Laboratorio 6 - ciencia de datos

Manuel Felipe Pineda L 1093223607

November 24, 2016

1 Realizar 10 experimentos con 2 bases de datos de weka.

1.1 Base de datos 'segment'

Se utilizó la base de datos 'segment' con las versiones provistas para entrenamiento y clasificación.

Con esta base de datos se realizaron algunos experimentos de clasificación para comparar diferentes algoritmos.

También, se usaron algoritmos de clustering para analizar si éstos podían brindar más información que los algoritmos de clasificación.

Finalmente se usó una selección de atributos para encontrar los más significativos y hacer proyecciones con menor dimensionalidad.

1.1.1 Clasification - multi layer perceptron

```
   a   b   c   d   e   f   g   <-- classified as
 123   0   0   2   0   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   1   0 109   1  11   0   0 |   c = foliage
   0   0   1 107   1   1   0 |   d = cement
   3   0   8   9 106   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   1   0   0   2 120 |   g = grass
```

1.1.2 Clasification - logistic

```
   a   b   c   d   e   f   g   <-- classified as
 123   0   0   0   2   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   1   0 112   1   8   0   0 |   c = foliage
   2   0   0 102   5   1   0 |   d = cement
   1   0  15  11  98   0   1 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   2   0   0   2 119 |   g = grass
```

1.1.3 Clasification - k star

```
   a   b   c   d   e   f   g   <-- classified as
```

```
 125   0   0   0   0   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0   0 117   1   4   0   0 |   c = foliage
   2   0   0 103   3   2   0 |   d = cement
   1   0   8   5 112   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   0   0   3   0   0 120 |   g = grass
```

1.1.4 Classification Random Tree

```
   a   b   c   d   e   f   g   <-- classified as
 124   0   0   0   1   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0   0 101   4  17   0   0 |   c = foliage
   2   0   0 101   2   5   0 |   d = cement
   2   0  13   8 103   0   0 |   e = window
   0   0   0   1   1  92   0 |   f = path
   1   0   2   0   1   0 119 |   g = grass
```

1.1.5 Classification Decision Stump

```
   a   b   c   d   e   f   g   <-- classified as
   0 125   0   0   0   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   0 122   0   0   0   0   0 |   c = foliage
   0 106   0   0   0   4   0 |   d = cement
   0 126   0   0   0   0   0 |   e = window
   0   0   0   0   0  94   0 |   f = path
   0   4   0   0   0 119   0 |   g = grass
```

1.1.6 Cluster - EM

```
=== Evaluation on test set ===
Clustered Instances
0      148 ( 18%)
1      120 ( 15%)
2       71 (  9%)
3       72 (  9%)
4       88 ( 11%)
5      109 ( 13%)
6      202 ( 25%)
```

1.1.7 Cluster - Kmeans

```
=== Evaluation on test set ===
Clustered Instances
```

```
0       205 ( 25%)
1       123 ( 15%)
2       110 ( 14%)
3        44 (  5%)
4        50 (  6%)
5       154 ( 19%)
6       124 ( 15%)
```

1.1.8 Cluster - hierarchical

```
=== Evaluation on test set ===
Clustered Instances
0        94 ( 12%)
1       123 ( 15%)
2       122 ( 15%)
3       126 ( 16%)
4       110 ( 14%)
5       125 ( 15%)
6       110 ( 14%)
```

1.1.9 Attribute selection - PCA

```
eigenvalue proportion cumulative
  7.62679   0.42371   0.42371 -0.357rawblue-mean-0.354value-mean-0.351intensity-mean-0.348
  3.0237    0.16798   0.59169 -0.489hedge-sd-0.475vegde-sd-0.472hedge-mean-0.467vedge-mean
  1.7551    0.09751   0.6892  0.597hue-mean+0.434exgreen-mean-0.363saturation-mean+0.358re
  1.05213   0.05845   0.74765 0.709short-line-density-5-0.645region-centroid-col+0.216shor
  0.94124   0.05229   0.79994 0.675region-centroid-col+0.653short-line-density-5-0.192exgr
  0.88844   0.04936   0.8493  -0.726short-line-density-2-0.379region-centroid-row+0.337veg
  0.71934   0.03996   0.88926 -0.508short-line-density-2+0.448region-centroid-row-0.43satu
  0.55697   0.03094   0.92021 -0.611saturation-mean+0.495exred-mean-0.395region-centroid-r
  0.50727   0.02818   0.94839 0.57 vedge-mean-0.491hedge-mean-0.444hedge-sd+0.301vegde-sd-
  0.39418   0.0219    0.97029 -0.508hedge-mean+0.486vegde-sd-0.414vedge-mean+0.365region-c
```

