

CAP6776 Homework 2

1. Converting .txt files to .arff files

- The provided webkb-train-stemmed.txt and webkb-test-stemmed.txt documents cannot be read by Weka.
- The attached Python script, TxtToArff.py, converts these files into webkb-train-stemmed.arff and webkb-test-stemmed.arff.
- The given files are formatted as <class>\t<document>\n, so they are converted to <.arff header>\n'<document>,<class> in the output files.
- Here is the header and a few documents in the training set:

```

1 @relation 'WebKB train'
2
3 @attribute Text string
4 @attribute class-att {course, faculty, project, student}
5
6 @data
7
8 'brian comput scienc depart univers wisconsin dayton street madison offic email wisc offic phone home phone advisor david wood tabl content interest schedul summer public
interest profession comput architectur oper system compil high speed network distribut parallel system secur account high perform person bicycl walk hike camp travel home
brew cook comput electron read schedul mondai wwt meet wednesday meet david cow meet milwaukee brian heidi wed madison comput architectur affili meet chicago base public
journal articl foster perform massiv parallel comput spectral atmospher model atmospher ocean technolog byte drake foster design perform scalabl parallel commun climat
model parallel comput decemb byte proceed paper foster algorithm comparison benchmark parallel spectral transform water model sixth workshop parallel process meteorolog
ed world scientif singapor byte drake foster hack williamson adapt scalabl parallel comput proceed global chang symposium american meteorolog societi byte foster load
balanc algorithm climat model proceed scalabl high perform comput confer ed walker ieee comput societi press byte technic report user guid technic report juli byte foster
load balanc algorithm parallel commun climat model anl technic report anl mc januari byte poster present foster sutton wagner harrison kendal calcul librari gordon
research confer high perform comput nation inform infrastructur juli foster sutton wagner calcul librari high perform comput chemistri workshop hilton california august
earth belong man man belong earth thing connect blood unit man web life strand web chief seattl man sat ground medit life mean accept creatur acknowledg uniti univers
thing true essenc civil stand bear modifi mon aug cdt',student
9 'denni swanson web page mail pop uki offic hour comput lab offic anderson quadrangl mailbox anderson hall lab resum dilbert comic sport data mine web imag web
yahoo',student
10 'russel impagliazzo depart comput scienc engin univers california san diego jolla offic appli physic mathemat build apm phone fax email russel ucsc assist professor
special complex theori research circuit lower bound theori cryptographi comput random cours fall cse algorithm cse algorithm offic hour student prioriti mondai wednesday
student prioriti tuesday thursday research paper beam cook edmond impagliazzo rel complex search problem beam impagliazzo improv depth lower bound small distanc connect
beam impagliazzo lower bound hilbert proposit proof impagliazzo reachabl problem finit cellular automata edmond impagliazzo commun complex lower bound circuit depth gupta
impagliazzo comput planar impagliazzo levin construct pseudo random gener function impagliazzo person view averag case complex impagliazzo distribut hard problem
impagliazzo effici cryptograph scheme provabl secur subset sum impagliazzo effect random boolean formula impagliazzo paturi size depth trade off threshold circuit
impagliazzo upper lower bound tree cut plane proof impagliazzo wigderson network algorithm impagliazzo limit provabl consequ permut beam impagliazzo exponenti lower bound
constant depth proof principl',faculty

```

2. Filtering the .arff files

- Using the .arff files as-is, without filtering will result in incompatible classifiers.
- Online tutorials recommend using the command line to apply filters to the .arff files, as shown:

```

> java weka.filters.unsupervised.attribute.StringToWordVector -b -i
/Users/Nick/OneDrive/Documents/CAP6776_Info_Retr/hw2/TxtToArff/arff/webkb-train-stemmed.arff -o
/Users/Nick/OneDrive/Documents/CAP6776_Info_Retr/hw2/TxtToArff/arff/webkb-train-stemmed-vector.arff -c last -r
/Users/Nick/OneDrive/Documents/CAP6776_Info_Retr/hw2/TxtToArff/arff/webkb-test-stemmed.arff -s
/Users/Nick/OneDrive/Documents/CAP6776_Info_Retr/hw2/TxtToArff/arff/webkb-test-stemmed-vector.arff -C

```

This filter applies the word count parameter with the '-C' argument.

