# ECLT5810/SEEM5750

Lab Class 1: Weka - installation, feature engineering, decision trees, explain Assignment 1

Shuaiyi Li sli@se.cuhk.edu.hk

#### Arrangement

Weka Demo (25 min)

Practice & QA & Break (20 min)

Weka Demo & Assignment 1 (25 min)

Practice & QA (20 min)

#### 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: <u>https://www.cs.waikato.ac.nz/ml/weka/</u>

#### 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.

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",
- 0

"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.



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

🖉 Weka Explorer				- 🗆 X
Preprocess Classify Cluster Associate	Select attributes Visualize			
Open file Open URL	Open DB Gene	erate Un	ndo Edit	Save
ilter				
Choose None				Apply Stop
urrent relation		Selected attribute		
Relation: None Instances: None	Attributes: None Sum of weights: None	Name: None Missing: None	Weight: None Distinct: None	Type: None Unique: None
ttributes				
All None	invert			Visualize All
Remove	•			
tatus				
Welcome to the Weka Explorer				Log X0

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

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



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

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



The statistics for each variable are also shown here.

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



## Feature Engineering

Feature engineering 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.

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.

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.

In this tutorial, we will use the Unsupervised Filters.

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 instances or one instance at a time.

In this tutorial, we will mostly deal with the Attribute Filters.

#### **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.

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

select Discretize

Preprocess Classify Cluster Associate Select attribute	es Visualize		
Open file Open URL Open	DB Ger	undo	Edit Save
filter			
🔻 📇 weka	1		Apply Stop
▼ 🚔 filters		Selected attribute	
AllFilter	10.2		
BenameBelation	Attributes: 17 um of weights: 4521	Missing: 0 (0%) Distinct:	67 Unique: 4 (0%)
► È supervised		Statistic	Value
Insupervised		Minimum	19
T attribute	Detter	Maximum	87
Add	Pattern	Mean	41.17
AddExpression		SIUDEV	10.570
AddID			
AddNoise			
AddUserFields			
AddValues			
CartesianProduct		Class: y (Nom)	<ul> <li>Visualize</li> </ul>
ChangeDateFormat			
ClassAssigner		_	
ClusterMembership		I <b>I</b>	
Copy			
EirstOrder			_
FixedDictionaryStringToWordVector			
InterquartileRange			
		20 68	70
Filter Remove filter Close		14	16 16 7 19 10 9 8 13 1 5
<u>Filter</u> <u>R</u> emove filter <u>C</u> lose			

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

Preprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB G	enerate Undo	Edit Save
er		
Choose Discretize -B 10 -M -1.0 -R first-last-precision 6		Apply Stop
rrent relation	Selected attribute	
Relation: bank Attributes: 17 Instances: 4521 Sum of weights: 4521	Name: age Missing: 0 (0%) Di:	Type: Numeric stinct: 67 Unique: 4 (0%)
ributes	Statistic	Value
	Minimum	19
	Maximum	87
All None Invert Pattern	Mean	41.17
Na Mana	StaDev	10.576
No. Name		
2 Job		
4 education		
5 default		
6 🔲 balance		
7 📃 housing	Class: y (Nom)	Visualize /
8 loan		
11 month		
12 duration		
13 Campaign		
14 📃 pdays		
15 📃 previous		
16 poutcome		
1/ 🛄 y		
	68	70
Remove	14 29	15 15 7 19 10 0 0 13 .
Itus	19	53

Here is the configuration of Discretize

attributes Indices means specify range of attributes to act on

binRangePrecison means the number of decimal places for cut points to use

bins means numbers of bin

About		_
An instance filter that discretizes a r the dataset into nominal attributes.	ange of numeric attributes in More Capabilities	
attributeIndices	first-last	
binRangePrecision	6	
bins	10	
debug	False	•
esiredWeightOfInstancesPerInterval	-1.0	
doNotCheckCapabilities	False	▼
findNumBins	False	▼
ignoreClass	False	▼
invertSelection	False	•
makeBinary	False	▼
spreadAttributeWeight	False	•
useBinNumbers	False	•
useEqualFrequency	False	

