

---

---

# SEEM4630 Tutorial

## Weka Demo on Data Preprocessing

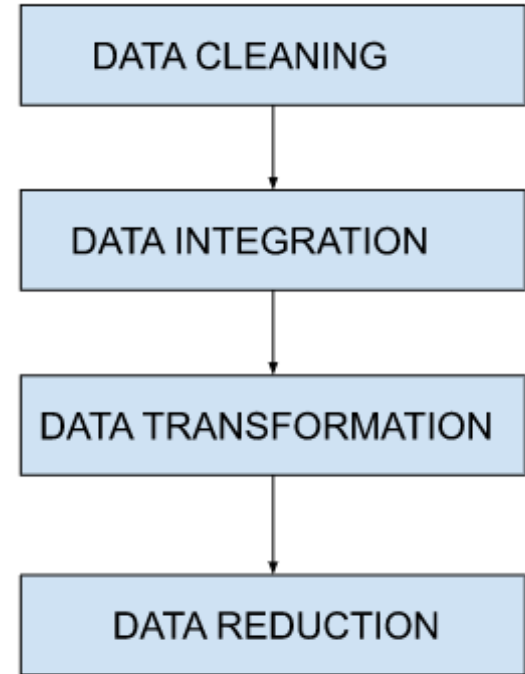
Chaojun Wang  
cj.wang@link.cuhk.edu.hk

---

---

# Data Preprocessing

- ❑ No quality data, no quality mining results!
- ❑ **Data cleaning**
  - ❑ Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- ❑ **Data integration**
  - ❑ Integration of multiple databases, data cubes, or files
- ❑ **Data transformation**
  - ❑ Normalization, standardization, discretization
- ❑ **Data reduction**
  - ❑ Obtains reduced representation in volume but produces the same or similar analytical results



# Overview

- ❑ Weka introduction
- ❑ Weka demo on data preprocessing
  - ❑ Data loading
  - ❑ Data preprocessing
    - ❑ Data transformation
    - ❑ Data reduction
  - ❑ Data splitting
- ❑ Assignment 1 introduction

# What is Weka?

- ❑ **W**aikato **E**nvironment for **K**nowledge **A**nalysis
  - ❑ It's a data mining tool developed by Department of Computer Science, University of Waikato, New Zealand.
  - ❑ Weka is also a bird found only on the islands of New Zealand.



# What is in Weka?

- ❑ WEKA provides implementations of data mining algorithms that you can easily apply to your dataset.
- ❑ It also includes a variety of tools for transforming datasets, such as the algorithms for discretization and sampling. You can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance—all without writing any program code at all.

