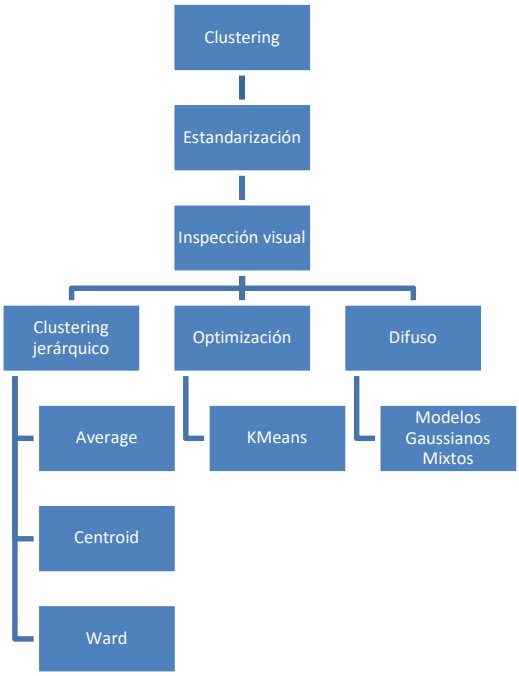
Análisis Clúster

Clustering (análisis de conglomerados) se refiere a la tarea de agrupar un conjunto de objetos de tal forma que se obtengan grupos (clústeres) lo más *similares* entre sí. Al detenernos un momento a reflexionar lo anterior se nos presenta la pregunta ¿qué es similar? Más adelante intentaremos clarificar esta pregunta, de momento revisemos la definición propuesta por Everitt:

"*El análisis clúster es un conjunto de métodos para construir (con suerte) una clasificación sensible e informativa de un conjunto de datos sin clasificar usando los valores de las variables observados en cada individuo*"

El objetivo del análisis clúster es partir un conjunto de datos en grupos, si bien esto suena simple, llevarlo a cabo es una compleja tarea.

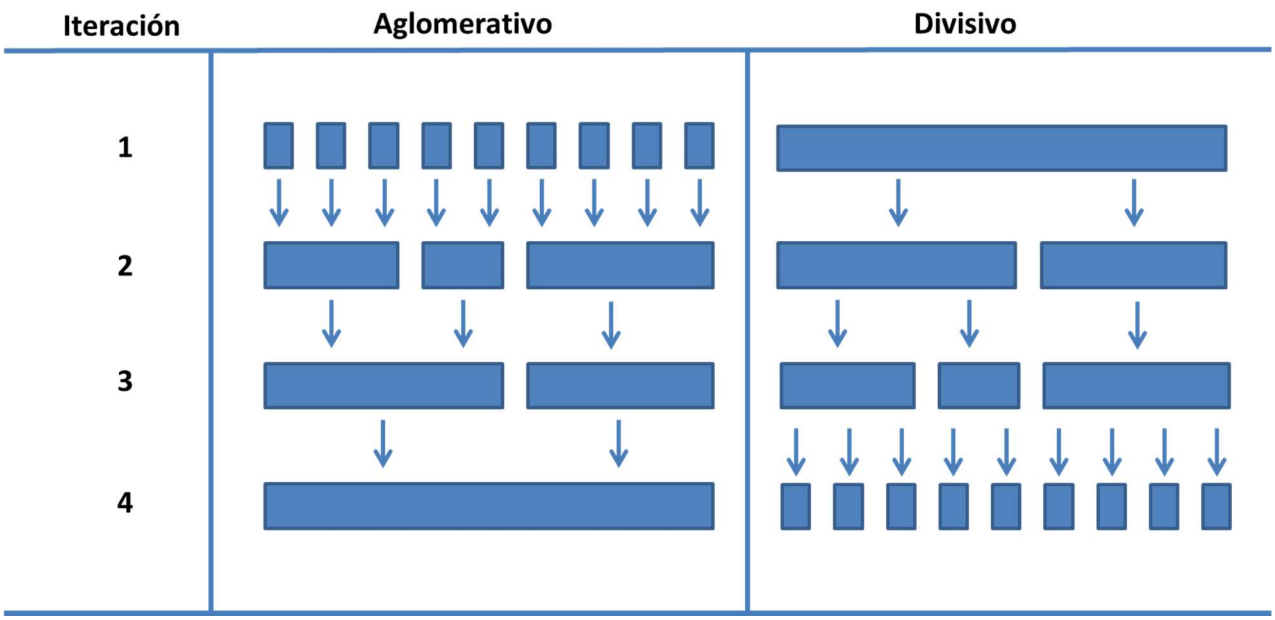
La estructura de nuestro estudio en el tema para este curso puede resumirse como sigue:



