Artificial Intelligence Problems **(2017–2018)**
Third year – Grado en Ingeniería Informática, Universidad de Sevilla

Inductive learning

Entropy table $Ent(X, Y)$

		Y									
		0	1	2	3	4	5	6	7	8	9
X	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1	0.000	1.000	0.918	0.811	0.722	0.650	0.592	0.544	0.503	0.469
	2	0.000	0.918	1.000	0.971	0.918	0.863	0.811	0.764	0.722	0.684
	3	0.000	0.811	0.971	1.000	0.985	0.954	0.918	0.881	0.845	0.811
	4	0.000	0.722	0.918	0.985	1.000	0.991	0.971	0.946	0.918	0.890
	5	0.000	0.650	0.863	0.954	0.991	1.000	0.994	0.980	0.961	0.940
	6	0.000	0.592	0.811	0.918	0.971	0.994	1.000	0.996	0.985	0.971
	7	0.000	0.544	0.764	0.881	0.946	0.980	0.996	1.000	0.997	0.989
	8	0.000	0.503	0.722	0.845	0.918	0.961	0.985	0.997	1.000	0.998
	9	0.000	0.469	0.684	0.811	0.890	0.940	0.971	0.989	0.998	1.000

Question 1. ¿How many possible hypotheses can we represent using a decision tree, if we have n binary attributes?

Question 2. The *size* of a tree is its number of nodes, including the leaves. Give two decision trees of different sizes representing the same classifier.

Question 3. Assume we want to apply ID3 to a trainig set D with N examples. Suppose we have an attribute $Attr_1$ with N possible values and that each of the examples in D has a different value in $Attr_1$. Calculate $Gain(D, Attr_1)$.

Question 4. Answer if the following statements are TRUE or FALSE, in a reasoned way. If the answer is FALSE, give an example where the statement does not hold.

(a) Let $D_1 = \{e_1, \ldots, e_n\}$ be a training set, a let D_2 be a training set obtained from D_1 considering that every example appears twice. Then we obtain the same decision tree if we apply ID3 to D_1 and to D_2.

(b) We have a set with $4 \cdot n$ examples, and we split it in a a training set with $3 \cdot n$ examples, and a test set with n examples. In the training set, we have $2 \cdot n$ positive examples and n negative examples. In the test set, all the examples are positive. If we obtain a decision tree applying ID3 to the training set, then that tree can always be pruned in such a way that the accuracy of the pruned tree on the test set is better than the accuracy of the original tree.

(c) If, instead of log_2 in the definition of entropy, we use log_b, with b any natural number, the the ID3 returns the same tree.

(d) The information gain criteria in the ID3 algorithm affects only to the size of the returned tree. If we apply the algorithm with a different selection criteria, we obtain a tree of a different size, but that it represents the same classifier.

Question 5. Suppose we modify the ID3 algorithm, and instead of using the *highest* information gain as the criteria to choose the best attribute, we use the *smallest* information gain.

- In that case, would we obtain a decision tree consistent with all the examples of the training set? (we assume absence of noise)
- Which inductive bias would this algorithm have? Explain your answer.

Question 6. Answer if the following statements are TRUE or FALSE, in a reasoned way. If the answer is FALSE, give an example where the statement does not hold.

(a) Let $C_1 = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $C_2 = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ be two *clusters*, obtained after applying the k-means algorithm and using the euclidean distance d. Let m_1 be the center of C_1 and m_2 the center C_2. Then FOR ALL $i \in \{1, \ldots, n\}$, $d(x_i, m_2) > d(y_i, m_2)$.

(B) SUPPOSE WE HAVE A SET WITH N POINTS $\{x_i\}_{i=1}^{i=N}$, AND THAT WE ARE USING THE EUCLIDENA DISTANCE. THEN IF $k = N - 1$ AND $\forall i \in \{1, \ldots, N\}\,(x_i \in \mathbb{R})$, THEN THE k-MEANS ALGORITHM RETURNS THE SAME ANSWER, REGARDLESS THE k POINTS TAKEN AS INITIAL CENTRES.

