Subject: Probability and Statistics

Class XX: Clustering: Gaussian Mixture Models

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Limitaciones de K-means

UdeA

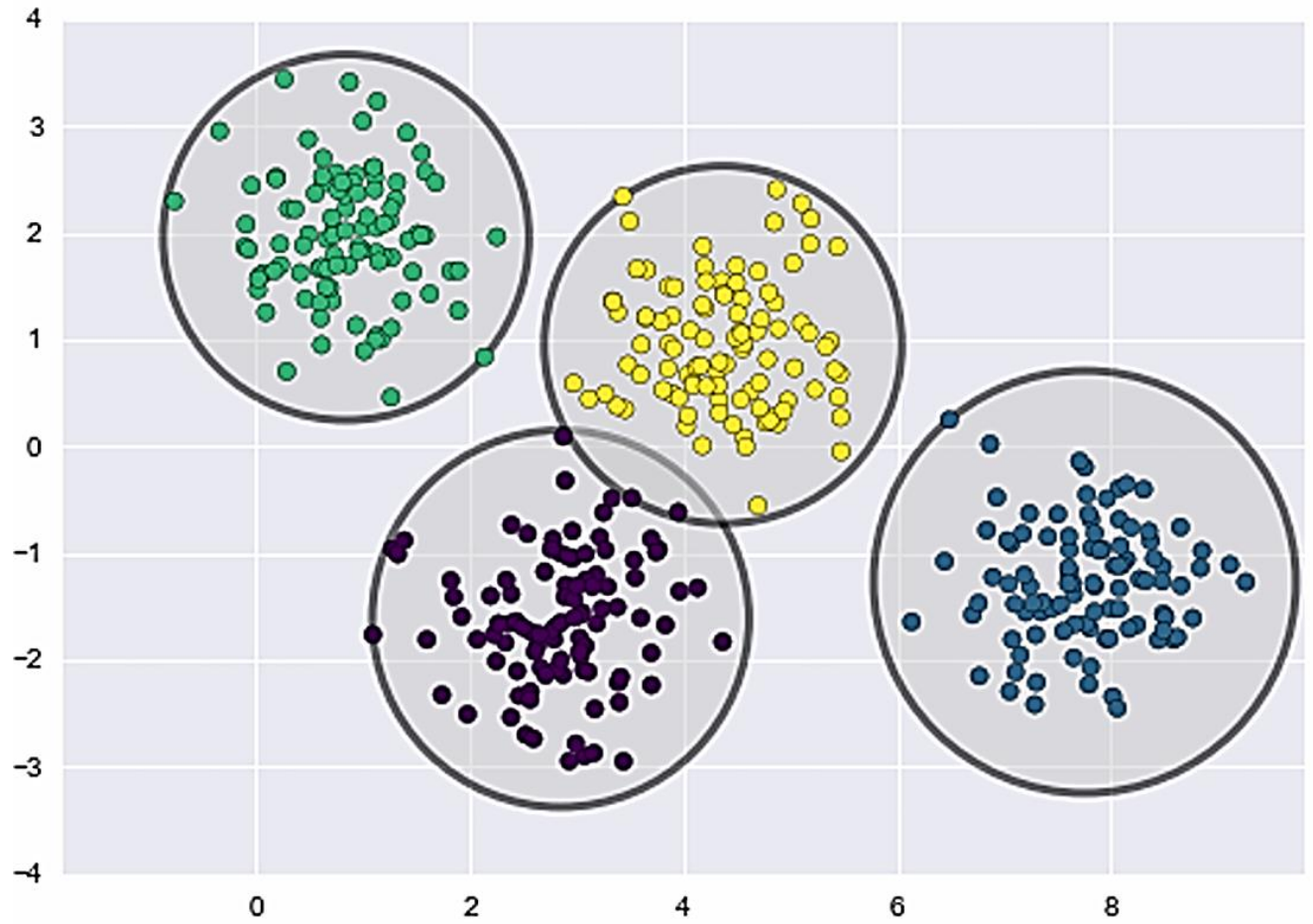
Por la manera en que K-means encuentra los *clusters*, este tiene ciertas limitaciones. Primero este no es un método probabilístico, lo cuál lo hace muy poco flexible, comparado con uno donde tuviéramos probabilidades y donde pudiéramos modificar el umbral de probabilidad de pertenecer a un *cluster* u otro

Por otra parte, dado que los *cluster* se forman minimizando la distancia al centro del grupo, la región de pertenencia al *cluster*, en 2D sería un círculo, y en más dimensiones en general una hiperesfera.