Dentro de la ciencia de datos, el análisis clúster ingresa en la categoría de aprendizaje no supervisado debido a la ausencia de variable objetivo.

Introducción a Clustering Jerárquico

La técnica consiste en crear clústeres que están jerárquicamente anidados dentro de los clústeres en iteraciones previas y se clasifican en dos tipos: Aglomerativo y divisivo.



Los métodos aglomerativos funcionan como sigue:

1. Asigna cada observación a su propio clúster
2. Computa la similitud entre cada clúster
3. Fusiona los dos clústeres más similares
4. Repite 2 y 3 hasta formar un solo clúster

Los divisivos:

1. Asignar todas las observaciones a un solo clúster
2. Computar la similitud entre cada uno de los clústeres
3. Particionar los dos clústeres menos similares
4. Repetir 2 y 3 hasta que cada observación pertenezca a su propio clúster

Si bien el clúster jerárquico es la columna vertebral del análisis clúster, no está exento de problemas:

- No escala bien con el número de observaciones (propaga el error al cuadrado o cubo del número de observaciones)
- Una vez que se ha hecho una división, ésta es irrevocable
- Ningún método en particular destaca de otro, todos tienen ventajas y desventajas

Ahora revisaremos la cuestión fundamental en el análisis clúster, la similitud.

Similitud

Es clave en nuestro estudio pensar en la similitud, aunque a menudo es muy difícil cuantificarla. Por ejemplo, ¿qué es más similar a un pato? ¿Un cuervo o un pingüino? La respuesta es tan ambigua como la pregunta: a ambos.

Para que una métrica de similitud sea satisfactoria, debe cumplir con los siguientes principios:

1. Simetría $d(x,y) = d(y,x)$
2. Distinguibilidad no idéntica, si $d(x,y) \neq 0 \Rightarrow x \neq y$
3. No distinguibilidad idéntica, si $d(x,y) = 0 \Rightarrow x = y$

Medidas de similitud

- Distancia euclídea

$$D_E = \sqrt{\sum_{i=1}^{d} (x_i - w_i)^2}$$

- Distancia cityblock (Manhattan)

$$D_M = |x_i - w_i|$$

- Métrica de Minkowski

$$\mathrm{D_{M\lambda}} = \left(\sum_{i=1}^{d} |x_i - w_i| \right)^{1/\lambda}$$

Además de las anteriores, revisaremos las métricas de similitud basadas en densidad.

Un clúster puede verse como un área de densidad de observaciones incrementada. La similitud es una función de la distancia entre las burbujas de densidad identificadas (hiper-esferas).

La densidad en el punto i está dada por:

$$\widehat{f}_i = \frac{n_i}{nV_i}$$

Donde n_i es el número de observaciones dentro del radio de la hiperesfera centrada en i, n es el número total de observaciones y V_i es el volumen de la i –ésima hiperesfera de radio r y d dimensiones dado por:

$$V(d) = \frac{\pi^{d/2} r^d}{\Gamma\left(d/2 + 1\right)}$$

Representación visual

Desafortunadamente, en ciencia de datos es muy poco probable encontrar un problema bidimensional por lo que recurriremos a dos artificios para encontrar la forma de los clústeres en el hiperespacio.

