Reporte del Proyecto 2, Minería de Datos

Dataset

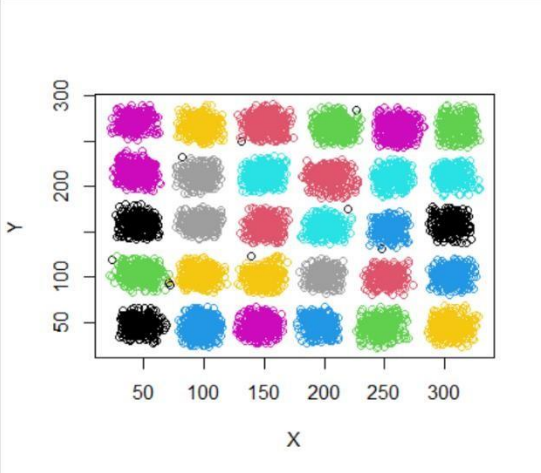
Se utilizarán tres datasets distintos del archivo proporcionado por el profesor, en particular será “Boxes”, “spiral2” e “isolation”.

Información del Dataset

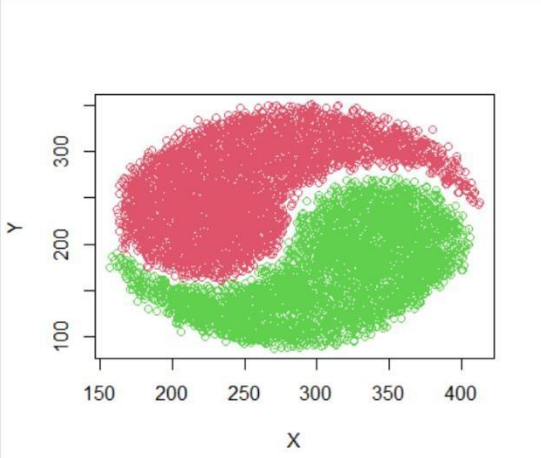
Los datos provenientes de los datasets comparten la estructura:

$$P = (x, y, color)$$

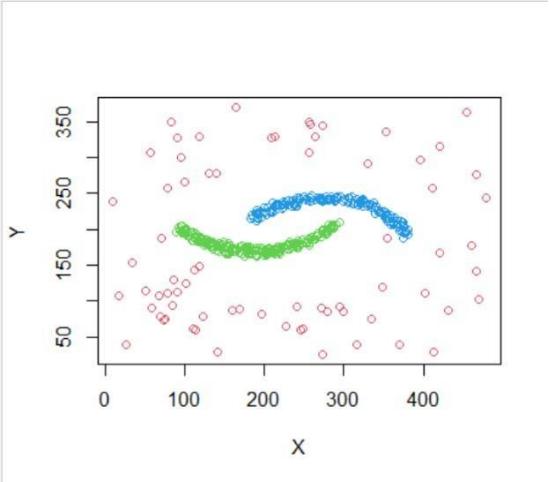
Para boxes: 8902 datos, se cuenta con 30 clusters.



Para spiral2: 9325 datos, se cuenta con 2 clusters.



Para isolation: 464 datos, se cuenta con 3 clusters.



Marco Teórico

Se utilizarán tablas de contingencia a las cuales se les obtendrá su medida F y su medida de entropía para realizar un criterio de validación en el cluster. Las medidas están definidas de la siguiente manera

Medida F

$$F = \frac{1}{r}\sum_{i=1}^{r} F_i$$

Donde:

$$F_i = \frac{2n_{ij_i}}{n_i + m_{j_i}}; \quad j_i = \max_{1 \le j \le k} n_{ij}$$

Medida de entropía:

$$H(T|C) = \frac{1}{n}\sum_{i=1}^{r} p_i \log_2(p_i)$$

Donde:

$$p_i = \frac{n_{ij}}{n_i}$$

Reporte del Proyecto 2, Minería de Datos

Hipótesis

Los datasets se pueden clusterizar óptimamente mediante un algoritmo basado en densidad.

Modelación

Mediante la librería dbscan de R se utilizará la función

$$dbscan(datos, eps, minPts)$$

para clusterizar los datos, para encontrar una buena clusterización se iterará el algoritmo en un intervalo de valores discretos con la función creada

$$DBSCAN\_Optimo(puntos, Clusters\ Reales, from, to, by)$$

Cabe destacar que el criterio de esta función para encontrar la mejor clusterización es maximizando la medida F, por consiguiente, se tuvieron que crear

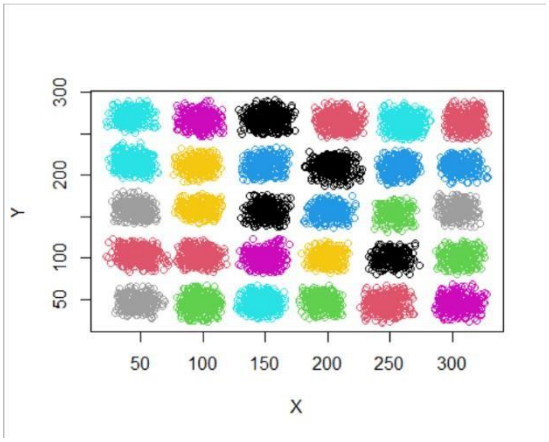
$\frac{to-from}{by}$ tablas de contingencia, esto es importante a la hora de tomar en cuenta la complejidad computacional.

