

Session 1

WEKA, datasets and ARFF files

The first session is about knowing your data: what WEKA is, how its Explorer is organised, what an ARFF file looks like, and how to describe a dataset before mining it.

Objectives

✕ Do not copy. Read for understanding and the viva

- Complete questions 1 to 4 of the manual: weka, datasets and arff files
- Prepare the deliverable before the lab and finish it during the session
- Be ready to explain every step in the viva

Questions Covered

✕ Do not copy. Read for understanding and the viva

Question	Requirement	Status
Q1	Download and install WEKA. Navigate the various options available in WEKA. Explore the...	Complete
Q2	Create your own EXCEL file. Convert the EXCEL file to .csv format and prepare it as...	Complete
Q3	Try to create your own datasets	Complete
Q4	Preprocess and classify Customer, Agriculture, Weather, Whole-sale Customers or the...	Complete

Preparation

✕ Do not copy. Read for understanding and the viva

- ARFF has three parts: `@relation`, one `@attribute` line per column with its type (numeric, nominal list, string, date), and `@data` rows. Write a five-row student.arff by hand.

- In Explorer's Preprocess tab, the right-hand panel gives count, distinct values, mean and standard deviation per attribute; the class attribute is chosen in the dropdown above the histogram.
- Save your Excel sheet as CSV, open it in WEKA with the CSV loader, then save as ARFF and inspect the generated header.

Question 1

Problem Statement

■ Write in lab record

Download and install WEKA. Navigate the various options available in WEKA. Explore the available datasets in WEKA. Load various datasets and observe the following:

1. List the attribute names and their types
2. No. of records in each dataset
3. Identify the class attribute (if any)
4. Plot Histogram
5. Determine the no. of records for each class.
6. Visualize the data in different dimensions.

Solution

■ Write in lab record

Steps

1. Download the stable release from the [WEKA download page](#). On Windows take the installer that bundles the Azul Zulu JDK (`weka-3-8-6-azul-zulu-windows.exe` , about 125 MB) so no separate Java install is needed. On Linux unzip `weka-3-8-6-azul-zulu-linux.zip` and run `./weka.sh` inside the `weka-3-8-6` folder.
2. Run the installer: Next, I Agree on the GPL licence, keep Full installation, keep the destination `C:\Program Files\Weka-3-8-6` , Install, Finish. The WEKA GUI Chooser opens.
3. Tour the GUI Chooser. It has five buttons and the sample data lives in the `data` sub-folder of the install directory.

Button	What it is for
Explorer	Load one dataset, preprocess it, then run Classify, Cluster, Associate, Select attributes and Visualize tabs on it
Experimenter	Run several algorithms on several datasets with repeated cross-validation and compare them with a statistical test
KnowledgeFlow	Drag-and-drop flow of Datasources, Filters, Classifiers, Evaluation and Visualization components
Workbench	All of the above in one window with tabs
Simple CLI	A shell where <code>help</code> lists commands and any class can be run, for example <code>java weka.associations.Apriori -t data/contact-lenses.arff</code>

- Click Explorer. In the Preprocess tab click Open file..., go to `C:\Program Files\Weka-3-8-6\data` and open `weather.nominal.arff`.
- Read the panel. Current relation shows Relation, Instances and Attributes. The Attributes list on the left gives every attribute name with a checkbox. Click a name: the Selected attribute box on the right shows Name, Type, Missing, Distinct and Unique, then a table of labels and counts for a nominal attribute or Minimum, Maximum, Mean and StdDev for a numeric one.
- The class attribute is the one selected in the dropdown above the histogram (WEKA defaults to the last attribute). The counts of that attribute are the number of records per class.
- Histogram: the bar chart under the dropdown is the histogram of the selected attribute, with the bars split by class colour. Click Visualize All to see every attribute at once.
- Click the Visualize tab for the scatter-plot matrix. Click any cell to open it full size, change the X and Y dropdowns to any pair of attributes, move the Jitter slider to separate overlapping nominal points, and use Select Instance to read a point.
- Repeat Open file for the other datasets and fill the table below. Attribute types are read from the Type field, counts from the label table.