Limitaciones de K-means

UdeA

Por ejemplo, para estos datos, que tienen una distribución circular, funciona muy bien el hecho de que así se generen los *clusters*. Adicionalmente aquí los *clusters* están bien diferenciados, pero cuando hay solapamiento sería conveniente conocer una probabilidad

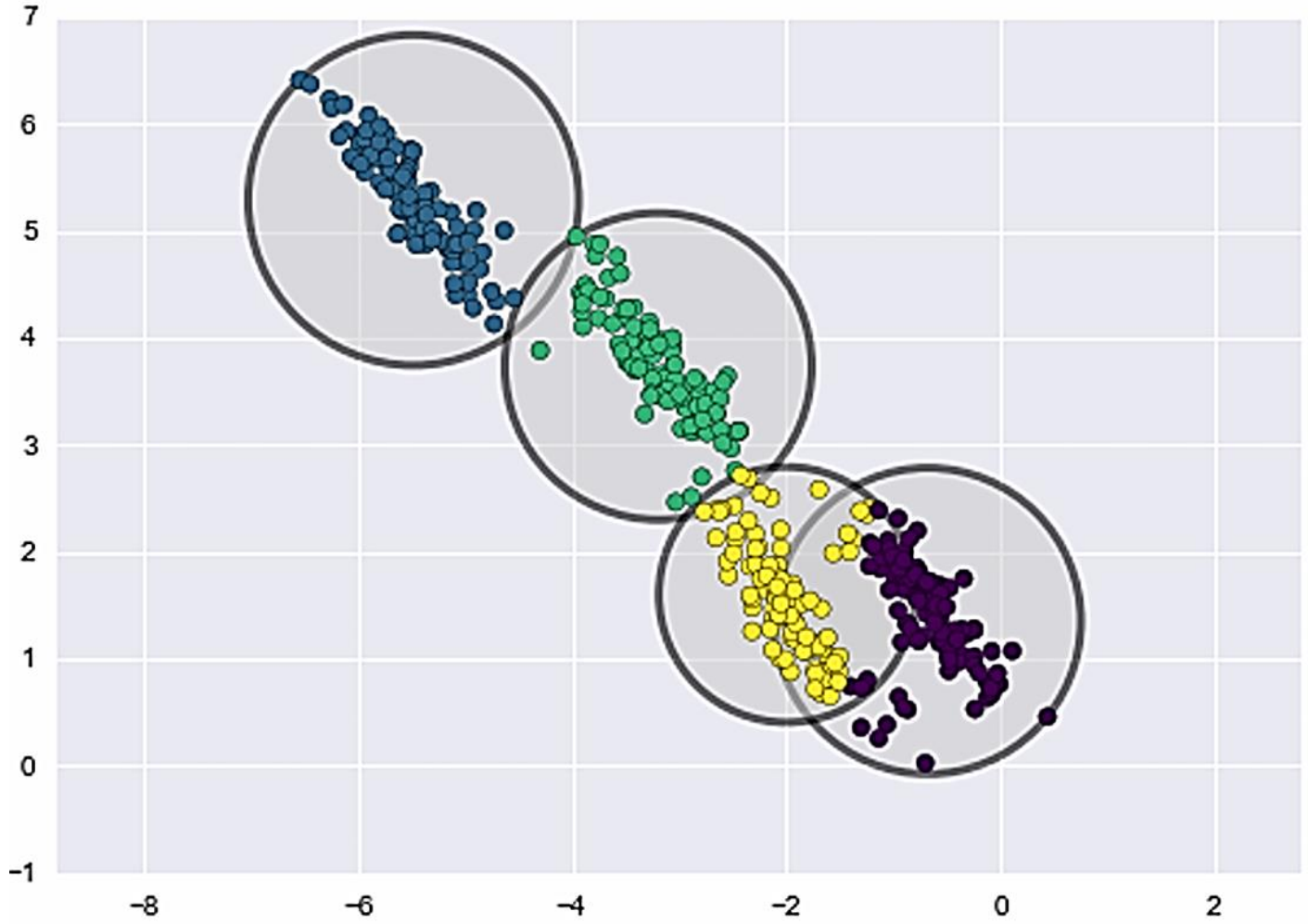


Limitaciones de K-means

UdeA

Ahora, si los datos son así, la forma circular de la región de pertenencia no es tan conveniente, veamos por ejemplo la case amarilla y la violeta.

Entonces si se hace una generalización del modelo, que permita conocer la probabilidad y además que las regiones sean elípticas tenemos los GMM.



Modelos de Mezcla Gausiana

UdeA

