

ECLT5810/SEEM5750

Weka – Installation, Data Pre-processing Introduction

What is Weka?

Weka is an open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research, and industrial applications, contains a lot of built-in tools for standard machine learning tasks.

Here is the official website: <https://ml.cms.waikato.ac.nz/weka>

Weka installation

Please follow the instruction here to install the stable version (3.8) of Weka

https://waikato.github.io/weka-wiki/downloading_weka/

It provides different versions to suit different OS. Please select the one you are using.

Dataset

- We will use the Bank Marketing Data Set. You can download the data set, known as “bank.csv”, via the link given on the course web site.
- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- It contains 4521 samples and 16 input variables. The target y is the client subscribed a term deposit or not. In machine learning terminology, it is a binary classification problem.

Dataset

Here is the information of the 16 input variables:

- 1 - age (numeric)
- 2 - job : type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")
- 3 - marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
- 4 - education (categorical: "unknown", "secondary", "primary", "tertiary")
- 5 - default: has credit in default? (binary: "yes", "no")
- 6 - balance: average yearly balance, in euros (numeric)
- 7 - housing: has housing loan? (binary: "yes", "no")
- 8 - loan: has personal loan? (binary: "yes", "no")

Dataset

- 9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
- 10 - day: last contact day of the month (numeric)
- 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 12 - duration: last contact duration, in seconds (numeric)
- # other attributes:
- 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- 15 - previous: number of contacts performed before this campaign and for this client (numeric)
- 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

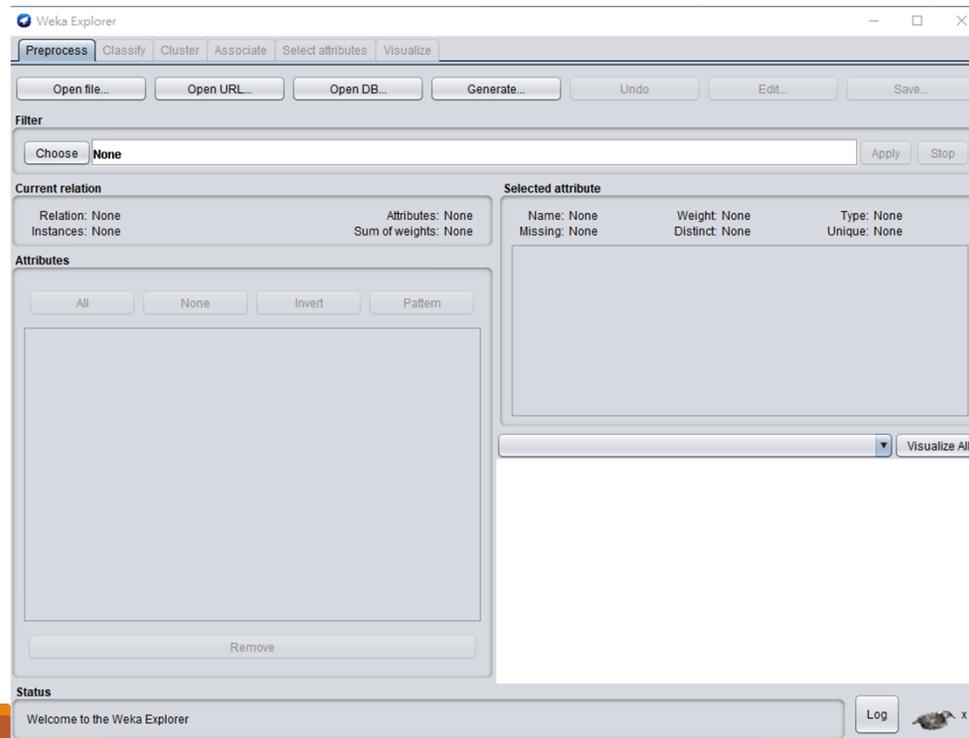
Weka GUI Chooser

If you open the Weka software, first is the Weka GUI Chooser like the following.