Output

What the Preprocess panel prints for `weather.nominal.arff` after clicking `outlook`:

1	Current relation	Selected attribute			
2	Relation: weather.symbolic	Name: outlook		Type: Nominal	
3	Instances: 14 Attributes: 5	Missing: 0 (0%)		Distinct: 3	Unique: 0 (0%)
4	Attributes	No.	Label	Count	Weight
5	1 outlook	1	sunny	5	5
6	2 temperature	2	overcast	4	4
7	3 humidity	3	rainy	5	5
8	4 windy				
9	5 play	Class: play (Nom)		[Visualize All]	

The built-in datasets, read the same way:

Dataset	Instances	Attributes	Attribute types	Class attribute	Records per class
weather.nominal	14	5	outlook, temperature, humidity, windy, play: all nominal	play	yes 9, no 5
weather.numeric	14	5	outlook, windy, play nominal; temperature, humidity numeric	play	yes 9, no 5
iris	150	5	sepalength, sepalwidth, petallength, petalwidth numeric; class nominal	class	Iris-setosa 50, Iris-versicolor 50, Iris-virginica 50
glass	214	10	RI, Na, Mg, Al, Si, K, Ca, Ba, Fe numeric; Type nominal (7 labels)	Type	build wind float 70, build wind non-float 76, vehic wind float 17, vehic wind non-float 0, containers 13, tableware 9, headlamps 29
contact-lenses	24	5	age, spectacle-prescrip, astigmatism, tear-prod-rate, contact-lenses: all nominal	contact-lenses	soft 5, hard 4, none 15
labor	57	17	8 numeric (duration, the three wage-increase attributes, working-hours, standby-pay, shift-differential, statutory-holidays); 9 nominal including class; many missing values	class	bad 20, good 37
zoo	101	18	animal nominal (one label per animal); 15 boolean nominal (hair ...	type	mammal 41, bird 20, reptile 5, fish 13, amphibian 4,

Dataset	Instances	Attributes	Attribute types	Class attribute	Records per class
			catsize); legs numeric; type nominal		insect 8, invertebrate 10
diabetes	768	9	preg, plas, pres, skin, insu, mass, pedi, age numeric; class nominal	class	tested_negative 500, tested_positive 268
credit-g	1000	21	7 numeric (duration, credit_amount, installment_commitment, residence_since, age, existing_credits, num_dependents); 14 nominal including class	class	good 700, bad 300
soybean	683	36	all 35 input attributes nominal; class nominal with 19 labels	class	brown-spot 92, alternarialeaf- spot 91, frog- eye-leaf-spot 91, phytophthora- rot 88, brown- stem-rot 44, anthracnose 44, nine diseases with 20 each, 2-4- d-injury 16, diaporthe-pod- and-stem- blight 15, cyst- nematode 14, herbicide- injury 8

Visualisation notes to record: on iris, the Visualize cell petallength against petalwidth separates the three classes almost perfectly, while sepalwidth against sepallength mixes versicolor and virginica. On glass, most Ba and Fe values are 0, so their histograms are one tall bar. On labor, the grey part of a histogram bar is the missing count.

Explanation

WEKA reads the attribute types from the ARFF header, not from the data, so the Type field is exactly what the `@attribute` line says. The class attribute is a choice, not a property of the file: WEKA guesses the last attribute and every classifier uses whatever the dropdown says. The label table under the histogram is the class distribution when the class attribute is selected, which is why the records-per-class column above comes straight from it. The Visualize tab plots two attributes at a time; picking the pair that separates the class colours best is the first hint of which attributes a classifier will use.

Question 2

Problem Statement

■ Write in lab record

Create your own EXCEL file. Convert the EXCEL file to .csv format and prepare it as .arff file.

