TEC | Tecnológico de Costa Rica

Aprendizaje Automático

Tecnológico de Costa Rica

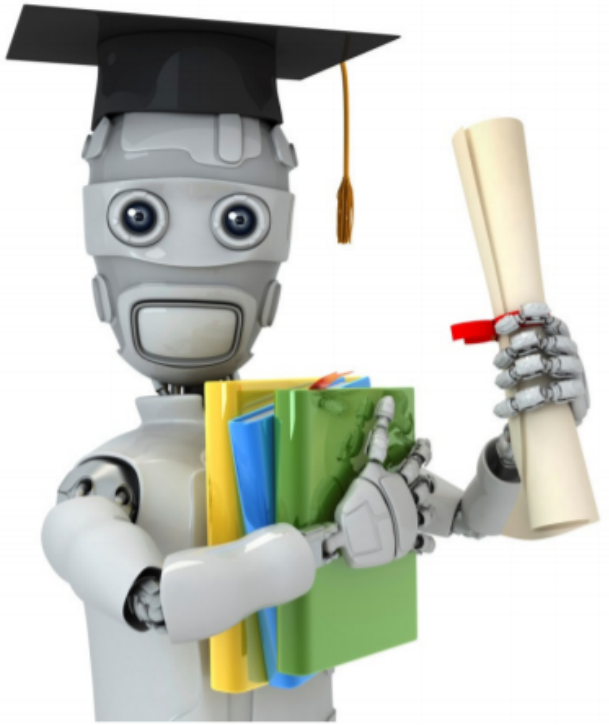
Programa de Ciencia de Datos

Frans van Dunné

Agenda

TEC | Tecnológico de Costa Rica

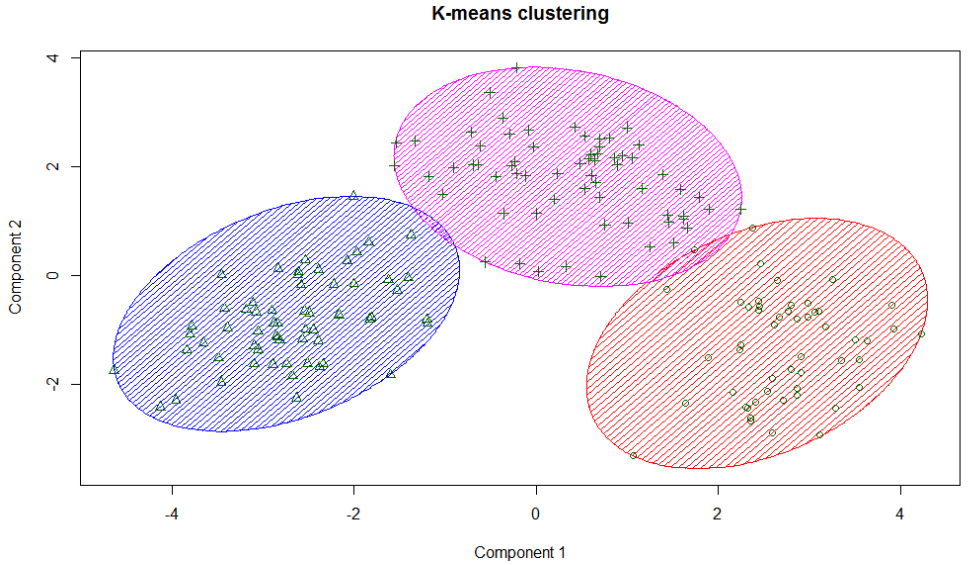
- **Aprendizaje Automático**
  - Métodos no supervisados.
    - Algoritmo PCA
    - Algoritmo K-medias.
    - Algoritmo Herencia



