Lab 02 - Classification & Clustering

Data Mining - Term I/2021-2022

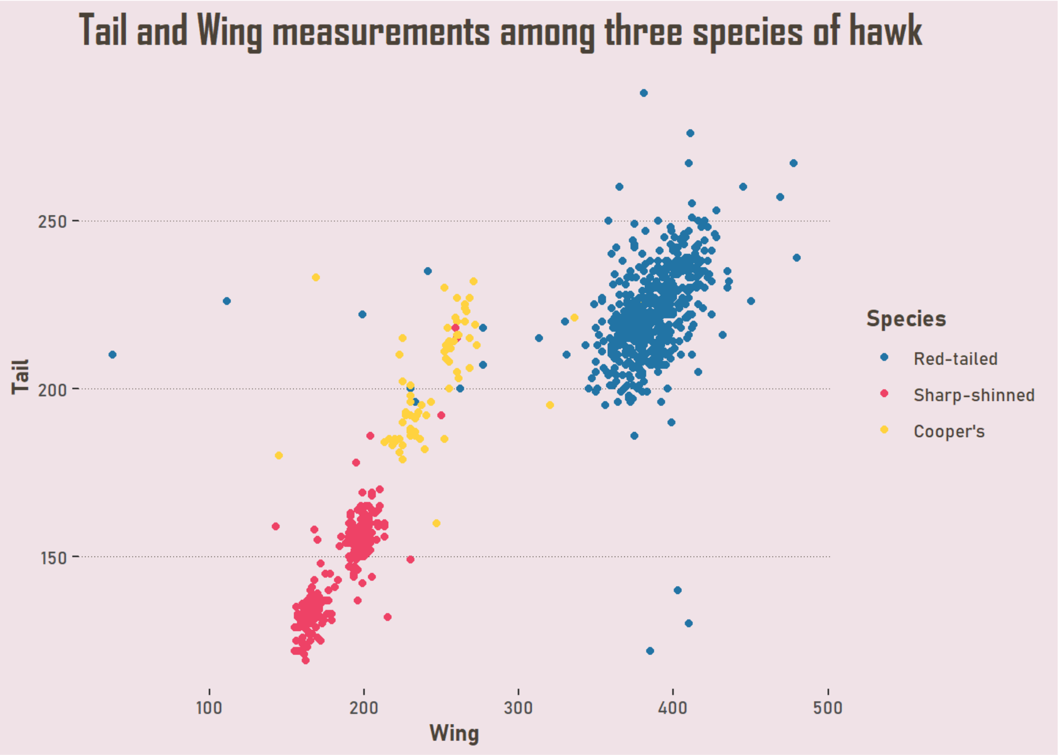
Objectives

- In this lab, students need to explore WEKA's data classification utilities by using two interfaces Explorer and Experimenter.
- Additionally, students get to improve programming skills by implementing basic clustering algorithms from scratch.

Rules

- Maximum 2 members/group.
- Duration: 3 weeks (more details on Moodle).
- Submission directory: **<StudentID>** or **<StudentID1>_<StudentID2>** depending on the number of members in your group, containing the following files:
  - preprocess.py: preprocess script (if necessary).
  - Results.xls: results summary of experiments A-D.
  - RawResults.csv: output of experiment D.
  - Observations.pdf: group's observations and answers for given questions.
  - Lab02-Clustering.ipynb: students' work on Clustering part.
- Compress the submission directory into a **zip** file then submit it on Moodle.
- **Similar works will be marked 0 for the whole course**.
- Cite references properly.

1 CLASSIFICATION



Specification for the dataset **hawks**: Hawks: Measurements on Three Hawk Species in Stat2Data: Datasets for Stat2. The target is to classify each data sample into one of the three species, namely red-tailed, sharp-shinned, and Cooper's.

1.1 Using Weka Explorer to classify data

For each of the experiments A-C below, use WEKA Explorer to perform classification by following methods (with default parameters): 1) NaiveBayesSimple; 2) Id3; and 3) J48. For each of the methods applied on the dataset, perform the following evaluation methods (see “Test options” in “Classify” window of Weka Explorer): a) “Use training set”; b) “Cross-validation” with 10 folds; and c) “Percentage split” with proportion of 66%. Write down the results of each run into an Excel file “Result.xls”, which includes the follow information:

(a) Experiment type (A-C);

(b) Name of the input dataset;

(c) Classification method;

(d) Evaluation strategy;

(e) The proportion of correctly classified samples (the total and for each specific class accuracy).

Experiments:

(A) Classify data points using methods above.

(B) Discretize all non-class attributes into 10 buckets with **equal widths**: use function "Filter" in window "Preprocess" of the Explorer interface, select ‘filters’ → ‘unsupervised’ → ‘attribute’ → ‘Discretize’. Use default parameters for ’Discretize’ filter. After ensuring all non-class attributes are discrete, perform classification on the newly processed data set with classification algorithms and evaluation strategies described earlier.

(C) Discretize all non-class attributes into 5 buckets with **equal depths** by using ’Discretize’ filter and appropriate parameters. After ensuring all non-class attributes are discrete, perform classification on the newly processed data set with classification algorithms and evaluation strategies described earlier.

1.2 Data classification using Weka Experimenter

(D) With this experiment, you must use WEKA Experimenter interface. Perform data classification with NaiveBayesSimple and J48 (default parameters). For each method, run 10-fold cross validation 10 times, and write down the results into a file called "RawResult.csv". From these results, compute the average accuracy of each method and add into the "Result.xls" file mentioned earlier: a) Experiment D; b) name of the data set, c) classification method; d) evaluation strategy; and e) average proportion of correctly classified samples after 10 × 10 runs.

Note: the given data set contains missing values, so you must handle it before classifying (you can use any reasonable method). If you use python to preprocess it, submit an extra **preprocess.py** file. Otherwise if you use Weka to preprocess, please report the steps and methods used.

1.3 Assess the experiment results

After successfully performing the experiments, you should spend time assessing the results. Specifically, you should at least be able to answer below questions:

- Which classification method typically has the best result?
- Which method does not work well and why?
- Why should we use the discretized version of the data set instead of the original one?

- Do the discretization process and method affect the classification results? If yes then how?
- Which evaluation strategy tends to overestimate the accuracy and why?
- Which evaluation strategy tends to underestimate the accuracy and why?

Answer those questions and make new observations into a “Observations.pdf” file.

2 CLUSTERING

In this section you must learn to use the **Jupyter notebook**, then do as instructed in the file **Lab02-Clustering.ipynb**.

If you have any questions related to the exercise, feel free to email me at *ktoan271199@gmail.com*.
Best regards.