Solution

■ Write in lab record

Steps

1. In Excel put the attribute names in row 1 (`sid` , `age` , `gender` , `stream` , `attendance` , `internal` , `sem_marks` , `hours_study` , `result`) and one student per row below it. Leave a cell blank where the value is unknown. Do not merge cells, do not add totals, do not put units in the numbers.
2. File, Save As, choose file type CSV (Comma delimited) (*.csv), name it `student.csv` , click Yes when Excel warns about losing workbook features. The result is the file below.
3. Open the CSV in WEKA: Explorer, Open file..., set Files of type to CSV data files (*.csv), select `student.csv` , Open. WEKA's `CSVLoader` reads row 1 as attribute names, makes a column numeric when every cell parses as a number, and nominal otherwise. A blank cell becomes a missing value.
4. Check the guessed types in the Attributes list. `sid` is numeric here, which is fine for an identifier you will remove later; if a numeric code should be nominal, apply Filter, unsupervised, attribute, `NumericToNominal` on that index.
5. Save as ARFF: click Save..., set Files of type to Arff data files (*.arff), name it `student.arff` . Open it in Notepad to see the header WEKA wrote.
6. Command-line alternative in Simple CLI: `java weka.core.converters.CSVLoader student.csv > student.arff` .

Program

The CSV as Excel writes it (row 9 has an empty `internal` cell, row 18 an empty `attendance` cell):

student.csv

```
1  sid,age,gender,stream,attendance,internal,sem_marks,hours_study,result
2  1,20,M,science,85,24,58,3.5,pass
3  2,21,F,commerce,72,18,42,2.0,pass
4  3,19,M,arts,55,12,28,1.0,fail
5  4,22,F,science,90,27,65,4.0,pass
6  5,20,M,commerce,60,15,35,1.5,fail
7  6,21,F,arts,78,20,47,2.5,pass
8  7,23,M,science,40,10,22,0.5,fail
9  8,20,F,science,88,26,61,3.0,pass
10 9,19,M,commerce,65,,38,2.0,fail
11 10,21,F,arts,82,22,50,2.5,pass
12 11,20,M,science,70,19,44,2.0,pass
13 12,22,F,commerce,50,11,30,1.0,fail
14 13,21,M,arts,93,28,66,4.5,pass
15 14,20,F,science,58,14,33,1.5,fail
16 15,19,M,commerce,75,21,49,2.5,pass
17 16,23,F,arts,45,9,25,0.5,fail
18 17,20,M,science,80,23,55,3.0,pass
19 18,21,F,commerce,,17,41,2.0,pass
20 19,22,M,arts,62,13,31,1.0,fail
21 20,20,F,science,95,29,68,5.0,pass
22 21,21,M,commerce,68,16,39,1.5,fail
23 22,19,F,arts,84,25,57,3.5,pass
24 23,20,M,science,52,12,29,1.0,fail
25 24,22,F,commerce,77,20,46,2.5,pass
26 25,21,M,arts,66,18,43,2.0,pass
27 26,20,F,science,48,10,26,0.5,fail
28 27,19,M,commerce,86,24,59,3.0,pass
29 28,23,F,arts,71,19,45,2.0,pass
30 29,20,M,science,57,13,34,1.5,fail
31 30,21,F,commerce,89,27,63,4.0,pass
```

The ARFF that WEKA produces, tidied by hand with comments and a blank line between sections:

student.arff

```
1 % student.arff - 30 MCA students, written by hand for MCSL-223 Session 1
2 % Class attribute: result (pass/fail). Two cells are missing ('?') on purpose.
3 @relation student
4
5 @attribute sid numeric
6 @attribute age numeric
7 @attribute gender {M, F}
8 @attribute stream {science, commerce, arts}
9 @attribute attendance numeric
10 @attribute internal numeric
11 @attribute sem_marks numeric
12 @attribute hours_study numeric
13 @attribute result {pass, fail}
14
15 @data
16 1,20,M,science,85,24,58,3.5,pass
17 2,21,F,commerce,72,18,42,2.0,pass
18 3,19,M,arts,55,12,28,1.0,fail
19 4,22,F,science,90,27,65,4.0,pass
20 5,20,M,commerce,60,15,35,1.5,fail
21 6,21,F,arts,78,20,47,2.5,pass
22 7,23,M,science,40,10,22,0.5,fail
23 8,20,F,science,88,26,61,3.0,pass
24 9,19,M,commerce,65,?,38,2.0,fail
25 10,21,F,arts,82,22,50,2.5,pass
26 11,20,M,science,70,19,44,2.0,pass
27 12,22,F,commerce,50,11,30,1.0,fail
28 13,21,M,arts,93,28,66,4.5,pass
29 14,20,F,science,58,14,33,1.5,fail
30 15,19,M,commerce,75,21,49,2.5,pass
31 16,23,F,arts,45,9,25,0.5,fail
32 17,20,M,science,80,23,55,3.0,pass
33 18,21,F,commerce,?,17,41,2.0,pass
34 19,22,M,arts,62,13,31,1.0,fail
35 20,20,F,science,95,29,68,5.0,pass
36 21,21,M,commerce,68,16,39,1.5,fail
37 22,19,F,arts,84,25,57,3.5,pass
38 23,20,M,science,52,12,29,1.0,fail
39 24,22,F,commerce,77,20,46,2.5,pass
40 25,21,M,arts,66,18,43,2.0,pass
41 26,20,F,science,48,10,26,0.5,fail
42 27,19,M,commerce,86,24,59,3.0,pass
43 28,23,F,arts,71,19,45,2.0,pass
44 29,20,M,science,57,13,34,1.5,fail
45 30,21,F,commerce,89,27,63,4.0,pass
```