Cuando los datos presentan más de dos dimensiones, los datos multivariados pueden ser convertidos a un sistema bidimensional que fungirá como sistema coordenado. Para tal propósito, nos valdremos de dos técnicas:

- Análisis de componentes principales (PCA)
- Escalamiento multidimensional (MDS)

PCA

Como habíamos revisado anteriormente, el análisis de componentes principales es un método para transformar un conjunto de variables en un nuevo conjunto de variables ortogonales conservando la mayor cantidad de varianza posible contenida en los datos. Cuando trabajamos en el hiperespacio, las primeras 2 componentes principales pueden ser utilizadas como ejes para ayudarnos a localizar visualmente los clústeres.

MDS

Es una técnica matemática de reducción de dimensiones que mapea las distancias entre las observaciones en un espacio de más alta dimensión en otro de más baja. MDS localiza las n observaciones un espacio dimensional reducido tal que las diferencias entre pares de puntos en el espacio dimensional más bajo coinciden lo más cercano posible. MDS no hace ningún supuesto estadístico pero es vulnerable a mínimos locales. En particular MDS minimiza la siguiente función de error:

$$E = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1}\frac{\left(D_{ij}-d_{ij}\right)^2}{D_{ij}}}{\sum_{i=1}^{n}\sum_{j=1}^{i-1}D_{ij}}$$

Donde D_{ij} es la distancia entre la i-ésima y la j-ésima observación en el espacio original mientras de d_{ij} es la distancia entre el mismo par de puntos pero en el espacio reducido.

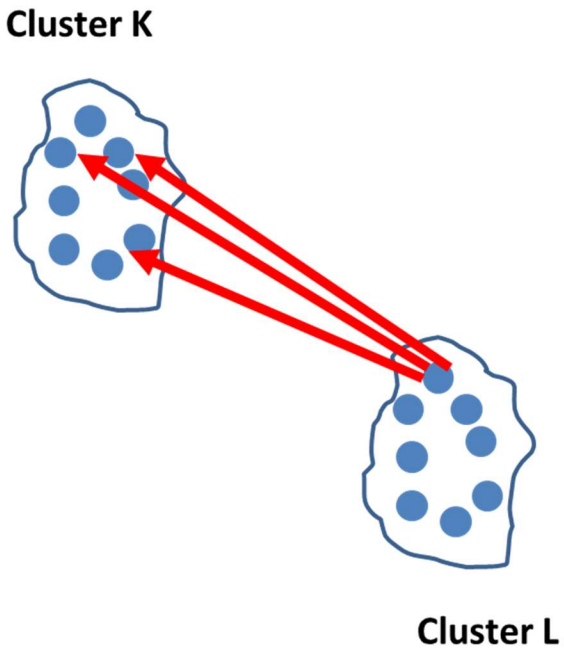
Clustering Jerárquico

Los métodos jerárquicos son la columna vertebral del análisis clúster. Parte de su popularidad recae en que no son afectados por el llamado "efecto de orden" aunado a que no se necesita intuir a priori cuántos clústeres están presentes en los datos. Los métodos jerárquicos no escalan bien con el número de observaciones (el error/carga de cómputo escala al cuadrado o cubo). No existe un método considerado como el "mejor", esto no quiere decir que todos los métodos sean equivalentes, simplemente poseen distintas características.

Ahora centrémonos en los métodos de clasificación, en nuestro estudio revisaremos 3: Average, Centroid y Ward.

