ECLT5810/SEEM5750 E-Commerce Data Mining Technique

2024-2025 First Term

Assignment 1

Due: Oct 14 (Mon) 17:00

Problem Statement

This assignment involves using Weka to predict client subscription to term deposits. The prediction process consists of two phases: 1) data preprocessing, and 2) modeling and deployment for operation. In phase 1, you will preprocess the data and split the dataset into training and testing sets. In phase 2, you will use Weka to build and evaluate a decision tree model for predicting client subscription to term deposits. **This is an individual assignment, and you must follow the requirements listed below in order.**

1) Data preprocessing:

1. Use the "Bank Marketing (with social/economic context)" dataset, specifically the "**bank-additional.csv**" file, to build the model.
2. The target variable is *"y"* which indicates whether the client subscribed to a term deposit or not.
3. Group clients into 6 equal width age groups.
4. For the *"duration"* variable, use min-max normalization with a minimum of 0 and a maximum of 1.
5. Standardize other numeric variables not involved in Step 4 to have a mean of 0 and a standard deviation of 1.
6. Create dummy variables for the *"job"* and *"education"* attributes.
7. Eliminate variables that are not related to the target variable by employing both correlation-based and information gain-based selection methods. Retain the top 10 attributes from each method, which means that **you should maintain the union set of top 10 attributes derived from both methods.** Moreover, please include the "age" attribute if it doesn't appear in the top 10 for either method.
8. Divide the dataset into two subsets: a training set comprising 80% of the original dataset and a testing set consisting of the remaining 20%.
9. Save the training and testing sets as two separate CSV files for submission. Please name the training set as "**bank-train-set**" and the testing set as "**bank-test-set**".

It is important to note that during Steps 3-6, specific operations such as data discretization and normalization must be performed on **particular variables.** In Weka, filters are applied to all variables by default. To address these requirements, necessary modifications and operations should be implemented.

2) Modeling and deployment for operation:

10. Build a decision tree model using the training set (**bank-train-set**) to predict the term deposit subscription of clients. Save the model as "Decision-Tree".
11. Evaluate the "Decision-Tree" model using the testing set (**bank-test-set**) and compute the accuracy.
12. Set minimum number of instances per leaf as **20** (which is 2 by default). Repeat Step 10.
13. Evaluate the modified model, as in Step 11.
14. Save the better model, i.e., the model with the highest accuracy.
15. The better model will be deployed for operation and the model will predict the term deposit subscription of new clients. To simulate this operational setting for a bunch of clients, we provide a new client data set, namely, **bank-new-clients.csv**. The aim is to make prediction

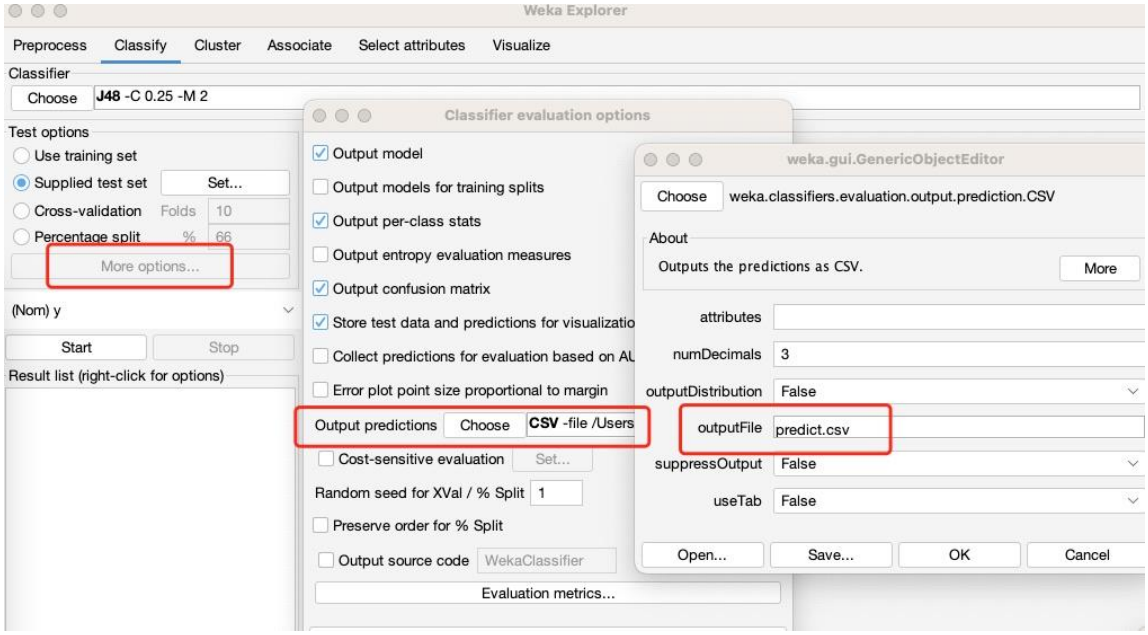
for each of the new clients.

