Introduction to Preprocessing in WEKA

In this exercise you will get to know some of the basic preprocessing that can be done using WEKA. Follow the steps below to create an .arff file.

An .arff file (Attribute-Relation File Format) consists of a header that looks like this:

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

And a data body:

```
@DATA
  5.1,3.5,1.4,0.2,Iris-setosa
  4.9,3.0,1.4,0.2,Iris-setosa
  4.7,3.2,1.3,0.2,Iris-setosa
  4.6,3.1,1.5,0.2,Iris-setosa
  …
```

The file/relation is named by the @relation definition. The attributes are defined by @attribute. They can be numeric, nominal-specifications (see attribute class), a string or a date. After the @data tag the samples follow - one sample per line - with the attributes separated by commas.

Tasks

1.) Go to http://www.cs.waikato.ac.nz/ml/weka/ and download and install WEKA.

2.) Start WEKA and choose the Explorer.

3.) Download the sample data set “bank-data.csv” from the iLearn and open it in WEKA. (If prompted convert the file into .arff format.) These are the attributes:

id	a unique identification number
age	age of customer in years (numeric)
sex	MALE / FEMALE
region	inner_city/rural/suburban/town
income	income of customer (numeric)
married	is the customer married (YES/NO)
children	number of children (numeric)
car	does the customer own a car (YES/NO)
save_acct	does the customer have a saving account (YES/NO)
current_acct	does the customer have a current account (YES/NO)
mortgage	does the customer have a mortgage (YES/NO)
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)

WEKA performs some basic statistic analysis when data is loaded.
Click on any attribute in the left panel to show the basic statistics of that attribute. For categorical attributes the frequency for each attribute value is shown, while for continuous attributes we can obtain min, max, mean, standard deviation, etc.
As an example, select the attributes “age” and “married” respectively.

4.) Now we can start with the preprocessing. As you can see the dataset contains an attribute “id” which identifies each sample uniquely. As you remember from the lecture such attributes have to be removed. Click on the “Choose” button in the “Filter” panel and navigate through the list and select the “weka.filters.unsupervised.attribute.Remove” filter.

5.) Now click on the text field on the right-hand side of the “Choose” button. Type in **the index** of the attribute (“id”) that you want to remove. Make sure that the “invertSelection” option is set to false. Then click “OK”. Apply the filter using the “Apply” button. (Alternatively you can select the checkbox of the “id” attribute on the bottom left and use the “Remove” button.)

6.) Since some techniques only work with nominal data, we have to perform discretization on some attributes (“age”, “income” and “children”).

- The attribute “children” has already only four possible values, so the easiest way to change that attribute is to save the data in an .arff file. Open it in a text editor (e.g. Notepad++) and change the attribute declaration from:
  @attribute children numeric
  to:
  @attribute children {0,1,2,3}
- For the other two attributes we can use WEKA's discretization filters. With these filters we can divide each attribute into 3 bins/intervals. We will rely on WEKA's ability to determine the ranges of these attributes automatically.
  - Open the recently manually changed .arff file. The attribute children should now be nominal.
  - Click on the “Choose” button in the filter panel and select “weka.filters.unsupervised.attribute.Discretize” from the list.
  - Click on the textbox to enter the index of the attributes to discretize (“1, 4”) and set the number of bins to 3. All other options should be “false”.
  - Apply the filter and save the file as an .arff file.

7.) Open the .arff file in Notepad++ (or an editor of your choice that has a find and replace function). As you can see the attributes “age” and “income” have been converted. But the labels WEKA chose are not well readable. Select each label and perform a find and replace operation on each so that they are changed as follows:

- '\'(-inf-34.333333]\'' → 0_34
- '\'(34.333333-50.666667]\'' → 35_51
- '\'(50.666667-inf)\'' → 52_max
- '\'(-inf-24386.173333]\'' → 0_24386
- '\'(24386.173333-43758.136667]\'' → 24387_43758
- '\'(43758.136667-inf)\'' → 43759_max

8.) Save the .arff file as “bank-data-final.arff”.