Clustering: K-means y Gaussian Mixture Model

Martín Gutiérrez

August 7, 2022

Aprovechando lo que llevamos...

Como les mencioné la clase pasada, vamos a aprovechar lo que hemos aprendido sobre métodos estadísticos para enunciar los primeros algoritmos de clustering que veremos en el curso... wait... Qué es clustering?

Clustering significa literalmente "agrupación". Consiste en asociar datos de características similares en grupos.

En esta presentación hablaremos de dos algoritmos de clustering: K-Means y Gaussian Mixture Models (GMMs).

Enunciado del algoritmo K-means

El algoritmo de clasificación **K-means** consiste en la agrupación de datos bajo lo que se llama un "Cluster" o grupo.

Consta de los pasos siguientes:

1. Colocar K puntos en el espacio de datos representado por los objetos a ser clasificados. Estos puntos se denominarán los *centroides inciales*.
2. Asignarle a cada objeto (dato, tupla) un grupo, representado por el *centroide* más cercano.
3. Una vez asignados todos los objetos a sus grupos respectivos, se recalcula la posición del *centroide* del grupo.
4. Se repiten los pasos 2 y 3 hasta que los *centroides* no cambien de posición.

Limitaciones

Desgraciadamente, el algoritmo no asegura una clasificación óptima de los objetos, vale decir un mínimo global de la función objetivo (puede ser varianza, distancia, etc.).

Asímismo, el algoritmo es muy sensible a la distribución inicial de los *centroides* por lo que una recomendación al ocupar este algoritmo es que se ejecute múltiples veces seleccionando distintos *centroides iniciales*.

A continuación se mostrará un ejemplo de cómo funciona el algoritmo.

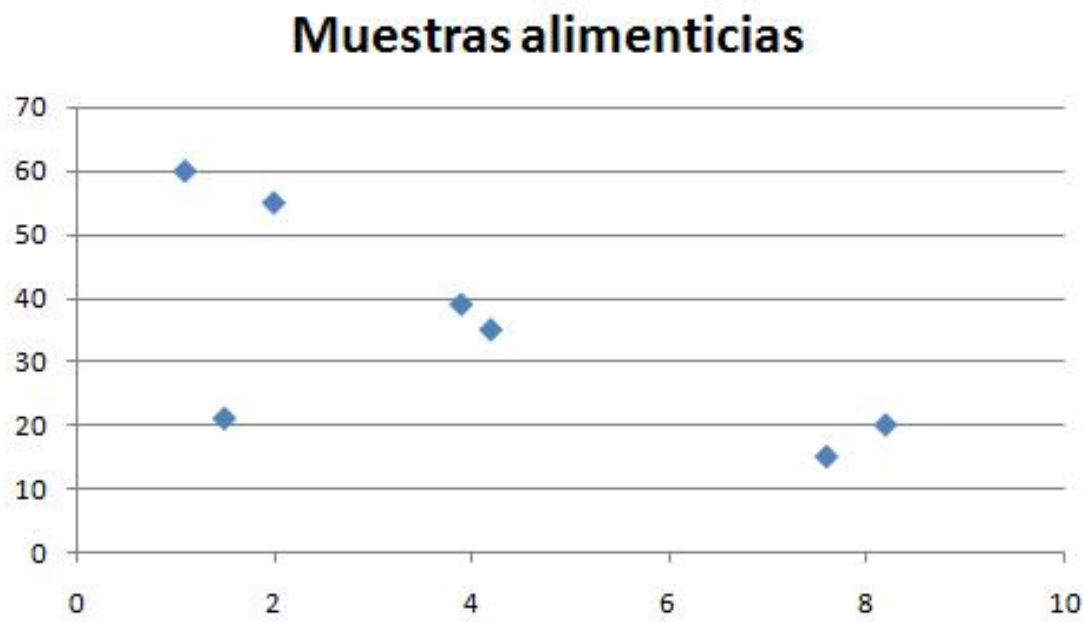
Ejemplo (I)

Supongamos los siguientes datos:

Muestra alimenticia	Contenido protéico, P	Contenido graso, G
Muestra 1	1.1	60
Muestra 2	8.2	20
Muestra 3	4.2	35
Muestra 4	1.5	21
Muestra 5	7.6	15
Muestra 6	2.0	55
Muestra 7	3.9	39

Ejemplo (II)

Representados en el plano, los datos se verían:



Del gráfico, vemos que las distancias son las mayores entre el dato 1 y los datos 2, 3, 4 y 5, entre el dato 2 y los datos 3 y 4 y entre los datos 3 y 4.