Average

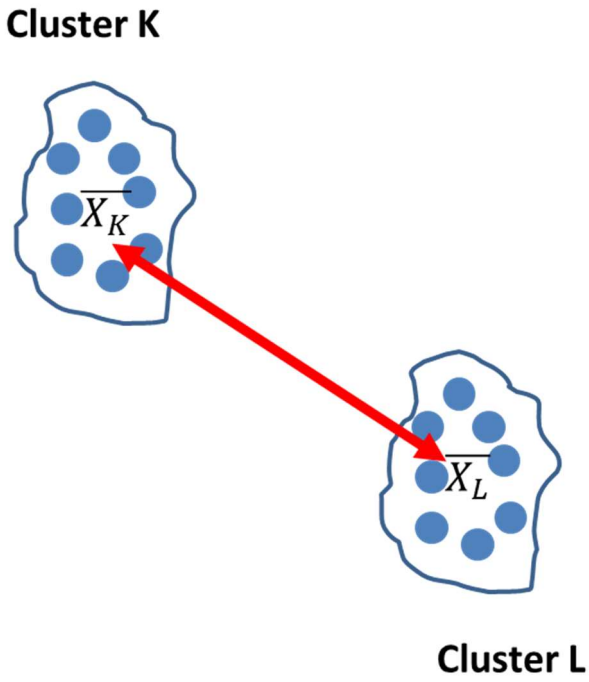
La distancia entre clústeres es la distancia promedio entre pares de observaciones. El método tiende a unir clústeres con varianzas pequeñas y sesga a producir clústeres con varianza igual. Otra característica es que es menos influenciado por valores atípicos que la mayoría de los métodos, además computacionalmente es más rápido que la mayoría de los métodos.



$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

Centroid

La distancia entre los clústeres es la distancia euclídea cuadrada de cada uno de los centroides de los clústeres.

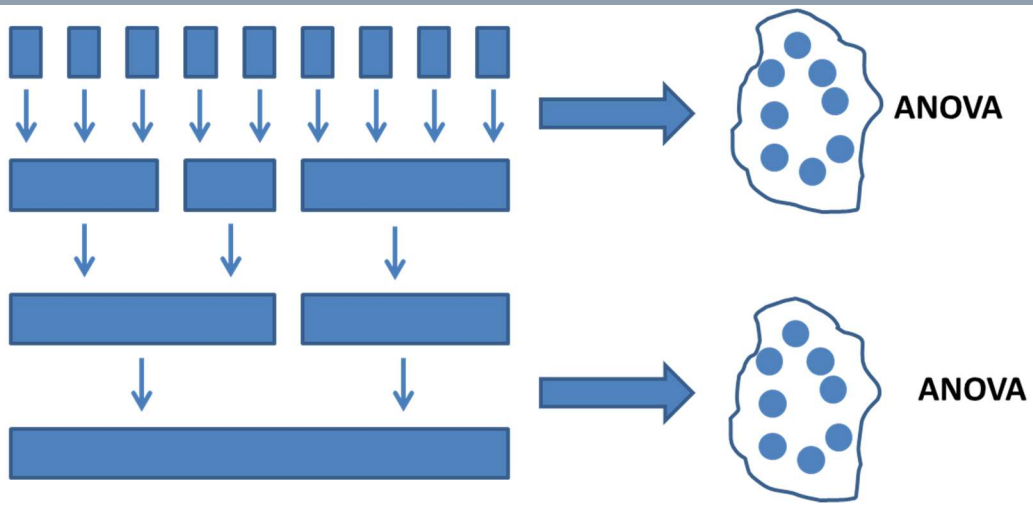


$$D_{KL} = \|\overline{X_K} - \overline{X_L}\|^2$$

En este método los datos extremos tampoco afectan demasiado, se trabaja directamente con los datos coordenados y si los grupos a fusionar son de tamaños muy distintos, el centroide del nuevo grupo estará muy cerca del centroide del grupo más grande.