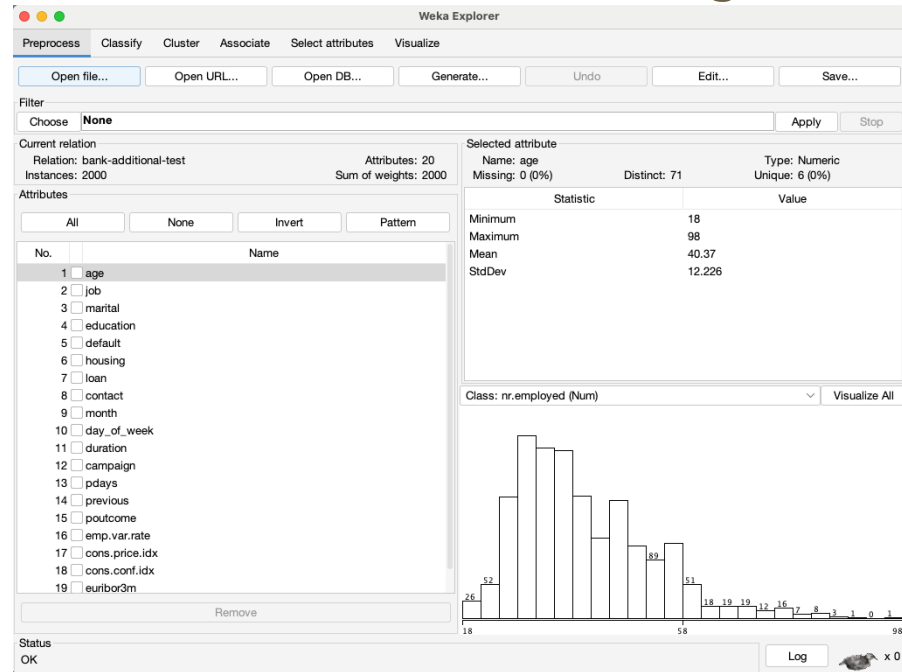
# Download and Install Weka

❏ WEKA is available from <http://www.cs.waikato.ac.nz/ml/weka>

Project	Software	Book	Courses	Publications	People	Related
<h2>Weka 3: Machine Learning Software in Java</h2> <p>Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.</p> <p>Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like <a href="#">this</a>, and the bird sounds like <a href="#">this</a>.</p> <p>Weka is open source software issued under the <a href="#">GNU General Public License</a>.</p> <p>We have put together several <a href="#">free online courses</a> that teach machine learning and data mining using Weka. The videos for the courses are available <a href="#">on Youtube</a>.</p> <p>Weka supports <a href="#">deep learning</a>!</p>						
Getting started		Further information		Developers		
<ul style="list-style-type: none"><li>• <a href="#">Requirements</a></li><li>• <a href="#">Download</a></li><li>• <a href="#">Documentation</a></li><li>• <a href="#">FAQ</a></li><li>• <a href="#">Getting Help</a></li></ul>		<ul style="list-style-type: none"><li>• <a href="#">Citing Weka</a></li><li>• <a href="#">Datasets</a></li><li>• <a href="#">Related Projects</a></li><li>• <a href="#">Miscellaneous Code</a></li><li>• <a href="#">Other Literature</a></li></ul>		<ul style="list-style-type: none"><li>• <a href="#">Development</a></li><li>• <a href="#">History</a></li><li>• <a href="#">Subversion</a></li><li>• <a href="#">Contributors</a></li><li>• <a href="#">Commercial licenses</a></li></ul>		

# Weka Interface

- ❑ WEKA's main graphical user interface, the Explorer, gives access to all its facilities using menu selection and form filling.



# Data Loading

The image shows the Weka Explorer application window. A red arrow points to the 'Open file...' button in the 'Preprocess' tab. The interface displays the 'bank-additional-test' dataset with 2000 instances. The 'Attributes' list on the left includes 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', and 'euribor3m'. The 'Selected attribute' panel on the right shows statistics for 'age', including Minimum (18), Maximum (98), Mean (40.37), and StdDev (12.226). The 'Class: nr.employed (Num)' is selected, and a histogram is displayed at the bottom right, showing the distribution of the number of employees.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

File Choose **None** Apply Stop

Current relation  
Relation: bank-additional-test  
Instances: 2000  
Attributes: 20  
Sum of weights: 2000

Attributes  
All None Invert Pattern

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"/> housing
7	<input type="checkbox"/> loan
8	<input type="checkbox"/> contact
9	<input type="checkbox"/> month
10	<input type="checkbox"/> day_of_week
11	<input type="checkbox"/> duration
12	<input type="checkbox"/> campaign
13	<input type="checkbox"/> pdays
14	<input type="checkbox"/> previous
15	<input type="checkbox"/> poutcome
16	<input type="checkbox"/> emp.var.rate
17	<input type="checkbox"/> cons.price.idx
18	<input type="checkbox"/> cons.conf.idx
19	<input type="checkbox"/> euribor3m

Remove

Status  
OK

Selected attribute  
Name: age  
Missing: 0 (0%)  
Distinct: 71  
Type: Numeric  
Unique: 6 (0%)

Statistic	Value
Minimum	18
Maximum	98
Mean	40.37
StdDev	12.226

Class: nr.employed (Num) Visualize All

Log x 0



# Data Loading

- ❑ Data can be imported from a file in various formats: ARFF, CSV, C4.5, JSON, etc.

