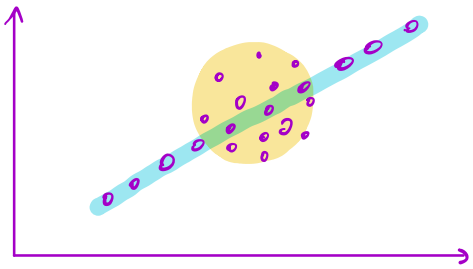
GMM

Supongamos que tenemos el siguiente dataset:

¿Cuáles son mis clusters? ¿Cómo funcionaría K-means en este caso?

Uno podría argumentar que los clusters son:

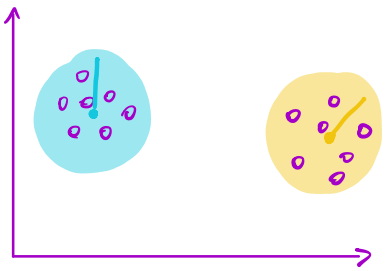


En donde los elementos del medio pertenecerían a dos Clusters.

Ahora vamos a ver una técnica llamada Gaussian Mixture Models (GMM) que la podemos entender como una generalización de K-means que permite:

- Generar clusters de formas "elipsoidales".
- Tener "soft clustering", esto es, tener una probabilidad de pertenencia a cada cluster.
- Detectar outliers.

Ojo cuando hacemos K-means, al computar un centroide, tenemos que ver la distancia de los puntos a los centroides. Intuitivamente, cada centroide "genera clusters esféricos":



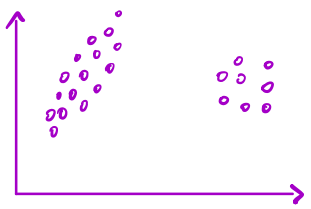
La idea de GMM

Supongamos que tenemos K clusters. GMM plantea que cada cluster esté representado por una distribución gaussiana, y que cada punto fue "creado" de la siguiente manera:

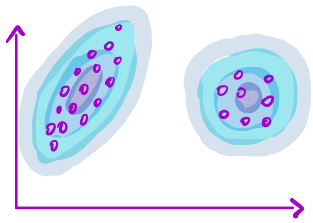
1. Escogemos uno de los K clusters de forma aleatoria. La probabilidad de escoger el cluster j será π_j. Este parámetro se conoce como el "cluster's weight".
2. Tomamos la distribución asociada al cluster j y vemos su media μ_j y su matriz de covarianza Σ_j. Recordemos que en más dimensiones, necesitamos esta matriz para describir la forma de la distribución.
3. Generamos un sample aleatorio x_i donde $x_i \sim N(\mu_j, \Sigma_j)$
4. Hacemos esto para cada x_i con $1 \leq i \leq n$.

Así, para hacer clustering tenemos que aprender las distribuciones que "generaron" nuestros datos.

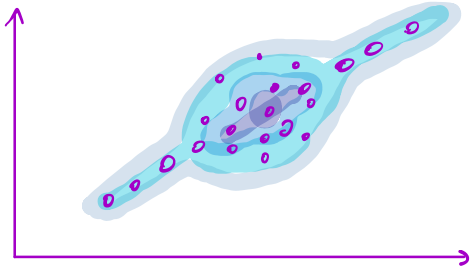
Ej. Consideremos este dataset:



GMM debería encontrar dos distribuciones gausianes con la siguiente forma:



Como vemos, este modelo puede entenderse como une generalización de K-Means. Ahora, respecto al ejemplo del inicio:



Tendríamos dos gausianas que se superponen, y podemos saber la probabilidad de pertenencia de cada punto a cada cluster.

Otros detalles

Al igual que en K-means, necesitamos

entregar el número de clusters que queremos encontrar. Podemos usar las métricas:

- BIC (Bayesian information criterion)
- AIC (Akaike information criterion)

Para encontrar el número de clusters, en donde nos quedamos con el número de clusters que minimice estas métricas.

Además, los puntos que están en zonas de baja densidad de probabilidad pueden ser clasificados como outliers.

Finalmente, este es un método generativo, donde podemos "samplear" nuevos puntos después de conocer las distribuciones.

Sobre el entrenamiento

Para entrenar el modelo (i.e. aprender los valores de μ_j, Σ_j y π_j) se usa el algoritmo EM (Expectation Maximization), que al igual que en K-means, parte con valores aleatorios y va convergiendo en distintas iteraciones.