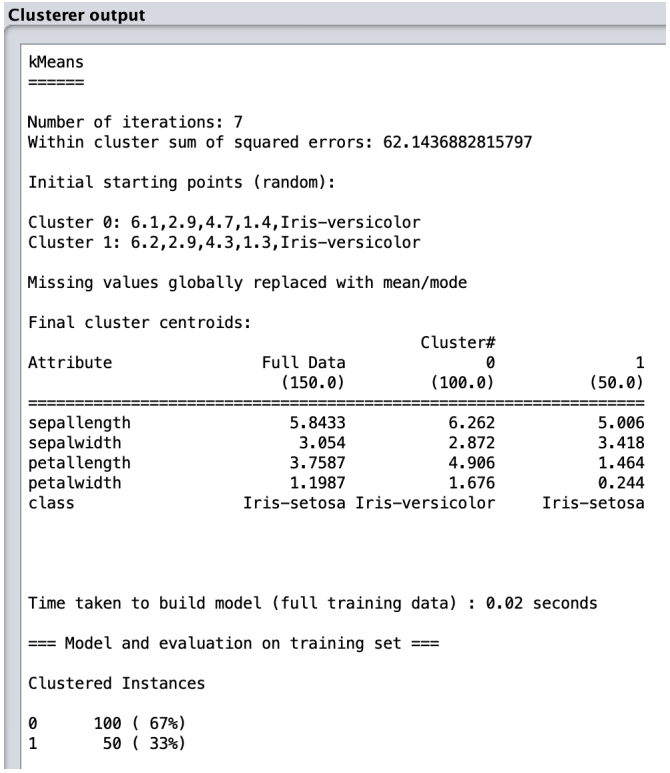
Clustering en Weka

Laura Rodríguez Navas
rodrigueznavas@posgrado.uimp.es

July 9, 2020

En esta práctica se realiza un estudio acerca de la base de datos Iris. Esta base de datos se distribuye junto a la herramienta Weka.

1. Ejecuta el algoritmo SimpleKMeans usando la herramienta Weka con las distancias Euclídea y Manhattan.



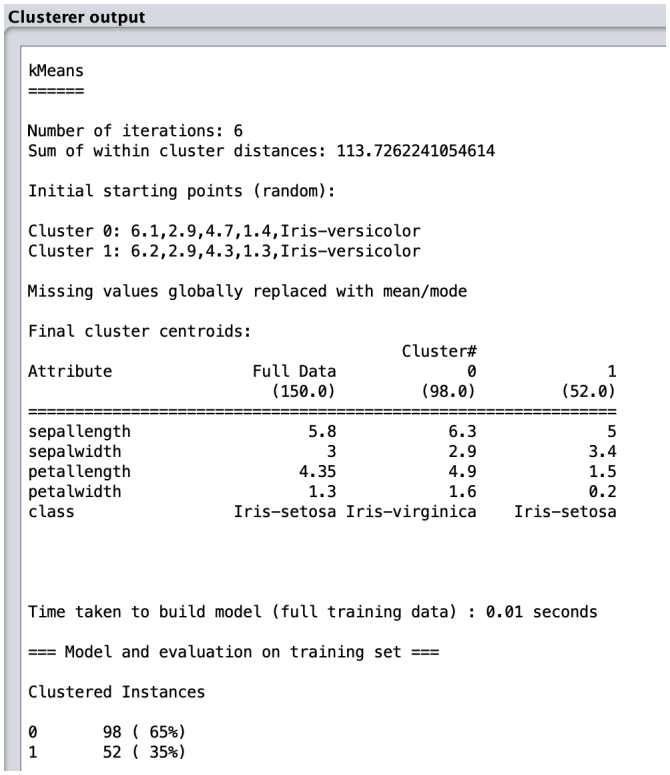
Clusterer output

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0              1
                          (150.0)         (100.0)         (50.0)
=============================================================
sepallength                5.8433           6.262           5.006
sepalwidth                  3.054           2.872           3.418
petallength                3.7587           4.906           1.464
petalwidth                 1.1987           1.676           0.244
class                 Iris-setosa Iris-versicolor     Iris-setosa



Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0      100 ( 67%)
1       50 ( 33%)
```

Figura 1: KMeans con distancia Euclídea.

(a) ¿Cuántas instancias contiene cada grupo?

En la ejecución del algoritmo KMeans con distancia Euclídea (ver Figura 1) se han formado dos grupos: 0 y 1. El grupo 0 contienen 100 instancias (el 67% de las instancias del conjunto de datos) y el grupo 1 contiene 50 instancias (el 33% de las instancias del conjunto de datos).