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

Weka Explorer			×
Preprocess Classify Cluster Associ	🥥 weka.gui.GenericObjectEditor	×	
Open file Open URL	weka.fliters.unsupervised.attribute.Disc About	retize	Edit Save
Filter			
Choose Discretize - B 10 - M - 1.0 - R firs	An instance filter that discretizes a r the dataset into nominal attributes.	ange of numeric attributes in More Capabilities	Apply Stop
Current relation			
Relation: bank Instances: 4521	attributeIndices	first-last	Type: Numeric Unique: 4 (0%)
Attributes	binRangePrecision	6	Value
	The number	of decimal places for cut points to use when gener	ting big lobele
All None	The number ons	to decimal places for cut points to use when genera	41.17
			41.1/ 10.576
No. Name	debug	False	
1 🗖 age			
2 📃 job	desiredWeightOfInstancesPerInterval	-1.0	
3 marital			
4 default	doNotCheckCapabilities	False	
6 balance	for this way Direct	Felse	
7 housing	TindivumBins	Faise	Visualize Al
8 🔲 Ioan	in a second second	False	
9 📃 contact	IgnoreClass	Faise	
10 🔄 day	invertCale diam	Falsa	
11 month	Invenselection	raise	
12 duration	makeDinery	Falsa	
14 ndays	makebinary	raise	
15 previous	cproadattribute\Moint	False	
16 poutcome	spreau-unouteweight	1 4150	
17 🗌 y	useBinNumbers	False	
	do com da morta		20
Ren	useEqualFrequency	False	16 16 7 19 10 9 8 13
Status	Onen Save	OK Cancel	
ОК			Log 🛷 X

Let's leave the setting as default and click OK

Then, click apply.

🖓 Weka Explorer	- D X
Preprocess Classify Cluster Associate Select attributes Visualize	
Open file Open URL Open DB G	Senerate Undo Edit Save
iter	
Choose Discretize -B 10 -M -1.0 -R first-last -precision 6	Apply Stop
urrent relation	Selected attribute
Relation: bank     Attributes: 17       Instances: 4521     Sum of weights: 4521	Name: age         Type: Numeric           Missing: 0 (0%)         Distinct: 67         Unique: 4 (0%)
ttributes	Statistic Value
All None Invert Pattern	Minimum         19           Maximum         87           Mean         41.17           StdDay         10.576
No. Name	
1         age           2         job           3         marital           4         education           5         default           6         balance	
7 housing	Class: y (Nom) Visualize All
8 loan 9 contact 10 day 11 month 12 duration 13 campaign	
14 pdays 15 previous 16 poutcome 17 y	a series and a series of the s
Remove	
itatus	19 53 87
ОК	Log 🛷 XC

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

Weka Explorer Preprocess Classify Cluster Associate Select attributes Visualize			- D >
Open file Open URL Open DB Gener	rate Undo	Edit.	. Save
Choose Discretize -B 10 -M -1.0 -R first-last-precision 6			Apply Stop
Current relation	Selected attribute		~
Relation: bank-weka.filters.unsupervised.attribute.Dis Attributes: 17 Instances: 4521 Sum of weights: 4521	Name: age Missing: 0 (0%)	Distinct: 10	Type: Nominal Unique: 0 (0%)
Attributes	No. Label	Count	Weight
	1 '(-inf-25.8]'	111	111.0
	2 '(25.8-32.6]'	944	944.0
All None Invert Pattern	3 '(32.6-39.4]'	1235	1235.0
	4 '(39.4-46.2]'	869	869.0
No. Name	5 '(46.2-53]'	706	706.0
1 age	6 '(53-59.8]'	482	482.0
2 job	7 '(59.8-66.6]'	100	100.0
3 marital	8 '(66.6-73.4]'	36	36.0
4 ducation	9 '(73.4-80.2]'	30	30.0
5 🗌 default	10 1/00 0 infl	0	
6 balance	Class; (Mam)		Vieweller
7 housing	Class. y (400m)		visualize
8 📃 Ioan			
9 📃 contact	1225		
10 🔲 day	1255		
11 month			
12 duration	944		
13 campaign	86	9	
14 poays		706	
15 previous			
		492	
· / 🗆 y		702	
Pamava			
Remove	111		100
) Natur			36 30 8
itatus			
ОК			Log

#### **Convert Nominal Attributes to Dummy Variables**

Some machine learning algorithms prefer to use real valued inputs and 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.

