# Homework 3

(Advanced) Data Mining: Algorithms and Applications-Winter 2023

**Due on Mar 13, 11.59PM**

**Important**

- Please type your answers for the calculations this time.
- Only submissions through Canvas will be accepted.
- Please submit your code with its own extension, i.e., `HW3.R` along with the document (`HW3.pdf`) which would include your calculations. In other words, you will be submitting two different files. Please do NOT zip.

1. Suppose that we have age data including the following numbers in sorted order. Then answer the questions below. (5 points each)

   age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

   (a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

   (b) Use IQR measure to determine if there are any outliers in this data.

   (c) Use *min-max* normalization to transform the value 35 for age onto the range [0.0, 1.0].

   (d) Use z-score normalization to transform the value 35 for age? (you need to compute mean and standard deviation first)

   (e) Use normalization by decimal scaling to transform the value 35 for age.

2. Write a function in your preferred language which can take a data vector and do min-max normalization by transforming data onto a desired range. For example, it should be able get the age data above and map it between any two numbers. (`foo (a, min_new, max_new`) where a is an one-dimensional array) (25 points)

| department | age | salary | status | count |
|---|---|---|---|---|
| sales | 31_35 | 46K_50K | senior | 30 |
| sales | 26_30 | 26K_30K | junior | 40 |
| sales | 31_35 | 31K_35K | junior | 40 |
| systems | 21_25 | 46K_50K | junior | 20 |
| systems | 31_35 | 66K_70K | senior | 5 |
| systems | 26_30 | 46K_50K | junior | 3 |
| systems | 41_45 | 66K_70K | senior | 3 |
| marketing | 36_40 | 46K_50K | senior | 10 |
| marketing | 31_35 | 41K_45K | junior | 4 |
| secretary | 46_50 | 36K_40K | senior | 4 |
| secretary | 26_30 | 26K_30K | junior | 6 |

Table 1: Data shows the count of each feature combination. For instance, There are 30 senior sales staff who are 31...35 years old and have 46...50K salary. Since each combination is unique, their corresponding groups are mutually exclusive which implies counts are not double counts for any of the cases. Notice that the status column is the class label to indicate whether someone is junior or senior.

3. Using **information gain** on the data in Table 1, do calculations for two levels of a decision tree which decides whether a person is senior or junior. Please show your calculations and clearly write down your junior and senior counts not to confuse yourself. Note that you need to calculate the information gain for all attributes (department, age, salary) and pick the one to start your tree. In your subsets of your data, you'll perform the same operation for the attributes available. (You can use a computing environment to write the mathematical expressions. i.e., $p*logp$, $(1/2)*log2(1/2)$ ) (25pts)

4. Using the decision tree you generate if-then rules. (25pts)