```
Clusterer output

kMeans
======

Number of iterations: 6
Sum of within cluster distances: 113.7262241054614

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1
                          (150.0)          (98.0)        (52.0)
============================================================
sepallength                  5.8             6.3             5
sepalwidth                     3             2.9           3.4
petallength                 4.35             4.9           1.5
petalwidth                   1.3             1.6           0.2
class                Iris-setosa Iris-virginica    Iris-setosa



Time taken to build model (full training data) : 0.01 seconds

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

Clustered Instances

0       98 ( 65%)
1       52 ( 35%)
```

Figura 2: KMeans con distancia Manhattan.

En la ejecución del algoritmo KMeans con distancia Manhattan (ver Figura 2) también se han formado los grupos 0 y 1. El grupo 0 contienen 98 instancias (el 65% de las instancias del conjunto de datos) y el grupo 1 contiene 52 instancias (el 35% de las instancias del conjunto de datos).

(b) ¿Cuáles son los centroides?
Si nos volvemos a fijar en la figura 1, podemos observar los centroides de la ejecución de Kmeans con distancia Euclídea. Se muestra en la siguiente figura:

```
Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1
                          (150.0)          (100.0)         (50.0)
============================================================
sepallength               5.8433            6.262         5.006
sepalwidth                 3.054            2.872         3.418
petallength               3.7587            4.906         1.464
petalwidth                1.1987            1.676         0.244
class                Iris-setosa Iris-versicolor    Iris-setosa
```

Figura 3: KMeans centroides con distancia Euclídea.

Si nos volvemos a fijar en la figura 2, podemos observar los centroides de la ejecución de Kmeans con distancia Manhattan. Se muestra en la siguiente figura:

(c) Analiza los centroides. ¿Hay algo destacable en esos centroides? ¿Están los centroides separados en el espacio? ¿Tienen componentes similares?

2. Ejecute el algoritmo HierarchicalClusterer con tipo de enlace completo y métrica de distancia