- ❑ WEKA's native data storage method is ARFF format

- ❑ ARFF= Attribute Relation File Format

- ❑ ARFF files include two main parts:

- ❑ Specification of the features

- ❑ The actual data

- ❑ When you specify a .csv file

- ❑ it is automatically converted into ARFF format.

Arff data files (\*.arff)

Arff data files (\*.arff.gz)

C4.5 data files (\*.names)

C4.5 data files (\*.data)

CSV data files (\*.csv)

JSON Instances files (\*.json)

JSON Instances files (\*.json.gz)

libsvm data files (\*.libsvm)

Matlab ASCII files (\*.m)

svm light data files (\*.dat)


Binary serialized instances (\*.bsi)

XRFF data files (\*.xrff)


XRFF data files (\*.xrff.gz)

---

# Weka ARFF Example

@relation heart-disease-simplified  This defines dataset name

@attribute age numeric

@attribute sex { female, male}  Nominal features must list values

@attribute chest\_pain\_type { typ\_angina, asympt, non\_anginal, atyp\_angina}

@attribute cholesterol numeric

@attribute exercise\_induced\_angina { no, yes}

@attribute class { present, not\_present}

@data

63,male,typ\_angina,233,no,not\_present

67,male,asympt,286,yes,present

67,male,asympt,229,yes,present

38,female,non\_anginal,?,no,not\_present

...

# Data Preprocessing

- ❑ Data pre-processing tools in WEKA are called “**filters**”
- ❑ Supervised vs Unsupervised
  - ❑ 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.
  - ❑ **Supervised Filters “require a class attribute”, while unsupervised filters do not**

# Data Preprocessing

- ❑ Data pre-processing tools in WEKA called “filters”
- ❑ Attribute vs Instance
  - ❑ Attribute Filters: Apply an operation on attributes or one attribute at a time.
  - ❑ Instance Filters: Apply an operation on instances or one instance at a time.

Weka Explorer

PreprocessClassifyClusterAssociateSelect attributesVisualize

Open file...Open URL...Open DB...Generate...UndoEdit...Save...

Filter

ChooseNone

ApplyStop

Current relation:  
Relation: bank-additional-test  
Instances: 2000

Selected attribute

Attributes

AllNone

No.		Name	Mean	StdDev
1	<input type="checkbox"/>	age	40.37	12.226
2	<input type="checkbox"/>	job		
3	<input type="checkbox"/>	marital		
4	<input type="checkbox"/>	education		
5	<input type="checkbox"/>	default		
6	<input type="checkbox"/>	housing		
7	<input type="checkbox"/>	loan		
8	<input type="checkbox"/>	contact		
9	<input type="checkbox"/>	month		
10	<input type="checkbox"/>	day_of_week		
11	<input type="checkbox"/>	duration		
12	<input type="checkbox"/>	campaign		
13	<input type="checkbox"/>	pdays		
14	<input type="checkbox"/>	previous		
15	<input type="checkbox"/>	poutcome		
16	<input type="checkbox"/>	emp.var.rate		
17	<input type="checkbox"/>	cons.price.idx		
18	<input type="checkbox"/>	cons.conf.idx		
19	<input type="checkbox"/>	euribor3m		

Remove

Class: nr.employed (Num)

Visualize All

Bin Range	Frequency
18-20	26
20-22	52
22-24	58
24-26	55
26-28	54
28-30	45
30-32	48
32-34	35
34-36	30
36-38	89
38-40	51
40-42	18
42-44	19
44-46	19
46-48	12
48-50	16
50-52	7
52-54	8
54-56	3
56-58	1
58-60	0
60-62	1

Status  
OK

Log x 0

Clicking on this will bring up a list of all of the filters, organized into a hierarchy. Click on each folder to expand the list. There are dozens of choices

