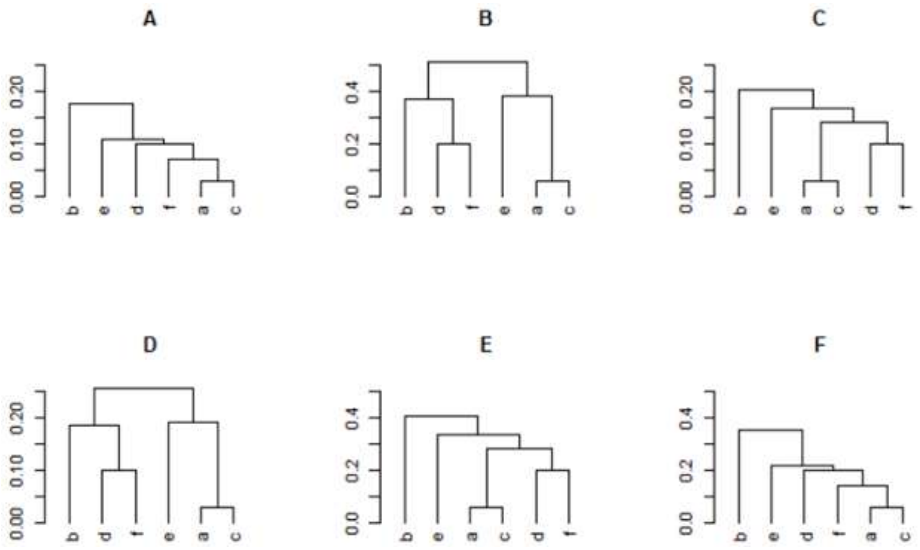
9. Aprendizaje no supervisado: Clustering

Ejercicio 9.1. La figura que se muestra abajo contiene 6 dendrogramas, de los cuales se sabe lo siguiente:



Tres de los seis fueron obtenidos a partir de una matriz de distancias dada a la que llamamos "L", mientras que los restantes tres fueron obtenidos a partir de la matriz de distancias 2 * L (o sea cada elemento de L multiplicado por 2).

Dos de los seis fueron obtenidos utilizando complete linkage como método de aglomeración, dos de los seis fueron obtenidos utilizando single linkage como método de aglomeración y los restantes dos utilizando average linkage como método de aglomeración.

Teniendo esto en cuenta, responda los siguientes puntos:

1. Indique cuáles de los dendrogramas fueron creados tomando como input la matriz de distancias L y cuáles fueron creados tomando como input la matriz de distancias 2 * L. Justifique brevemente su respuesta.
2. Sabiendo que L es la matriz que se muestra debajo y que los dendrogramas obtenidos son los que se mostraron previamente, indique qué método de aglomeración (single, complete o average) se utilizó en cada uno de los seis dendrogramas. Justifique brevemente su respuesta.

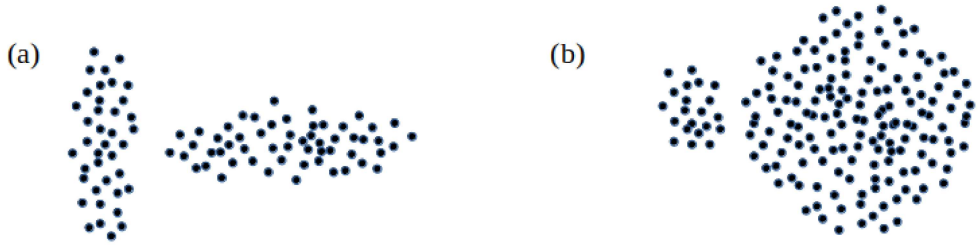
	a	b	c	d	e	f
a	0.000	0.255	0.030	0.120	0.190	0.170
b	0.255	0.000	0.205	0.185	0.195	0.175
c	0.030	0.205	0.000	0.200	0.115	0.070
d	0.120	0.185	0.200	0.000	0.250	0.100
e	0.190	0.195	0.115	0.250	0.000	0.110
f	0.170	0.175	0.070	0.100	0.110	0.000