```
Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0             1
                          (150.0)        (98.0)        (52.0)
=============================================================
sepallength                   5.8           6.3             5
sepalwidth                      3           2.9           3.4
petallength                  4.35           4.9           1.5
petalwidth                    1.3           1.6           0.2
class                 Iris-setosa Iris-virginica    Iris-setosa
```

Figura 4: KMeans centroides con distancia Manhattan.

euclídea, y visualice las gráficas de los puntos agrupados. ¿Alguno de ellos produce grupos bien diferenciados y con fronteras claras?

Nota: Compara que el eje X instance_number y el eje Y vaya variando y muestra cada una de las variables (debes adjuntar las imágenes).

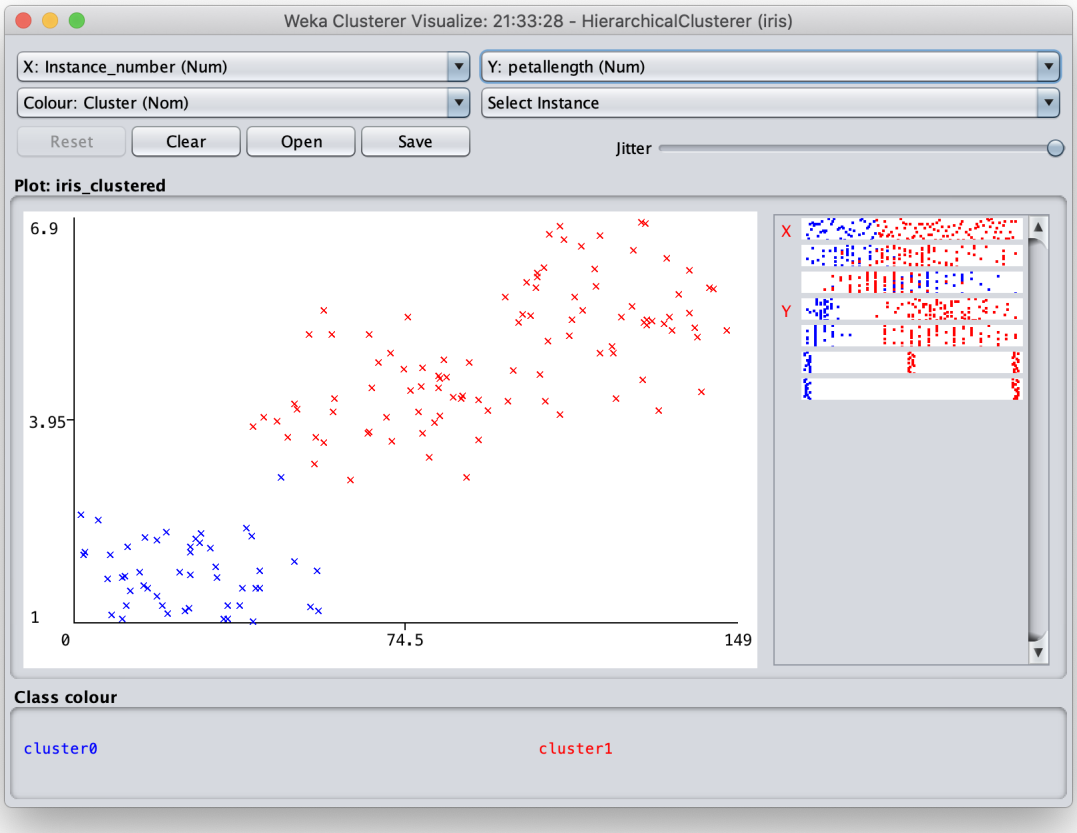