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

Preprocess Classify Cluster Associate Select attributes Visualize Open file.. Open URL Open DB. Generate. Edit. Save... Filter maanvoide Apply Stop 14 AddUserFields AddValues Selected attribute CartesianProduct Attributes: 17 Name: age Type: Numeric Center um of weights: 4521 Missing: 0 (0%) Distinct: 67 Unique: 4 (0%) ChangeDateFormat ClassAssigner Statistic Value ClusterMembership Minimum 19 Maximum 87 Copy Pattern 41.17 Mean DateToNumeric 10,576 StdDev Discretize FirstOrder FixedDictionaryStringToWordVector InterguartileRange KernelFilter MakeIndicator MathExpression Visualize All Class: y (Nom) MergeInfrequentNominalValues MergeManyValues Merce Two Val NominalToBinary NominalTeStrin Normalize NumericCleaner NumericToBinary NumericToDate NumericToNominal Remove filter Filter.... Close 16 16 7 19 10 9 8 13 1 5 2 53 Status Log OK

\_

 $\times$ 

In the configuration, set the attributeIndices to 2 as this time we only want to transform the job attirbute, then Click OK

**Click Apply** 

🕝 weka.gui.GenericObje	ctEditor	×
weka.filters.unsupervised.a	attribute.NominalToBinary	
About		
Converts all nominal a	ttributes into binary numeric attributes.	More Capabilities
		]
attributeIndices	2	
binaryAttributesNominal	False	T
debug	False	•
doNotCheckCapabilities	False	•
invertSelection	False	•
spreadAttributeWeight	False	•
transformAllValues	False	•
Open	Save OK	Cancel

Now, you can see the job attribute is split into 12 new attributes where each of it denotes an instance belongs to that occupation or not

For example, job=unemployed, 0 denotes not unemployed, 1 denotes unempolyed. The count of 0 is 4393 and the count of 1 is 128



#### Gradient Descent with and without feature scaling



The MSE cost function for a Linear Regression model has the shape of a bowl, but it can be an elongated bowl if the features have very different scales.

The figure above shows Gradient Descent on a training set where features 1 and 2 have the same scale (on the left), and on a training set where feature 1 has much smaller values than feature 2 (on the right).

**Standardization** 

Standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1.

Standardization assumes that your data has a Normal distribution. This does not strictly have to be true, but the technique is more effective if your attribute distribution is Normal.

The formula is given below.

$$z=rac{x_i-\mu}{\sigma}$$

Click Choose, under

filters->unsupervised->attributes,

select Standardize

Teprocess Classify Cluster Associate Select ato	ibutes Visualize			
Open file Open URL	)pen DB Ger	erate Und	o Edit	Save
er				
	121			Annly Ston
Obfuscate	Ê			Comb
Ordinal I oNumeric		Selected attribute		
PartitionedMultiFilter				
PRIDISCIENZE	Attributes: 17	Name: balance	Distinct 0252	Type: Numeric
PrincipalComponents     PandamProjection	um or weights: 4521	Missing. 0 (0%)	Distinct 2355	Unique: 1495 (55%)
RandomProjection		Statistic	Value	
- Randomsdoset		Minimum	-3313	
Remove	Pattern	Maximum	71188	
RemoveByName	- data	StdDev	1422.058	
Removel isolass	10	0.0000	0000.000	
Penameáttribute				
RenameNominalValues				
Reorder				
ReplaceWissingValues				
ReplaceMissingVithUserConstant				
ReplaceWithMissinoValue		Class: y (Nom)		<ul> <li>Visualize</li> </ul>
Soll abels				
Standardize				
Shine ToNominal				
StringToWordVector				
SwapValues				
TimeSeriesDelta				
TimeSeriesTranslate				
Transpose				
instance	T			
—				
	e			
Eilter Remove filter				
<u>Filter</u> <u>R</u> emove filter <u>C</u> los		-3313	33937.5	711:

#### No setting is required in the configuration, simply click OK

Then, click Apply

🥥 weka.gui.GenericObje	ctEditor	>
weka.filters.unsupervised.a	attribute.Standardize	
About		
Standardizes all nume zero mean and unit val set).	ric attributes in the given dataset to have riance (apart from the class attribute, if	More Capabilities
(		
debug	False	<b></b>
doNotCheckCapabilities	False	<b></b>
ignoreClass	False	<b></b>
Open	Save OK	Cancel

Now, you can see the age attribute is rescaled. The mean is 0 and standard deviation is 1.

Weka Explorer     Preprocess Classify Cluster Associate Select attributes Visualize	- 🗆 X
Open file Open URL Open DB Generate Undo	Edit Save
Choose Standardize	Apply Stop
Current relation Selected attribute	
Relation: bank-weka.filters.unsupervised.attribute.No       Attributes: 17       Name: age         Instances: 4521       Sum of weights: 4521       Missing: 0 (0%)       Distinct	Type: Numeric 67 Unique, 4 (0%)
Attributes	Value
All None Invert Pattern Minimum Mean StdDev	-2.096 4.333 -0 1
No. Name	
1 age 2 job 3 marital 4 education 5 default	
7 housing Class: y (Nom)	Visualize All
8 loan 9 contact 10 day 11 month 12 duration	
12 Gualdon 13 Grampaign 14 Dedays 15 Derevious	
16 poutcome 17 y	
Remove	70 <u>15 15 7 <sup>19</sup> 10 9 8 13</u> 1 <u>5 2</u>
Status	4.33
ОК	Log 🛷 x0

#### Normalization

Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Normal distribution. Also, some machine learning algorithms are sensitive so the scale of the data, rescale the data into a range of 0 to 1 can lower the effect of scale.

This is also called min-max normalization, the formula is given below.

$$x_{scaled} = rac{x-x_{min}}{x_{max}-x_{min}}$$

Click Choose, under

filters->unsupervised->attributes,

select Normalize

	outes Visualize			
Open file Open URL Op	pen DB Gene	rate Und	o Edit.	
er				
ClusterMembership				Apply Stop
Сору		Colorial attribute		
DateToNumeric		Selected attribute		
Discretize	Attributes: 28	Name: balance		Type: Numeric
FirstOrder	um of weights: 4521	Missing: 0 (0%)	Distinct: 2353	Unique: 1493 (33%)
FixedDictionaryStringToWordVector		Statistic	Value	
lnterquartileRange		Minimum	-3313	
🕒 KernelFilter		Maximum	71188	
MakeIndicator	Pattern	Mean	1422.6	58
MathExpression		StdDev	3009.6	38
MergeInfrequentNominalValues				
MergeManyValues	A			
MergeTwoValues				
NominalToBinary				
NominalToString		C		
Normalize		Class: v (Nom)		Visualize
Normalize		Class: y (Nom)		<ul> <li>Visualize</li> </ul>
Normalize		Class: y (Nom)	eksaan olitikuta villalas k	Visualize
Normalize NumericCleaner NumericToBinary NumericToDate		Class: y (Nom)	chosen attribute will also b	Visualize     visualize     visualize     visualize
Normalize NumericCleaner NumericToBinary NumericToDate NumericToDate		Class: y (Nom)	chosen attribute will also b	Visualize     Visualize     visualize     visualize
Normalize NumericOleaner NumericToBinary NumericToDate NumericToNominal NumericTransform		Class: y (Nom)	chosen attribute will also b	Visualize     Visualize     e used as the class attribute who
Normalize Normalize NumericCleaner NumericToBinary NumericToDate NumericToNominal NumericTransform Optiscate Optiscate Optiscate		Class: y (Nom)	chosen attribute will also b	Visualize     Visualize     e used as the class attribute who
Normalize NumericOleaner NumericToBinary NumericToDate NumericToNominal NumericTransform Obfuscate OrdinalToNumeric RadiationedMultificities		Class: y (Nom)	chosen attribute will also b	Visualize     e used as the class attribute who
Normalize NumericOleaner NumericToBinary NumericToDate NumericToNominal NumericTransform Obfuscate OrdinalToNumeric PartitionedMultiFilter RK/Discretze		Class: y (Nom)	chosen attribute will also b	Visualize     e used as the class attribute who
Normalize NumericOleaner NumericToBinary NumericToDate NumericToNominal NumericTransform Obfuscate OrdinalToNumeric PartitionedMultiFilter PKIDIscretize RidicialComponents		Class: y (Nom)	chosen attribute will also b	Visualize the used as the class attribute who
Normalize NumericOleaner NumericToBinary NumericToDate NumericToNominal NumericTransform Obfuscate OrdinalToNumeric PartitionedMultiFilter PKIDIscretize PrincipalComponents		Class: y (Nom)	chosen attribute will also b	Visualize we used as the class attribute who
Normalize  NumericToBinary  NumericToDate  NumericToNominal  NumericTransform  Obfuscate  OrdinalToNumeric  PartitionedMultiFilter  PKIDiscretize  FincipalComponents <u>Filter</u> <u>Remove filter</u> Close		Class: y (Nom)	chosen attribute will also b	▼ Visualize