Ejemplo (III)

Así pues, seleccionamos los datos 1, 2, 3 y 4 como centroides iniciales, que forman 4 clusters:

Cluster	Contenido protéico, P	Contenido graso, G
C1	1.1	60
C2	8.2	20
C3	4.2	35
C4	1.5	21

De aquí, se asignan los puntos restantes al grupo cuyo centroide esté más cercano a ellos.

Ejemplo (IV)

En particular, el dato 6 está cercano al dato 1, el dato 5 está cercano al dato 2 y el dato 7 está cercano al dato 3. El centroide del cluster C4 queda solo puesto que ningún dato "restante" está cerca de él.

Entonces, se recalculan los centroides:

El cluster C1 pasará a llamarse ahora C16 debido a que está ahora compuesto del dato 1 y el dato 6. El nuevo centroide de C16 es:
$P = (1.1 + 2.0)/2 = 1.55$, $G = (60 + 55)/2 = 57.50$

Del mismo modo aparecen los clusters C25 y C37 con los respectivos centroides (7.9, 17.50) y (4.05, 37).

Ejemplo (V)

Finalmente, se ha llegado a un punto fijo y los clusters quedan:

Cluster	Contenido protéico, P	Contenido graso, G
C16	1.55	57.50
C25	7.9	17.50
C37	4.05	37
C4	1.5	21

Aquí finaliza el algoritmo, dado que no hay cambios al hacer asignación de puntos a los nuevos centroides.

Este ejemplo era bastante simple y no requirió un gran análisis de la cercanía de los datos. No obstante, se recomienda que cuando el análisis no sea simple, se use la distancia Euclidiana entre datos.

Otras limitaciones de K-means

Es importante notar que K-Means no es un algoritmo probabilístico. Además, siguiendo la recomendación, se usa muchas veces la distancia euclidiana entre datos. Si bien esto presenta un escenario simple, a veces puede ser poco práctico para situaciones de la vida real.

Además, intuitivamente y de forma geométrica, otra limitación más de K-Means es que ubica una hiper-esfera centrada en el centroide y alrededor de los datos que pertenecen a un grupo. Esto es rígido, en el sentido que solo admite hiper-esferas y no otras formas geométricas (ejemplo de grupos alargados). Esto eventualmente puede llevar a traslape en los grupos.

GMMs

Las GMMs pueden verse como una extensión de K-Means que incluye una componente probabilística y ataca la limitación de la forma del grupo. Específicamente, enuncia que todo dato se genera a partir de una mezcla de un número finito de distribuciones Gaussianas con parámetros desconocidos (suena a un déjà vu?).

Concretamente, ese supuesto implica que además se conoce información referente a la estructura de covarianza en los datos y el centro asociado a las distribuciones Gaussianas latentes. Una interpretación de la información adicional es que se pueda categorizar datos con cierta incertidumbre.

Propuestas:

1. En vez de hacer una comparación directa del dato con un solo centroide, comparar con todos los centroides.
2. Alterar las formas que engloban los grupos.

EM, el retorno...

Si nos fijamos bien en las propuestas, si es que decimos que cada centroide es un μ y que la forma que engloba los datos viene dada por una combinación de su centro (μ) y la distancia a él (σ), estaríamos en presencia de distribuciones normales desconocidas con etiquetas inciertas para los datos (pertenencia al grupo). Eso huele a EM, no?

Concretamente, EM se efectúa en este escenario como:

1. Adivinar ubicaciones y formas de los grupos
2. E: Para cada dato, encontrar pesos que codifiquen la probabilidad de pertenencia a cada grupo
3. M: Para cada grupo, actualizar su ubicación, normalización y forma basado en los datos y haciendo uso de los pesos encontrados previamente
4. Repetir pasos 2 y 3 hasta que ocurra convergencia

GMMs es K-Means on crack?

Sí. Se le tilda de algoritmo de clustering, aunque en estricto rigor es un estimador de densidad. Esto es, busca representar una mezcla de distribuciones que se ajuste bien a los datos de input del algoritmo.

Así entonces, es posible ejecutar múltiples veces el algoritmo con distintos K y revisar métricas que indiquen la mejor combinación de distribuciones para representar los datos.

Más clustering incoming...

Durante la siguiente clase, vamos a ver otros algoritmos más de clustering como AHC, Mean Shift y DBSCAN.