# Data Transformation

- ❑ Data transformation techniques
  - ❑ Data discretization
  - ❑ Data normalization
  - ❑ Data standardization
  - ❑ Convert nominal attributes to dummy variables

# Data Transformation

## ❑ Data discretization

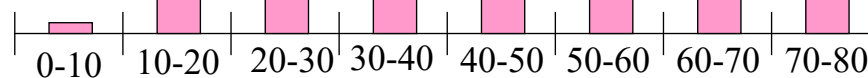
- ❑ This technique converts continuous data into discrete data by assigning each **observation to a specific category or class**. This can be useful in cases where the relationship between the variables is non-linear or when using algorithms that require categorical inputs, such as decision trees and Naive Bayes classifiers.
- ❑ Common methods:
  - ❑ **Equi-width binning**: Divide the data range into equal intervals
  - ❑ **Equi-frequency binning**: Divide the data into intervals with equal numbers of observations

# Data Discretization

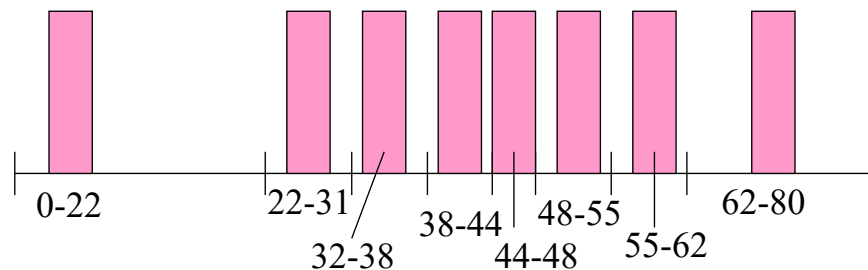
Example: customer ages

number  
of values

Equi-width binning



Equi-frequency binning





Weka Explorer

Preprocess    Classify    Cluster    Associate    Select attributes    Visualize

Open file...    Open URL...    Open DB...    Generate...    Undo    Edit...    Save...

Filter

weka

- filters
  - AllFilter
  - MultiFilter
  - RenameRelation
  - supervised
  - unsupervised
    - attribute
      - Add
      - AddCluster
      - AddExpression
      - AddID
      - AddNoise
      - AddUserFields
      - AddValues
      - CartesianProduct
      - Center
      - ChangeDateFormat
      - ClassAssigner
      - ClusterMembership
      - Copy
      - DateToNumeric
      - Discretize**
      - FirstOrder
      - FixedDictionaryStringToWordVector
      - InterquartileRange

Filter...    Remove filter    Close

Attributes: 20  
Sum of weights: 2000

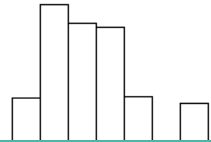
Pattern

Selected attribute

Name: age  
Missing: 0 (0%)  
Distinct: 71  
Type: Numeric  
Unique: 6 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.28
StdDev	0.153

Class: nr.employed (Num)    Visualize All



19 | eunbor3m

Remove

Status  
OK

Log    x 0

Clicking choose, under filters->unsupervised->attribute, select Discretize

# Data Transformation

## ❑ Data normalization

- ❑ **This technique scales the data to a specific range, usually between 0 and 1.**

This helps in reducing the impact of outliers and allows for a better comparison between different variables.

- ❑ min-max normalization (usually new\_max=1, new\_min=0)

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

- ❑ e.g. convert age=30 to range 0-1, when min=10, max=80.  $\text{new\_age} = (30 - 10) * (1 - 0) / (80 - 10) + 0 = 2/7$

Weka Explorer

Preprocess   Classify   Cluster   Associate   Select attributes   Visualize

Open file...   Open URL...   Open DB...   Generate...   Undo   Edit...   Save...

