Exercises of Machine Learning and Inductive Inference

See Toledo: "Machine Learning and Inductive Inference: **Lecture**" [H02C1a], folder "Exercise sessions".

Questions?

Vladimir Dzyuba
E-Mail: Vladimir.Dzyuba@cs.kuleuven.be
Office: 200A 04.034

Davide Nitti
E-Mail: davide.nitti@kuleuven.be
Office: 200A 02.65

Jérôme Renaux
E-Mail: Jerome.Renaux@cs.kuleuven.be
Office: 200A 04.002

1 Exercise Session 1: Concept learning and decision trees

1.1 Generality

Order the following concepts over the boolean attributes A and B according to generality ($\geq_g$): $A \wedge B$, $A \vee B$, A xor B, T (true), F (false).

Solution

See Figure 1.

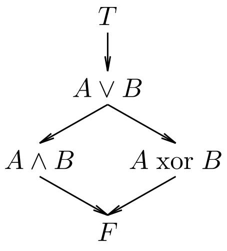


Figure 1: Solution for exercise 1.1.

1.2 Version Spaces

Consider the hypothesis space shown in Figure 2 (ordered in a lattice using $\geq_g$). A Version Space VS is defined by its most general border $G = \{h_1, h_2\}$ and its most specific border $S = \{h_3, h_4\}$.

a. Mark all hypotheses that belong to VS (on Figure 2).

b. We say that a version space classifies an instance as positive if all the hypotheses in the version space predict it to be positive, negative if all the hypotheses in the version space predict negative, and "don't know" in all other cases. Let P_i be the instances predicted positive by h_i, and N_i the instances predicted negative by h_i (for $i = 1, 2, 3, 4$). Answer the following questions in terms of P_i and N_i.

   - Which instances are classified positive by VS?

- Which instances are classified negative by VS?
- Which instances are classified "don't know" by VS?

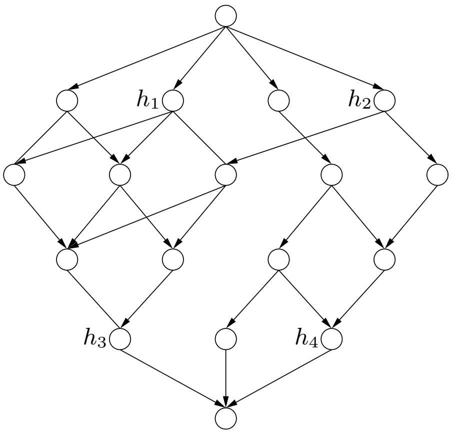


Figure 2: Hypothesis space for exercise 1.2.

Solution

a. See Figure 3.

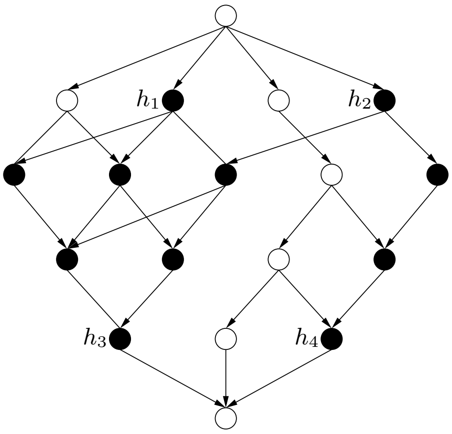


Figure 3: Solution for exercise 1.2.

b. By definition, the generality relationship between hypotheses corresponds to the subset-superset relationship between the instances classified as positive or negative by these hypotheses: $h_i >_g h_j \Leftrightarrow P_i \supset P_j,\ N_i \subset N_j$.

- $P_3 \cap P_4$: An instance is positive if it is classified as positive by both h_3 and h_4.

  Indeed, for these instances, since h_3 and h_4 classify it as positive, and since each hypothesis in VS is more general than either h_3 or h_4, each hypothesis in VS classifies the instances as positive, which means "VS classifies them as positive".

  See Figure 4 for one possible Venn diagram of P_{1-4}. Instances in the region marked by the black circle are classified as positive.
- $N_1 \cap N_2$: An instance is negative if it is classified as negative by both h_1 and h_2.