Aprendizaje no supervisado

TEC | Tecnológico de Costa Rica

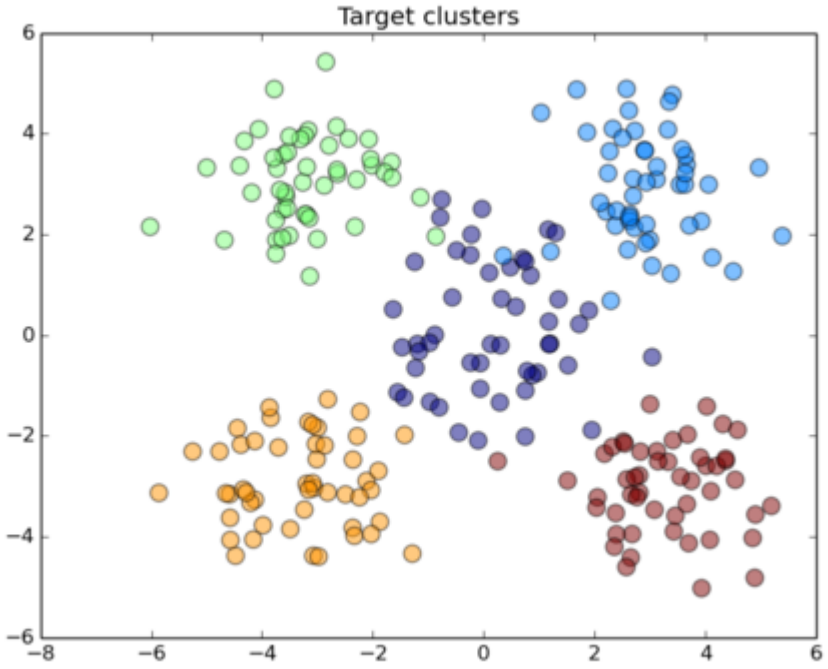
- **Algoritmos no supervisados**
  - Solo se le otorgan las características, sin proporcionarle al algoritmo ninguna etiqueta. Su función es la **agrupación**, por lo que el algoritmo debería catalogar por similitud y poder crear grupos, sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes del grupo.



El desafío del aprendizaje no supervisado

TEC | Tecnológico de Costa Rica

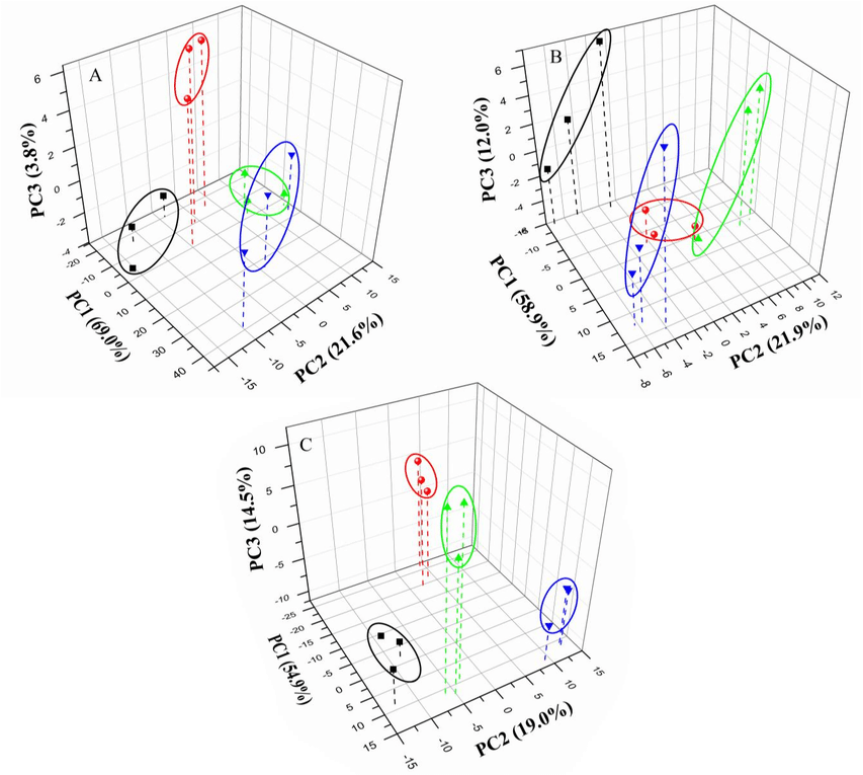
- Las técnicas para el aprendizaje no supervisado son cada vez más importantes en un sin numero de campos.
  - Un investigador del cáncer podría analizar los niveles de expresión genética en 100 pacientes con cáncer de mama. Él o ella podría buscar subgrupos entre las muestras de cáncer de mama, o entre los genes, para obtener una mejor comprensión de la enfermedad. Un sitio de compras en línea podría intentar para identificar grupos de compradores con historiales similares de navegación y compras, así como artículos que son de particular interés para los compradores dentro cada grupo.
  - Entonces se puede mostrar preferencialmente a un comprador individual artículos en los que es más probable que él o ella estén interesados, según
  - Los historiales de compra de compradores similares.
  - Un motor de búsqueda podría elegir qué resultados de búsqueda mostrar a un individuo en particular en función del clic con las historias de otras personas con patrones de búsqueda similares.



