# Association Rule Mining on Iris Dataset

IMT Atlantique – FIL A3

Project – UE Apprentissage Automatique

## 1 Iris dataset discretization

Consider the Iris dataset which has $n = 150$ examples, over one categorical attribute (`class`), and four numeric attributes (`sepal length`, `sepal width` , `petal length` , and `petal width`). We want in this project to apply the different algorithms seen in the previous lab to extract representative association rules that better characterizes the class attribute. To generate association rules we first need to discretize the numeric attributes as shown in Table 1.

| Attribte | Range or value | Label |
|---|---|---|
| Sepal length | 4.30–5.55 | $sl_1$ |
| | 5.55–6.15 | $sl_2$ |
| | 6.15–7.90 | $sl_3$ |
| Sepal width | 2.00–2.95 | $sw_1$ |
| | 2.95–3.35 | $sw_2$ |
| | 3.35–4.40 | $sw_3$ |
| petal length | 1.00–2.45 | $pl_1$ |
| | 2.45–4.75 | $pl_2$ |
| | 4.75–6.90 | $pl_3$ |
| Petal width | 0.10–0.80 | $pw_1$ |
| | 0.80–1.75 | $pw_2$ |
| | 1.75–2.50 | $pw_3$ |
| Class | Iris-setosa | $c_1$ |
| | Iris-versicolor | $c_2$ |
| | Iris-virginica | $c_3$ |

Table 1 – Iris dataset discretization and labels employed.

We want to determine representative class-specific rules that characterize each of the three Iris classes : `iris versicolor`, `iris setosa` and `iris virginica`, that is, we generate rules of the form $X \rightarrow y$, where $X$ is an itemset over the discretized numeric attributes, and $y$ is a single item representing one of the Iris classes.

## 2 Work to do

1. Provide a binary representation of the Iris data in such a way that all attribute values of each row will be described by the labels corresponding to the discretization schema of table 1. Thereby, the set of all items of the resulting transaction dataset will be constituted by all the values of the label column.
2. Generate all class-specific association rules using $minsup = 10$ (absolute frequency value) and a minimum lift value of 0.1.
3. To look for the most surprising rules, plot the graphics $lift$ vs. $conviction$ and $rsup$ vs. $conf$ for the set of extracted rules. Represent all the rules having the same conclusion (i.e. the same class $y$) with the same color (or same specific symbol).
4. For each class, select the most specific (i.e., with maximal antecedent) rule with the highest relative support and then confidence, and also those with the highest conviction and then lift. Highlight the selected rules in your graphics with specific colour or symbol.
5. According to these results what is the best rule for each class.

**Note**

You must provide a python code that runs on Jupyter netbook + a report (in pdf format) presenting the work carried out, the results obtained and an analysis of these results.