Exercise 1. A team of biologists who were exploring the Amazon river just discovered a new species of insects, which they decided to call *lepistos*. Unfortunately, they are all missing, and the only information we have about such insects consists on the following set of examples found in a notebook. They classify a series of observed insects, according to attributes such as their COLOR, whether they have WINGS, their SIZE, and their SPEED:

Example	COLOR	WINGS	SIZE	SPEED	LEPISTO
E_1	black	yes	small	high	yes
E_2	yellow	no	big	average	no
E_3	yellow	no	big	low	no
E_4	white	yes	medium	high	yes
E_5	black	no	medium	high	no
E_6	red	yes	small	high	yes
E_7	red	yes	small	low	no
E_8	black	no	medium	average	no
E_9	black	yes	small	average	no
E_{10}	yellow	yes	big	average	no

Answer the following questions:

- Which is the entropy value of this set of examples, according to their classification with respect to the attribute LEPISTO?
- Which attribute provides the greatest information gain?
- Apply (explaining in detail each step) the **ID3 algorithm** to build, from this training set, a decision tree that allows us to decide if any observed insect is a lepisto or not.
- Obtain a set of rules translating the branches of the tree learned in the previous item.
- According to the decision tree, is there any irrelevant attribute (with respect to deciding whether an insect is a lepisto)?

Exercise 2. The following table shows the classification concerning the existence of health risk for 10 touristic locations having different characteristics. Let us use this table as training set D for the concept *Health Risk*.

Ej.	Situation	Hospital	Country	Health Risk
1	Beach	Near	Mozambique	YES
2	Beach	Near	Tanzania	YES
3	City	Near	Tanzania	YES
4	Beach	Far	Mozambique	YES
5	Beach	Far	Mozambique	YES
6	City	Far	Mozambique	NO
7	Countryside	Near	Tanzania	NO
8	Countryside	Near	Mozambique	NO
9	City	Far	Tanzania	NO
10	Countryside	Near	Mozambique	NO

Build a decision tree by means of ID3 algorithm, and use it in order to classify the following instance:

(City, Near, Mozambique)

Exercise 3. A bank office decides if they should grant a loan to a client depending on a collection of attributes: his/her AGE (`young`, `middle` o `mature`), his/her INCOME (`high`, `average` or `low`), a REPORT over their finantial activities (either `positive` or `negative`) and, finally, if they already have ANOTHER LOAN already granted to them. The following table presents a set of examples sshowing whether the loan was granted or not:

Example	Age	Income	Report	Another loan	Granted
E_1	young	high	negative	no	no
E_2	young	high	negative	yes	no
E_3	middle	high	negative	no	yes
E_4	mature	average	negative	no	yes
E_5	mature	low	positive	no	yes
E_6	mature	low	positive	yes	no
E_7	middle	low	positive	yes	yes
E_8	young	average	negative	no	no
E_9	young	low	positive	yes	yes
E_{10}	mature	average	positive	no	yes
E_{11}	young	average	positive	yes	yes
E_{12}	middle	average	negative	yes	yes
E_{13}	middle	high	positive	no	yes
E_{14}	mature	average	negative	yes	no

Let us suppose that we modify the ID3 pseudocode and create another algorithm in such a way that the criterion to select the "best" attribute classifying a set of examples is the one which provides the *least* information gain. In this situation, provide a motivated answer to the following questions:

- If the data are free of noise, would this modified algorithm obtain a decision tree consistent with the examples of the training set?
- Which is the inductive biass for the modified algorithm?
- Apply (explaining in detail each step) the modified algorithm to build, for this training set, a tree that help us to decide whether the loans should be granted or not.

Exercise 4. Apply **ID3 algorithm** to build a decision tree consistent with the following examples, so that we can use it to decide whether to buy or not a CD.