Análisis de componentes principales

TEC | Tecnológico de Costa Rica

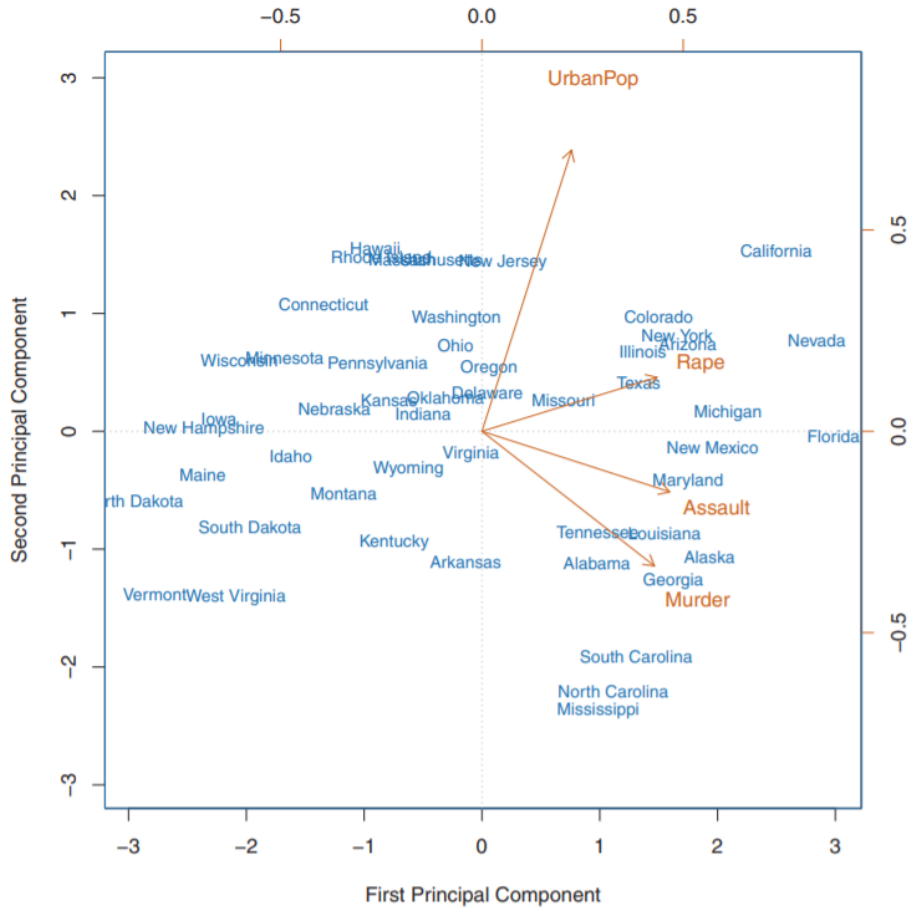
- PCA proporciona una herramienta para hacer precisamente esto. Encuentra una representación de baja dimensión de un conjunto de datos que contiene la mayor cantidad posible de la variación.
- La idea es que cada una de las n observaciones vive en el espacio p-dimensional, pero no todas estas dimensiones son igualmente interesantes.
- PCA busca un pequeño número de dimensiones que son tan interesantes como sea posible, donde el concepto de interesante se mide por la cantidad que las observaciones varían a lo largo de cada dimensión. Cada una de las dimensiones encontradas por PCA es una combinación lineal de las características p.