1.1.10 Attribute selection - Info Gain

```
average merit      average rank  attribute
 1.676 +- 0.02      1   +- 0      11 rawred-mean
 1.613 +- 0.011     2.1 +- 0.3    13 rawgreen-mean
 1.596 +- 0.01      3.4 +- 0.66   10 intensity-mean
 1.582 +- 0.021     4   +- 1.1    19 hue-mean
 1.57  +- 0.013     4.6 +- 0.49   17 value-mean
 1.545 +- 0.013     5.9 +- 0.3    12 rawblue-mean
 1.339 +- 0.018     7.6 +- 0.66   16 exgreen-mean
 1.326 +- 0.015     7.6 +- 0.66    2 region-centroid-row
 1.307 +- 0.007     9.2 +- 0.6    15 exblue-mean
 1.291 +- 0.022     9.6 +- 0.66   18 saturation-mean
 1.088 +- 0.017    11   +- 0      14 exred-mean
 0.52  +- 0.01     12.1 +- 0.3     8 hedge-mean
 0.5   +- 0.023    12.9 +- 0.3     6 vedge-mean
```

```
 0.433 +- 0.015    14   +- 0       9 hedge-sd
 0.391 +- 0.01     15   +- 0       7 vegde-sd
 0.1   +- 0.011    16   +- 0       1 region-centroid-col
 0.028 +- 0.002    17   +- 0       5 short-line-density-2
 0     +- 0        18.2 +- 0.4     3 region-pixel-count
 0.003 +- 0.008    18.8 +- 0.4     4 short-line-density-5
```

1.2 Base de datos 'glass'

1.2.1 cluster - canopy

```
Class attribute: Type
Classes to Clusters:

  0  1  2  3  <-- assigned to cluster
 70  0  0  0 | build wind float
 65 10  1  0 | build wind non-float
 17  0  0  0 | vehic wind float
  0  0  0  0 | vehic wind non-float
  2  8  1  2 | containers
  5  4  0  0 | tableware
  4 13 12  0 | headlamps

Cluster 0 <-- build wind float
Cluster 1 <-- build wind non-float
Cluster 2 <-- headlamps
Cluster 3 <-- containers
```

1.2.2 cluster - coweb

```
Class attribute: Type
Classes to Clusters:

  1  3  4  7 10 11  <-- assigned to cluster
 70  0  0  0  0  0 | build wind float
 68  2  0  1  4  1 | build wind non-float
 17  0  0  0  0  0 | vehic wind float
  0  0  0  0  0  0 | vehic wind non-float
  7  3  1  2  0  0 | containers
  7  1  1  0  0  0 | tableware
  6  1 22  0  0  0 | headlamps

Cluster  1 <-- build wind float
Cluster  3 <-- containers
Cluster  4 <-- headlamps
Cluster  7 <-- No class
Cluster 10 <-- build wind non-float
Cluster 11 <-- No class
```

1.2.3 cluster - EM

```
Class attribute: Type
Classes to Clusters:

  0  1  2  3  4  5  6  <-- assigned to cluster
  0 17  0  0 19 32  2 | build wind float
  0 23 10  1  2 35  5 | build wind non-float
  0  4  0  0  5  8  0 | vehic wind float
  0  0  0  0  0  0  0 | vehic wind non-float
  3  0  8  1  0  0  1 | containers
  1  0  3  0  0  0  5 | tableware
 21  1  0  5  0  0  2 | headlamps

Cluster 0 <-- headlamps
Cluster 1 <-- build wind non-float
Cluster 2 <-- containers
Cluster 3 <-- No class
Cluster 4 <-- vehic wind float
Cluster 5 <-- build wind float
Cluster 6 <-- tableware
```

1.2.4 cluster - Hierarchical

```
Class attribute: Type
Classes to Clusters:

  0  1  2  3  4  5  6  <-- assigned to cluster
 70  0  0  0  0  0  0 | build wind float
 74  1  0  0  1  0  0 | build wind non-float
 17  0  0  0  0  0  0 | vehic wind float
  0  0  0  0  0  0  0 | vehic wind non-float
 10  0  1  0  0  0  2 | containers
  8  0  0  1  0  0  0 | tableware
 28  0  0  0  0  1  0 | headlamps

Cluster 0 <-- build wind non-float
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- tableware
Cluster 4 <-- No class
Cluster 5 <-- headlamps
Cluster 6 <-- containers

Incorrectly clustered instances : 136.0  63.5514 %
```