Explorer

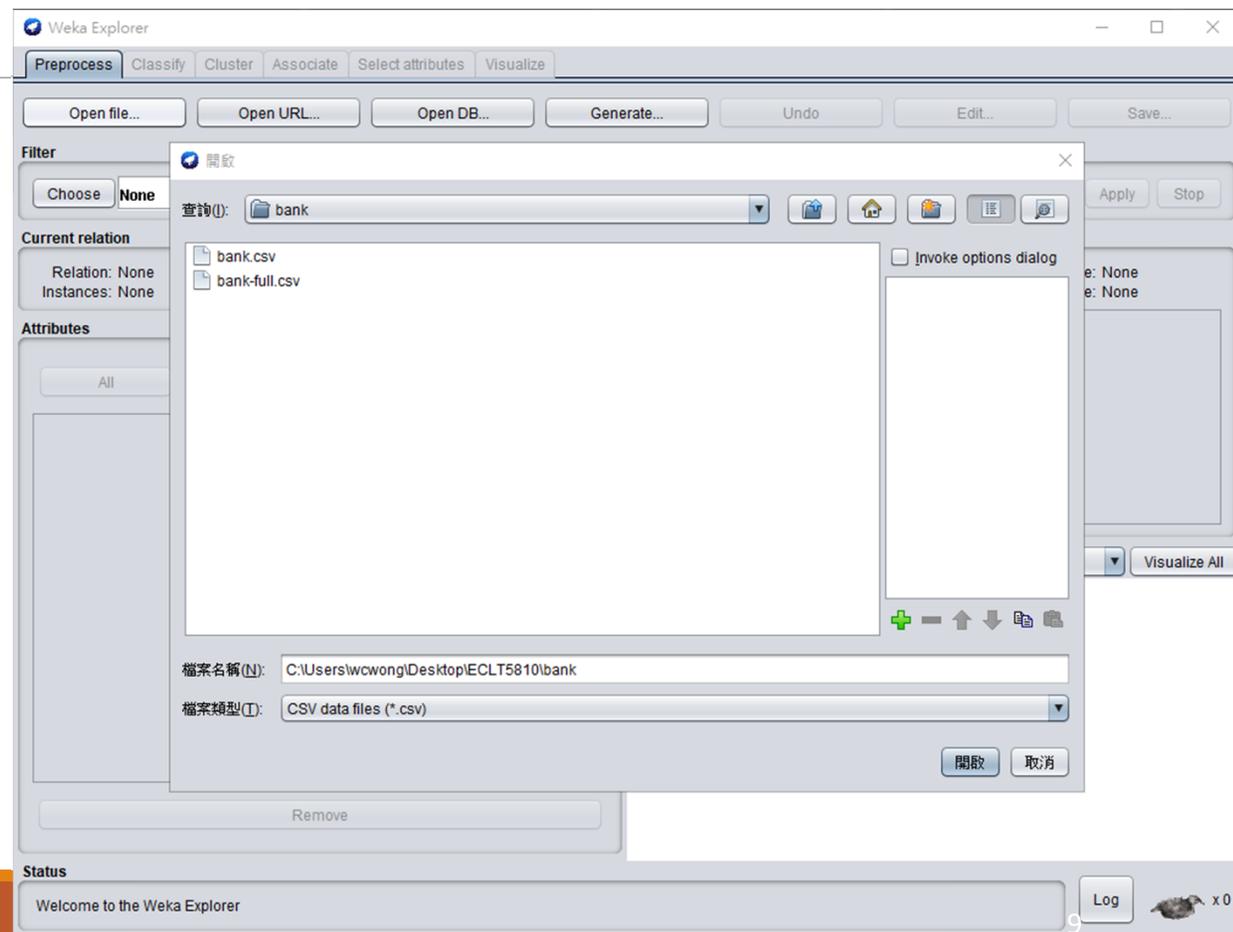
Then, select the Explorer, you will enter to this interface.



Explorer

Click **Open file**, then open the bank.csv saved in your computer.

Please remember to change to **CSV data files (*.csv)** in file type.



Explorer

Now, you can see the data is loaded into Explorer.

You can check out each variable by click on it in this panel.

The screenshot shows the Weka Explorer interface. The 'Current relation' panel displays 'Relation: bank' and 'Instances: 4521'. The 'Attributes' panel lists 17 variables, with 'age' selected. The 'Selected attribute' panel shows statistics for 'age', including Minimum (19), Maximum (87), Mean (41.17), and StdDev (10.576). A histogram for 'age' is visible at the bottom right.

Current relation
Relation: bank
Instances: 4521
Attributes: 17
Sum of weights: 4521

Attributes

No.	Name
1	<input checked="" type="checkbox"/> age
2	<input type="checkbox"/> job
3	<input type="checkbox"/> marital
4	<input type="checkbox"/> education
5	<input type="checkbox"/> default
6	<input type="checkbox"/> balance
7	<input type="checkbox"/> housing
8	<input type="checkbox"/> loan
9	<input type="checkbox"/> contact
10	<input type="checkbox"/> day
11	<input type="checkbox"/> month
12	<input type="checkbox"/> duration
13	<input type="checkbox"/> campaign
14	<input type="checkbox"/> pdays
15	<input type="checkbox"/> previous
16	<input type="checkbox"/> poutcome
17	<input type="checkbox"/> y

Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 67
Type: Numeric
Unique: 4 (0%)

Statistic	Value
Minimum	19
Maximum	87
Mean	41.17
StdDev	10.576

Class: y (Nom) Visualize All

Status: OK Log x 0

Explorer

The statistics for each variable are also shown here.