Output

Header that `CSVLoader` generates (nominal labels appear in the order they are first met in the file, missing cells become `?`):

```
1 @relation student
2
3 @attribute sid numeric
4 @attribute age numeric
5 @attribute gender {M,F}
6 @attribute stream {science,commerce,arts}
7 @attribute attendance numeric
8 @attribute internal numeric
9 @attribute sem_marks numeric
10 @attribute hours_study numeric
11 @attribute result {pass,fail}
12
13 @data
14 1,20,M,science,85,24,58,3.5,pass
15 ...
16 9,19,M,commerce,65,?,38,2,fail
```

Preprocess panel after loading: Instances 30, Attributes 9, `attendance` Missing 1 (3%), `internal` Missing 1 (3%), class `result` pass 18, fail 12.

Explanation

CSV only carries names and values, so the loader has to guess types; ARFF carries the types explicitly, which is why WEKA prefers it. The nominal value list in the header is a contract: a row containing a label not in the list is rejected when the file is loaded. The `?` token is the only way to say missing in ARFF, so the blank Excel cell must become `?`, which `CSVLoader` does for you.

Question 3

Problem Statement

■ Write in lab record

Try to create your own datasets.

Solution

■ Write in lab record

Steps

1. Decide the relation, the attributes with their types, and which attribute is the class. Here: 30 employees, class `promoted`.
2. Type the header: `@relation employee`, one `@attribute name type` line per column. Numeric columns get `numeric`; categorical columns get the label list in braces. Lines starting with `%` are comments.
3. Type `@data` and one comma-separated row per employee, values in the same order as the attributes, no spaces needed.
4. Save as `employee.arff` (plain text, UTF-8, `.arff` extension) and open it in Explorer. If the file has a typo (a label not in the list, a missing comma) WEKA reports the line number in the error dialog.
5. Also check the file with Tools, ArffViewer in the GUI Chooser, which shows it as a grid and lets you edit cells.

Program

employee.arff

```
1 % employee.arff - 30 employees, written by hand for MCSL-223 Session 1
2 % Class attribute: promoted (yes/no)
3 @relation employee
4
5 @attribute eid numeric
6 @attribute age numeric
7 @attribute gender {M, F}
8 @attribute dept {HR, IT, Sales, Finance}
9 @attribute experience numeric
10 @attribute education {UG, PG, PhD}
11 @attribute salary numeric
12 @attribute performance {poor, average, good}
13 @attribute promoted {yes, no}
14
15 @data
16 1,28,M,IT,4,PG,45000,good,yes
17 2,35,F,HR,10,PG,52000,average,no
18 3,42,M,Sales,18,UG,60000,good,yes
19 4,25,F,IT,2,UG,32000,average,no
20 5,31,M,Finance,7,PG,48000,good,yes
21 6,45,F,HR,20,PhD,70000,good,yes
22 7,29,M,Sales,5,UG,35000,poor,no
23 8,38,F,IT,13,PG,65000,good,yes
24 9,50,M,Finance,25,PG,80000,average,no
25 10,27,F,Sales,3,UG,30000,average,no
26 11,33,M,IT,9,PG,55000,good,yes
27 12,40,F,Finance,15,PhD,72000,good,yes
28 13,24,M,HR,1,UG,28000,poor,no
29 14,36,F,Sales,12,PG,50000,average,no
30 15,48,M,IT,22,PhD,90000,good,yes
31 16,30,F,Finance,6,UG,40000,average,no
32 17,39,M,HR,14,PG,58000,good,yes
33 18,26,F,IT,3,UG,33000,poor,no
34 19,44,M,Sales,19,PG,62000,average,no
35 20,32,F,Finance,8,PG,47000,good,yes
36 21,37,M,IT,11,PG,60000,good,yes
37 22,23,F,Sales,1,UG,26000,poor,no
38 23,41,M,HR,16,PhD,68000,average,no
39 24,34,F,IT,10,PG,54000,good,yes
40 25,46,M,Finance,21,PG,75000,good,yes
41 26,28,F,Sales,4,UG,31000,average,no
42 27,52,M,HR,27,PhD,85000,average,no
43 28,35,F,Finance,9,UG,44000,good,yes
44 29,30,M,IT,6,PG,46000,poor,no
45 30,43,F,Sales,17,PG,59000,good,yes
```