Los Modelos de Mezcla Gausiana, lo cuales buscan una mezcla de modelos multidimensionales Gausianos que mejor se ajuste a la distribución de unos datos. En realidad, estos modelos no son modelos de agrupamiento, sino modelos que buscan describir la distribución de ciertos datos multidimensionales como la mezcla de muchos modelos Gausianos. Sin embargo, cada modelo Gausiano independiente se puede entender como un grupo, y de este modo puede ser empleado como un método de agrupamiento.

UdeA

Permite conocer la probabilidad de pertenencia aun grupo.

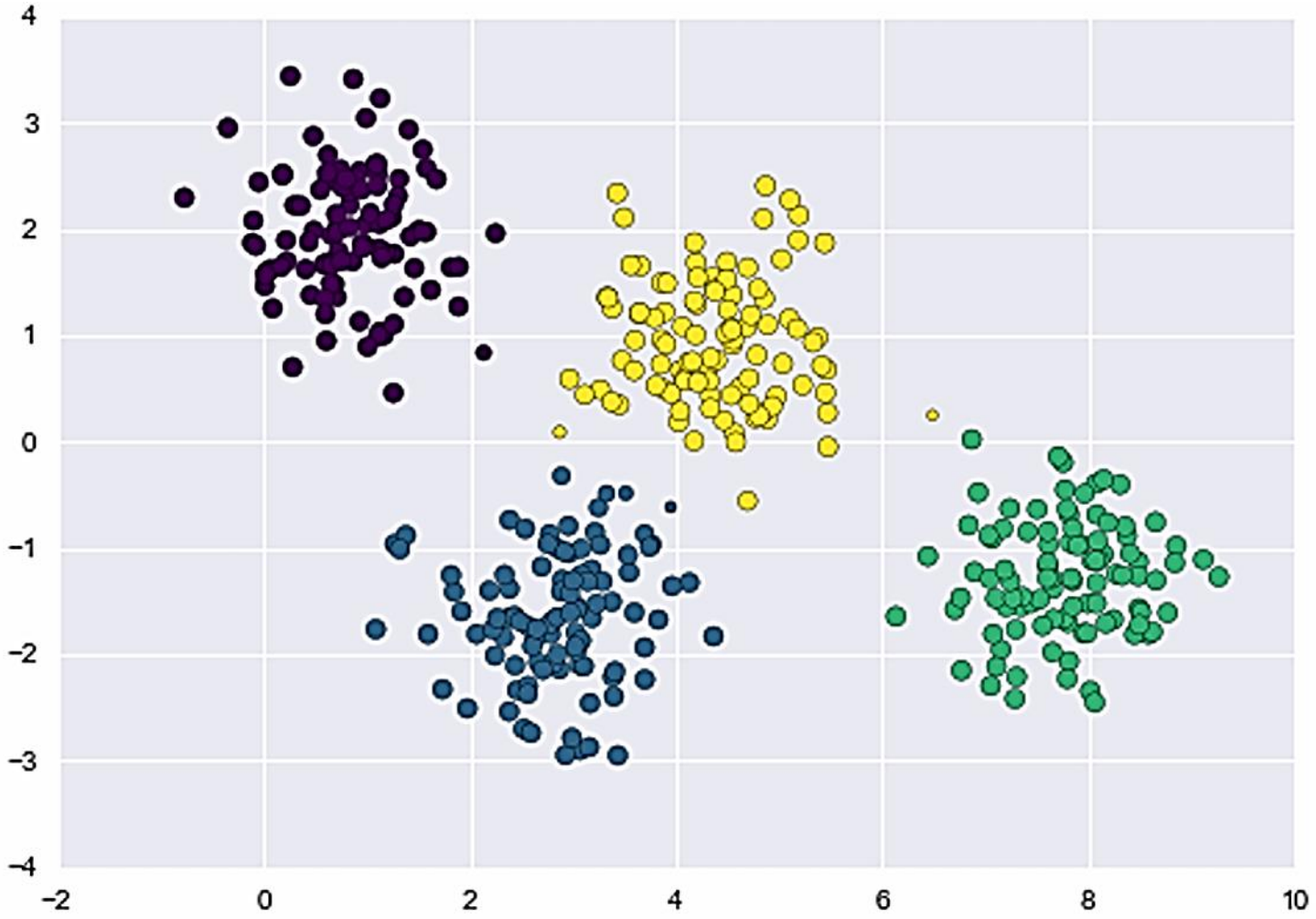
Una ventaja importante frente al agrupamiento k-means, es la posibilidad de conocer la probabilidad que tiene un record de pertenecer a un grupo u otro.