  Similar reasoning as above. Note that $N_1 \cap N_2 = (X \setminus P_1) \cap (X \setminus P_2) = X \setminus (P_1 \cup P_2)$, where X denotes the entire instance space, i.e. the universal set of all possible instances.

  In Figure 4, instances in the region marked by the black square are classified as negative.
- $(P_1 \cup P_2) \setminus (P_3 \cap P_4)$: All remaining instances are classified "don't know".

  The remaining instances are classified as negative by h_3 or h_4, and since this h_i might in principle be the correct one (since it is in VS), each such instance might be negative. Similarly, these instances are classified as positive by h_1 or h_2, so each such instance might also be positive.

  For example, in Figure 4, instances in the region marked by the black triangle are classified as positive by h_1, h, h_2, and h_4, but as negative by h_3.

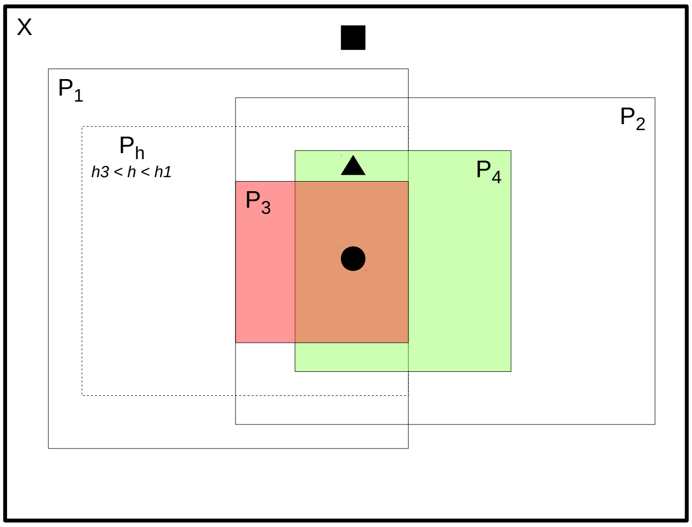


Figure 4: One possible Venn diagram for exercise 1.2.

1.3 Decision Trees Representing Logical Concepts

Represent as a decision tree:

a. $A \wedge \neg B$

b. $\neg A \vee B$

c. $A \vee (B \wedge C)$

d. $(\neg A \wedge B) \vee (A \wedge \neg B)$

e. $(A \vee B) \wedge (C \vee D \vee \neg E)$

f. $(A \vee B \vee C) \wedge (D \vee E \vee F)$

Solution

See Figure 5.

1.4 Decision Trees and Generality

Decision tree D2 is an *elaboration* of decision tree D1 if D2 can be constructed out of D1 by replacing a leaf of D1 with a subtree in D2.

1. Which of the trees a and d from Exercise 1.3 is the elaboration of the other?
2. True or false? If a boolean decision tree $D2$ is an elaboration of $D1$ then $D1$ is more general than $D2$.