Ejercicio 9.2. Suponga que para un determinado conjunto de datos se lleva adelante un análisis de clustering jerárquico. Primero calcula la matriz de distancias (en la cual no se observan valores duplicados excepto por los 0 de la diagonal principal). Después, por un lado se obtiene un dendrograma utilizando average linkage y por el otro se obtiene otro dendrograma utilizando single linkage. Responda las siguientes dos preguntas:

a) Suponga que en cierto paso en el dendrograma que utiliza average linkage, un cluster formado únicamente por las observaciones {A, B} y un cluster formado únicamente por las observaciones {C, D, E} se fusionan. A su vez, en el dendrograma que utiliza single linkage, también ocurre que en cierto paso un cluster formado únicamente por las mismas observaciones {A, B} y otro formado únicamente por las mismas observaciones {C, D, E} se fusionan. ¿Cuál de las dos fusiones ocurrirá más arriba en el dendrograma? ¿O ocurrirán a la misma altura? ¿O no se dispone de información suficiente para responder? Justifique su respuesta

b) Suponga que en cierto punto en el dendrograma que utiliza average linkage, un cluster formado únicamente por la observación {L} y un cluster formado únicamente por la observación {M} se fusionan a una altura X. Suponga también que en el dendrograma que utiliza single linkage ocurre que en cierto punto un cluster formado únicamente por la observación {R} y un cluster formado por las observaciones {S, T} se fusionan a una altura Y. A su vez, asuma que X es menor a Y ($X < Y$). Si llamamos $d(i, j)$ a la distancia entre la observación i y la observación j, ¿es cierto o falso que $d(R, S) < d(L, M)$? Justifique su respuesta. Si considera que no se dispone de la información suficiente para responder esta pregunta indique qué información adicional se requiere para responderla.

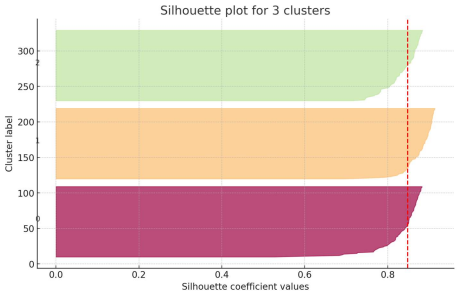
Ejercicio 9.3. ¿Qué resultará de ejecutar K-Means y GMM (en ambos casos con K=2) para cada uno de los siguientes datasets? Justificar.



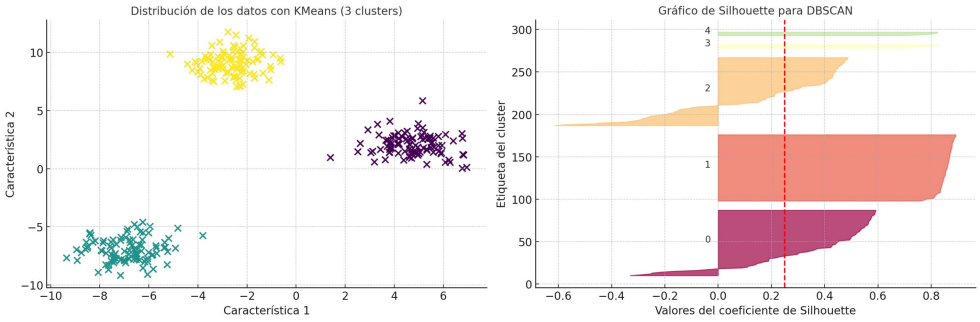
Ejercicio 9.4. Dado un conjunto de datos dividido en clusters, queremos medir qué tan bien están agrupados los puntos en cada cluster. Una forma de hacerlo es utilizando el **Silhouette Score**. Este score mide cuán cerca está un punto de los puntos de su propio cluster comparado con los de otros clusters. La fórmula para el Silhouette Score $s(i)$ para un punto i es:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde:
- $a(i)$ es la distancia promedio entre i y todos los demás puntos en su propio cluster.
- $b(i)$ es la distancia promedio entre i y los puntos del cluster más cercano al que no pertenece.



