Actividades resueltas

Weka: EM – Clústering probabilístico

Descripción de la actividad

Se ha obtenido la siguiente salida al aplicar Weka a un conjunto de datos.

```
=== Clustering model (full training set) ===

Number of clusters selected by cross validation: 3
Number of iterations performed: 16

                Cluster
Attribute        0       1       2
                (0.33)  (0.36)  (0.31)
=======================================
preg
  mean          4.0284  1.4999  6.3804
  std. dev.     3.1712  1.2122  3.3682

plas
  mean        115.4268 115.1168 133.5004
  std. dev.    27.8245  30.5879  33.9917

pres
  mean         62.9638  68.2572  76.6921
  std. dev.    27.6978   12.157  10.8051

skin
  mean          7.1784  28.5425  25.5598
  std. dev.     12.292  10.9223   15.177

insu
  mean               0 126.3568 111.2879
  std. dev.     0.0001 116.7935 129.2923

mass
  mean         29.8164  33.0587  33.0884
  std. dev.     8.9628   7.8967   5.8762
```



```
pedi
 mean       0.3491   0.5238   0.5432
 std. dev.  0.2085   0.3844   0.3345

age
 mean      32.3749  24.5802  44.2618
 std. dev.  9.7378   3.0226  11.1922
Time taken to build model (full training data) : 2.62 seconds

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

Clustered Instances

0     369 ( 48%)
1     250 ( 33%)
2     149 ( 19%)


Log likelihood: -23.10594
```

Contesta a las siguientes cuestiones:

- ¿Qué algoritmo se ha aplicado? ¿Qué tipo de agrupamiento proporciona este algoritmo? ¿A qué categoría de aprendizaje pertenece este algoritmo?

- ¿Cuantos clústeres se han generado? ¿Cuantas instancias de los datos de prueba conforman cada uno de los clústeres?

- ¿Qué información proporciona Weka para cada uno de los clústeres? ¿Por qué? Indica los valores que definen el Cluster 0.

Resolución de la actividad

- ¿Qué algoritmo se ha aplicado?
  EM (Expectation Maximization)

¿Qué tipo de agrupamiento proporciona este algoritmo?
Agrupamiento probabilista

Explicación de la respuesta:
Weka presenta el valor de la media y la desviación estándar para cada atributo en cada clúster. Con estos dos valores se puede definir la distribución normal que modela la probabilidad de que las instancias en el clúster tengan unos determinados valores de sus atributos. Por lo tanto se trata de un agrupamiento probabilista en el cual los clústeres se generan mediante un método probabilístico. Debemos tener en cuenta que no puede ser el algoritmo k-means que genera un agrupamiento exclusivo porque éste define los clústeres en base al valor de los centroides que son los puntos centrales de cada clúster. En este caso no tenemos un único valor para cada atributo, que pudiera ser el centroide, sino que tenemos dos valores la media y la desviación estándar para cada atributo. Por ejemplo, para el atributo preg se ha obtenido la siguiente información:

```
               Cluster
Attribute        0       1       2
                (0.33)  (0.36)  (0.31)
=====================================
preg
  mean          4.0284  1.4999  6.3804
  std. dev.     3.1712  1.2122  3.3682
```

Si se hubiera aplicado el algoritmo k-means para el mismo conjunto de datos de entrenamiento se habría obtenido la siguiente información:

```
                  Cluster#
Attribute    Full Data       0         1         2
                (768.0)   (208.0)   (229.0)   (331.0)
=========================================================
preg            3.8451    2.0144    7.7031    2.3263
```

¿A qué categoría de aprendizaje pertenece este algoritmo?
Aprendizaje no supervisado

Explicación de la respuesta:

Todos algoritmos de clústering, también EM, pertenecen a la categoría de aprendizaje no supervisado porque a partir de las instancias que conforman los datos de entrenamiento se pueden identificar agrupamientos con instancias similares entre sí, es decir instancias que tienen valores similares para sus atributos.

- ¿Cuantos clústeres se han generado?
  3 clústeres

  Explicación de la respuesta:

  Nos viene indicado en la descripción del modelo Number of clusters selected by cross validation: 3. Es importante remarcar que para el algoritmo EM no es necesario establecer el número de clústeres deseados sino que el número óptimo se puede calcular a través de un proceso de validación cruzada.

  ¿Cuantas instancias de los datos de prueba conforman cada uno de los clústeres?
  Cluster 0: 369 instancias.
  Cluster 1: 250 instancias.
  Cluster 2: 149 instancias.

  Explicación de la respuesta:

  El número de instancias del conjunto de datos de prueba que conforman cada uno de los clústeres al evaluar el modelo se visualiza tal que así:

  ```
  Clustered Instances

  0       369 ( 48%)
  1       250 ( 33%)
  2       149 ( 19%)
  ```

- ¿Qué información proporciona Weka para cada uno de los clústeres?
  Para cada atributo de entrada en cada clúster Weka proporciona el valor de la media y la desviación estándar que definen la distribución normal que modela la probabilidad de que las instancias en el clúster tengan unos determinados

valores del atributo. Además para cada clúster también presenta la probabilidad de población, es decir la probabilidad apriori de que una instancia pertenezca a un clúster (o la proporción de instancias de los datos de entrenamiento que se estima que pertenecen a cada clúster).

¿Por qué?
EM es un algoritmo de agrupamiento probabilista por tanto hace las agrupaciones en base a la probabilidad de que las instancias pertenezcan a un clúster. El algoritmo EM utiliza una distribución normal definida a través de la media y la desviación estándar para modelar el comportamiento de cada atributo en cada clúster. Además no todos los clústeres son igualmente probables, no todos contienen el mismo número de instancias, por lo que se necesita modelar la probabilidad de población.

Indica los valores que definen el Cluster 0.
La probabilidad de población del Clúster 0 es de 0.33.
La distribución normal del atributo preg tiene una media de 4.0284 y una desviación estándar de 3.1712.
La distribución normal del atributo plas tiene una media de 115.4268 y una desviación estándar de 27.8245.
La distribución normal del atributo pres tiene una media de 62.9638 y una desviación estándar de 27.6978.
La distribución normal del atributo skin tiene una media de 7.1784 y una desviación estándar de 12.292.
La distribución normal del atributo insu tiene una media de 0 y una desviación estándar de 0.0001.
La distribución normal del atributo mass tiene una media de 29.8164 y una desviación estándar de 8.9628.
La distribución normal del atributo pedi tiene una media de 0.3491 y una desviación estándar de 0.2085.
La distribución normal del atributo age tiene una media de 32.3749 y una desviación estándar de 9.7378.

Explicación de la respuesta:

Los valores que definen el Clúster 0 se observan en:

```
            Cluster
Attribute       0
                (0.33)
=====================================
preg
  mean        4.0284
  std. dev.   3.1712

…
```