Figura 5: X instance_number con distancia Manhattan.

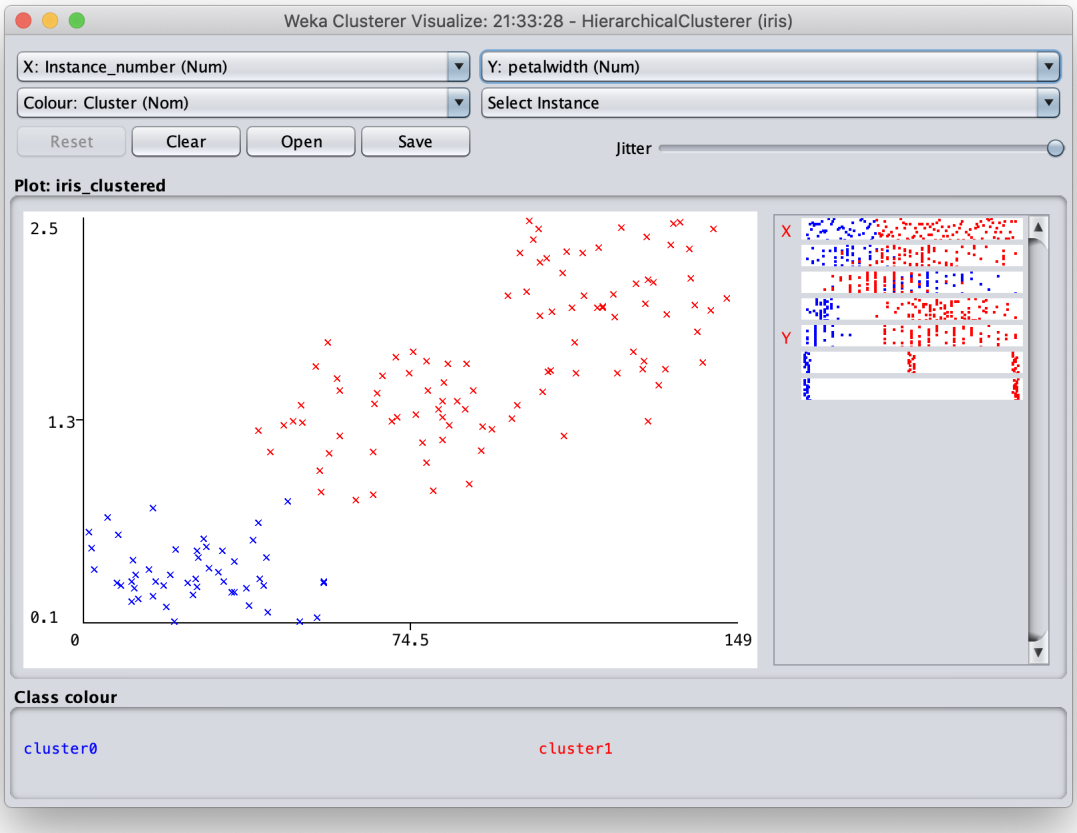


Figura 6: X instance_number con distancia Manhattan.

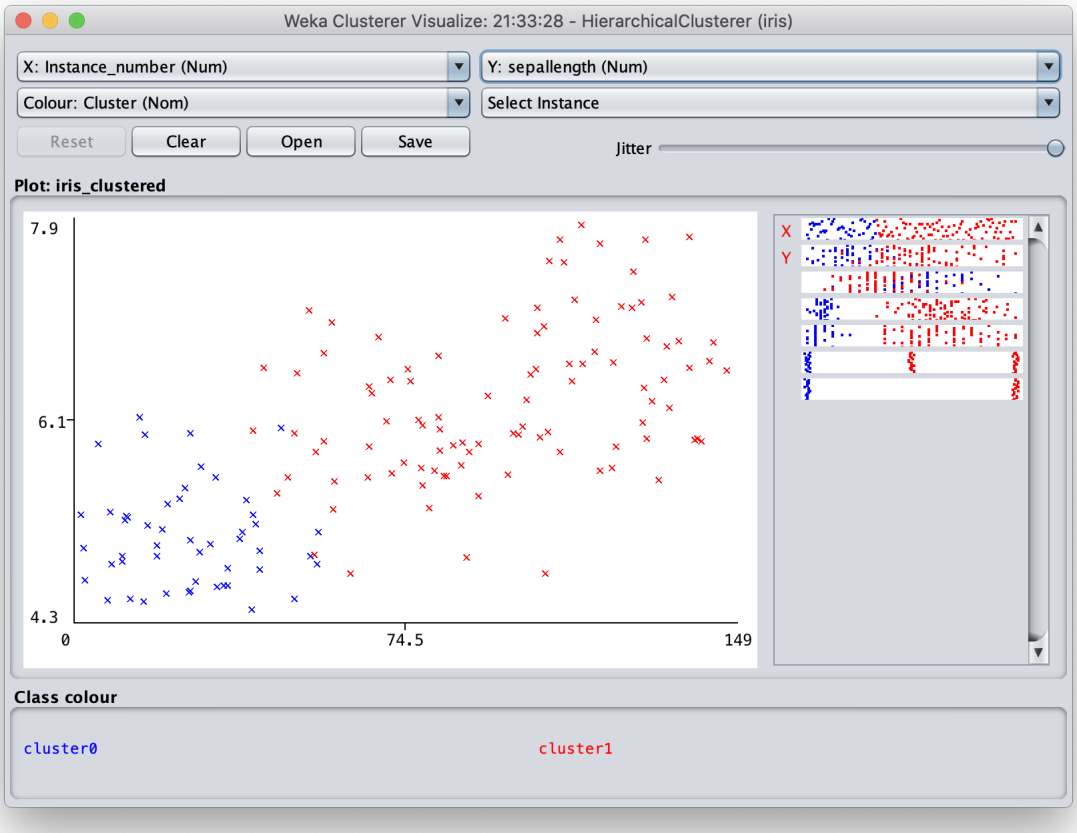


Figura 7: X instance_number con distancia Manhattan.

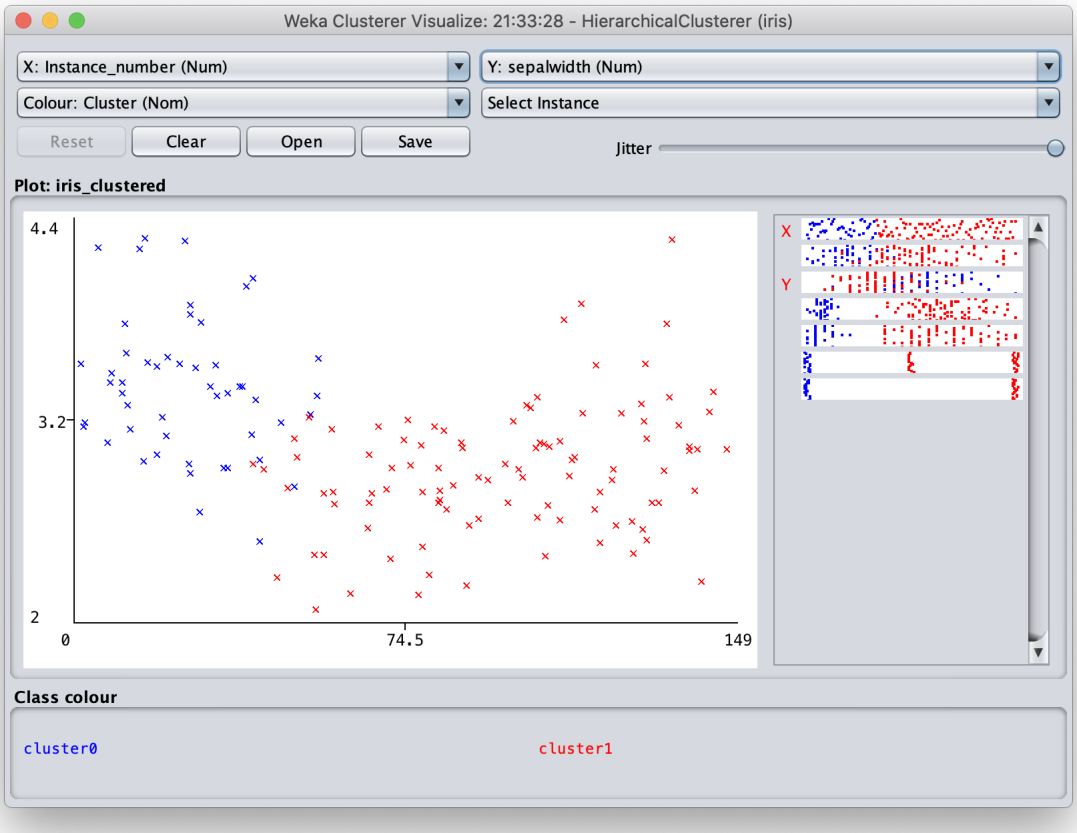


Figura 8: X instance_number con distancia Manhattan.

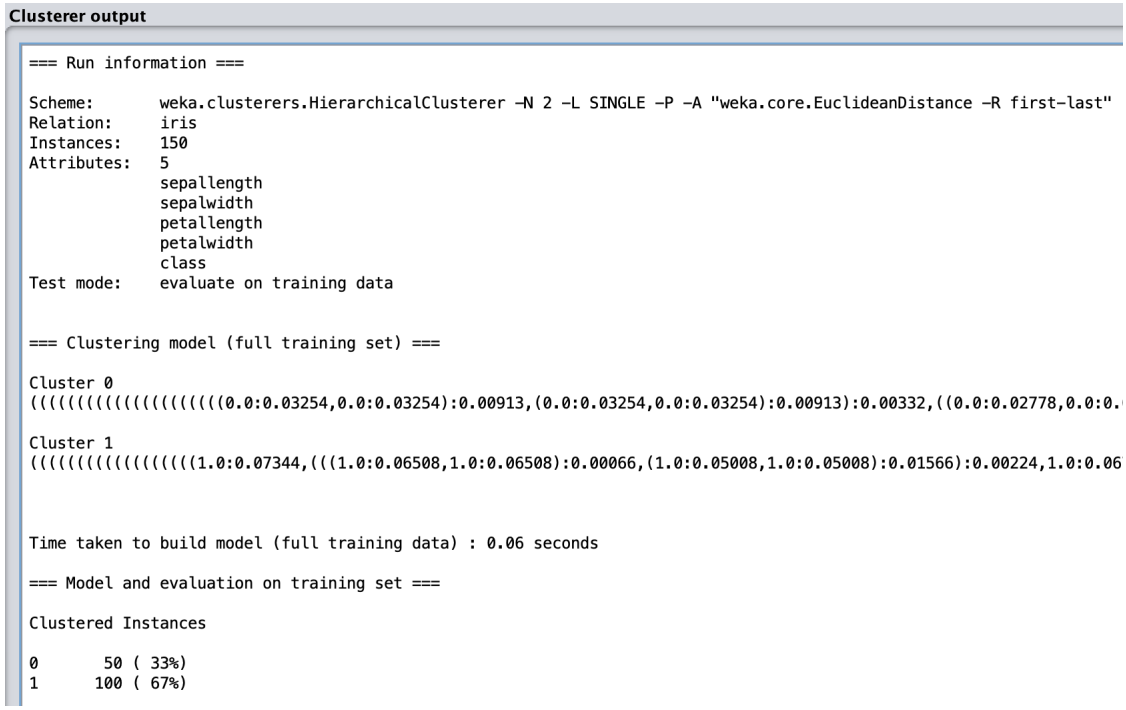


Figura 9: Hierarchical Clustering con distancia Euclídea.