For example, the maximum and minimum value of age is 87 and 19 respectively.

The screenshot shows the Weka Explorer interface. The 'Selected attribute' panel is circled in red and displays the following statistics for the 'age' variable:

Statistic	Value
Minimum	19
Maximum	87
Mean	41.17
StdDev	10.576

Below the statistics, a histogram for the 'age' variable is displayed, showing a distribution of values from 19 to 87. The x-axis is labeled 'Class: y (Nom)' and the y-axis represents frequency. The histogram bars are blue with red outlines. The x-axis has tick marks at 19, 53, and 87. The y-axis has tick marks at 14, 29, 58, 70, 15, 16, 7, 19, 10, 9, 8, 13, 1, 5, 2.

Data Pre-processing/Feature Engineering

Data pre-processing/Feature engineering in Weka mostly contains two components.

- The variable transformation and,
- The variable selection

Variable transformation can be applied to the inputs for improving the precision of the predictive models.

Variable selection is useful when you want to make an initial selection of inputs or eliminate irrelevant inputs. It can also help identify non-linear relationships between the inputs and the target.

Variable Transformation

- There are several variable transformation methods that can be applied to the input variables such that the precision of the predictive models can be improved.
- However, we cannot know which variable transformation methods will produce the most accurate models.
- Therefore, it is a good idea to try a number of different variable transformation methods techniques on the data and in turn create many different models to test it.

Variable Transformation

In Weka, it provides **filters** for variable transformation.

- Supervised Filters: That can be applied but require user control or make use of the class information in some way. Such as rebalancing instances for a class.
- Unsupervised Filters: That can be applied in an undirected manner. For example, discretize the numerical attributes or rescale all values in the attribution to the range 0 to 1.

We will show the Unsupervised Filters.

Variable Transformation

Under these two filters, there are two groups:

- Attribute Filters: Apply an operation on attributes or one attribute at a time.
- Instance Filters: Apply an operation on instance or one instance at a time.

We will mostly deal with the Attribute Filters.

Variable Transformation

Discretize Numerical Attributes

- Some machine learning algorithms prefer to work with discrete attributes rather than real-valued attributes.
- For example, decision tree algorithms can choose split points in real-valued attributes but are much cleaner when split points are chosen between bins or predefined groups.
- Discrete attributes are those that describe a category, called **nominal attributes**. Those attributes that describe a category that where there is a meaning in the order for the categories are called **ordinal attributes**. The process of converting a real-valued attribute into an ordinal attribute or bins is called discretization.

Variable Transformation

Click Choose, under filters->unsupervised->attributes, select Discretize

The screenshot shows the Weka Explorer interface. The 'Filter' panel on the left displays a tree view of filters. The 'unsupervised' folder is expanded, and the 'Discretize' filter is selected and circled in red. The 'Selected attribute' panel on the right shows the 'age' attribute with the following statistics:

Statistic	Value
Minimum	19
Maximum	87
Mean	41.17
StdDev	10.576

Below the statistics, a histogram shows the distribution of the 'age' attribute. The x-axis represents age values from 19 to 87, and the y-axis represents frequency. The histogram bars are blue with red outlines, showing a distribution that peaks around age 40-50.