Since the data in this new client data set was not available during the training stage, this new data has not been pre-processed. Therefore, in order to make it suitable for use, you should perform the same variable transformation and variable selection techniques on this new client data, using the same data pre-processing statistics as in the training stage when the model was trained. Precisely the pre-processing steps have been specified in Phase 1. But the pre-processing manipulation for this new client data set are different in order to make it suitable. For example, for min-max normalization, use the min and max values adopted when the model was trained.

Some hints for this step are described below.

Save the pre-processed file in .arff format.

16. Load the decision tree model obtained in Step 14. Then use it to conduct prediction on the pre-processed file in Step 15. Output the prediction results to a CSV file as follows:



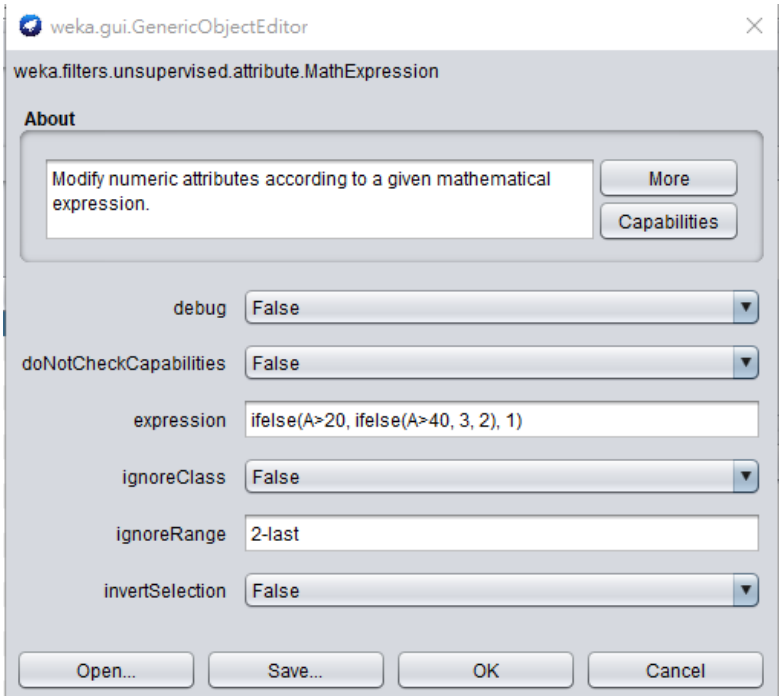
Hints for pre-processing the new client data

In Step 15, you are required to perform the same variable transformation and variable selection techniques on the new client dataset. Since discretization depends on the data, the resulted bins in the training process and the operational process will be different, although you applied the same configuration. In practice, we should use the ranges obtained from the training data and apply the same ranges to the operational client data. As a result, we can perform manual discretization on operational client data.