Después de estimar una clusterización óptima se creará una tabla de contingencia para estimar la medida F y la de entropía.

Resultados

A continuación, las clusterizaciones obtenidas:

Para boxes:

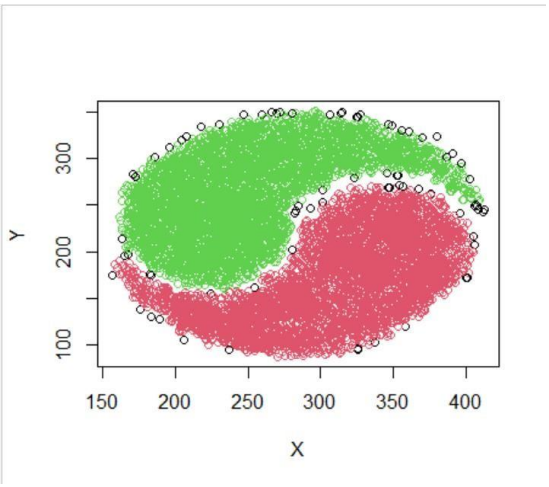


$$Clusters = 31\,, eps = 8,\ minPts = 12$$

$$F = 0.968,\ Entropia = 0.004$$

Para spiral2:

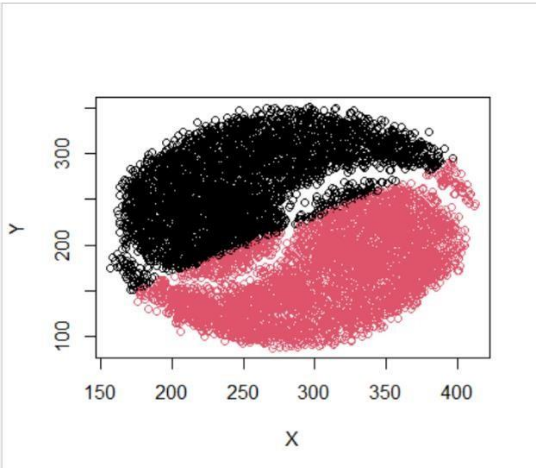
Por algoritmo de densidad:



$$Clusters = 3, eps = 6, minPts = 7$$

$$F = 0.671,\ \ Entropia = 0.009$$

Puesto que la medida F fue mala, también se realizó por medio de k-means con 2 centros y 1000 iteraciones, el resultado fue el siguiente:

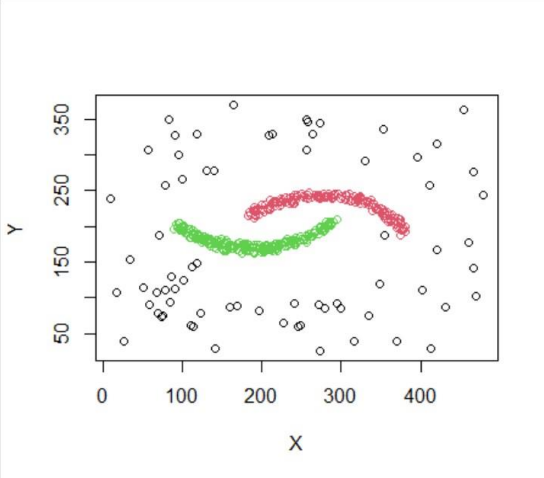


$$Clusters = 2$$

$$F = 0.948,\ Entropia = 0.293$$

Reporte del Proyecto 2, Minería de Datos

Para isolation:



$Clusters = 3, eps = 8, minPts = 4$

$F = 1, Entropia = 0$

Conclusión

En general un algoritmo basado en densidad es bueno para clusterizar, aunque el algoritmo a usar no es lo único a tomar en cuenta, por ejemplo en el dataset “spiral2”, cuando se utilizó dbscan se consiguió una mala medida F sin embargo una entropía más baja que cuando se empleó k-means, esto quiere decir que nuestra medida de desempeño si influye en la decisión de cuál clusterización utilizar, también se puede concluir que la medida F penaliza mucho un cluster mal estimado, como se hizo en “spiral2” donde se estimó un cluster de más mientras que la medida de entropía premia la ausencia de ruido en la clusterización.