Example	Singer	Record label	Genre	Price	Shop	Buy
E_1	Queen	Emi	rock	30	Mixup	yes
E_2	Mozart	Emi	classical	40	Virgin	no
E_3	Anastacia	Universal	soul	20	Virgin	yes
E_4	Queen	Sony	rock	20	Virgin	yes
E_5	Anastacia	Universal	soul	30	Mixup	yes
E_6	Queen	Sony	rock	30	Virgin	yes
E_7	Wagner	Sony	classical	30	Mixup	no
E_8	Anastacia	Universal	soul	30	Virgin	no
E_9	Queen	Emi	rock	40	Virgin	no
E_{10}	Mozart	Sony	classical	40	Mixup	yes

Let us consider the following examples as a test set. Obtain the performance of the learned tree.

Example	Singer	Record label	Genre	Price	Shop	Buy
E_{11}	Queen	Emi	rock	30	Virgin	yes
E_{12}	Anastacia	Universal	soul	20	Virgin	no
E_{13}	Queen	Sony	rock	20	Virgin	no
E_{14}	Anastacia	Universal	soul	30	Virgin	no
E_{15}	Queen	Sony	rock	40	Virgin	no
E_{16}	Mozart	Sony	classical	40	Mixup	yes

Exercise 5. The following table shows examples of plants, indicating whether or not they survived more than one year after been bought. The attributes that are taken into account are their size (`big`, `medium` or `small`), their type (suitable for `indoor` or `outdoor` environment), if they have flowers, and the season when they were bought.

Example	Size	Flowers	Type	Season	Survives
E_1	big	yes	indoor	summer	no
E_2	big	yes	indoor	summer	no
E_3	big	yes	outdoor	spring	no
E_4	big	yes	outdoor	winter	no
E_5	big	no	indoor	autumn	no
E_6	big	no	outdoor	spring	no
E_7	medium	yes	indoor	summer	yes
E_8	medium	yes	indoor	summer	yes
E_9	medium	no	indoor	spring	yes
E_{10}	medium	no	outdoor	autumn	no
E_{11}	medium	no	outdoor	summer	no
E_{12}	small	yes	indoor	winter	no
E_{13}	small	yes	outdoor	summer	yes
E_{14}	small	no	indoor	spring	no
E_{15}	small	no	indoor	summer	yes
E_{16}	small	no	outdoor	autumn	no

1. Apply (explaining in detail each step) the **ID3 algorithm** to build a decision tree consistent with the training set $\{E_1, \ldots, E_{16}\}$, so that it can decide whether a plant will survive more than one year after being bought. You should assume that the attribute size is selected for the root node, and continue the execution from there on.

2. Let us consider the following table of examples as test set

Example	SIZE	FLOWERS	TYPE	SEASON	SURVIVES
E_{17}	big	no	outdoor	summer	no
E_{18}	medium	no	indoor	autumn	yes
E_{19}	medium	no	outdoor	spring	no
E_{20}	medium	yes	outdoor	summer	no
E_{21}	small	yes	indoor	summer	no
E_{22}	small	yes	indoor	winter	no
E_{23}	small	no	indoor	summer	no
E_{24}	small	no	outdoor	autumn	no

(a) Calculate the performance of the decision tree obtained in the previous item

(b) Apply the pruning algorithm over this tree.

Exercise 6.

