

### 3.4 Homework: Bird associations

**Recall that this task is homework that is done in groups of 2–3 students – you cannot do the task alone or in a larger group.** You can search collaborators in Zulip, exercise sessions, or ask help from the TAs.

*Learning goals: Mining association rules in practice; making efficient data mining pipelines.*

This is an explorative task, where you should invent good features to extract from the extended bird data and then search and analyze association rules. The pattern discovery process is iterative, and you will very likely experiment with multiple versions of feature extraction. Therefore, it is recommended to do the preparations well and make a shell script that speeds up the process. You can find instructions and hints in MyCourses (**instructionsforkingfisher.pdf**). You can find an extended version of the bird species data, **birds2025ext.csv**, and its description in MyCourses.

- a) Extract good features for association discovery from the bird data. You can find interesting associations only if the involved properties are captured by features! It is suggested to proceed iteratively, from easier to more difficult features:
  - Group, habitat and diet can be used as such (just list group and all elements of habitat and diet in the transaction). If the subfamily (3rd level category) is given, consider adding also the family, which is a more general attribute (e.g., *lari* for *laridae* and *sternidae*).
  - For most binary features (like long-billed), you can use only the Yes-values (list attribute “long-billed” in the transaction, but forget its opposite, “non-long-billed”). The only exception is field *sim*, where both values are interesting (if genders look similar or different).
  - For multi-valued categorical features, you can create one attribute for each value.
  - Invent some informative features from the spring and autumn migrations times (fields “arrives” and “leaves”), e.g., describing that migration starts early or ends late.
  - Invent how to handle numerical features. Usually, only the extremes are interesting, like laying relatively few eggs or many eggs. Here you can utilize the indices that you derived in Exercise 2

(e.g., attribute “robust” for high BMI or “short-winged” for low WSI).

- b) Search association rules with Kingfisher. You may need to search quite many rules (e.g., 300) to find more versatile rules, since there will be many variants of similar associations. Try to find rules that describe different aspects of the data, like different groups, appearance, diet, habits, environment, etc, but remember that all attributes do not necessarily participate any significant associations.
- c) Report the most significant and interesting rules. The idea is not to list all rules, but group rules and describe the information they reveal (e.g., what things are associated to scolopacidae or plunge-divers). Tell also which features seem to be irrelevant (did not occur in any rules).

### **Parts of the report:**

1. Cover page: title (course name and assignment number), names and student ids of all participants of the team.
2. Section 1 “Methods”: Describe very briefly the methods: what programming language you used for feature extraction, what were the parameter settings for Kingfisher, if you used any constraints etc. However, do not describe the feature extraction here.
3. Section 2 “Feature extraction”: Describe compactly but carefully what features you extracted. You can, e.g., use a list or a table that tells the original feature, new attributes, and how they were extracted. Describe carefully non-trivial extraction (like handling numerical values or migration month ranges).
4. Section 3: “Results”: Describe the most significant and interesting associations you discovered (see above).
5. Section “Appendix”: Include here the code of your feature extraction program.

**Produce a PDF report containing all parts as described above and submit it in MyCourses before the deadline.** You can find a L<sup>A</sup>T<sub>E</sub>X template for the report in MyCourses, under section “Exercises”. One submission per group!

## Appendix A: Required equations of mutual information

Mutual information of rule  $\mathbf{X} \rightarrow C=c$  is

$$MI = \log \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)} P(\neg \mathbf{X}C)^{P(\neg \mathbf{X}C)} P(\neg \mathbf{X}\neg C)^{P(\neg \mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})} P(\neg \mathbf{X})^{P(\neg \mathbf{X})} P(C)^{P(C)} P(\neg C)^{P(\neg C)}}$$

Conditional mutual information for evaluating rule  $\mathbf{XQ} \rightarrow C=c$  given  $\mathbf{X}$  in the value-based interpretation is

$$MI_C = \log \frac{P(\mathbf{X})^{P(\mathbf{X})} P(\mathbf{XQC})^{P(\mathbf{XQC})} P(\mathbf{XQ}\neg C)^{P(\mathbf{XQ}\neg C)} P(\mathbf{X}\neg QC)^{P(\mathbf{X}\neg QC)} P(\mathbf{X}\neg Q\neg C)^{P(\mathbf{X}\neg Q\neg C)}}{P(\mathbf{XQ})^{P(\mathbf{XQ})} P(\mathbf{X}\neg Q)^{P(\mathbf{X}\neg Q)} P(\mathbf{XC})^{P(\mathbf{XC})} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)}}$$

The base of the logarithm is not fixed (usually 2 or  $e$ ), but in this task you are asked to use the 2-based logarithm for better comparison of results. Note that in the task you should report  $n \cdot MI$  and the thresholds are also given for  $n \cdot MI$ , where  $n$  = data size.

Note also that mutual information doesn't differentiate between positive and negative dependencies. Therefore you need other means to find out if (conditional) dependence is positive or negative.