Output

Preprocess panel values for `employee.arff` (mean and standard deviation computed with `preprocess.py` from Session 2, which uses the same sample standard deviation as WEKA):

```

1 Instances: 30   Attributes: 9
2
3 eid           Numeric  Missing: 0 (0%)  Distinct: 30  Min: 1  Max: 30  Mean: 15.5  Std
4 age           Numeric  Missing: 0 (0%)  Distinct: 27  Min: 23  Max: 52  Mean: 35.7  St
5 gender        Nominal  Missing: 0 (0%)  M: 15  F: 15
6 dept         Nominal  Missing: 0 (0%)  HR: 6  IT: 9  Sales: 8  Finance: 7
7 experience     Numeric  Missing: 0 (0%)  Distinct: 24  Min: 1  Max: 27  Mean: 11.267  S
8 education     Nominal  Missing: 0 (0%)  UG: 10  PG: 15  PhD: 5
9 salary        Numeric  Missing: 0 (0%)  Distinct: 29  Min: 26000  Max: 90000  Mean: 53
10 performance  Nominal  Missing: 0 (0%)  poor: 5  average: 10  good: 15
11 promoted     Nominal  Missing: 0 (0%)  yes: 15  no: 15

```

Explanation

The dataset was designed so that later sessions can learn something from it: `promoted` is `yes` exactly when `performance` is `good`, and `salary` grows with `experience`. A classifier in Session 7 should find the first rule with 100 percent accuracy, and a regression in Session 6 should find a positive slope of salary on experience. Both `student.arff` and `employee.arff` are the files the manual asks for in Sessions 7 and 8.

Question 4

Problem Statement

■ Write in lab record

Preprocess and classify Customer, Agriculture, Weather, Whole-sale Customers or the datasets of your own choice from the [UCI Machine Learning Repository](#).

Solution

■ Write in lab record

Two datasets: the 14-row weather data (in the brief, so every number can be checked by hand) and iris, which WEKA ships and UCI hosts.

Steps

1. Preprocess weather: Open file `weather.nominal.arff`. Every attribute is nominal with no missing values, so no filter is needed. Confirm class `play` with 9 yes and 5 no.
2. Classify tab, Choose, trees, J48. Keep the defaults (`confidenceFactor` 0.25, `minNumObj` 2). Under Test options pick Use training set, click Start. Then pick Cross-validation, Folds 10, Start again.
3. Right-click the entry in the Result list and choose Visualize tree to see the decision tree.
4. Preprocess iris: Open file `iris.arff`. Four numeric attributes, no missing values. Apply Filter, unsupervised, attribute, `Normalize` if you also plan to run IBk (k-nearest neighbour); J48 does not need it because it compares one attribute with a threshold at a time.
5. Classify iris with J48, Cross-validation 10 folds, Start. Record Correctly Classified Instances, Kappa and the confusion matrix.