1.2.5 cluster - Kmeans

```
Class attribute: Type
Classes to Clusters:
```

```
  0  1  2  3  4  5  6  <-- assigned to cluster
 15  0  0 17 38  0  0 | build wind float
 19  0  0  2 42  4  9 | build wind non-float
  3  0  0  2 12  0  0 | vehic wind float
  0  0  0  0  0  0  0 | vehic wind non-float
  1  1  2  0  0  3  6 | containers
  0  1  0  0  0  6  2 | tableware
  0 22  0  1  2  3  1 | headlamps

Cluster 0 <-- vehic wind float
Cluster 1 <-- headlamps
Cluster 2 <-- No class
Cluster 3 <-- build wind float
Cluster 4 <-- build wind non-float
Cluster 5 <-- tableware
Cluster 6 <-- containers

Incorrectly clustered instances : 118.0  55.1402 %
```

1.2.6 class - logistic

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

Correctly Classified Instances         137               64.0187 %
Incorrectly Classified Instances        77               35.9813 %
Kappa statistic                          0.5052
Mean absolute error                      0.1217
Root mean squared error                  0.2766
Relative absolute error                 57.4847 %
Root relative squared error             85.227  %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.643    0.201    0.608      0.643    0.625      0.435    0.821     0.632
                 0.658    0.217    0.625      0.658    0.641      0.436    0.760     0.613
                 0.118    0.025    0.286      0.118    0.167      0.140    0.796     0.294
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?
                 0.692    0.030    0.600      0.692    0.643      0.620    0.807     0.451
                 0.778    0.010    0.778      0.778    0.778      0.768    0.985     0.707
                 0.828    0.027    0.828      0.828    0.828      0.801    0.981     0.839
Weighted Avg.    0.640    0.151    0.625      0.640    0.629      0.487    0.825     0.619

=== Confusion Matrix ===
```

```
  a  b  c  d  e  f  g   <-- classified as
 45 19  5  0  1  0  0 |  a = build wind float
 19 50  0  0  3  2  2 |  b = build wind non-float
  9  6  2  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  3  0  0  9  0  1 |  e = containers
  0  0  0  0  0  7  2 |  f = tableware
  1  2  0  0  2  0 24 |  g = headlamps
```

1.2.7 class - multilayer perceptron

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

Correctly Classified Instances         145               67.757  %
Incorrectly Classified Instances        69               32.243  %
Kappa statistic                          0.5528
Mean absolute error                      0.1114
Root mean squared error                  0.2627
Relative absolute error                 52.5915 %
Root relative squared error             80.958  %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.743    0.181    0.667      0.743    0.703      0.548    0.862     0.643
                 0.684    0.217    0.634      0.684    0.658      0.460    0.815     0.741
                 0.059    0.005    0.500      0.059    0.105      0.151    0.654     0.161
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?
                 0.692    0.030    0.600      0.692    0.643      0.620    0.957     0.604
                 0.778    0.020    0.636      0.778    0.700      0.689    0.926     0.514
                 0.828    0.011    0.923      0.828    0.873      0.856    0.931     0.889
Weighted Avg.    0.678    0.141    0.671      0.678    0.659      0.537    0.847     0.665

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 52 16  1  0  0  0  1 |  a = build wind float
 17 52  0  0  4  3  0 |  b = build wind non-float
  8  8  1  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  3  0  0  9  0  1 |  e = containers
  0  1  0  0  1  7  0 |  f = tableware
  1  2  0  0  1  1 24 |  g = headlamps
```

1.2.8 class - decision table

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

Correctly Classified Instances         146               68.2243 %
Incorrectly Classified Instances        68               31.7757 %
Kappa statistic                          0.5507
Mean absolute error                      0.1724
Root mean squared error                  0.2768
Relative absolute error                 81.4177 %
Root relative squared error             85.2945 %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.886    0.264    0.620      0.886    0.729      0.585    0.839     0.670
                 0.671    0.167    0.689      0.671    0.680      0.507    0.792     0.720
                 0.176    0.015    0.500      0.176    0.261      0.264    0.570     0.163
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?
                 0.538    0.010    0.778      0.538    0.636      0.629    0.873     0.532
                 0.667    0.010    0.750      0.667    0.706      0.695    0.853     0.514
                 0.586    0.000    1.000      0.586    0.739      0.742    0.926     0.813
Weighted Avg.    0.682    0.148    0.702      0.682    0.669      0.560    0.815     0.652