En sklearn, los modelos tienen el método *predict_proba()* para este fin. Lo cual arroja una matriz de [numero de muestras, numero de cluester]

Modelos de Mezcla Gausiana

UdeA

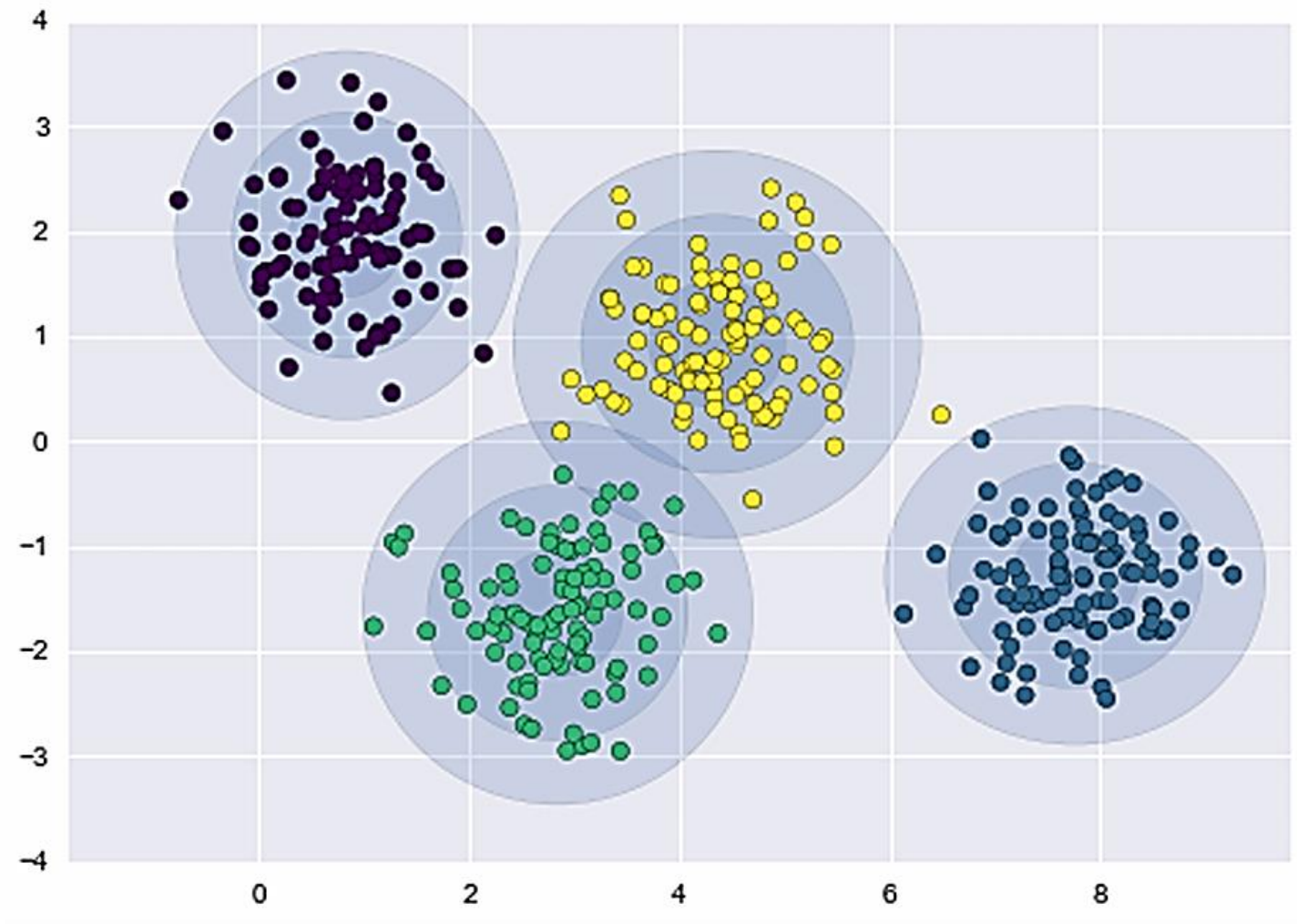
En este gráfico hacemos que el tamaño de los marcadores sea proporcional a la probabilidad, para ilustrar que donde hay cierto solapamiento hay menor certeza.