Output

Information gain of each weather attribute at the root, from the formula sheet with

$$H(S) = -\frac{9}{14}\log_2 \frac{9}{14} - \frac{5}{14}\log_2 \frac{5}{14} = 0.940:$$

Attribute	Gain
outlook	0.247
humidity	0.152
windy	0.048
temperature	0.029

So J48 splits on outlook first, and WEKA prints:

```

1  ≡≡ Classifier model (full training set) ≡≡
2
3  J48 pruned tree
4  -----
5
6  outlook = sunny
7  |   humidity = high: no (3.0)
8  |   humidity = normal: yes (2.0)
9  outlook = overcast: yes (4.0)
10 outlook = rainy
11 |   windy = TRUE: no (2.0)
12 |   windy = FALSE: yes (3.0)
13
14 Number of Leaves   :      5
15 Size of the tree   :      8
16
17 ≡≡ Evaluation on training set ≡≡
18 Correctly Classified Instances      14           100      %
19 Incorrectly Classified Instances      0             0      %
20 Kappa statistic              1
21
22 ≡≡ Confusion Matrix ≡≡
23  a b   ← classified as
24  9 0 | a = yes
25  0 5 | b = no

```

The number in brackets at each leaf is the count of training rows reaching it ($3 + 2 + 4 + 2 + 3 = 14$). With 10-fold cross-validation on 14 rows the accuracy drops to about 50 percent: the folds are tiny and the tree changes from fold to fold.

Iris, J48, 10-fold cross-validation (expected from WEKA, not run here):

```

1 J48 pruned tree
2 -----
3 petalwidth ≤ 0.6: Iris-setosa (50.0)
4 petalwidth > 0.6
5 |   petalwidth ≤ 1.7
6 | |   petallength ≤ 4.9: Iris-versicolor (48.0/1.0)
7 | |   petallength > 4.9
8 | | |   petalwidth ≤ 1.5: Iris-virginica (3.0)
9 | | |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
10 |   petalwidth > 1.7: Iris-virginica (46.0/1.0)
11
12 Number of Leaves   :    5
13 Size of the tree   :    9
14
15 ≡≡ Stratified cross-validation ≡≡
16 Correctly Classified Instances      144           96      %
17 Incorrectly Classified Instances      6            4      %
18 Kappa statistic                   0.94
19
20 ≡≡ Confusion Matrix ≡≡
21   a  b  c   ← classified as
22  49  1  0 |   a = Iris-setosa
23   0 47  3 |   b = Iris-versicolor
24   0  2 48 |   c = Iris-virginica

```

Explanation

Preprocessing here is inspection: types, missing values, class balance. Classification is J48 (C4.5) picking the attribute with the highest gain ratio at every node and pruning leaves that do not help. The weather tree reproduces the textbook ID3 tree because the gain ordering above is the same at every level. On iris the tree uses only the petal measurements, which agrees with the Visualize tab in Question 1: petal length and width separate the classes, sepal width does not. Accuracy is $\frac{144}{150} = 0.96$ and Kappa 0.94 means the agreement is far above the 0.33 expected by chance for three equal classes.

Viva Questions

✗ Do not copy. Read for understanding and the viva

Q: What does ARFF stand for and what are its three sections? **A:** Attribute-Relation File Format: @relation , the @attribute declarations, and @data .

Q: How does WEKA decide which attribute is the class? **A:** By the dropdown above the histogram in Preprocess; it defaults to the last attribute. Classifiers use that choice.

Q: What is the difference between Explorer and Experimenter? **A:** Explorer works on one dataset interactively. Experimenter runs many algorithm and dataset combinations with repeated cross-validation and compares the results statistically.

Q: What does Distinct mean in the Selected attribute box, and what does Unique mean? **A:** Distinct is the number of different values present. Unique is the number of values that occur exactly once.

Q: How is a missing value written in ARFF and what does a blank CSV cell become? **A:** `?` in both cases; `CSVLoader` converts the blank cell for you.

Q: Why is `weather.numeric` different from `weather.nominal` for Apriori? **A:** Apriori needs nominal attributes; temperature and humidity are numeric in `weather.numeric` and must be discretised first.

Q: Why did J48 choose outlook at the root of the weather tree? **A:** Its information gain 0.247 is the largest of the four attributes.