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 62  7  1  0  0  0  0 |  a = build wind float
 21 51  1  0  1  2  0 |  b = build wind non-float
 11  3  3  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  1  5  0  0  7  0  0 |  e = containers
  0  3  0  0  0  6  0 |  f = tableware
  5  5  1  0  1  0 17 |  g = headlamps
```

1.2.9 class - one r

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

Correctly Classified Instances         124               57.9439 %
Incorrectly Classified Instances        90               42.0561 %
Kappa statistic                          0.3946
Mean absolute error                      0.1202
Root mean squared error                  0.3466
Relative absolute error                 56.7438 %
Root relative squared error            106.8083 %
```

```
Total Number of Instances              214

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.786    0.299    0.561      0.786    0.655      0.459    0.744     0.511
                 0.579    0.268    0.543      0.579    0.561      0.307    0.655     0.464
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.079
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?
                 0.231    0.000    1.000      0.231    0.375      0.469    0.615     0.277
                 0.000    0.000    0.000      0.000    0.000      0.000    0.500     0.042
                 0.759    0.054    0.688      0.759    0.721      0.676    0.852     0.554
Weighted Avg.    0.579    0.200    0.530      0.579    0.534      0.379    0.690     0.432

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 55 15  0  0  0  0  0 |  a = build wind float
 26 44  0  0  0  0  6 |  b = build wind non-float
 12  5  0  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  7  0  0  3  0  3 |  e = containers
  3  5  0  0  0  0  1 |  f = tableware
  2  5  0  0  0  0 22 |  g = headlamps
```

1.2.10 class - random forest

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

Correctly Classified Instances         171               79.9065 %
Incorrectly Classified Instances        43               20.0935 %
Kappa statistic                          0.7229
Mean absolute error                      0.1004
Root mean squared error                  0.2123
Relative absolute error                 47.4315 %
Root relative squared error             65.4186 %
Total Number of Instances              214

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.871    0.118    0.782      0.871    0.824      0.734    0.934     0.861
                 0.816    0.116    0.795      0.816    0.805      0.696    0.926     0.874
                 0.353    0.020    0.600      0.353    0.444      0.426    0.931     0.514
                 0.000    0.000    0.000      0.000    0.000      0.000    ?         ?
                 0.846    0.010    0.846      0.846    0.846      0.836    0.985     0.907
                 0.778    0.005    0.875      0.778    0.824      0.818    0.998     0.963
                 0.828    0.016    0.889      0.828    0.857      0.836    0.969     0.912
Weighted Avg.    0.799    0.084    0.794      0.799    0.793      0.720    0.941     0.852
```

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 61  7  2  0  0  0  0 |  a = build wind float
  8 62  2  0  2  1  1 |  b = build wind non-float
  8  3  6  0  0  0  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  1  0  0 11  0  1 |  e = containers
  0  1  0  0  0  7  1 |  f = tableware
  1  4  0  0  0  0 24 |  g = headlamps
```

2 Experimentos con la base de datos del proyecto.

2.1 clasificacion

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

Correctly Classified Instances        8373               83.7467 %
Incorrectly Classified Instances      1625               16.2533 %
Kappa statistic                          0.5671
Mean absolute error                      0.2148
Root mean squared error                  0.3454
Relative absolute error                 53.8714 %
Root relative squared error             77.3624 %
Total Number of Instances             9998

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.611    0.077    0.751      0.611    0.674      0.572    0.871     0.756
                 0.923    0.389    0.862      0.923    0.892      0.572    0.871     0.936
Weighted Avg.    0.837    0.303    0.832      0.837    0.832      0.572    0.871     0.887

=== Confusion Matrix ===

    a    b   <-- classified as
 1678 1070 |    a = true
  555 6695 |    b = false
```

2.2 Seleccion de atributos

634 a partir de 54

2.3 Clustering

```
=== Model and evaluation on training set ===
```

```
Clustered Instances

0       774 (  8%)
1       408 (  4%)
2      1307 ( 13%)
3      6070 ( 61%)
4       439 (  4%)
5       560 (  6%)
6       440 (  4%)


Log likelihood: -137.40306


Class attribute: opened
Classes to Clusters:

    0    1    2    3    4    5    6  <-- assigned to cluster
  194  181  394 1547  122  188  122 | true
  580  227  913 4523  317  372  318 | false

Cluster 0 <-- No class
Cluster 1 <-- No class
Cluster 2 <-- true
Cluster 3 <-- false
Cluster 4 <-- No class
Cluster 5 <-- No class
Cluster 6 <-- No class

Incorrectly clustered instances : 5081.0  50.8202 %
```