# CMP4336 – Introduction to Data Mining

## Homework 2

**Deadline:** September 9, 2020 till 23:59 (strict deadline, no extension!)

You are asked to implement Naïve Bayes classifier on abalone dataset. You can use scikit-learn or any other library.

The dataset (input features and class labels of the samples) is provided as a separate text file (abalone_dataset.txt):

Detailed information about abalone dataset can be found at
http://archive.ics.uci.edu/ml/datasets/Abalone

The aim of the dataset is to predict the age of abalone from physical measurements. Originally it is a regression problem in which the output is age in years. However, we will use it as a classification problem. The age value is discretized as young, middle-aged, and old. The dataset with class labels is provided as a separate text file (abalone_dataset.txt):

**input:** Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight

**output:** class label which is the last column of the dataset (less than 8 in age belongs to class 1 (young), between 8 and 12 to class 2 (middle-aged), greater than 12 to class 3 (old))

Hyper-parameter optimization is not required in Naive Bayes classification. So, the dataset will be divided into training and validation sets only (there will not be a test set). Assume gaussian distribution for continuous features.

1) Apply naive bayes classifier using all features as input, and

1.1) 100 samples for training, and rest for validation set

1.2) 1000 samples for training, and rest for validation set

2) Apply bi-directional search feature selection algorithm to the dataset using Naive-Bayes as the baseline classification algorithm.

2.1) Report the order of features selected by the algorithm.

2.2) Using top 3 selected features and 100 samples for training, apply naïve bayes classifier (the rest of the samples will be used for validation).

2.3) Using top 3 selected features and 1000 samples for training, apply naïve bayes classifier (the rest of the samples will be used for validation).

**For each** of the **above cases**,

- Report how many total misclassification errors are there on the training and validation sets, together with the confusion matrices.

(Note: A confusion matrix is a 3x3 matrix (if # of classes is 3) where entry (i,j) contains the number of instances belonging to i but are assigned to j; ideally it should be a diagonal matrix.)

- Report the case in which highest accuracy is obtained. Write your **comments about the results**.

**Guidelines**

1. Use Python.
2. Submit **a single pdf** file which includes
    a. the **required output for each of the cases given** above,
    b. your comments about the results,
    c. and the **source code** you have written.
3. Submissions that include more than one pdf file will NOT be evaluated.
4. Submission will be made through itslearning, NOT e-mail.