- Loading the training set into Weka:

Current relation
Relation: WebKB train-weka.filters.unsupervised.attribute.StringToWordV...
Instances: 2803
Attributes: 1905
Sum of weights: 2803

Attributes

| No. | Name |
|-----|-----------|
| 1 | class-att |
| 2 | abstract |
| 3 | academ |
| 4 | accept |
| 5 | access |
| 6 | account |
| 7 | acm |
| 8 | acrobat |
| 9 | activ |
| 10 | actual |
| 11 | ad |
| 12 | ada |
| 13 | adapt |
| 14 | add |
| 15 | addison |
| 16 | addit |
| 17 | address |

Selected attribute

Name: class-att
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|-----|---------|-------|--------|
| 1 | course | 620 | 620.0 |
| 2 | faculty | 750 | 750.0 |
| 3 | project | 336 | 336.0 |
| 4 | student | 1097 | 1097.0 |

Class: yale (Num)

Visualize All

Bar chart showing counts for each class: course (620), faculty (750), project (336), student (1097).

Nick Petty

CAP6776 Homework 2

d. Setting the class attribute as class gives a document-word matrix, which is partially shown here:

| No. | 1: abstract Numeric | 2: academ Numeric | 3: accept Numeric | 4: access Numeric | 5: account Numeric | 6: acm Numeric | 7: acrobat Numeric | 8: activ Numeric | 9: actual Numeric | 10: ad Numeric | 11: ada Numeric |
|-----|------------------------|----------------------|----------------------|----------------------|-----------------------|-------------------|-----------------------|---------------------|----------------------|-------------------|--------------------|
| 1 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 8 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

3. Building the classifiers

a. Using the training set with Naïve Bayes and 10-fold cross validation gives the following results:

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

23:52:27 - rules.ZeroR

23:52:37 - bayes.NaiveBayes

Classifier output

Time taken to build model: 1.06 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 1830 65.2872 %

Incorrectly Classified Instances 973 34.7128 %

Kappa statistic 0.506

Mean absolute error 0.1735

Root mean squared error 0.4155

Relative absolute error 48.7457 %

Root relative squared error 98.4809 %

Total Number of Instances 2803

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|---------|
| | 0.760 | 0.041 | 0.841 | 0.760 | 0.798 | 0.746 | 0.942 | 0.829 | course |
| | 0.485 | 0.142 | 0.556 | 0.485 | 0.518 | 0.359 | 0.770 | 0.528 | faculty |
| | 0.500 | 0.062 | 0.522 | 0.500 | 0.511 | 0.446 | 0.854 | 0.443 | project |
| | 0.754 | 0.257 | 0.653 | 0.754 | 0.700 | 0.487 | 0.790 | 0.637 | student |
| Weighted Avg. | 0.653 | 0.155 | 0.653 | 0.653 | 0.650 | 0.505 | 0.826 | 0.627 | |

=== Confusion Matrix ===

| a | b | c | d | <-- classified as |
|-----|-----|-----|-----|-------------------|
| 471 | 36 | 8 | 105 | a = course |
| 33 | 364 | 87 | 266 | b = faculty |
| 13 | 87 | 168 | 68 | c = project |
| 43 | 168 | 59 | 827 | d = student |

Nick Petty

CAP6776 Homework 2

b. The same set with SVM:

Classifier

Choose **SMO** -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.fun

Test options

☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

- 23:52:27 - rules.ZeroR
- 23:52:37 - bayes.NaiveBayes
- 00:01:31 - functions.SMO

Classifier output

Time taken to build model: 2.72 seconds

=== Stratified cross-validation ===
=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 2462 | 87.8345 % |
| Incorrectly Classified Instances | 341 | 12.1655 % |
| Kappa statistic | 0.8272 | |
| Mean absolute error | 0.2639 | |
| Root mean squared error | 0.3327 | |
| Relative absolute error | 74.1159 % | |
| Root relative squared error | 78.8519 % | |
| Total Number of Instances | 2803 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|---------|
| | 0.918 | 0.013 | 0.952 | 0.918 | 0.934 | 0.916 | 0.986 | 0.936 | course |
| | 0.832 | 0.040 | 0.884 | 0.832 | 0.857 | 0.808 | 0.912 | 0.801 | faculty |
| | 0.708 | 0.021 | 0.821 | 0.708 | 0.760 | 0.733 | 0.909 | 0.667 | project |
| | 0.940 | 0.104 | 0.853 | 0.940 | 0.894 | 0.823 | 0.928 | 0.833 | student |
| Weighted Avg. | 0.878 | 0.057 | 0.879 | 0.878 | 0.877 | 0.829 | 0.934 | 0.827 | |

=== Confusion Matrix ===

| a | b | c | d | <-- classified as |
|-----|-----|-----|------|-------------------|
| 569 | 7 | 7 | 37 | a = course |
| 11 | 624 | 34 | 81 | b = faculty |
| 10 | 28 | 238 | 60 | c = project |
| 8 | 47 | 11 | 1031 | d = student |

c. This Naïve Bayes classifier was about 65% accurate, while the Support Vector Machine classifier was about 88% accurate.

d. For this training set, the SVM classifier has much better results.