Variable Transformation

Click on the text near Choose, you can configure the setting of the method

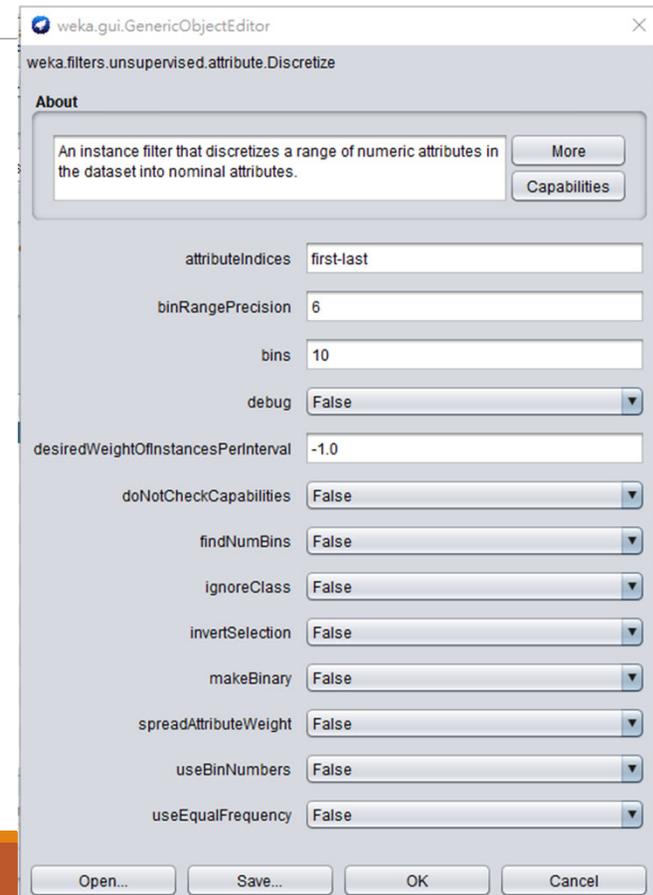
The screenshot shows the Weka Explorer interface with the 'Discretize' method selected for the 'age' attribute. The 'Choose' button is circled in red. The 'Current relation' shows 17 attributes and 4521 instances. The 'Selected attribute' table shows statistics for 'age'.

Statistic	Value
Minimum	19
Maximum	87
Mean	41.17
StdDev	10.576

The histogram below shows the distribution of the 'age' attribute, with a peak around 40-50 years old. The x-axis ranges from 19 to 87, and the y-axis shows frequency.

Variable Transformation

- Here is the configuration of Discretize
- attributesIndices means specify range of attributes to act on
- binRangePrecision means the number of decimal places for cut points to use
- bins means numbers of bin



Variable Transformation

You can move the mouse cursor on the configuration's attributes to see the meaning of it.

The screenshot shows the Weka Explorer interface with the 'Discretize' filter configuration dialog open. The dialog is titled 'weka.filters.unsupervised.attribute.Discretize' and contains the following settings:

- attributeIndices: first-last
- binRangePrecision: 6 (with a tooltip: 'The number of decimal places for cut points to use when generating bin labels')
- debug: False
- desiredWeightOfInstancesPerInterval: -1.0
- doNotCheckCapabilities: False
- findNumBins: False
- ignoreClass: False
- invertSelection: False
- makeBinary: False
- spreadAttributeWeight: False
- useBinNumbers: False
- useEqualFrequency: False

The background shows the 'Attributes' list in Weka Explorer, with 'age' selected. A histogram is visible at the bottom right of the dialog, showing the distribution of the 'age' attribute.

Variable Transformation

Let's leave the setting as default and click OK

Then, click apply.

The screenshot shows the Weka Explorer interface with the 'Discretize' filter applied to the 'age' attribute. The 'Apply' button is circled in red. The interface includes a menu bar (Preprocess, Classify, Cluster, Associate, Select attributes, Visualize), a toolbar (Open file..., Open URL..., Open DB..., Generate..., Undo, Edit..., Save...), and a Filter section with a dropdown menu set to 'Discretize - B 10 - M - 1.0 - R first-last - precision 6'. The 'Current relation' section shows 'Relation: bank' and 'Instances: 4521'. The 'Attributes' section lists 17 attributes, with 'age' selected. The 'Selected attribute' section shows 'Name: age', 'Missing: 0 (0%)', 'Distinct: 67', and 'Type: Numeric'. A histogram of the 'age' attribute is displayed at the bottom right, showing a distribution of values from 19 to 87. The 'Status' bar at the bottom shows 'OK' and a 'Log' button.

Variable Transformation

Now, you can see the attribute age is discretized to 10 bins and you can see the range of each bin in this panel

The screenshot shows the Weka Explorer interface with the 'Discretize' filter applied to the 'age' attribute. The 'Selected attribute' panel displays the following data:

No.	Label	Count	Weight
1	'(-inf-25.8]	111	111.0
2	'(25.8-32.6]	944	944.0
3	'(32.6-39.4]	1235	1235.0
4	'(39.4-46.2]	869	869.0
5	'(46.2-53]	706	706.0
6	'(53-59.8]	482	482.0
7	'(59.8-66.6]	100	100.0
8	'(66.6-73.4]	36	36.0
9	'(73.4-80.2]	30	30.0
10	'(80.2-inf)	8	8.0

The bar chart below the table shows the distribution of instances across the bins, with the highest count in the third bin (1235).