Discussion Board 4

Week 5

Daniel Prusinski

Discretize the continuous variable of a given itemset, and recalculate the support index.

I chose to use the labor data from Weka.

What is the difference before and after the discretization? Why?

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes

Relation: labor-neg-data

Instances: 57

Attributes: 17

duration

wage-increase-first-year

wage-increase-second-year

wage-increase-third-year

cost-of-living-adjustment

working-hours

pension

standby-pay

shift-differential

education-allowance

statutory-holidays

vacation

longterm-disability-assistance

contribution-to-dental-plan

bereavement-assistance

contribution-to-health-plan

class

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute bad good

(0.36) (0.64)

=================================================

duration

mean 2 2.25

std. dev. 0.7071 0.6821

weight sum 20 36

precision 1 1

wage-increase-first-year

mean 2.6563 4.3837

std. dev. 0.8643 1.1773

weight sum 20 36

precision 0.3125 0.3125

wage-increase-second-year

mean 2.9524 4.447

std. dev. 0.8193 0.9805

weight sum 15 31

precision 0.3571 0.3571

wage-increase-third-year

mean 2.0344 4.5795

std. dev. 0.1678 0.7893

weight sum 4 11

precision 0.3875 0.3875

cost-of-living-adjustment

none 10.0 14.0

tcf 2.0 8.0

tc 6.0 3.0

[total] 18.0 25.0

working-hours

mean 39.4887 37.5491

std. dev. 1.8903 2.9266

weight sum 19 32

precision 1.8571 1.8571

pension

none 12.0 1.0

ret\_allw 3.0 3.0

empl\_contr 6.0 8.0

[total] 21.0 12.0

standby-pay

mean 2.5 11.2

std. dev. 0.866 2.0396

weight sum 4 5

precision 2 2

shift-differential

mean 2.4691 5.6818

std. dev. 1.5738 5.0584

weight sum 9 22

precision 2.7778 2.7778

education-allowance

yes 4.0 8.0

no 10.0 4.0

[total] 14.0 12.0

statutory-holidays

mean 10.2 11.4182

std. dev. 0.805 1.2224

weight sum 20 33

precision 1.2 1.2

vacation

below\_average 12.0 8.0

average 8.0 11.0

generous 3.0 15.0

[total] 23.0 34.0

longterm-disability-assistance

yes 6.0 16.0

no 9.0 1.0

[total] 15.0 17.0

contribution-to-dental-plan

none 8.0 3.0

half 8.0 9.0

full 1.0 14.0

[total] 17.0 26.0

bereavement-assistance

yes 10.0 19.0

no 4.0 1.0

[total] 14.0 20.0

contribution-to-health-plan

none 9.0 1.0

half 3.0 8.0

full 7.0 15.0

[total] 19.0 24.0

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 51 89.4737 %

Incorrectly Classified Instances 6 10.5263 %

Kappa statistic 0.7741

Mean absolute error 0.1042

Root mean squared error 0.2637

Relative absolute error 22.7763 %

Root relative squared error 55.2266 %

Total Number of Instances 57

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.9 0.108 0.818 0.9 0.857 0.965 bad

0.892 0.1 0.943 0.892 0.917 0.965 good

Weighted Avg. 0.895 0.103 0.899 0.895 0.896 0.965

=== Confusion Matrix ===

a b <-- classified as

18 2 | a = bad

4 33 | b = good

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes

Relation: labor-neg-data-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 57

Attributes: 17

duration

wage-increase-first-year

wage-increase-second-year

wage-increase-third-year

cost-of-living-adjustment

working-hours

pension

standby-pay

shift-differential

education-allowance

statutory-holidays

vacation

longterm-disability-assistance

contribution-to-dental-plan

bereavement-assistance

contribution-to-health-plan

class

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute bad good

(0.36) (0.64)

===============================================

duration

'All' 21.0 37.0

[total] 21.0 37.0

wage-increase-first-year

'(-inf-2.65]' 14.0 3.0

'(2.65-inf)' 8.0 35.0

[total] 22.0 38.0

wage-increase-second-year

'(-inf-3.25]' 12.0 4.0

'(3.25-inf)' 5.0 29.0

[total] 17.0 33.0

wage-increase-third-year

'(-inf-3.25]' 5.0 2.0

'(3.25-inf)' 1.0 11.0

[total] 6.0 13.0

cost-of-living-adjustment

none 10.0 14.0

tcf 2.0 8.0

tc 6.0 3.0

[total] 18.0 25.0

working-hours

'All' 20.0 33.0

[total] 20.0 33.0

pension

none 12.0 1.0

ret\_allw 3.0 3.0

empl\_contr 6.0 8.0

[total] 21.0 12.0

standby-pay

'(-inf-6]' 5.0 1.0

'(6-inf)' 1.0 6.0

[total] 6.0 7.0

shift-differential

'(-inf-3.5]' 9.0 6.0

'(3.5-inf)' 2.0 18.0

[total] 11.0 24.0

education-allowance

yes 4.0 8.0

no 10.0 4.0

[total] 14.0 12.0

statutory-holidays

'(-inf-10.5]' 13.0 5.0

'(10.5-inf)' 9.0 30.0

[total] 22.0 35.0

vacation

below\_average 12.0 8.0

average 8.0 11.0

generous 3.0 15.0

[total] 23.0 34.0

longterm-disability-assistance

yes 6.0 16.0

no 9.0 1.0

[total] 15.0 17.0

contribution-to-dental-plan

none 8.0 3.0

half 8.0 9.0

full 1.0 14.0

[total] 17.0 26.0

bereavement-assistance

yes 10.0 19.0

no 4.0 1.0

[total] 14.0 20.0

contribution-to-health-plan

none 9.0 1.0

half 3.0 8.0

full 7.0 15.0

[total] 19.0 24.0

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 52 91.2281 %

Incorrectly Classified Instances 5 8.7719 %

Kappa statistic 0.8138

Mean absolute error 0.098

Root mean squared error 0.2458

Relative absolute error 21.4201 %

Root relative squared error 51.4693 %

Total Number of Instances 57

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.95 0.108 0.826 0.95 0.884 0.982 bad

0.892 0.05 0.971 0.892 0.93 0.982 good

Weighted Avg. 0.912 0.07 0.92 0.912 0.913 0.982

=== Confusion Matrix ===

a b <-- classified as

19 1 | a = bad

4 33 | b = good

Overall Discretization produced a result that better classified one instance, which was a false ‘bad’. The variable I would like to focus on is wage increase second year:

Before Discretization

wage-increase-second-year

mean 2.9524 4.447

std. dev. 0.8193 0.9805

weight sum 15 31

precision 0.3571 0.3571

After

wage-increase-second-year

'(-inf-3.25]' 12.0 4.0

'(3.25-inf)' 5.0 29.0

[total] 17.0 33.0

With this variable there is both a positive and negative support index. 3.25 appears to be the support and given its ‘bucket’ it has an overall positive outcome on the classification.

On a side note, this Discussion Board has not yet been posted to. Given that it is Wednesday, please post if this response is not correct so that we can head in the right direction before the discussion boards are due.