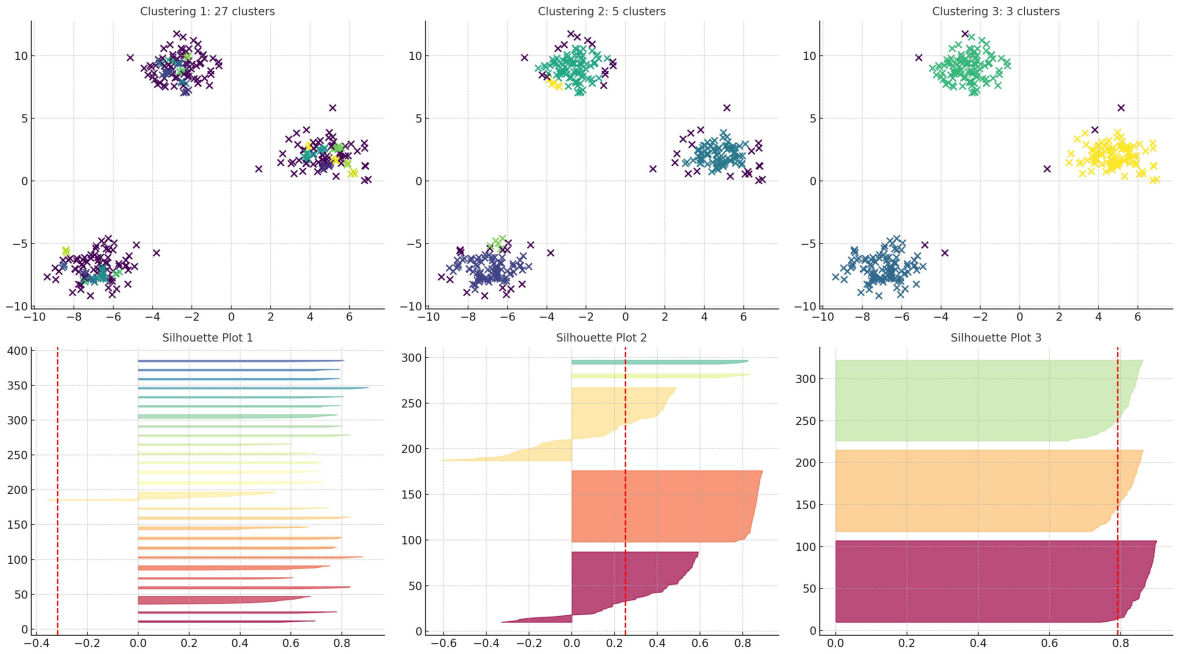
1. Interpreta qué significa cuando el Silhouette Score es cercano a 1, 0, o negativo.
2. En el gráfico se representa el resultado de aplicar K-Means a un dataset, cada barra representa el Silhouette Score de un punto dentro de su cluster. La línea roja punteada es el *Silhouette Score promedio*. ¿Que conclusiones puedes sacar del gráfico?
3. Ahora al mismo conjunto de datos se le aplicó DBSCAN y se obtiene el siguiente gráfico de Silhouette Score, ¿cuál te parece que fue el mejor clustering?



4. Finalmente, se quiere mejorar el resultado obtenido por DBSCAN, y para eso se prueban tres combinaciones diferentes y se reportan los siguientes resultados para los valores de `epsilon` y `min_samples`:

   - `epsilon` $= 1$, `min_samples` $= 10$
   - `epsilon` $= 0.1$, `min_samples` $= 3$
   - `epsilon` $= 0.5$, `min_samples` $= 5$

   ¿Qué rol jugaban dichos hiperparámetros en DBSCAN? ¿Cuál combinación corresponde a cuál asignación de clustering y por qué?



Ejercicio 9.5. Dada la densidad de un modelo de mezcla de gaussianas p-dimensional (i.e. $x \in \mathbb{R}^p$):

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x)$$

donde $g_k = \mathcal{N}(\mu_k, \mathbf{I} \cdot \sigma^2)$ y $\pi_k \geq 0$ con $\sum_{k=1}^{K} \pi_k = 1$. Aquí $\{\mu_k, \pi_k\}$, $k = 1, \ldots, K$ y σ^2 son parámetros desconocidos. Supongamos que tenemos datos $x_1, x_2, \ldots, x_N \sim g(x)$ y queremos ajustar el modelo de mezcla.

