

CS6301 – MACHINE LEARNING - WEEK-1 EXPLORING WEKA TOOL

22-02-2021 MONDAY

NAME: SRIHARI S

ROLLNO: 2018103601

COMPUTER SCIENCE ENGINEERING

BATCH – P

DATASET USED: LABOR

ATTRIBUTES IN THE DATASET:

No.		Name
1	<input checked="" type="checkbox"/>	duration
2	<input type="checkbox"/>	wage-increase-first-year
3	<input type="checkbox"/>	wage-increase-second-year
4	<input type="checkbox"/>	wage-increase-third-year
5	<input type="checkbox"/>	cost-of-living-adjustment
6	<input type="checkbox"/>	working-hours
7	<input type="checkbox"/>	pension
8	<input type="checkbox"/>	standby-pay
9	<input type="checkbox"/>	shift-differential
10	<input type="checkbox"/>	education-allowance
11	<input type="checkbox"/>	statutory-holidays
12	<input type="checkbox"/>	vacation
13	<input type="checkbox"/>	longterm-disability-assistance
14	<input type="checkbox"/>	contribution-to-dental-plan
15	<input type="checkbox"/>	bereavement-assistance
16	<input type="checkbox"/>	contribution-to-health-plan
17	<input type="checkbox"/>	class

Selecting the attributes which have the major impact on the final classification using filters in WEKA.

Upon applying the filter AttributeSelection under Filters->Supervised->Attributes, we get the following attributes.

No.		Name
1	<input checked="" type="checkbox"/>	wage-increase-first-year
2	<input type="checkbox"/>	wage-increase-second-year
3	<input type="checkbox"/>	cost-of-living-adjustment
4	<input type="checkbox"/>	statutory-holidays
5	<input type="checkbox"/>	vacation
6	<input type="checkbox"/>	longterm-disability-assistance
7	<input type="checkbox"/>	contribution-to-dental-plan
8	<input type="checkbox"/>	class

Task for Level-1: Classification on the dataset using Logistic Regression

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **Logistic -R 1.0E-8 -M -1 -num-decimal-places 4**

Test options

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) class

Result list (right-click for options)

15:22:09 - functions.Logistic

Classifier output

```

cost-of-living-adjustment=tc 5.1967806786819718E17
statutory-holidays 0
vacation=below_average 2.436596543990006E43
vacation=average 0
vacation=generous 0
longterm-disability-assistance=no 3.059000167638578E55
contribution-to-dental-plan=none 4.2083169306851606E57
contribution-to-dental-plan=half 0.0225
contribution-to-dental-plan=full 0

```

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	49	85.9649 %
Incorrectly Classified Instances	8	14.0351 %
Kappa statistic	0.6919	
Mean absolute error	0.1333	
Root mean squared error	0.3588	
Relative absolute error	29.148 %	
Root relative squared error	75.1418 %	
Total Number of Instances	57	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.800	0.108	0.800	0.800	0.800	0.692	0.959	0.927	bad
	0.892	0.200	0.892	0.892	0.892	0.692	0.959	0.981	good
Weighted Avg.	0.860	0.168	0.860	0.860	0.860	0.692	0.959	0.962	

=== Confusion Matrix ===

```

a b  <-- classified as
16 4 | a = bad
 4 33 | b = good


```

Using logistic regression, we group the employees into two classes – good and bad.

As we can see 85.9% accuracy is established.

The confusion matrix is also displayed in the bottom justifying the accuracy.

Task for Level-2: Using SVM to the same dataset

 Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **SMOreg** -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -"

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds

☐ Percentage split %

More options...