Ward

El método de la mínima varianza de Ward une clústeres tales que en cada generación es minimizada la suma de cuadrados dentro del clúster sobre todas las particiones obtenidas al fusionar dos clústeres de cada generación previa. En este método se realiza un análisis de varianza (ANOVA) en cada fusión de la jerarquía y tiende a unir clústeres con un número pequeño de observaciones sesgando a producir clústeres esféricos con aproximadamente el mismo tamaño. Es muy sensible a datos atípicos.



$$D_{KL} = \frac{\|\overline{x_K} - \overline{x_L}\|^2}{\left(\frac{1}{n_k} + \frac{1}{n_L}\right)}$$

Donde $\overline{x_K}$ es el vector de medias y n_k es el número de observaciones en del clúster k respectivamente.

Clustering de optimización

El clustering de optimización particiona un conjunto de datos al optimizar algún criterio específico (función de error). A diferencia de los métodos jerárquicos, el error escala linealmente con el número de observaciones por lo que estos métodos son útiles en grandes conjuntos de datos. El número de particiones debe especificarse a priori.

En teoría se puede generar y calificar cada posible partición de tamaño n en g grupos, asegurando así la obtención de un mínimo global. Problemas no triviales ocasionarían una explosión combinatoria de particiones insostenible, por tanto, se desarrollaron algoritmos de búsqueda heurística que realizan el trabajo de forma eficiente en el espacio de particiones potenciales usando la siguiente metodología:

1. Encontrar una partición inicial de n observaciones en g grupos (clústeres)
2. Calcular el cambio en el criterio de clusterización producido al mover cada observación de su grupo actual a otro grupo haciendo el cambio que proporcione la máxima mejora en el criterio
3. Repetir 2 hasta que los movimientos no produzcan mejora.

Desafortunadamente, lo anterior no garantiza la obtención de un óptimo global, en la práctica, aplicar lo anterior a una muestra diferente o incluso al mismo conjunto en distinto orden puede resultar en soluciones clúster completamente diferentes.

El objetivo de la clusterización de optimización es maximizar o minimizar algún criterio específico tal como:

- Separación
- Homogeneidad intra-clúster

La variación entre clústeres se define como la suma de las disimilitudes entre las observaciones en el clúster y las observaciones fuera de él. De manera más formal, dada alguna medida δ_{m_i,m_j} de la disimilitud entre las observaciones i y j en el clúster m es:

$$\sum_{i=1}^{n_m} \sum_{k \neq m} \sum_{j=1, j \neq i}^{n_k} \left(\delta_{m_i,m_j} \right)^r$$

Donde n_m es el número de casos en el clúster m y n_k es el número de casos en el clúster k.

La distancia inter-clúster define la separación entre los clústeres como la mínima disimilitud entre un caso en el clúster m y uno fuera

$$min \left(\delta_{m_i,m_j} \right)^r$$

Homogeneidad

El criterio está basado en grupos que tienen una estructura relativamente cohesiva midiendo la similitud de observaciones en el mismo clúster, dos ejemplos son:

- Variación dentro del clúster: suma de disimilitudes entre dos observaciones en el mismo clúster, puede usarse para reflejar la cohesión del clúster

$$h_1(m) = \sum_{i=1}^{n_m} \sum_{j=1, j \neq i}^{n_m} \left(\delta_{m_i,m_j} \right)^r \quad ; \; r \in \{1,2\}$$

- Diámetro del clúster: la máxima disimilitud entre dos observaciones del mismo clúster, puede ser usada para medir homogeneidad.

$$h_2(m) = max \left(\delta_{m_i,m_j} \right)^r \quad ; \; r \in \{1,2\}$$

Variabilidad

La variabilidad total del clúster puede ser partida en variabilidad dentro del clúster y variabilidad entre clústeres, es decir: $T = W + B$, donde:

$$T = \sum_{i=1}^{g} \sum_{j=1}^{n_m} (x_{ij} - \bar{x})(x_{ij} - \bar{x})^T$$

$$B = \sum_{i=1}^{g} n_i (\overline{x_i} - \bar{x})(\overline{x_i} - \bar{x})^T$$

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_m} (x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})^T$$

En las ecuaciones previas, x_{ij} es el j-ésimo miembro del clúster i, $\overline{x_i}$ es la media del clúster i y $\bar{x}$ es el vector de las medias de cada variable.