Modelos de Mezcla Gausiana

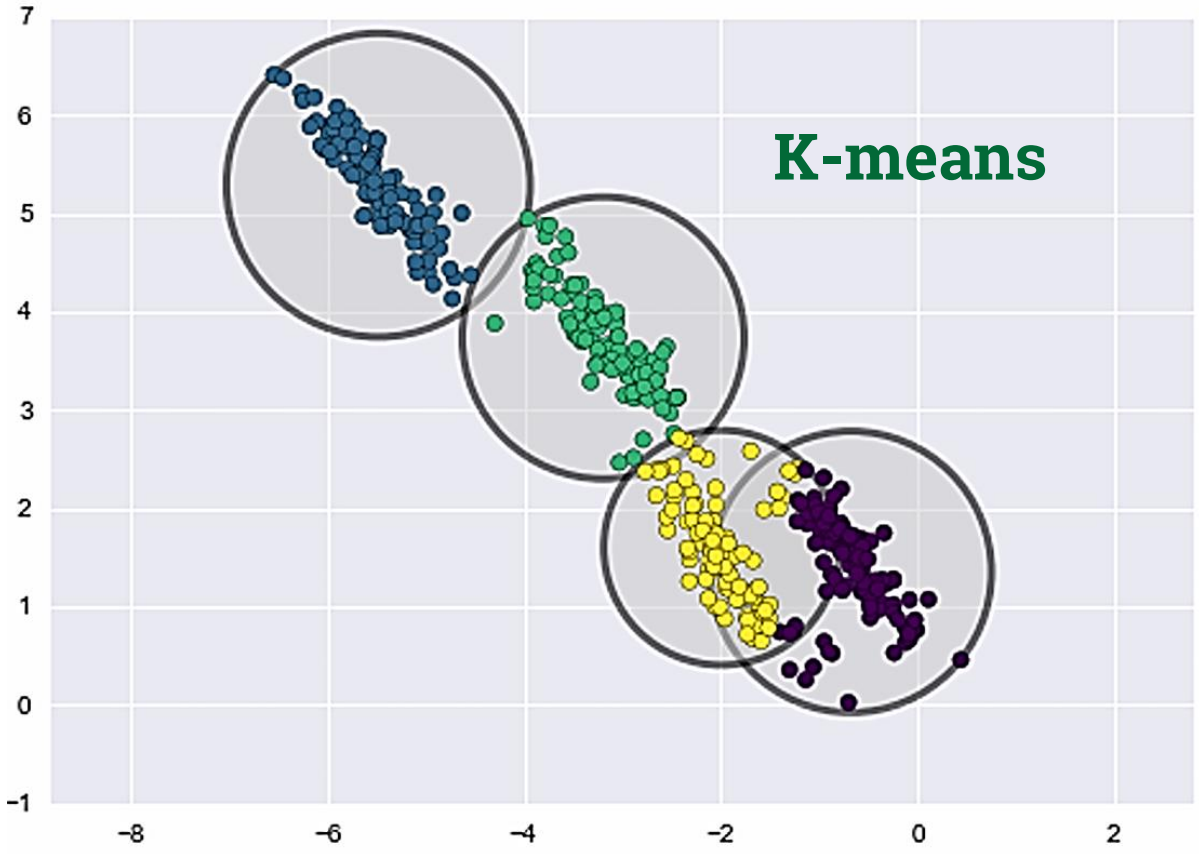
UdeA

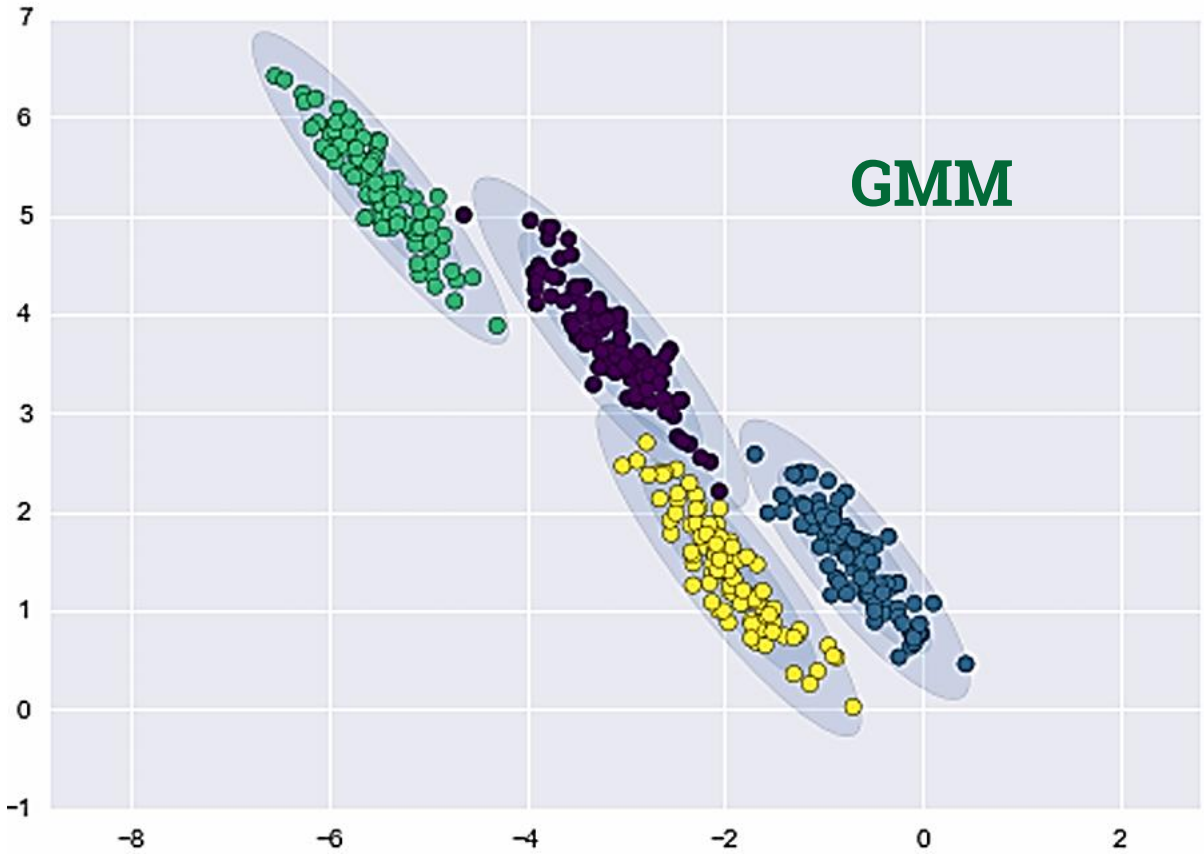
En general, a diferencia de *k-means*, donde cada *cluster* está asociado a una hiperesfera, con bordes bien definidos, en los GMM, cada *cluster* está asociado a un modelo gausiano probabilístico.