In the configuration, the scale and translation mean the maximum and minimum value after normalization where the default value is 1 and 0 respectively. We leave it as default. Click OK

Then, click Apply

🧔 weka.gui.GenericObje	ctEditor	$\times$
weka.filters.unsupervised.a	ttribute.Normalize	
About		_
Normalizes all numerion the class attribute, if se	c values in the given dataset (apart from More et). Capabilities	
debug	False	
doNotCheckCapabilities	False	•
ignoreClass	False	•
scale	1.0	
translation	0.0	
Open	Save OK Cancel	

Now, you can see the balance attribute is rescale into the range of 0 to 1

Open file Open URL Open DB G	Generate Undo Edit Save
ilter	
Choose Normalize S10 T00	Apply
Current relation	Selected attribute
Relation: bank-weka.filters.unsupervised.attribute.No Attributes: 28 Instances: 4521 Sum of weights: 4521	Name:         balance         Type:         Numeric           1         Missing:         0 (0%)         Distinct:         2353         Unique:         1493 (33%)
Attributes	Statistic Value
	Minimum 0
All None Invert Dettern	Maximum 1
	Mean 0.064
No Name	SidDev 0.04
12 job=retired	
13 job=unknown	
15 education	
16 default	
17 📃 balance	Class: v (Nom) Visualize
18 📃 housing	
19 loan	
20 contact	
21 Oay	
22 duration	
24 campaign	
25 ndavs	
27 DOUTCOME	÷
27 poutcome	
27 poutcome 28 y	
27 D poutcome 28 y Remove	
27 Doutcome 28 y Remove	
27 Doutcome 28 y Remove	0.5
27 D poutcome 28 y Remove	0.8

This normalization does not affect the distribution of the data. For example, look at the graph of balance attribution, the shape is the same after performing the normalization.



## What is Decision Tree?

A decision tree is decision making tool using a tree-like graph or model of decisions and their possible consequences such as event outcomes, resource costs, and utility.

All the conditional control statements used in the decision tree can be displayed for easily understand the logic behind it.

## What is Decision Tree?

A decision tree is a flowchart-like structure contains these components:

- each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails)
- each branch represents the outcome of the test
- each leaf node represents a class label (decision taken after computing all attributes).

The paths from root to leaf represent classification rules.

#### What is Decision Tree?

This is an example of a decision tree for the target variable response. This variable has two labels: 1 for response and 0 for no response.

Each node determine which attribute should be used for splitting the dataset based on the information gain. In this example, Node 1 uses Income as splitting attribute, <\$25k go to Node 2 and >= \$25k go to Node 3.

There are 4 leaf nodes (Node 4-7) for determine the predicted label.