Q: Where are the sample datasets on disk? **A:** In the `data` folder under the WEKA install directory, for example `C:\Program Files\Weka-3-8-6\data`.

Common Mistakes

✗ Do not copy. Read for understanding and the viva

- Saving the Excel sheet as `.xlsx` and trying to open it in WEKA; only CSV or ARFF are loaded by Open file.
- Writing a nominal label in the data that is not in the `@attribute` list, or using spaces inside labels without quotes; WEKA refuses the file with a line number.
- Leaving the identifier column (`sid` , `eid`) in the data before classification; a tree can memorise it and the accuracy on the training set becomes meaningless.
- Reading the histogram counts with the wrong class selected in the dropdown and reporting them as the class distribution.
- Reporting training-set accuracy as the model's accuracy; use cross-validation for the figure you write in the record.

Formula Sheet

✗ Do not copy. Read for understanding and the viva

Association rules

For a rule $X \Rightarrow Y$ over N transactions:

$$\text{support}(X \Rightarrow Y) = \frac{|X \cup Y|}{N}, \quad \text{confidence}(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|}, \quad \text{lift}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{support}(Y)}$$

WEKA's Apriori starts at the upper bound of minimum support and lowers it by the delta each pass until the requested number of rules is found or the lower bound is reached.

Entropy, information gain and Gini

For a set S with class proportions $p_1, \dots, p_c$:

$$H(S) = - \sum_{i=1}^c p_i \log_2 p_i, \quad \text{Gain}(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v), \quad \text{Gini}(S) = 1 - \sum_{i=1}^c p_i^2$$

ID3 splits on the attribute with the highest gain; J48 (C4.5) uses the gain ratio $\text{Gain}(S, A) / \text{SplitInfo}(S, A)$ where $\text{SplitInfo}(S, A) = - \sum_v \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$.

Classifier evaluation

From the confusion matrix with true positives TP , false positives FP , false negatives FN , true negatives TN :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Kappa compares observed agreement p_o (accuracy) with the agreement expected by chance p_e :

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_e = \sum_i \frac{(\text{row}_i \text{ total}) (\text{column}_i \text{ total})}{N^2}$$

The ROC curve plots true positive rate $TP / (TP + FN)$ against false positive rate $FP / (FP + TN)$; the area under it (AUC) is 0.5 for guessing and 1.0 for a perfect classifier.

Naive Bayes and k-nearest neighbour

$$P(C | x_1, \dots, x_n) \propto P(C) \prod_{i=1}^n P(x_i | C)$$

k-NN assigns the majority class among the k nearest training records under Euclidean distance

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

after normalising each attribute to $[0, 1]$ with $x' = (x - x_{\min}) / (x_{\max} - x_{\min})$.

Linear regression

$$\hat{y} = \beta_0 + \beta_1 x, \quad \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

WEKA reports the correlation coefficient, mean absolute error and root mean squared error $\sqrt{\frac{1}{N} \sum (y_i - \hat{y}_i)^2}$.

Clustering

k-means minimises the within-cluster sum of squared errors over clusters $C_1, \dots, C_k$ with centroids μ_j :

$$SSE = \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|^2, \quad \mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

Hierarchical (agglomerative) clustering merges the two closest clusters each step; linkage defines closeness: single $\min d(a, b)$, complete $\max d(a, b)$, average $\frac{1}{|A \cup B|} \sum d(a, b)$.

DBSCAN calls a point a core point when at least minPts points lie within radius ϵ ; clusters grow from core points, and points reachable from none are noise.

Session Summary

Write in lab record

- Question 1: WEKA 3.8 installed, GUI Chooser tour, and a table of ten built-in datasets with attribute types, instance counts, class attribute and class counts read from the Preprocess panel
- Question 2: `student.csv` saved from Excel, loaded with CSVLoader, saved as `student.arff`, header checked
- Question 3: `employee.arff` written by hand (30 rows, 9 attributes) and its Preprocess panel statistics recorded
- Question 4: weather.nominal and iris preprocessed and classified with J48; information gains, trees, accuracy and confusion matrices recorded

[Edit this page](#)

[< Previous](#)
Section 2 · Data Mining Lab (WEKA)

Next [> Session 2](#)