A sports equipment company wants to do a market research to find the main characteristics of its potential customers. In a first phase, the characteristics to be studied are as follows: the AGE (`young` or `adult`), being a PROFESSIONAL sportsman, the level of INCOME (`high`, `average` or `low`) and the SEX. To this end, a questionnaire is carried out for 21 people, obtaining the results reflected in the following table:

Example	AGE	PROFESSIONAL	INCOME	SEX	INTERESTED
E_1	young	yes	low	male	yes
E_2	young	yes	high	male	yes
E_3	young	no	high	female	no
E_4	young	yes	low	female	yes
E_5	young	no	average	female	no
E_6	adult	yes	high	male	no
E_7	adult	no	high	female	no
E_8	adult	yes	high	female	no
E_9	adult	no	average	female	no
E_{10}	adult	yes	low	female	no
E_{11}	adult	no	average	female	no
E_{12}	adult	yes	average	male	no
E_{13}	adult	no	high	female	yes
E_{14}	young	yes	high	female	yes
E_{15}	young	yes	average	male	yes
E_{16}	adult	no	average	male	no
E_{17}	adult	no	low	male	no
E_{18}	young	no	average	male	no
E_{19}	young	no	low	female	no
E_{20}	adult	yes	average	female	no
E_{21}	young	yes	average	female	yes

- Apply the **ID3 algorithm** (showing all the steps) to obtain a decison tree classifier. Take as training set, the first fifteen examples.
- Now take as test set the remaining six examples and calculate the accuracy of the previous tree on it. Apply the *reduced error pruning algorithm* on it and give the accuracy of the obtained tree both to the test set and to the training set

- Give an equivalent set of rules

Exercise 7. The table below contains examples of situations when it is recommended to buy a computer, according to its PRICE (`high` and `medium` or `low`), its PROCESSOR (`AMD` or `Intel`), if it has BLUETOOTH and if it has a GPU (we assume that the rest of technical specificatiosn are equal).

Example	PRICE	BLUETOOTH	PROCESSOR	GPU	BUY
E_1	high	yes	AMD	yes	no
E_2	high	yes	AMD	no	no
E_3	high	yes	Intel	yes	no
E_4	high	yes	Intel	no	no
E_5	high	no	AMD	no	no
E_6	high	no	Intel	no	no
E_7	medium	yes	AMD	yes	yes
E_8	medium	yes	AMD	no	yes
E_9	medium	no	AMD	yes	yes
E_{10}	medium	no	Intel	yes	no
E_{11}	medium	no	Intel	no	no
E_{12}	low	yes	AMD	yes	no
E_{13}	low	yes	Intel	no	yes
E_{14}	low	no	AMD	yes	no
E_{15}	low	no	AMD	no	yes
E_{16}	low	no	Intel	yes	no

Apply the sequential covering algorithm to this training set, in order to obtain a set of rules recommending when to buy or not to buy a computer. According to the rules learned: Should we buy a computer with a GPU if the price is low? Is there any irrelevant attribute?

Exercise 8. Consider the following training set, describing when to buy a flight ticket, depending on its PRICE, the travel CLASS and if the company has INTERNATIONAL flights.

Example	PRICE	CLASS	INTERNATIONAL	BUY
E_1	high	business	no	no
E_2	high	business	yes	no
E_3	high	tourist	no	no
E_4	high	tourist	yes	no
E_5	medium	business	no	yes
E_6	medium	business	yes	no
E_7	medium	tourist	no	yes
E_8	medium	tourist	yes	no
E_9	low	business	no	yes
E_{10}	low	business	yes	no
E_{11}	low	tourist	no	yes
E_{12}	low	tourist	yes	no

- Apply the **sequential covering algorithm** to learn a set of rules to aid in the decision of baying or not a flight ticket.
- Analyzing the result obtained, can we say that there is some irrelevant attribute?

Exercise 9. Consider the following training set:

$Ex.$	$Attr_1$	$Attr_2$	$Attr_3$	$Class$
E_1	1	1	0	SI
E_2	1	0	0	SI
E_3	1	0	1	SI
E_4	0	0	1	NO

(a) Apply k-NN with $k = 1$ to classify $P = (0.75, 0, 0)$, using the Manhattan distance.

(b) Now consider the above training set, but ignoring the class column and let $m_1 = (1, 1, -1)$ and $m_2 = (0, -1, 1)$. Apply k-means (k=2) with m_1 and m_2 as initial centres *until the first updating of the centres.*