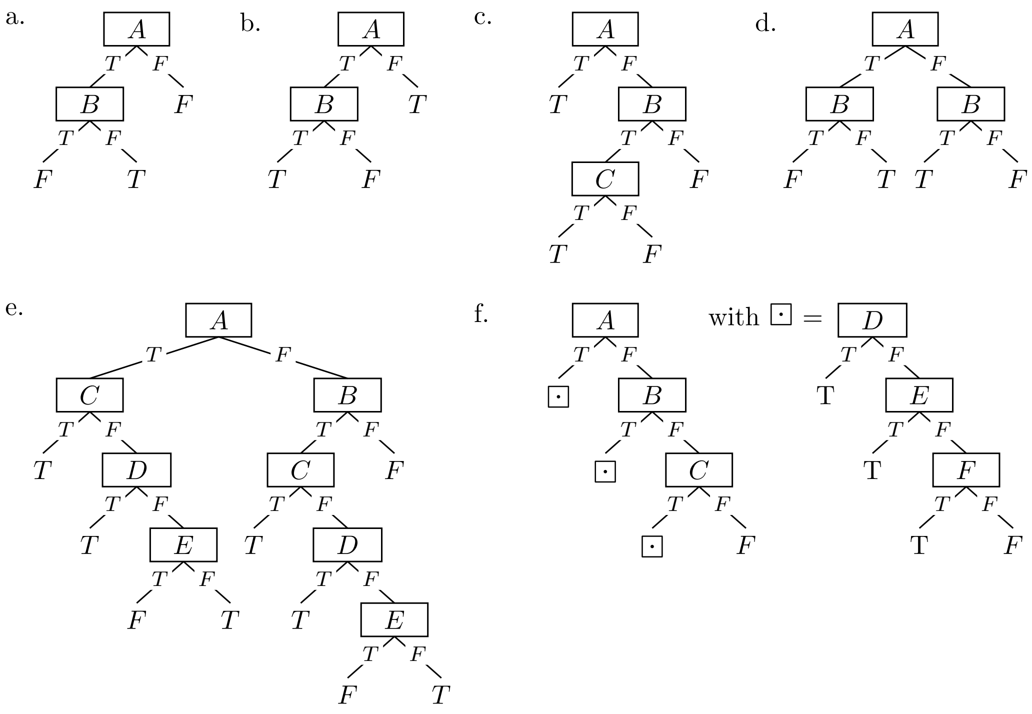


Figure 5: Solution for exercise 1.3.

Solution

1. d is an elaboration of a: the right leaf of a is replaced by a subtree in d.
2. *False.* Refining a "false" leaf into an internal node with one "false" and one "true" leaf makes the tree more general. For example, more instances belong to the concept represented by the tree d than to the one represented by the tree a (more paths return T).

1.5 Decision Surface

Consider a data set with two *numeric* attributes a_1 and a_2 and one nominal target attribute c with two possible values: $\oplus$ and $\ominus$. The training examples are shown in Figure 6.

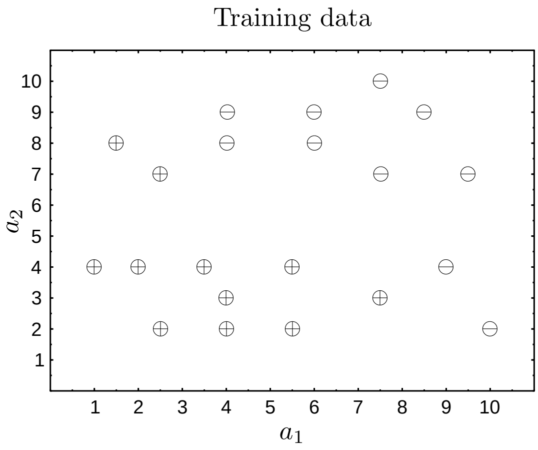


Figure 6: Training data for exercise 1.5

a. Find a decision tree that classifies all training examples correctly.

b. Draw the decision surface of this tree on Figure 6.

Solution

a. The decision tree is shown in Figure 7.a.

b. The decision surface is shown in Figure 7.b.

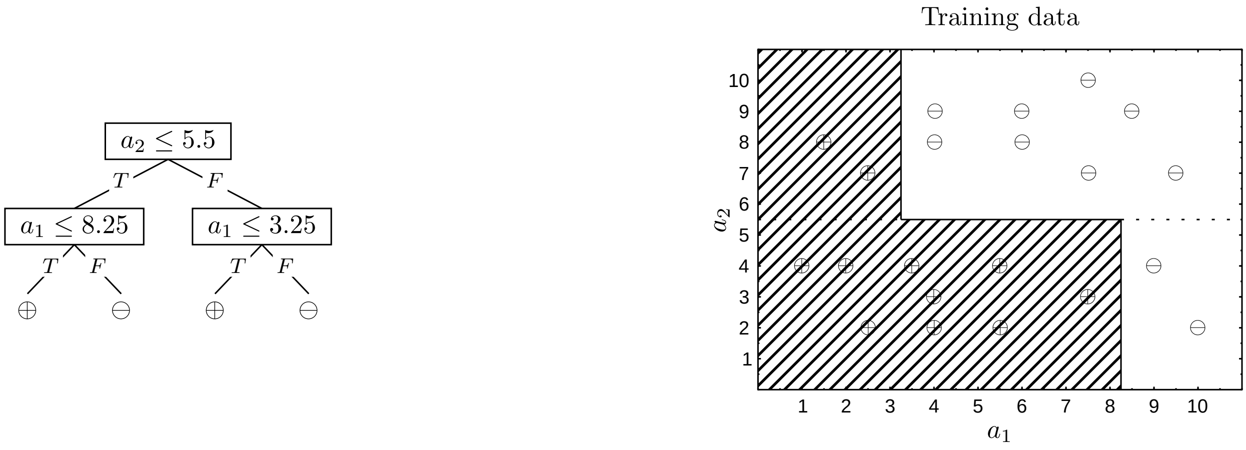


Figure 7: Decision tree and surface for exercise 1.5

1.6 Entropy and Information Gain

Consider the following table of training examples:

Instance	Classification	a_1	a_2
1	⊕	T	T
2	⊕	T	T
3	⊖	T	F
4	⊕	F	F
5	⊖	F	T
6	⊖	F	T

a. What is the entropy with respect to the “classification” attribute? (Answer without a calculator.)

b. What is the information gain of a_2 relative to these training examples? (Answer without a calculator.)

c. What is the information gain of a_1 relative to these training examples? (Use a calculator.)

Solution

a. Equal probabilities → entropy = 1.

b. Entropy stays 1 → gain = 0.

c. $E(D_{a_1=T}) = -\frac{2}{3} \cdot log_2(\frac{2}{3}) - \frac{1}{3} \cdot log_2(\frac{1}{3}) = 0.5283208 + 0.389975 = 0.9182958 (= E(D_{a_1=F})$
$G(D, a_1) = 1 - \frac{1}{2}0.9182958 - \frac{1}{2}0.9182958 = 0.0817042$

1.7 The ID3 Algorithm

a. Show a decision tree that could be learned by ID3 assuming it gets the following examples:

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	sunny	warm	normal	strong	warm	same	yes
2	sunny	warm	high	strong	warm	same	yes
3	rainy	cold	high	strong	warm	change	no
4	sunny	warm	high	strong	cool	change	yes

b. Add this example:

5	sunny	warm	normal	weak	warm	same	no

then show how ID3 would induce a decision tree for these 5 examples.

Compute the class entropy for the entire dataset S:

	$+$	$-$	Entropy
S	3	2	

Compute the split heuristic for each attribute and select the attribute for the root node:

Attribute	Values	$+$	$-$	Entropy	IG
Sky	Sunny	3	1	0.811	0.32
	Rainy	0	1	0.000	
AirTemp	Warm				
	Cold				
Humidity	Normal	1	1	1.000	0.02
	High	2	1	0.918	
Wind	Strong				
	Weak				
Water	Warm				
	Cool				
Forecast	Same				
	Change				

Follow the ID3 algorithm until you obtain a complete decision tree.

Solution

a. Sky : sunny → yes ; rainy → no
Alternative: AirTemp : warm → yes ; cold → no

b. S = [3+,2-], entropy(S) = 0.97

Attribute	Values	$+$	$-$	Entropy	IG
Sky	Sunny	3	1	0.811	$0.97 - \frac{4}{5} \cdot 0.811 - \frac{1}{5} \cdot 0 = 0.32$
	Rainy	0	1	0.000	
AirTemp	Warm	3	1	0.811	0.32
	Cold	0	1	0.000	
Humidity	Normal	1	1	1.000	0.02
	High	2	1	0.918	
Wind	Strong	3	1	0.811	0.32
	Weak	0	1	0.000	
Water	Warm	2	2	1.000	0.17
	Cool	1	0	0.000	
Forecast	Same	2	1	0.918	0.02
	Change	1	1	1.000	

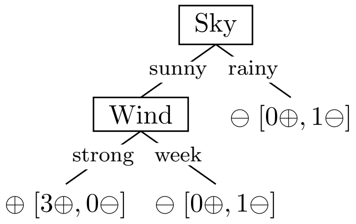
Three attributes have equal information gain: *Sky*, *AirTemp*, and *Wind*. Any of them can be selected for the root node of the decision tree.

Assume *Sky* is selected. The right node corresponding to *Sky* = *rainy* is pure, i.e. it contains examples of one class only ($-$). It becomes a leaf node.

Computing the information gain for the left node ($Sky = sunny$) yields the following values:

Attribute	IG
$AirTemp$	0.00
Hum	0.31
$Wind$	0.81
$Water$	0.12
$Forecast$	0.12

$Wind$ has the highest IG. Splitting on $Wind$ results in two pure leaves, and ID3 terminates.