PCA

TEC | Tecnológico de Costa Rica

- Los primeros dos componentes principales para el dataset USArrests.
- Los nombres de los estados azules representan los puntajes de los dos primeros componentes principales.
- Las flechas naranjas indican los dos primeros vectores de carga de componentes principales (con ejes en la parte superior y derecha).
- Por ejemplo, la carga para violación en el primer componente es 0.54, y su carga en el segundo componente principal 0.17 (la palabra **RAPE** se centra en el punto (0.54, 0.17)).
- Esta cifra se conoce como biplot, porque muestra las puntuaciones del componente principal y peso.



Métodos de agrupamiento

TEC | Tecnológico de Costa Rica

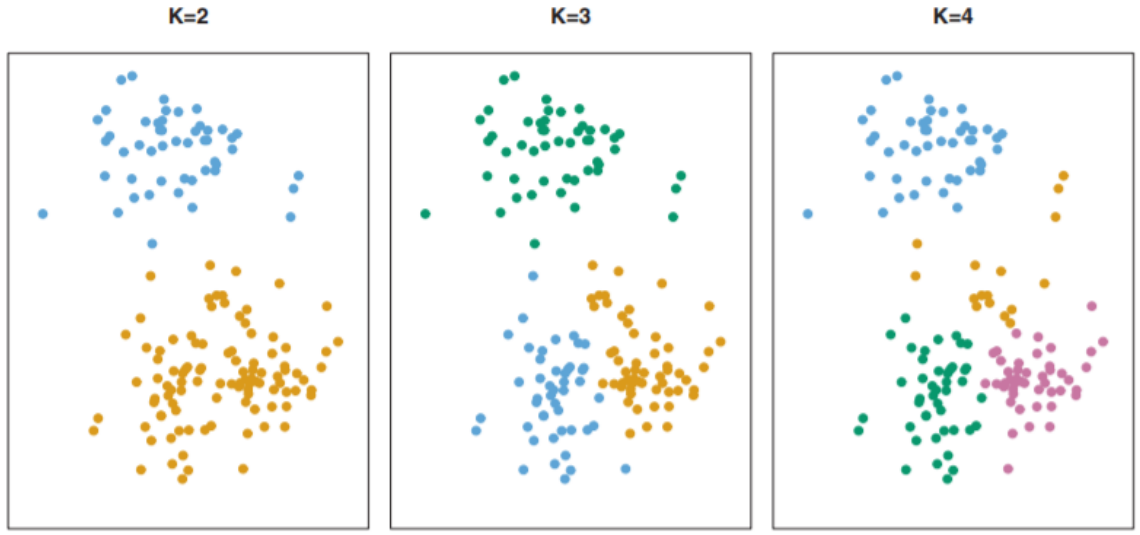
- **K-Means**
- **Hierarchical Clustering**