Filter

- ChangeDateFormat
- ClassAssigner
- ClusterMembership
- Copy
- DateToNumeric
- Discretize
- FirstOrder
- FixedDictionaryStringToWordVector
- InterquartileRange
- KernelFilter
- MakeIndicator
- MathExpression
- MergeInfrequentNominalValues
- MergeManyValues
- MergeTwoValues
- NominalToBinary
- NominalToString
- Normalize**
- NumericCleaner
- NumericToBinary
- NumericToDate
- NumericToNominal
- NumericTransform
- Obfuscate
- OrdinalToNumeric
- PartitionedMultiFilter
- PKIDiscretize

Attributes: 20  
Sum of weights: 2000

Pattern

Selected attribute

Name: age  
Missing: 0 (0%)  
Distinct: 71  
Type: Numeric  
Unique: 6 (0%)

Statistic	Value
Minimum	18
Maximum	98
Mean	40.37
StdDev	12.226

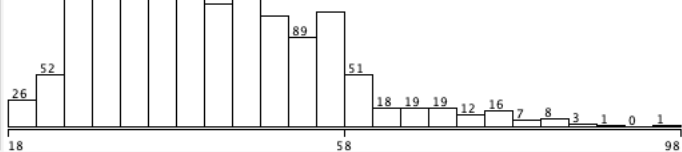
Class: nr.employed (Num)   Visualize All

19   eurbor3m   Remove

Status  
OK

Log   x 0

Clicking choose, under filters->unsupervised->attribute, select Normalize



# Data Transformation

## □ Data standardization

- This technique transforms the data to have a mean of 0 and a standard deviation of 1. It helps in comparing features with different units or scales and is particularly useful for algorithms that are sensitive to the scale of input features

$$v' = \frac{v - \text{mean}_A}{\text{stand\_dev}_A}$$

- e.g. convert age=30, when mean=20, stand\_dev=10. new\_age=(30-20) / 10 = 1

Weka Explorer

Preprocess   Classify   Cluster   Associate   Select attributes   Visualize

Open file...   Open URL...   Open DB...   Generate...   Undo   Edit...   Save...

Filter

Obfuscate  
OrdinalToNumeric  
PartitionedMultiFilter  
PKIDiscretize  
PrincipalComponents  
RandomProjection  
RandomSubset  
Remove  
RemoveByName  
RemoveType  
RemoveUseless  
RenameAttribute  
RenameNominalValues  
Reorder  
ReplaceMissingValues  
ReplaceMissingWithUserConstant  
ReplaceWithMissingValue  
SortLabels  
**Standardize**  
StringToNominal  
StringToWordVector  
SwapValues  
TimeSeriesDelta  
TimeSeriesTranslate  
Transpose

> instance

Filter...   Remove filter   Close

Attributes: 21  
Sum of weights: 4119

Pattern

Selected attribute

Name: age  
Missing: 0 (0%)  
Distinct: 67  
Type: Numeric  
Unique: 3 (0%)

Statistic	Value
Minimum	18
Maximum	88
Mean	40.114
StdDev	10.313

Class: y (Nom)   Visualize All

19   eunbor3m   Remove

Status  
OK

Log   x 0

Clicking choose, under filters->unsupervised->attribute, select Standardize

# Data Transformation

## ❑ Convert Nominal Attributes to Dummy Variables

- ❑ Some machine learning algorithms prefer to use real valued inputs or do not support nominal or ordinal attributes.
- ❑ Nominal attributes can be converted to real values. **This is done by creating one new binary attribute for each category.** For a given instance that has a category for that value, the binary attribute is set to 1 and the binary attributes for the other categories is set to 0. **This process is called creating dummy variables.**

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Attributes: 21  
Sum of weights: 4119

Selected attribute  
Name: age  
Missing: 0 (0%)  
Distinct: 67  
Type: Numeric  
Unique: 3 (0%)

Statistic	Value
Minimum	18
Maximum	88
Mean	40.114
StdDev	10.313

Class: y (Nom) Visualize All

Clicking choose, under filters->unsupervised->attribute, select NominalToBinary

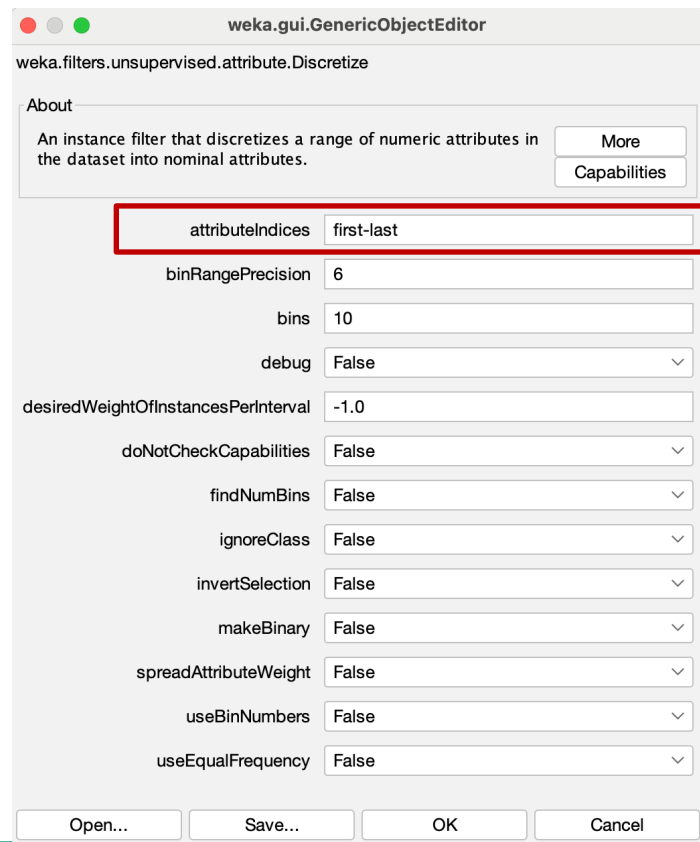
19 eunbor3m

Status OK

Log x 0

# How to apply filters on particular variables?

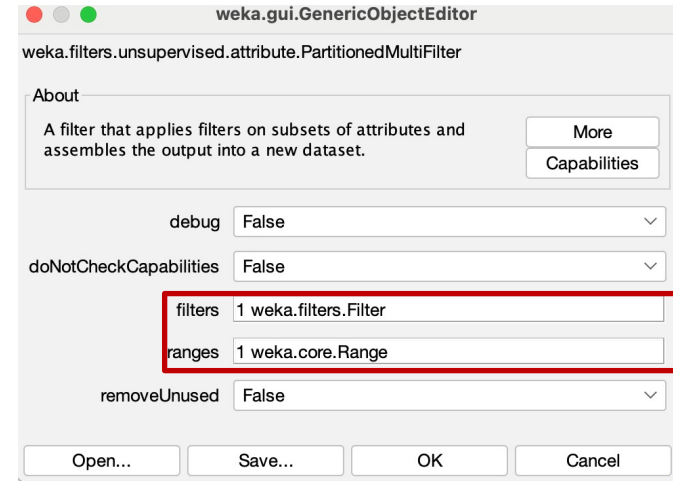
- ❑ Some filters in Weka such as Discretize naturally support this
  - ❑ Use the parameter “attributeIndices” to indicate the desired attributes
  - ❑ By default, this parameter is set to “first-last”, which means that applying the filter on all variables






# How to apply filters on particular variables?

- ❑ Some filters in Weka such as Normalize do not support this
  - ❑ We can use **unsupervised.attribute.PartitionedMultiFilter** to apply a filter on selected variables!
  - ❑ Use the “filters” parameter to indicate the filters used and the “ranges” parameter to indicate the selected variables



# Data Reduction

## Problem:

-  Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

## Solution:

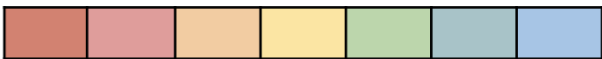
-  Data reduction ...