Lo anterior conduce a dos criterios naturales de agrupación:

- Minimizar la suma de cuadrados dentro del grupo
- Maximizar la suma de cuadrados entre grupos

Para minimizar la variabilidad dentro del clúster se puede usar la función de la traza:

$$Tr(w) = \sum_{i=1}^{filas} w_{ii}$$

Minimizar $Tr(w)$ es equivalente a minimizar la suma de las distancias euclídeas al cuadrado entre un grupo de observaciones y la media del grupo. $Tr(w)$ es dependiente de la escala.

K-means clustering

Es el más común de los algoritmos de clustering, consiste en los siguientes pasos:

1. Un conjunto de puntos conocido como “semillas” es seleccionado como las medias de los clústeres finales
2. Cada observación es asignada a la semilla más cercana formando clústeres temporales. Las semillas son reemplazadas por las medias de los clústeres temporales y se repite el proceso hasta que ocurra un cambio no significativo en la posición de los centros.
3. Forma los clústeres finales asignando cada observación al centroide más cercano.

Los primeros dos pasos son el algoritmo de búsqueda heurística. El paso 3 es la iteración extra que realiza las asignaciones finales a cada clúster.

Clustering Difuso

El clustering difuso se basa en criterios de densidad, son una poderosa alternativa a los métodos jerárquicos o de optimización ya que sobrepasan algunas de las debilidades de éstos, por ejemplo, pueden generar clústeres de cualquier forma y tamaño además de no ser afectados por valores extremos. Al ser nuestro estudio de carácter introductorio, nos centraremos en los modelos gaussianos mixtos. Los modelos gaussianos mixtos son un tipo de modelo de densidad que comprenden cierto número de funciones componente usualmente gaussianas. Revisemos su construcción:

Sean $\vec{x} \in \mathbb{R}^n$, podemos definir un modelo gaussiano mixto haciendo cada uno de los componentes una densidad gaussiana con parámetros μ_k y Σ_k. Cada componente es entonces una densidad gaussiana multivariante:

$$N\left(\vec{x}\middle|\vec{\mu},\Sigma\right)=\frac{1}{\left(2\pi\right)^{d/2}\sqrt{\left|\Sigma\right|}}e^{-\frac{1}{2}\left(\vec{x}-\vec{\mu}\right)'\Sigma^{-1}\left(\vec{x}-\vec{\mu}\right)}$$

Con sus propios parámetros $\theta_k=\left\{\mu_k,\Sigma_k\right\}$. La probabilidad dado un modelo gaussiano mixto sería:

$$p\left(x\right)=\sum_{i=1}^{N}w_i\ N\left(\vec{x}\middle|\vec{\mu}_i,\Sigma_i\right)$$

Donde N es el número de gaussianas y w_i es el peso de la $i-$ésima gaussiana tal que:

$$\sum_{i}^{N}w_i=1\ \ w_i\geq 0$$

Los modelos gaussianos mixtos pueden entrenarse mediante máxima verosimilitud mediante el algoritmo EM (Expectation-Maximization), esto es, maximizar la verosimilitud $p\left(X\middle|\theta\right)$ de los datos X extraídos de una distribución desconocida, dado el modelo parametrizado por θ:

$$\theta^*=\arg\max_{\theta}p\left(X\middle|\theta\right)=\arg\max_{\theta}\prod_{p=1}^{n}p\left(x_p\middle|\theta\right)$$

Los pasos básicos del algoritmo son los siguientes:

- Introducir una variable oculta tal que *su conocimiento* simplificaría la maximización de $p\left(X\middle|\theta\right)$
- En cada iteración:

- Paso E: estimar la distribución de la variable oculta dados los datos y los valores paramétricos actuales.
- Paso M: modificar los parámetros para maximizar la distribución conjunta de los datos y de la variable oculta.