- La agrupación se refiere a un conjunto muy amplio de técnicas para encontrar subgrupos, o agrupaciones, en un conjunto de datos. Los grupos son bastante diferentes entre sí. Por supuesto, para hacer esto concreto, debemos definir qué significa que dos o más observaciones sean similares o diferente De hecho, esto es a menudo una consideración específica del dominio que debe hacerse en base al conocimiento de los datos que se estudian.
- Por ejemplo, supongamos que tenemos un conjunto de n observaciones, cada una con p características.
  - Las n observaciones podrían corresponder a muestras de tejido para pacientes con cáncer de mama, y las características p podrían corresponder a mediciones recolectado para cada muestra de tejido; Estas podrían ser mediciones clínicas, como como estadio o grado tumoral, o podrían ser medidas de expresión génica.
  - Podemos tener una razón para creer que existe cierta heterogeneidad entre las n muestras de tejido; por ejemplo, tal vez hay algunos subtipos desconocidos diferentes de cáncer de seno. La agrupación podría usarse para encontrar estos subgrupos Este es un problema no supervisado porque estamos tratando de descubrir la estructura, en este caso, grupos distintos, sobre la base de un conjunto de datos.
- Tanto la agrupación en clúster como PCA buscan simplificar los datos a través de un pequeño número de resúmenes, pero sus mecanismos son diferentes:
  - PCA busca encontrar una representación de baja dimensión de las observaciones que expliquen una buena fracción de la varianza;
  - La agrupación busca encontrar subgrupos homogéneos entre las observaciones.

Agrupamiento de medias K

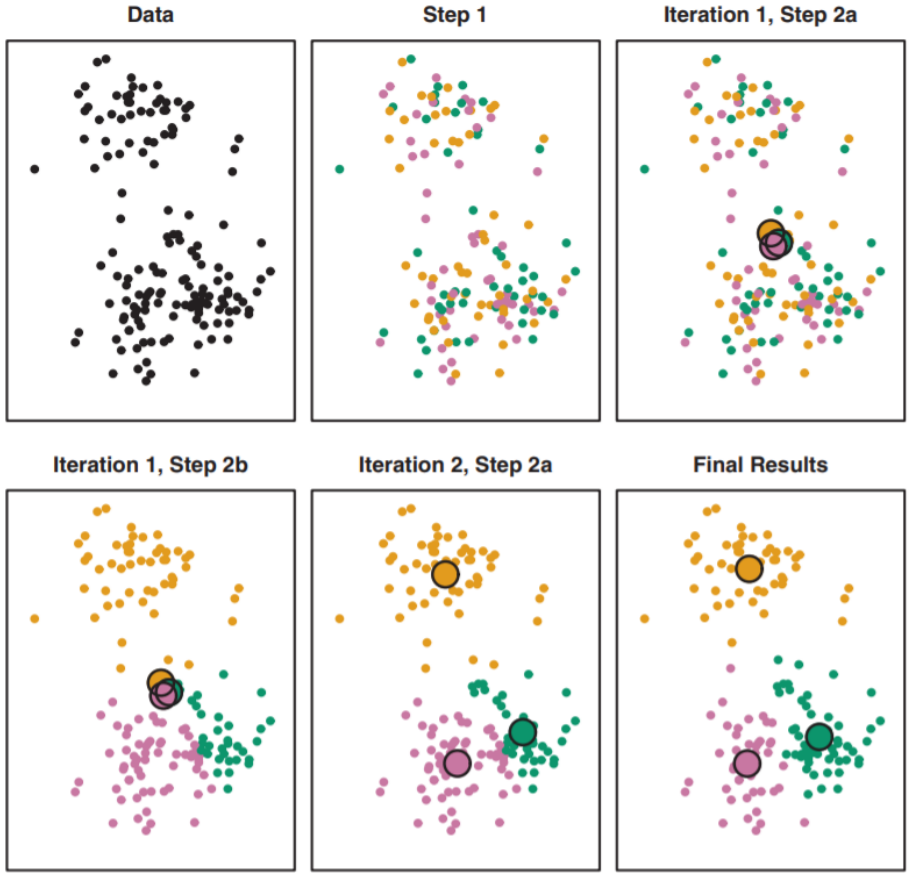
TEC | Tecnológico de Costa Rica

- La agrupación de K-means es un enfoque simple y elegante para particionar un conjunto de datos en K grupos distintos, no superpuestos. Para realizar K-means agrupación, primero debemos especificar el número deseado de agrupaciones K;



Agrupamiento de medias K

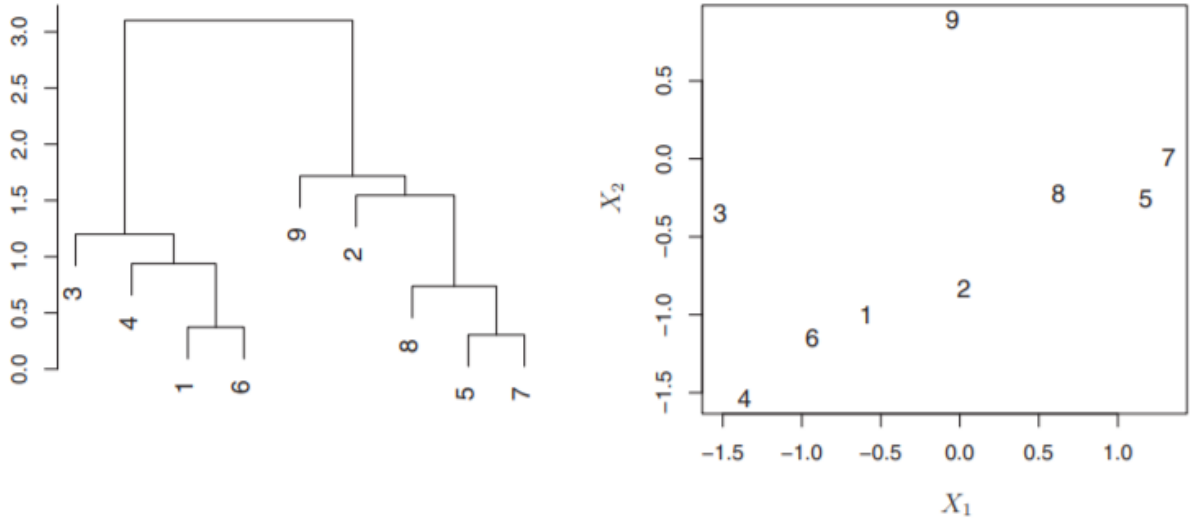
TEC | Tecnológico de Costa Rica



Agrupación jerárquica

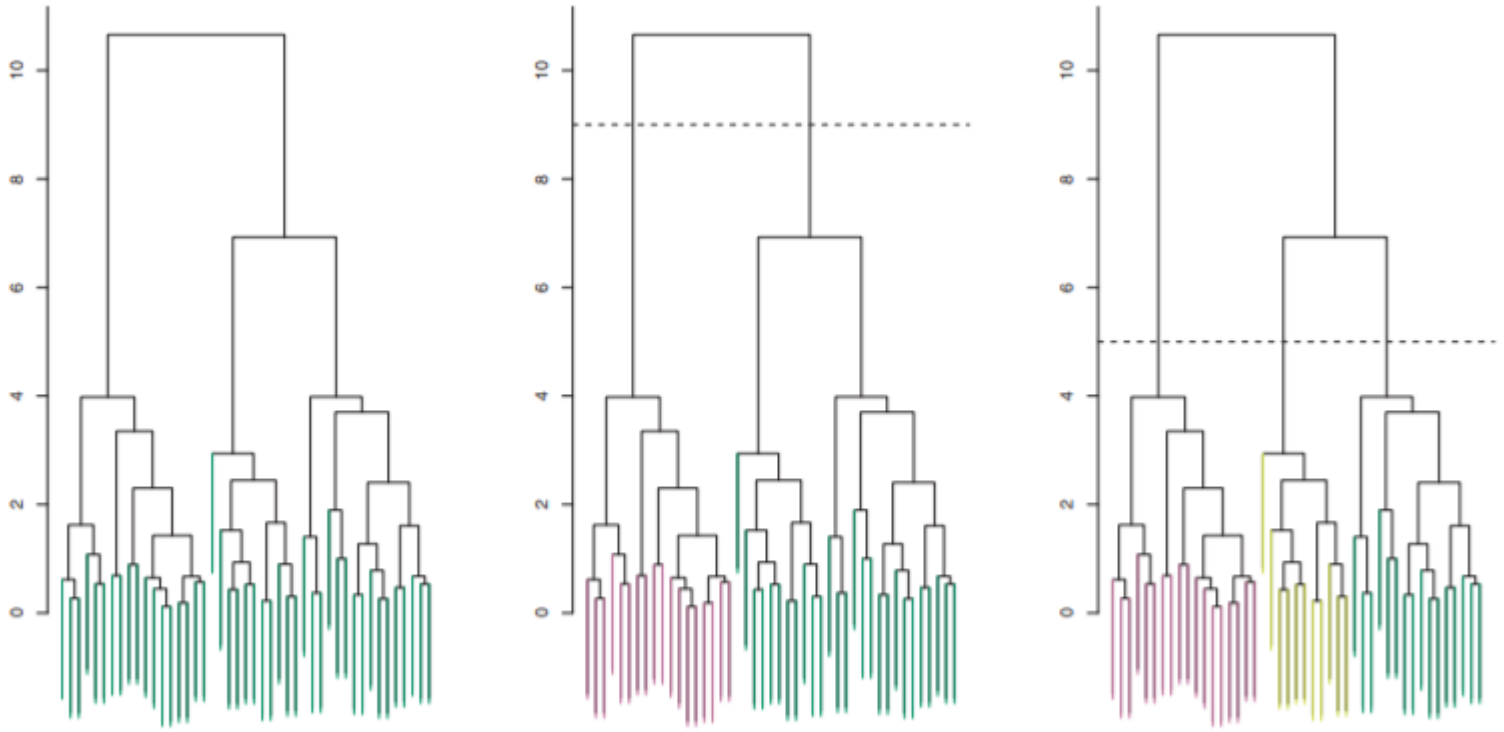
TEC | Tecnológico de Costa Rica

- Una desventaja potencial de la agrupación de K-means es que nos exige especificar previamente el número de grupos K. El agrupamiento jerárquico es un enfoque alternativo que no requiere que nos comprometamos con un determinado la elección de K. La agrupación jerárquica tiene una ventaja adicional sobre K-means agrupación en que da como resultado una atractiva representación basada en el árbol de la observaciones, llamadas dendrogramas



Agrupación jerárquica

TEC | Tecnológico de Costa Rica



Pequeñas decisiones con grandes consecuencias

TEC | Tecnológico de Costa Rica

- Pequeñas decisiones con grandes consecuencias
- Para realizar la agrupación, se deben tomar algunas decisiones.
  - ¿Deben las observaciones o características primero ser estandarizadas de alguna manera?
  - Por ejemplo, quizás las variables deberían estar centradas para tener una media cero y escalado para tener una desviación estándar uno.
- En el caso de agrupamiento jerárquico,
  - - ¿Qué medida de disimilitud se debe utilizar?
  - - ¿Qué tipo de enlace se debe usar?
  - - ¿Dónde debemos cortar el dendrograma para obtener racimos?
- En el caso de la agrupación de K-means, ¿cuántos grupos deberíamos buscar?

Bibliografía

TEC | Tecnológico de Costa Rica

- https://www.displayr.com/what-is-dendrogram/
- http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/
- https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html
- https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
- https://www.coursera.org/lecture/mineria-de-datos-introduccion/ejemplo-algoritmo-k-means-d0fgs