# Data Reduction: Feature Selection

## ❑ Feature selection (i.e., attribute subset selection): :

- ❑ Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
- ❑ Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)

All Features



Feature Selection



Final Features



# Feature Selection in Weka

- ❑ Feature Selection is divided into two parts
  - ❑ **Attribute Evaluator:** The attribute evaluator is the evaluation method for evaluating each attribute in the dataset based on the class
  - ❑ **Search Method:** The search method is the method for trying different combinations of attributes in the dataset in order to arrive on a short list of chosen variables.
- ❑ Some Attribute Evaluator techniques require the use of specific Search Methods.
  - ❑ For example, the CorrelationAttributeEval can only be used with a Ranker Search Method, it evaluates each attribute and rank the results.

Weka Explorer

Preprocess   Classify   Cluster   Associate   **Select attributes**   Visualize

Attribute Evaluator  
Choose **InfoGainAttributeEval**

Search Method  
Choose **Ranker -T -1.7976931348623157E308 -N -1**

Attribute Selection Mode  
☒ Use full training set  
☐ Cross-validation   Folds: 10   Seed: 1

No class

Start   Stop

Result (right-click for options)  
22:43:21 - Ranker + InfoGainAttributeEval

Attribute selection output

=== Attribute Selection on all input data ===

Search Method:  
Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 21 y):  
Information Gain Ranking Filter

Ranked attributes:

0.1096656	11	duration
0.0951712	18	cons.conf.idx
0.0874463	19	euribor3m
0.0873909	20	nr.employed
0.0860484	17	cons.price.idx
0.0736447	16	emp.var.rate
0.0469643	15	poutcome
0.0461121	13	pdays
0.0355763	9	month
0.031243	14	previous
0.0151493	8	contact
0.011311	2	job
0.0102173	1	age
0.0058546	12	campaign
0.0048049	5	default
0.0040156	4	education
0.0017391	3	marital
0.0002066	7	loan
0.0001175	6	housing
0.0000897	10	day_of_week

Selected attributes: 11,18,19,20,17,16,15,13,9,14,8,2,1,12,5,4,3,7,6,10 : 20

Status  
OK

Log   x 0

After choosing the attribute evaluator and search method, clicking Start to conduct feature selection

We can retain top K (e.g. K=10) attributes after feature selection

# Data Splitting

- ❑ For developing models, we need to split the dataset into training and testing sets
  - ❑ For example, a training set comprising 80% of the original dataset and a test set consisting of the remaining 20%
- ❑ Data splitting using Weka
  - ❑ Use the RemovePercentage filter (An instance filter)

# Data Splitting

## ❑ Training set

- ❑ Load the full dataset
- ❑ select the RemovePercentage filter in the preprocess panel
- ❑ set the correct percentage for the split
- ❑ apply the filter
- ❑ save the generated data as a new file

## ❑ Test set

- ❑ Load the full dataset (or just use undo to revert the changes to the dataset)
- ❑ select the RemovePercentage filter if not yet selected
- ❑ set the invertSelection property to true
- ❑ apply the filter
- ❑ save the generated data as a new file

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

- weka
  - filters
    - AllFilter
    - MultiFilter
    - RenameRelation
    - supervised
    - unsupervised
      - attribute
      - instance
        - NonSparseToSparse
        - Randomize
        - RemoveDuplicates
        - RemoveFolds
        - RemoveFrequentValues
        - RemoveMisclassified
        - RemovePercentage**
        - RemoveRange
        - RemoveWithValues
        - Resample
        - ReservoirSample
        - SparseToNonSparse
        - SubsetByExpression

Attributes: 21  
Sum of weights: 4119

Pattern

Selected attribute

Name: age	Type: Numeric
Missing: 0 (0%)	Distinct: 67
Unique: 3 (0%)	

Statistic	Value
Minimum	18
Maximum	88
Mean	40.114
StdDev	10.313

Class: y (Nom) Visualize All

19 eurbor3m

Remove

Status OK

Log x 0

Clicking choose, under filters->unsupervised->instance, select RemovePercentage



Q&A