4. Testing the classifiers

a. The training set is run against the test set. This is with the Naïve Bayes classifier on default parameters:

Classifier

Choose **NaiveBayes**

Test options

☐ Use training set
☒ Supplied test set Set...
☐ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) class-att

Start Stop

Result list (right-click for options)

- 23:52:27 - rules.ZeroR
- 23:52:37 - bayes.NaiveBayes
- 00:01:31 - functions.SMO
- 00:04:45 - misc.InputMappedClassifier

Classifier output

Time taken to test model on supplied test set: 3.99 seconds

=== Summary ===

| | | |
|----------------------------------|------------|----------|
| Correctly Classified Instances | 886 | 63.467 % |
| Incorrectly Classified Instances | 510 | 36.533 % |
| Kappa statistic | 0.4808 | |
| Mean absolute error | 0.1827 | |
| Root mean squared error | 0.4264 | |
| Relative absolute error | 51.3108 % | |
| Root relative squared error | 101.0238 % | |
| Total Number of Instances | 1396 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|---------|
| | 0.697 | 0.033 | 0.857 | 0.697 | 0.769 | 0.717 | 0.918 | 0.804 | course |
| | 0.551 | 0.168 | 0.545 | 0.551 | 0.548 | 0.381 | 0.794 | 0.540 | faculty |
| | 0.423 | 0.069 | 0.455 | 0.423 | 0.438 | 0.365 | 0.802 | 0.362 | project |
| | 0.722 | 0.255 | 0.644 | 0.722 | 0.681 | 0.460 | 0.773 | 0.624 | student |
| Weighted Avg. | 0.635 | 0.160 | 0.642 | 0.635 | 0.636 | 0.485 | 0.814 | 0.610 | |

=== Confusion Matrix ===

| a | b | c | d | <-- classified as |
|-----|-----|----|-----|-------------------|
| 216 | 20 | 10 | 64 | a = course |
| 7 | 206 | 43 | 118 | b = faculty |
| 8 | 54 | 71 | 35 | c = project |
| 21 | 98 | 32 | 393 | d = student |

Nick Petty
CAP6776 Homework 2

b. The training set is run against the test set with SVM and default parameters:

Classifier

Choose **SMO** -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.func

Test options

☐ Use training set
☒ Supplied test set **Set...**
☐ Cross-validation Folds **10**
☐ Percentage split % **66**
More options...

(Nom) class-att **▼**

Start **Stop**

Result list (right-click for options)

23:52:27 - rules.ZeroR
 23:52:37 - bayes.NaiveBayes
 00:01:31 - functions.SMO
 00:04:45 - misc.InputMappedClassifier
 00:09:34 - misc.InputMappedClassifier

Classifier output

Time taken to test model on supplied test set: 1.05 seconds

=== Summary ===

| | | |
|----------------------------------|-----------|-----------|
| Correctly Classified Instances | 1224 | 87.6791 % |
| Incorrectly Classified Instances | 172 | 12.3209 % |
| Kappa statistic | 0.8257 | |
| Mean absolute error | 0.2644 | |
| Root mean squared error | 0.3336 | |
| Relative absolute error | 74.2497 % | |
| Root relative squared error | 79.032 % | |
| Total Number of Instances | 1396 | |

=== Detailed Accuracy By Class ===

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|-------|----------|----------|---------|
| | 0.881 | 0.008 | 0.968 | 0.881 | 0.922 | 0.903 | 0.983 | 0.930 | course |
| | 0.861 | 0.044 | 0.877 | 0.861 | 0.869 | 0.822 | 0.925 | 0.808 | faculty |
| | 0.744 | 0.028 | 0.786 | 0.744 | 0.765 | 0.734 | 0.904 | 0.648 | project |
| | 0.926 | 0.099 | 0.857 | 0.926 | 0.890 | 0.818 | 0.922 | 0.832 | student |
| Weighted Avg. | 0.877 | 0.055 | 0.879 | 0.877 | 0.877 | 0.828 | 0.934 | 0.825 | |

=== Confusion Matrix ===

| a | b | c | d | <-- classified as |
|-----|-----|-----|-----|-------------------|
| 273 | 5 | 9 | 23 | a = course |
| 3 | 322 | 17 | 32 | b = faculty |
| 2 | 12 | 125 | 29 | c = project |
| 4 | 28 | 8 | 504 | d = student |

- c. Again, the Naïve Bayes classifier was much less accurate than the Support Vector Machine, at 64% accuracy to 88% accuracy.
- d. With the provided training and test data sets, Weka shows the SVM classification is better than Naïve Bayes at determining which documents refer to courses, students, projects, or faculty.