# Practical Exercise on Finding Association Rules

You are provided with a dataset which contains some data resulting from Mammography tests.

**Description of dataset.** Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last years.These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short term follow-up examination instead. The dataset we provided can be used to predict the severity (benign or malignant) of a mammographic mass lesion from the attributes of the mass lesion (size, shape, etc.) and the patient's age. Additionally, the dataset contains for each patient the BI-RADS assessment, which is a score ranging from 0 (definitely benign) to 6 (highly suggestive of malignancy) given by the radiologist upon checking the results of the test. The ground truth is also provided (the severity field), which specifies whether the corresponding mass lesion is benign (0) or malign (1). In the dataset, there are 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg (Germany) between 2003 and 2006.

In particular the dataset contains the following attributes (separated by a ","):

- BI-RADS assessment: score from 0 (definitely benign) to 6 (highly suggestive of malignancy), (**ordinal**)
- Age in years of the patient, (**ordinal**)
- Shape: mass shape: round=1 oval=2 lobular=3 irregular=4, (**nominal**)
- Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (**nominal**)
- Density: mass density high=1 iso=2 low=3 fat-containing=4 (**ordinal**)
- Severity: benign=0 or malignant=1 (**nominal**)

**Nominal vs. Ordinal.** Observe that for each attribute we also specified its type: "nominal" or "ordinal". "Nominal" specifies that there is no intrinsic

order on the values that the attributes can take (e.g. the attribute "country" can take the following values: 'China', 'France','Italy', 'USA', with no order, i.e. no country is more important than the others). "Ordinal" specifies that there is an intrinsic order on the values that the attributes can take (e.g. the shirt size = 'S','L','XL', the age, etc.) Some data mining techniques work well only if the attribute type is either "nominal" or "ordinal".

**Notes.** Observe that some values might be missing. In that case there will be the placeholder '?'. Those instances should not be deleted. For questions 1-4, you can safely ignore the missing values, however, in question 5 you should be a little bit more careful on how to handle that.

**Questions.** You should answer the following questions.

1. Report 3 rules with support at least 0.2 and confidence at least 0.9. Specify for each of them the support and the confidence.

2. This task consists of determining some attributes and their values that help us to find out whether a given instance is benign (severity =0) or malign (severity=1). We are looking for rules of the kind:

$$A_1 = a_1, \ldots, A_k = a_k \rightarrow \text{Severity='0'}$$
$$\text{or}$$
$$A_1 = a_1, \ldots, A_k = a_k \rightarrow \text{Severity='1'}$$

   where $A_i = a_i$ denotes an attribute and its value. For example the following rule:

$$\text{Shape='4', Margin='1'} \rightarrow \text{Severity='0'}$$

   tells us that instances with the specified shape and margin are usually benign. Remember that only rules with support at least 0.1, (i.e. their frequency is at least 10%) are relevant for us. Rules with lower support are usually not informative, as there is no much evidence they are true or not. In our exercise we consider relevant any rule with confidence at least 0.9 (i.e. they are true 90% of times). Report one or two rules with the specified requirements that you think might help us predicting whether a given instance is benign or malign. You should not report rules with the attribute BI-RADS for this question. Which insights did you get from those rules? (e.g. the margin of the lesion can help us determining whether a lesion is benign or malign).

3. As discussed above, the BI-RADS assessment is not always accurate and it might lead to unnecessary breast biopsy. Provide one or two rules that might give some evidence that the BI-RADS assessment is not always accurate. Explain your answer.

4. Write a script in Python to find the confidence and support of the following rule: Age=35 $\Rightarrow$ Severity=0. Report its support and confidence. Do you think this rule tells us something valuable or that we should ignore it as there is not enough evidence to support this rule?

5. The attribute "Age" is ordinal which makes the rule mining approach not ideal. In particular, one would like to obtain rules of the kind

$$\text{Age} \geq n, A_1 = a_1, \ldots, A_k = ak -> \text{Severity='1'}$$

(where n is an integer), as the age is an important factor in determining whether a given instance is malign or benign. However, this issue can be circumvented in our case by modifying the input file (the 'csv' file) accordingly. Be careful on how you handle the missing values (i.e. those with a '?'). Provide at least one rule of that kind with support at least 0.1 and confidence at least 0.9.

**What to submit.** You should send us your Jupyter notebook containing your code in Python as well as the answers to the previous questions.