1. Escribir la función de log-verosimilitud de los datos.
2. Derivar un algoritmo EM para calcular las estimaciones de máxima verosimilitud.
3. Mostrar que si σ tiene un valor conocido y tomamos $\sigma \to 0$, entonces en un sentido este algoritmo EM coincide con K-means.

Ejercicio 9.6. En el `notebook_7_clustering` se construye un dataset de ejemplo, ejecuta K-Means y grafica los resultados:

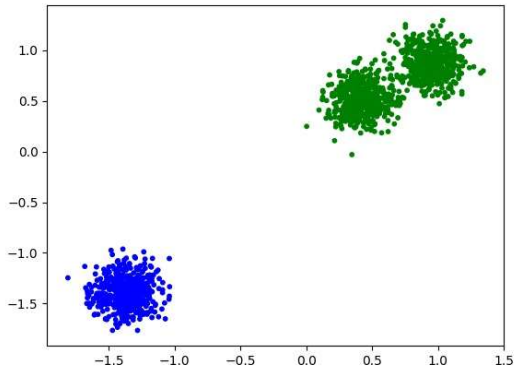
(a) Experimentar con diferentes valores para `cluster_std` (desvío estándar de las nubes de puntos) y `n_clusters` (valor de K en K-Means).

(b) Generar datasets con otras formas, como por ejemplo:[4]

```
datasets.make_circles(n_samples=N, factor=.5, noise=.05)
datasets.make_moons(n_samples=N, noise=.05)
```

Ejecutar K-means, y también DBSCAN con el comando `cluster.DBSCAN(eps=.2)`.
Experimentar con diferentes valores para `noise` y `eps`.

[4] Ver más ejemplos en `http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html`



Ejercicio 9.7. Completar las implementaciones del `notebook_7_clustering` de:

(a) K-Means.

(b) Clustering Jerárquico Aglomerativo.

Comparar cómo funcionan sus implementaciones sobre los datos del ejercicio anterior.

Ejercicio 9.8. Completar la implementación de Expectation Maximization para GMMs de dimensión 1 del notebook `notebook_7_clustering`. Guiar el algoritmo en las fórmulas detalladas en el notebook.

(a) Completar la función `inicializacion`. Para este punto recomendamos utilizar K-Means (de sklearn o propio) como algoritmo para encontrar las medias iniciales.

(b) Completar la función `e`.

(c) Completar la función `m`.

Ejercicio: Ajuste de GMM en $\mathbb{R}^1$

Completar el pseudocódigo del método $.fit(X)$ (ajustar) para Mezclas de Gaussianas (GMMs) para instancias en R^1 (ej: altura de las personas). Suponer definida $pdf(x; \mu, \sigma)$ que devuelve la probabilidad de x para una normal con media μ y desvío σ.

Algorithm: Ajuste de GMM en R^1

Input: X $\in R^{N\times 1}$ y cantidad de componentes: K, tolerancia: tol, max iteraciones: max_iter
Output: Parámetros $\{\pi_k, \mu_k, \sigma_k\}$ para $k = 1, \ldots, K$

Inicializar $\mu_k =$ **{(a) COMPLETAR, indicar 2 maneras}** para $k = 1, \ldots, K$
Inicializar $\sigma_k = \text{std}(X)/K$ para $k = 1, \ldots, K$
Inicializar $\pi_k =$ **{(b) COMPLETAR}** para $k = 1, \ldots, K$
while iteración $<$ max_iter **and** Δlog-likelihood $>$ tol **do**
  paso E:
    for $i = 1, \ldots, N$ y $k = 1, \ldots, K$ **do**
      Actualizar las responsabilidades $\gamma_{ik} =$ **{(c) COMPLETAR}**
    end for
  paso M:
    $N_k = \sum_{i=1}^{N} \gamma_{ik}$
    Actualizar $\pi_k =$ **{(d) COMPLETAR}**
    Actualizar $\mu_k =$ **{(e) COMPLETAR}**
    Actualizar $\sigma_k = \sqrt{\frac{\sum_{i=1}^{N} \gamma_{ik} \cdot (x_i - \mu_k)^2}{N_k}}$
  Calcular log-likelihood: $\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \cdot \text{pdf}(x_i; \mu_k, \sigma_k)\right)$
end while
Retornar $\{\pi_k, \mu_k, \sigma_k\}$