UdeA

Veamos la comparación para los datos que no eran esféricos





Modelos de Mezcla Gausiana

UdeA

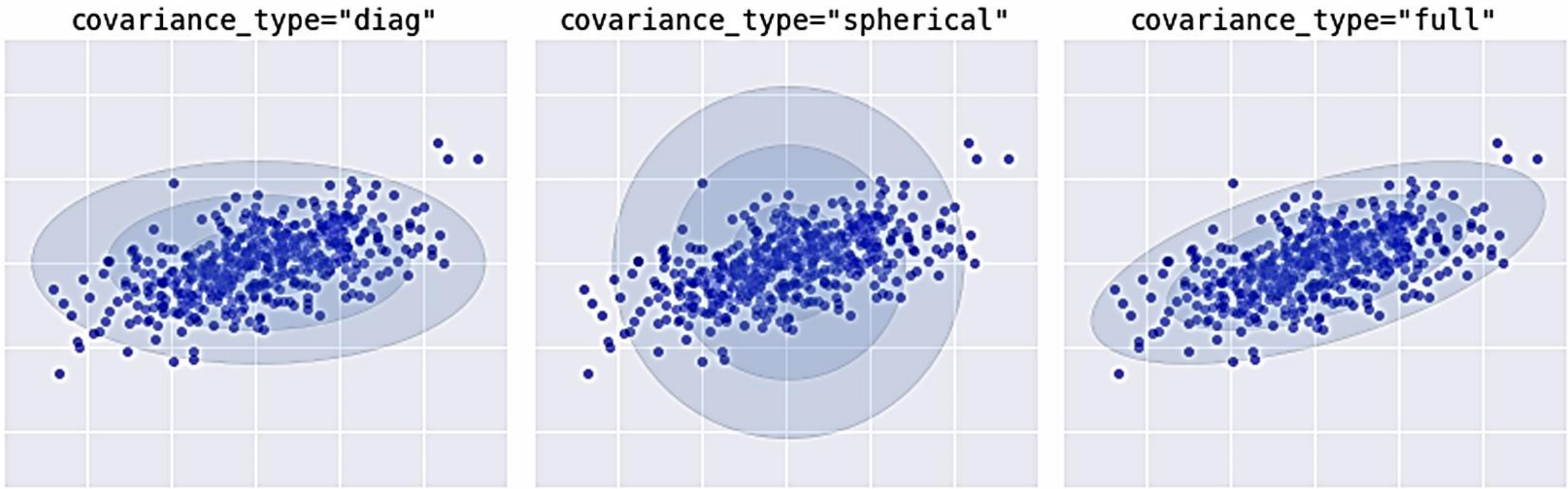
En los modelos GMM hay un hiperparámetro muy importante que en el modelo de sklearn se llama covariance_type

Este hiperperámetro básicamente guía la orientación de los modelos gausianos. Vemos que las proyecciones de estos modelos en 2D son como elipses, entonces en ese caso, este hiperparámetro me restringe la dirección de esas elpises.
Si covariance_type = 'diag', esto quiere decir que los ejes de las "elipses" serán paralelos a los ejes del sistema.
Si covariance_type = 'spherical', tendremos un modelo muy similar a k-means, donde las regiones son hiper-esferas.
Si covariance_type = 'full', da libertad a la orientación de las diagonales de las "elipeses" a costa de un gasto computacional considerablemente mayor.

Modelos de Mezcla Gausiana

UdeA

Veamos los 3 casos



Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering