**Resumen de la sección 12.4: Métodos de Clustering**

El clustering es un conjunto de técnicas de aprendizaje no supervisado que tiene como objetivo agrupar observaciones en subgrupos o clusters. Dentro de cada cluster, las observaciones son más similares entre sí que con aquellas de otros clusters. Las aplicaciones incluyen identificación de subgrupos en datos médicos, segmentación de mercado, entre otras​​.

**1. K-Means Clustering**

K-means es un método simple para dividir un conjunto de datos en KKK clusters no solapados. El algoritmo funciona de la siguiente manera:

1. Se eligen KKK clusters iniciales aleatoriamente.
2. Se calculan los centroides (medias) de los clusters.
3. Cada observación se asigna al cluster con el centroide más cercano (distancia euclidiana).
4. Los pasos 2 y 3 se repiten hasta que las asignaciones no cambien.

La función objetivo minimiza la variación dentro de cada cluster, expresada como la suma de distancias cuadradas de cada punto al centroide de su cluster. Dado que el algoritmo depende de la inicialización, se recomienda ejecutarlo varias veces​​.

**2. PAM (Partitioning Around Medoids)**

PAM es similar a K-means, pero usa los **medoides** (observaciones reales) en lugar de centroides. Esto lo hace más robusto a valores atípicos. PAM puede utilizar tanto distancias euclidianas como Manhattan para definir disimilitudes entre observaciones.

**3. Dendrogramas y Clustering Jerárquico**

El clustering jerárquico genera una representación en forma de árbol (dendrograma), que ilustra cómo se agrupan las observaciones paso a paso. Hay dos enfoques:

* **Ascendente (aglomerativo):** Cada observación empieza como un cluster individual y se fusionan iterativamente.
* **Descendente (divisivo):** Todo el conjunto de datos es un solo cluster, que se divide iterativamente.

**3.1 Función HCLUST**

Define el método para medir la disimilitud entre clusters:

* **Complete linkage:** Máxima disimilitud entre observaciones.
* **Single linkage:** Mínima disimilitud.
* **Average linkage:** Promedio de disimilitudes.
* **Centroid linkage:** Distancia entre centroides.
* **Ward’s method:** Minimiza la varianza total dentro de los clusters​​.

**3.2 Coeficiente de Cophenetic**

Mide qué tan bien representa el dendrograma las distancias originales entre observaciones. Es un criterio para ver los dendogramas con distintas distancias y evaluar que tan correlacionados están con las distancias originales de los datos. (esto se evalua de acuerdo a la distancia que usan los dendogramas – allí se observa si están o no claramente correlacionados)

**4. Cómo Elegir la Cantidad de Clusters**

1. **Coeficiente de Silhouette:** Evalúa la separación y cohesión de los clusters. Un valor más alto indica mejor calidad.
2. **Índice de Dunn:** Mide la separación mínima entre clusters y la dispersión máxima dentro de clusters. D= mínima distancia entre conjuntos / máximo diametro
3. **Método del Codo:** Evalúa la suma de cuadrados dentro de los clusters (inercia) para determinar un punto de saturación en la varianza explicada​​.

**5. Índices de Comparación – para comparar que tan parecidos me dan los agrupamientos al utilizar métodos distintos**

1. **Índice de Rand:** Evalúa la similitud entre dos particiones. Cuanto mas grande mejor
2. **Índice de Jaccard:** Basado en intersección y unión de conjuntos.

**6. Tendencias y Métodos Avanzados**

1. **Clustering Espectral:** Usa la descomposición espectral de matrices para encontrar clusters en datos no lineales.
2. **Fuzzy Clustering:** Permite que las observaciones pertenezcan parcialmente a múltiples clusters.

**7. Teorema de Imposibilidad de Kleinberg y Maldición de la Dimensionalidad (ver apuntes de clase)**

El teorema de Kleinberg establece que no hay un método de clustering que satisfaga simultáneamente ciertas propiedades deseables. La alta dimensionalidad de los datos puede dificultar la identificación de clusters significativos​.

KLEINBERG

Al aplicar una técnica de clustering deberíamos pedir 3 condiciones que se tienen que cumplir. Este teorema demuestra que es imposible que se cumplan las 3 condiciones a la vez.

Cuales son las condiciones?

* Invarianza de escala
* Riqueza o Alcance
* Consistencia

MALDICION DE DIMENSIONALIDAD

Para hacer buena estadística de modelos se necesitan muchos datos, y cuanto mayor es la dimensión mayor cantidad de datos preciso para poder hacer buena estadística.

- Es muy difícil hacer estadística de modelos en dimensión alta porque el N que tengamos (cantidad de datos) siempre nos va a quedar chico

Para un heredado, en la medida que la dimensión aumenta mas también aumenta el espacio de los datos en algún sentido, es decir, los datos van a estar mas lejos entre sí. Por ej., bayes es la medida optima, pero es inalcanzable por la maldición de la dimensionalidad.

**8. Estadístico de Hopkins**

Mide la tendencia al clustering. Si el valor es cercano a 1, los datos son altamente clusterizables; si es cercano a 0.5, los datos están distribuidos aleatoriamente.

**(RESUMEN EN BASE A MI PUNTEO DE ORGANIZACIÓN ESTUDIO)**

**Punto 2: Clustering – Análisis no Supervisado**

El clustering es una técnica de aprendizaje no supervisado utilizada para agrupar observaciones en subconjuntos llamados **clusters**, de forma que las observaciones dentro de cada cluster sean más similares entre sí que con las de otros clusters. Este proceso es fundamental cuando no se tiene una variable objetivo y se busca encontrar patrones o estructuras en los datos. A continuación, se describen los métodos principales con sus conceptos, definiciones y bases teóricas:

**2.1 K-Means Clustering**

K-means es uno de los algoritmos de clustering más comunes y se utiliza para dividir un conjunto de datos en KKK clusters no superpuestos.

**Definición y Método**

1. **Inicialización:** Se eligen K centroides iniciales, que pueden seleccionarse aleatoriamente o mediante métodos específicos como K-means++.
2. **Asignación:** Cada observación se asigna al cluster cuyo centroide está más cerca, utilizando la distancia euclidiana como métrica por defecto:

A black and white math equation

Description automatically generated

Donde:

* + xi​ es la i-ésima observación.
  + cj​ es el centroide del cluster j.

1. **Actualización:** Los centroides de los clusters se recalculan como la media de las observaciones asignadas a cada cluster:

A mathematical equation with numbers and symbols

Description automatically generated

Donde nj​ es el número de observaciones en el cluster j.

1. **Iteración:** Los pasos 2 y 3 se repiten hasta que no cambien las asignaciones o se alcance un criterio de convergencia.

**Función Objetivo**

El algoritmo minimiza la **suma de cuadrados dentro del cluster (WSS, Within-Cluster Sum of Squares):**

A number and symbols on a white background

Description automatically generated

Esto significa que K-means busca compactar los datos alrededor de sus centroides.

**Ventajas y Limitaciones**

* **Ventajas:**
  + Rápido y eficiente para grandes conjuntos de datos.
  + Fácil de interpretar y aplicar.
* **Limitaciones:**
  + Depende de la inicialización de los centroides (puede llevar a mínimos locales).
  + No funciona bien con clusters de formas no esféricas o de tamaños desiguales.

**2.2 PAM (Partitioning Around Medoids)**

PAM es un algoritmo similar a K-means, pero usa **medoides** en lugar de centroides. Un medoide es una observación real dentro del cluster que minimiza la distancia promedio a todas las demás observaciones del cluster.

**Definición y Método**

1. **Inicialización:** Se eligen KKK medoides iniciales.
2. **Asignación:** Cada observación se asigna al cluster cuyo medoid está más cerca, utilizando distancias como:

**A math equations with black text

Description automatically generated with medium confidence**

1. **Optimización:** Se intercambian los medoides con otras observaciones para reducir la función objetivo:

A black and white symbols

Description automatically generated with medium confidence

1. **Iteración:** Los pasos se repiten hasta que no haya mejoras en el costo total.

**Ventajas y Limitaciones**

* **Ventajas:**
  + Más robusto que K-means frente a valores atípicos porque utiliza puntos reales como representativos.
* **Limitaciones:**
  + Computacionalmente más costoso que K-means, especialmente para grandes conjuntos de datos.

**2.3 Dendrogramas**

Un dendrograma es una representación jerárquica en forma de árbol que muestra cómo se agrupan las observaciones a diferentes niveles.

**Clustering Jerárquico**

1. **Enfoque Aglomerativo (ascendente):**
   * Cada observación comienza como un cluster individual.
   * Se fusionan iterativamente los clusters más similares hasta que todos estén en un único cluster.
2. **Enfoque Divisivo (descendente):**
   * Comienza con un único cluster que incluye todas las observaciones.
   * Se divide recursivamente en clusters más pequeños.

**Métodos de Disimilitud (Función HCLUST)**

Para decidir qué clusters combinar o dividir, se utilizan medidas de disimilitud:

* **Complete linkage:** Máxima disimilitud entre observaciones de dos clusters.
* **Single linkage:** Mínima disimilitud.
* **Average linkage:** Promedio de las disimilitudes.
* **Centroid linkage:** Distancia entre los centroides de los clusters.
* **Ward’s method:** Minimiza la suma de varianzas dentro de los clusters.

**Coeficiente de Cophenetic**

Evalúa qué tan bien representa el dendrograma las distancias originales entre observaciones. Se calcula como la correlación entre las distancias originales y las distancias representadas por el dendrograma.

**2.4 Cómo Elegir la Cantidad de Grupos en Clustering**

1. **Coeficiente de Silhouette:**
   * Evalúa la calidad de un cluster midiendo la cohesión (distancia promedio dentro del cluster) y la separación (distancia promedio al cluster más cercano).
   * Valores cercanos a 1 indican buena calidad.
2. **Índice de Dunn:**
   * Relación entre la mínima distancia entre clusters y la máxima dispersión dentro de los clusters.
   * Un índice alto indica clusters compactos y bien separados.
3. **Método del Codo (Suma de Cuadrados Within):**
   * Grafica la WSS frente al número de clusters KKK.
   * El "codo" del gráfico indica el punto donde agregar más clusters no reduce significativamente la WSS.