Strasbourg University / French-Azerbaijani University

Artificial Intelligence

Computer Science track - Year 3

**Lab: Iris dataset classification using a Decision Tree**

The objective of this lab is to implement a Decision Tree in order to classify Irises (yes, the flowers[a]). A code template is available to you on Moodle.

**Specific objectives:**

- Observe the data, understand their nature and how to adapt them (if needed) so you can use them in a Decision Tree model.
- Understand how Decision Trees work so as to implement this model in a computer program.
- Evaluate the results and put them into perspective with what we know about the data.

[a] `https://en.wikipedia.org/wiki/Iris_(plant)`

# 1 The data

During this lab, we will work on a classification problem (*i.e.* assign labels to data so as to group these data into distinct caterories). The "legacy" dataset used in machine learning when we begin is the *Iris dataset*. This multi-variate dataset characterizes 3 different species of Iris. You will see it contains 50 instanes of each species. There are 4 attributes (or features) used to describe the datapoints and, hopefully, to discriminate them into the 3 species: petal width, petal length, sepal width and sepal length.
Fig. 1 illustrates a way (among others) to visualize these data. The fact they are low dimension and that only 150 instances total compose the dataset makes the visualization quite easy (this is not the case for the majority of "real world" data). Other methods, such as diagrams, histograms, etc. can be used to visualize data.

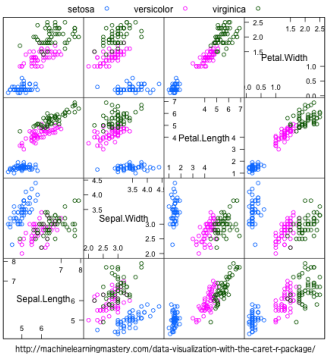


Figure 1: Visualizing data in a 2D scatterplot

We can see that data can effectively be grouped into different classes (the species of the datapoint). Interestingly, we can see that 2 main groups emerge: Setosa *v.s.* Versicolor Irises *and* Virginica Irises. The class "Setosa" is linearly separable from the other two classes. Discriminating between Versicolor and Virginica species is not this trivial (at least graphically): if we were to use an unsupervised clustering method, groups could end up being different that what you would expect.

★ *Is the Decision Tree model used for supervised or unsupervised classification? Explain your answer.*

# 2 A code template to help you

A code template (written in Java) is avaiable on Moodle. You do not *have* to use it, but if you do, then you start with a working data model (and you can begin the interesting part –implement the core of the Decision Tree model– right away).

★ *Get this code, have a look at it and localize the parts you need to implement the procedures to build a Decision Tree.*

# 3 Buidling a Decision Tree

Each instance of the dataset has 4 attributes. Building a Decision Tree is basically determining which of these attributes has *the highest discriminative power*. Once you have determined this attribute, you must determine which attribute has the second highest discriminative power, and so on and so forth. Once this process is set you will be able, upon the examination of a new instance, to predict the species (the class) of this "unknown" instance.

**★ Preliminary questions**

1. What is the nature of the attributes of the dataset?
2. Do you think it is necessary to transform the attributes (scaling, standardization, ...)?
3. How are you going to use real value attributes to build your Decision Tree?

What we suggest is defining three partitions (this is an arbitrary number that can be discussed...), each partition having the same number of instances (because this is easier, but it does not have to be the case). Since we deal with distances, we can attach the semantic "short", "average" and "long" to our partitions.

**★ Then, the procedure goes as follows:**

1. Pick an attribut $i$ to examine (we will compute its discriminative power).
2. Initially, this dataset is sorted in a lexicographic order (based on the label): the first 50 instances are Setosa flowers, then Versicolor flowers and then Virginica flowers: you must change this order and sort the instances based on the highest discriminative attribute (see Fig. 3 in appendix).
3. Now, partition the data into $G$ groups (here, $G = 3$, using the semantic representation set above). Each group $g$ will contain $150/G$ instances.
4. For each group $g$, enumerate the number of occurences of each species $\#s_g$
5. Compute a) the entropy of the whole dataset and b) the entropy of each group $g$.
6. Use these entropy values to compute the discriminative power of attribute $i$.