1.8 Regression Tree

Consider the following set S of training examples with a numeric target attribute.

Instance	Target	a_1	a_2
1	1.0	T	T
2	1.0	T	T
3	1.0	T	F
4	5.0	F	F
5	6.0	F	T
6	5.5	F	T

We are going to use *weighted average variance of the subsets* as the split heuristic H. For each attribute A, compute:

$$H(A) = \sum_{v \in \text{Values}(A)} \frac{|S_{A=v}|}{|S|} \cdot \text{Var}[S_{A=v}] \qquad \text{(heuristic value for attribute } A\text{), where}$$

$$\text{Var}[S_{A=v}] = \frac{1}{|S_{A=v}|} \cdot \sum_{i \in S_{A=v}} \left(Target_i - \overline{Target}[S_{A=v}]\right)^2 \qquad \text{(variance of } Target \text{ values of instances with } A = v\text{)}$$

$$\overline{Target}[S_{A=v}] = \frac{1}{|S_{A=v}|} \cdot \sum_{i \in S_{A=v}} Target_i \qquad \text{(average } Target \text{ value of instances with } A = v\text{)}$$

Which attribute (a_1 or a_2) will be put in the top node of the regression tree?

Solution

a. $H(a_1) = 0.084$, $(\text{Var}[S_{a_1=T}] = 0, \text{Var}[S_{a_1=F}] = 0.167)$:

$$\begin{aligned}
S_{a_1=T} &= \{Instances\ 1,2,3\}: \\
\overline{Target}[S_{a_1=T}] &= \frac{1}{3}\cdot(1.0+1.0+1.0) = 1.0 \\
Var[S_{a_1=T}] &= \frac{1}{3}\cdot((1.0-1.0)^2+(1.0-1.0)^2+(1.0-1.0)^2) = 0.0
\end{aligned}$$

$$\begin{aligned}
S_{a_1=F} &= \{Instances\ 4,5,6\}: \\
\overline{Target}[S_{a_1=F}] &= \frac{1}{3}\cdot(5.0+6.0+5.5) = 5.5 \\
Var[S_{a_1=F}] &= \frac{1}{3}\cdot((5.0-5.5)^2+(6.0-5.5)^2+(5.5-5.5)^2) = \frac{1}{3}\cdot(0.5^2+0.5^2) = 0.167
\end{aligned}$$

$$H(a_1) = \frac{3}{6}\cdot 0.0 + \frac{3}{6}\cdot 0.167 = 0.084$$

b. $H(a_2) = 5.11$ $(\text{Var}[S_{a_2=T}] = 5.67,\ \text{Var}[S_{a_2=F}] = 4)$.

c. $H(a_1) < H(a_2) \Rightarrow$ Attribute a_1 will be put in the top node.

1.9 Using Weka: Homework

The explanation of the Weka Tool in this exercise is kept to a minimum. For more information you can look online and download a manual or tutorial from `http://www.cs.waikato.ac.nz/ml/weka/` or you can consult the book: "Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten and Eibe Frank (ISBN 978-0123748560, available in the KU Leuven library).

1. Download the Excel document `weka-zoo.xls` from Toledo (`Course Documents/Exercise Sessions/Session 1/Exercise session 1/`). Study the animals in the document and without using a data mining tool, draw a decision tree of three to five levels deep that classifies animals into a mammal, bird, reptile, fish, amphibian, insect or invertebrate.

2. Download and install Weka from: `http://www.cs.waikato.ac.nz/ml/weka/`

3. Read about the ARFF-format at `http://www.cs.waikato.ac.nz/ml/weka/arff.html` and construct the header for the animal file.

4. Download and unzip the file `weka-datasets.zip` from Toledo (`Course Documents/Exercise Sessions/Session 1/Exercise session 1/`). Open the `zoo.arff` file in the Weka Explorer. Find out how many animals this dataset contains.

5. Go to the classifier tab and select the decision tree classifier j48. Click on the line behind the choose button. This shows you the parameters you can set and a button called 'More'. Which algorithm is implemented by j48?

6. Click the start button to run the j48 algorithm. Which percentage of instances is correctly classified by j48? Which families are mistaken for each other? (Hint: Take a look at the confusion matrix.)