(Num) statutory-holidays

Start Stop

Result list (right-click for options)

- 15:22:09 - functions.Logistic
- 15:30:14 - functions.SMO
- 15:31:24 - trees.J48
- 15:35:39 - functions.LinearRegression
- 15:39:37 - functions.SMOreg
- 15:40:40 - functions.SMOreg
- 15:42:35 - functions.SMOreg**

Classifier output

Number of kernel evaluations: 1431 (98.432% cached)

Time taken to build model: 0.02 seconds

=== Predictions on test data ===

inst#	actual	predicted	error
1	10	9.657	-0.343
2	?	12.72	?
3	11	11.427	0.427
4	12	10.563	-1.437
5	11	10.917	-0.083
6	13	11.335	-1.665
1	10	10.007	0.007
2	13	12.224	-0.776
3	11	10.903	-0.097
4	10	10.26	0.26
5	11	10.993	-0.007
6	15	12.206	-2.794
1	9	9.996	0.996
2	10	11	1
3	11	11.006	0.006
4	10	14.004	4.004
5	12	10.004	-1.996
6	11	12.002	1.002
1	11	10.488	-0.512
2	11	11.057	0.057
3	?	11.478	?
4	11	10.512	-0.488
5	11	11.297	0.297
6	12	11.297	-0.703
1	11	11.658	0.658
2	10	9.904	-0.096
3	11	11.751	0.751
4	11	11.862	0.862
5	11	11.77	0.77
6	10	11.8	1.8
1	10	9.399	-0.601

Classifier

Choose **SMOreg** -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1

Test options

☐ Use training set

☐ Supplied test set

Set...

☒ Cross-validation

Folds

10

☐ Percentage split

%

66

More options...

(Num) statutory-holidays

Start

Stop

Result list (right-click for options)

15:22:09 - functions.Logistic
 15:30:14 - functions.SMO
 15:31:24 - trees.J48
 15:35:39 - functions.LinearRegression
 15:39:37 - functions.SMOreg
 15:40:40 - functions.SMOreg
 15:42:35 - functions.SMOreg

Classifier output

3	12	11.265	-0.735
4	12	11.969	-0.031
5	10	9.445	-0.555
6	11	10.865	-0.135
1	12	10.337	-1.663
2	11	9.781	-1.219
3	11	11.359	0.359
4	11	11.608	0.608
5	10	10.423	0.423
6	12	11.23	-0.77
1	11	11.004	0.004
2	11	10.977	-0.023
3	9	9.986	0.986
4	12	11.023	-0.977
5	?	12.988	?
1	12	11.01	-0.99
2	10	10.01	0.01
3	9	11.04	2.04
4	10	10.968	0.968
5	9	11.074	2.074
1	11	11.046	0.046
2	?	12.136	?
3	13	12.118	-0.882
4	12	10.988	-1.012
5	12	11.916	-0.084

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.35
Mean absolute error	0.8526
Root mean squared error	1.2363
Relative absolute error	94.2849 %
Root relative squared error	96.7512 %
Total Number of Instances	53
Ignored Class Unknown Instances	4

Status

OK

Task for Level-3: Using Decision Tree to the same dataset

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **J48** -C 0.25 -M 2

Test options:

- ☐ Use training set
- ☐ Supplied test set
- ☐ Cross-validation Folds
- ☒ Percentage split %

(Nom) class

Result list (right-click for options):

- 15:22:09 - functions.Logistic
- 15:30:14 - functions.SMO
- 15:31:24 - trees.J48**

Classifier output:

11	1:bad	1:bad	0.876
12	2:good	2:good	0.952
13	1:bad	1:bad	0.567
14	2:good	2:good	0.82
15	1:bad	1:bad	0.876
16	2:good	1:bad	0.876
17	2:good	2:good	0.82
18	2:good	2:good	0.82
19	2:good	2:good	0.952

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances	17	89.4737 %
Incorrectly Classified Instances	2	10.5263 %
Kappa statistic	0.7564	
Mean absolute error	0.2381	
Root mean squared error	0.3419	
Relative absolute error	52.4444 %	
Root relative squared error	72.9745 %	
Total Number of Instances	19	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.833	0.077	0.833	0.833	0.833	0.756	0.814	0.725	bad
	0.923	0.167	0.923	0.923	0.923	0.756	0.814	0.865	good
Weighted Avg.	0.895	0.138	0.895	0.895	0.895	0.756	0.814	0.821	

=== Confusion Matrix ===

```
a b <-- classified as
5 1 | a = bad
1 12 | b = good
```

Tree Visualizer