**What's next?** You have to iterate over this procedure for each attribute. Once you have examined all 4 attributes, the highest discriminative attribute will be used to perform a first segmentation (used to answer the question "To which class does a datapoint belong?").

For instance, for $G = 3$, after a first iteration over the procedure, we determine that Disc(Iris/PL) $\geq$ Disc(Iris/PW) $>$ Disc(Iris/SL) $>$ Disc(Iris/SW). We can examine the instances belonging to each group for the attribute `PL` and note the following repartitions per class (Table 1):

We can observe that for the first group ("short"), if $\text{Min}(g_1/\text{PL}) \leq \text{PL}_{\text{instance}} \leq \text{Max}(g_1/\text{PL})$, then there is no doubt left about the class of the datapoint we are looking at: it is a `Setosa`. For group 2 (resp. 3), there is a 95% chance the datapoint is a `Versicolor` (resp. a `Virginica`).

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Setosa | 50 | 0 | 0 |
| Versicolor | 0 | 47 | 3 |
| Virginica | 0 | 3 | 47 |

Table 1: Class repartitions for each group, based on the attribute `PL`

### ★ Discussions

1. Have a look back at Fig. 2 p. 5: what relationship is there between this figure and the class repartitions of Table 1?
2. So basically, do you wish to continue adding branches in the tree? Explain your answer.

**Let's continue the segmentation!** Let us say we set `SW` as the first discriminative variable[1]. Table 2 presents the class repartitions for this attribute.

### ★ Questions

1. Does the use of this attribute allow discrimination between classes?
2. Explain (and implement if you still have time) the procedure to continue segmenting the data.

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Setosa | 2 | 16 | 32 |
| Versicolor | 29 | 16 | 5 |
| Virginica | 19 | 18 | 13 |

Table 2: Class repartitions for each group, based on the attribute `SW`

[1] which is NOT a good choice since it is the *least* discriminative attribute

# 4 Annexes

## 4.1 Entropy and discriminative power

**Dataset entropy**

$$H(\text{Iris}) = -\sum_{k=0}^{K} \left( \frac{\#k}{\#\text{Iris}} \right) log_2 \left( \frac{\#k}{\#\text{Iris}} \right) \tag{1}$$

where $\#k$ is the number of instances of a given class $K$ in the dataset (in *this* dataset, we already know that each class is represented by 50 instances) and #Iris is the total number of instance in the dataset (here, 150).

**Entropy of a group $g$**

$$H(g) = -\sum_{s=0}^{S} \left( \frac{\#s_g}{\#g} \right) log_2 \left( \frac{\#s_g}{\#g} \right) \tag{2}$$

**Discriminative power of attribute $i$**

$$\text{Disc}(\text{Iris}/i) = H(\text{Iris}) - \left( \sum_{g=0}^{G} \left( \frac{\#g}{\#\text{Iris}} \right) H(g) \right) \tag{3}$$

