Tim auf der Landwehr
t.auf.der.landwehr@student.rug.nl
S2548682
Last revision: November 29, 2013

Learning from Data
Assignment 3

Summary

This is the assignment 3 in the course Learning from Data at the University of Groningen.

Exercise 1: Mutual Information (letters)

The file `mutual_information_letters.py` contains the script to compute the 50 most worthy letters for the language classification task. The best letters according to mutual information are:

```
e: 0.003071    k: 0.000701    g: 0.000311    W: 0.000178    C: 0.000110
j: 0.002135    v: 0.000681    i: 0.000302    a: 0.000168    G: 0.000109
n: 0.001694    z: 0.000468    d: 0.000263    /: 0.000140    O: 0.000108
 : 0.001588    r: 0.000427    x: 0.000260    l: 0.000132    t: 0.000106
y: 0.000795    o: 0.000312    >: 0.000226    B: 0.000125    D: 0.000103
```

Exercise 2: Mutual Information (words)

The file `mutual_information_words.py` contains the script to compute the 50 most worthy words for the language classification task. The best words according to mutual information are:

```
ik:    0.005253   niet:  0.002582   maar:  0.001687   nog:   0.001339   te:    0.001088
je:    0.004725   van:   0.002435   als:   0.001599   ook:   0.001308
een:   0.003227   op:    0.002220   get:   0.001573   dan:   0.001267
en:    0.003090   met:   0.002214   Ik:    0.001548   I:     0.001197
de:    0.002796   voor:  0.001895   naar:  0.001475   wel:   0.001185
het:   0.002628   a:     0.001737   heb:   0.001348   echt:  0.001133
```

Exercise 3: kNN classification with WEKA

The file `ned_arff_generator.py` contains a script, to create an arff file from the given training data in `ned.train`. The script produces an output file `ned.train.arff`. This file can be run in WEKA using the following command:

```
java -cp weka.jar weka.classifiers.lazy.IBk -c 1 -t ../assignment3/ned.arff
```

The following 16 features are implemented:

```
entity_name
entity_prefix-2
entity_prefix-3
entity_prefix-4
entity_suffix-2
entity_suffix-3
entity_suffix-4
direct_preceding_word
direct_subsequent_word
contains_numbers
number_of_parts
number_of_capital_letters
number_of_whitespaces
contains_hyphen
contains_dot
preceding_word_suffix-4
```

The output of the classifier is as follows:

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances       10050               89.6601 %
Incorrectly Classified Instances      1159               10.3399 %
Kappa statistic                          0.8504
Mean absolute error                      0.0535
Root mean squared error                  0.2147
Relative absolute error                 15.3863 %
Root relative squared error             51.5048 %
Total Number of Instances            11209

=== Detailed Accuracy By Class ===

              TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                0.869     0.031      0.866     0.869     0.868      0.952    ORG
                0.689     0.018      0.825     0.689     0.751      0.885    MISC
                0.936     0.057      0.923     0.936     0.93       0.969    PER
                0.934     0.042      0.899     0.934     0.916      0.974    LOC
Weighted Avg.   0.897     0.043      0.895     0.897     0.895      0.958

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1809   56  118   99 |    a = ORG
  110  829  144  120 |    b = MISC
  109   74 4415  118 |    c = PER
   60   46  105 2997 |    d = LOC
```