shows the state transition function (\(\delta\)) and the reinforcement (\(r\)) obtained by each action:

```
\(\begin{array}{c}
A \quad B \quad C \quad D \\
\quad a, 10 \quad a, 20 \quad a, 60 \\
\quad b, 50 \quad b, 50 \\
\quad a, 70 \\
\end{array}\)
```
Using the Q-Learning algorithm with $\gamma = 0.9$ and $\alpha = 1$, the action value function $Q$ that is obtained with the sequence that begins in the state $A$ and performs the actions $\{a,a,b,a,b,a\}$ is:

(a) $Q(s,a)$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>b</td>
<td>20</td>
<td>0</td>
<td>60</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) $Q(s,a)$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>10</td>
<td>20</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>b</td>
<td>47</td>
<td>0</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) $Q(s,a)$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>b</td>
<td>20</td>
<td>50</td>
<td>60</td>
<td>50</td>
</tr>
</tbody>
</table>

(d) $Q(s,a)$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>78</td>
<td>10</td>
<td>47</td>
<td>30</td>
</tr>
<tr>
<td>b</td>
<td>79</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>

(1pt) **3.** Thinking about unsupervised learning and the k-Means and expectation maximization (EM) algorithms, which one of the following statements is true:

(a) K-Means algorithm fits clusters only in hyperspheres and EM algorithm fits data only in hyperellipsoids

(b) K-Means using euclidean distance is a particular case of the EM-algorithm when we are fitting K-gaussian distributions with the same variance for each attribute.

(c) EM algorithm has the same computational cost no matter the number of parameters that have to be estimated for the probability distribution that we are fitting to the attributes.

(d) K-Means assigns a probability to the membership of each example to each cluster

(1pt) **4.** Thinking about unsupervised learning, which ones of the following statements are true (multiple choice):

(a) Differently from partitional graph based algorithms and density estimation algorithms the K-means and the EM algorithm need as a parameter the number of clusters to find

(b) With hierarchical clustering algorithms based on graph theory we obtain a partition of a dataset in K different classes

(c) The K-Means algorithm obtains a global optimal solution for the partition of a dataset by minimizing the square distance between examples and their nearest centroid

(d) The EM algorithm assumes that the model of the data comes from a mixture of K n-dimensional probability distributions

(1pt) **5.** Thinking about the naive bayes algorithm and bayesian networks, which ones of the following statements are true (multiple choice):

(a) Naive Bayes is a probabilistic method that assumes that the statistical distribution of the examples is independent

(b) Naive Bayes is a particular case of a Bayesian Network where all the attributes are assumed to be independent and that their individual distributions are not independent of the class attribute

(c) A Bayesian Network represents the joint probability distribution of a set of variables

(d) Bayesian Networks are unsupervised methods and can not be used for classification

(1pt) **6.** Thinking about algorithms for rule induction, which ones of the following statements are true (multiple choice):

(a) The advantage of rule induction over learning a decision tree is that usually the rules obtained are shorter that the ones that can be extracted from a decision tree

(b) The sequential covering algorithm for rule induction is based on learning sequentially the rules with the lower accuracy until all the dataset is covered

(c) ILP methods learn propositional rules and decision trees and rule induction algorithms learn first order rules

(d) FOIL is an ILP method that learns rules by sequentially specializing a candidate rule
7. Thinking about learning a bayesian network topology, which ones of the following statements are true (multiple choice):
   (a) To learn a bayesian network for a dataset means to search in the space of Directed Acyclic Graphs that can be built with the attributes
   (b) Bayesian networks only can be applied to datasets where all the attribute are discrete
   (c) The heuristic functions used in learning a BN topology usually include a term that represents the a priori information about the network, another that evaluates the fitting of the data to the probability distribution function defined by the current topology and a penalization for the network complexity
   (d) The K2 algorithm uses a hill climbing algorithm to learn the topology of a bayesian network assuming that any order of precedence among the variables in the network is initially possible

8. Given a K-nn classifier which ones of the following statements are true (multiple choice):
   (a) The more examples are used for classifying an example, the higher accuracy we obtain
   (b) The more attributes we use to describe the examples the more difficult is to obtain high accuracy
   (c) The most costly part of this method is to learn the model
   (d) We can use K-nn for classification and regression

9. Thinking about Reinforcement Learning which ones of the following statements are true (multiple choice):
   (a) The maximization of the future cumulative reward allows to Reinforcement Learning to perform global decisions with local information
   (b) Q-learning is a temporal difference RL method that does not need a model of the task to learn the action value function
   (c) Reinforcement Learning only can be applied to problems with a finite number of states
   (d) In Markov Decision Problems (MDP) the future actions from a state depend on the previous states

10. Thinking about reinforcement learning which one of the following statements is true:
    (a) Estimation using Dynamic Programming is less computational costly than using Temporal Difference Learning
    (b) Estimating using Montecarlo methods has the advantage that it is not needed to have absorbent states in the problem
    (c) Temporal Difference learning allows on-line learning and Montecarlo methods need complete training sequences for estimation
    (d) Dynamic Programming and Montecarlo methods only work if we know the transitions probabilities for the actions and the reward function

11. Thinking about dimensionality reduction and attribute selection, which ones of the following statements are true (multiple choice):
    (a) PCA and ICA transform a dataset to a space with the same dimensionality optimizing a measure that preserves the distances among all the pairs of examples
    (b) PCA and multidimensional scaling are unsupervised methods for dimensionality reduction
    (c) Wrappers are feature selection methods that, given a classifier as a performance criteria, search in the space of subset of features for the minimal one that obtains the higher accuracy
    (d) Filters are unsupervised feature selection methods because they do not use a classifier as selection criteria
12. Thinking about Explanation Based Learning (EBL), which ones of the following statements are true (multiple choice):
   (a) The goal regression algorithm is used for the generalization of resolution traces
   (b) In EBL only an example is needed to perform the learning
   (c) EBL does not allow to discover new knowledge, it only makes explicit knowledge that we already have
   (d) EBL learning is not used in problem solving

13. Thinking about classifier ensembles (meta-learning), which ones of the following statements are true (multiple choice):
   (a) Classifier ensembles usually have better performance for weak classifiers because they combine the different points of view of each individual classifier
   (b) Boosting is an ensemble method that uses sampling with replacement of a dataset and learns independently several classifiers
   (c) Random subspaces methods (like random forest) are meta-learning methods that learn a set of classifiers using datasets that have subsets of attributes of the original data
   (d) For methods like Boosting or Bagging to perform well, it is necessary that the different subsamples from the dataset used in each classifier are as similar as possible

14. Given a dataset we first learn a decision tree and we apply after that a postprunning to the tree. We look at the pruned tree and we identify in the leaves what examples are in the majority class. We consider the examples not in the majority class of the leaves as misclassifications. We create a new dataset only with the examples in the majority classes of the leaves. We train a new decision tree only with this examples. Which one of this statements is true:
   (a) The new tree will be the same than the pruned original tree
   (b) The new tree will be different and, given a new test set, the error of this new tree will be the same than the one of the original pruned tree with this new test set
   (c) The new tree will be different, but, given a new test set, the error of this new tree with this new test set will be lower than one of the original pruned tree
   (d) None of the above