## 4.2 Visualizing data in 2D

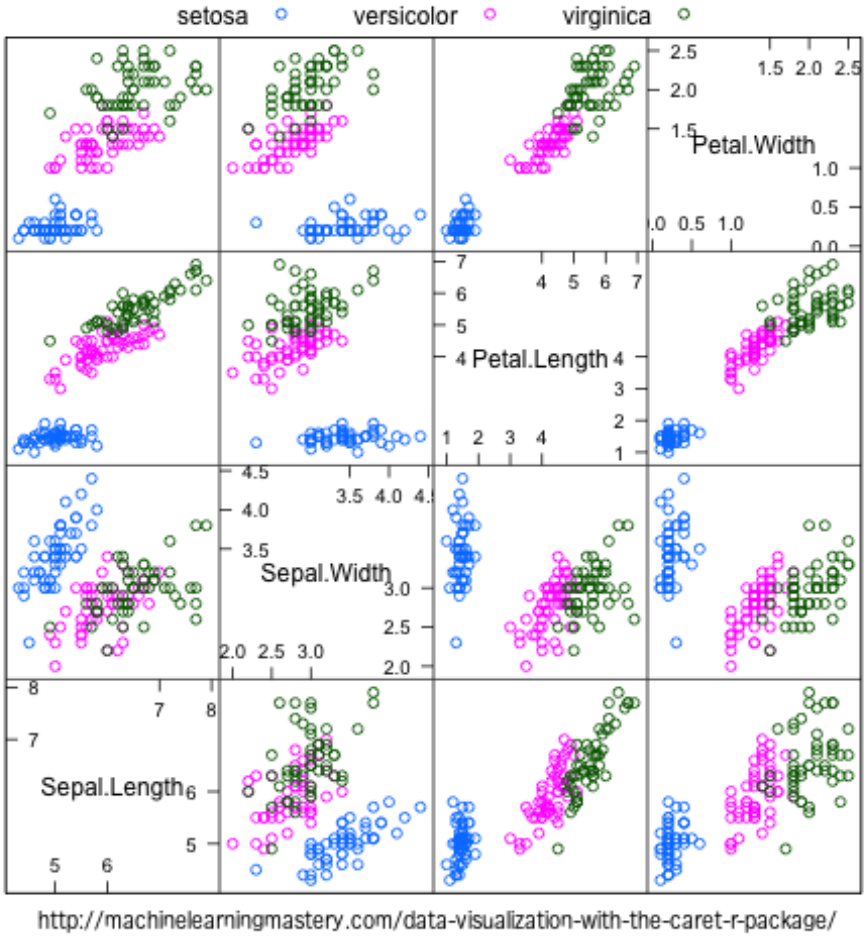


Figure 2: Data visualization in 2D

### 4.3 Sorting example for attribute $i = 0$

| | | | | | |
|---|---|---|---|---|---|
| 15 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 16 | 4.8 | 3.4 | 1.9 | 0.2 | Iris-setosa |
| 17 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 18 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 19 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 20 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 21 | 5 | 3 | 1.6 | 0.2 | Iris-setosa |
| 22 | 5 | 3.2 | 1.2 | 0.2 | Iris-setosa |
| 23 | 5 | 3.3 | 1.4 | 0.2 | Iris-setosa |
| 24 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 25 | 5 | 3.4 | 1.6 | 0.4 | Iris-setosa |
| 26 | 5 | 3.5 | 1.3 | 0.3 | Iris-setosa |
| 27 | 5 | 3.5 | 1.6 | 0.6 | Iris-setosa |
| 28 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 29 | 5.1 | 3.3 | 1.7 | 0.5 | Iris-setosa |
| 30 | 5.1 | 3.4 | 1.5 | 0.2 | Iris-setosa |

(a) Data sorted with their labels

| | | | | | |
|---|---|---|---|---|---|
| 15 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 16 | 4.8 | 3.4 | 1.9 | 0.2 | Iris-setosa |
| 17 | 4.9 | 2.4 | 3.3 | 1 | Iris-versicolor |
| 18 | 4.9 | 2.5 | 4.5 | 1.7 | Iris-virginica |
| 19 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 20 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 21 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 22 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 23 | 5 | 2 | 3.5 | 1 | Iris-versicolor |
| 24 | 5 | 2.3 | 3.3 | 1 | Iris-versicolor |
| 25 | 5 | 3 | 1.6 | 0.2 | Iris-setosa |
| 26 | 5 | 3.2 | 1.2 | 0.2 | Iris-setosa |
| 27 | 5 | 3.3 | 1.4 | 0.2 | Iris-setosa |
| 28 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 29 | 5 | 3.4 | 1.6 | 0.4 | Iris-setosa |
| 30 | 5 | 3.5 | 1.3 | 0.3 | Iris-setosa |

(b) Data sorted with the attribute $i = 0$

Figure 3: Sorting the data