Suppose the ranges for the first attribute obtained from training data were (-inf-20.0], (20.0-40.0] and (40.0-inf).

After the operational client data is loaded in Weka, in preprocess tab, choose MathExpression, it is under unsupervised->attribute.

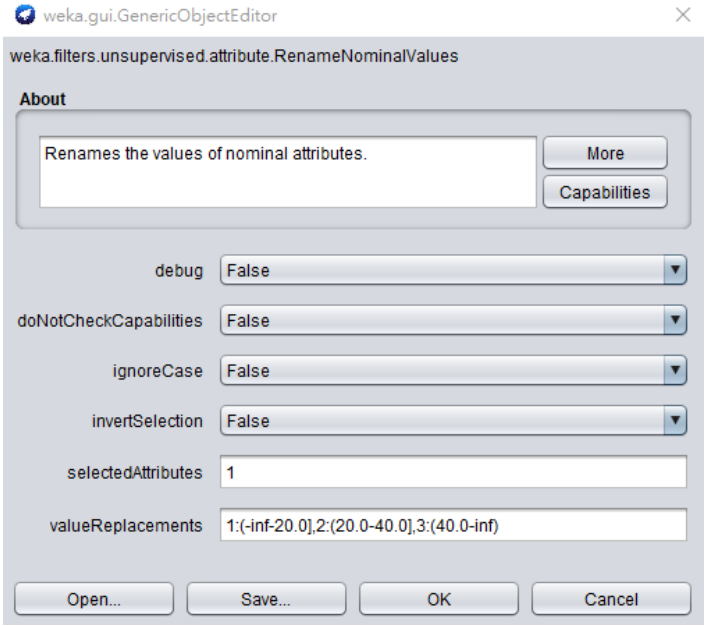
Enter this expression: **ifelse(A>20, ifelse(A>40, 3, 2), 1)**
Enter **2-last** in ignoreRange if we only want to apply it to the first attribute.



Then, we need to use NumericToNominal to turn it into a nominal attribute.

After applying NumericToNominal, we will have (-inf-20.0] as label 1, (20.0-40.0] as label 2 and (40.0-inf) as 3. Now, we need to rename the label to the range.

Choose RenameNominalValues, it is under unsupervised->attribute.
Enter **1:(-inf-20.0],2:(20.0-40.0],3:(40.0-inf)** in valueReplacements.
Enter **1** in selectedAttributes as we only want to apply on the first attribute.



We will obtain the same discretization as our training data.

For Normalization and Standardization, you should use the values obtained from the pre-processing stage of the training data to handle the operational client data, e.g. the min and max values for Normalization, mean and std for Standardization. You can use MathExpression to manually apply those methods on the operational client data.

Assignment Submission

1. There are five files for this assignment to be submitted:

a) One model object file representing the trained decision tree;

b) Three CSV files representing the training and testing sets after completing the specified operations, and the prediction results of your decision tree model;

c) One PDF file detailing the operations performed along with corresponding Weka screenshots. Use Weka screenshots to report the results.

2. Compress these five files into a single ZIP file. Ensure all the file names follow the "id-name" format. For example, if your name is Chan Tai Ming and your student ID is 00123456, the files should be named as follows:
   a) "00123456-CHANTaiMing-Assignment1-decision-tree.model" for the model object file;
   b) "00123456-CHANTaiMing-Assignment1-bank-train-set.csv" for the training set results;
   c) "00123456-CHANTaiMing-Assignment1-bank-test-set.csv" for the testing set results;
   d) "00123456-CHANTaiMing-Assignment1-predict.csv" for the prediction results;
   e) "00123456-CHANTaiMing-Assignment1-report.pdf" for the assignment report;
   f) "00123456-CHANTaiMing-Assignment1.zip" for the ZIP file.

3. Upload the ZIP file to Blackboard for submission.

4. You may resubmit your assignment before the deadline, with only the latest submission being assessed.

5. For any inquiries, please email sli@se.cuhk.edu.hk.