Ciencia de Datos Aplicada a Ciencias de la Tierra

Tema: Métodos de aprendizaje no supervisado

Martín A. Díaz-Viera[1], Farhid M. Elisea Guerrero[2]

[1)] *mdiazv@imp.mx*, [2)] *felisea@imp.mx*

27 de febrero de 2024

Contenido I

Contenido II

Métodos de Aprendizaje Automatizado (ML)

- Algunos de los métodos más comunes implementados para "hacer que las máquinas aprendan"son:
  - Aprendizaje supervisado
  - Aprendizaje no supervisado
  - Aprendizaje reforzado

Métodos de aprendizaje no supervisado

Métodos de Aprendizaje No Supervisado

Aprendizaje no supervisado

- Es un método de aprendizaje automátizado donde un modelo se ajusta a las observaciones.
- No hay un conocimiento a priori.
- Un conjunto de datos de entrada es tratado.
- Trata los datos de entrada como un conjunto de variables aleatorias.

Métodos de aprendizaje no supervisado

Algunos de los métodos más comunes utilizados en el aprendizaje no supervisado son:

- Redes neuronales no supervisadas
- Métodos de agrupamiento

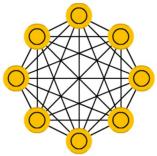
Redes neuronales no supervisadas

Redes Neuronales No Supervisadas

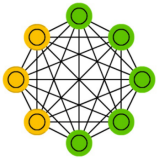
Redes neuronales no supervisadas

- Muestran cierto grado de auto-organización
- La red descubre de forma autónoma: características, regularidades, correlaciones y categorías
- Menores tiempos de entrenamiento que las supervisadas
- Arquitectura simple, habitualmente son de una sola capa
- Tipos fundamentales Kohonen y Grossberg

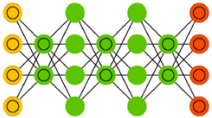
Redes neuronales no supervisadas



Hopfield

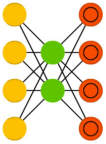


Boltzmann

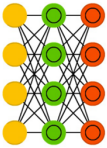


Deep Belief Network

RBM



Autoencoder



Variational Autoencoder

Métodos de agrupamiento

Métodos de Agrupamiento

Métodos de agrupamiento

- El análisis agrupamiento es la tarea de agrupar objetos por similitud.
- Se agrupan en clases o conjuntos de manera que los miembros del mismo grupo tengan características similares.
- Es una tarea principal del análisis exploratorio de datos.
- Es una técnica común en el análisis estadístico de datos.

Métodos de agrupamiento

Los criterios de agrupamiento pueden ser según:

- **Conectividad**: por ejemplo, agrupamiento jerárquico construye modelos basados en la distancia de las conexiones.
- **Centroide**: por ejemplo, el algoritmo k-means representa cada grupo por un solo vector medio.
- **Distribución**: los grupos son modelados utilizando distribuciones estadísticas, como la distribución normal multivariada.
- **Densidad**: por ejemplo, DBSCAN y OPTICS definen grupos como regiones densas conectadas en el espacio de los datos.

Algoritmos de agrupamiento

En base a los criterios anteriores se pueden formular los algoritmos de agrupamiento como:

- Agrupamiento basado en conectividad (agrupamiento jerárquico)
- Agrupamiento basado en centroide (agrupamiento de partición)
- Agrupamiento basado en distribuciones
- Agrupamiento basado en densidad

Métodos de agrupamiento jerárquico

Métodos de Agrupamiento Jerárquico

Métodos de agrupamiento jerárquico

- Está basado en la idea principal de que los objetos más cercanos están más relacionados que los que están alejados.
- Estos algoritmos conectan "objetos" para formar "los grupos" basados en su distancia.
- Un grupo puede ser descrito, en gran parte, por la distancia máxima que se necesitó para conectar todas las partes del grupo.
- Busca construir una jerarquía de grupos: a distancias diferentes se formarán grupos diferentes.

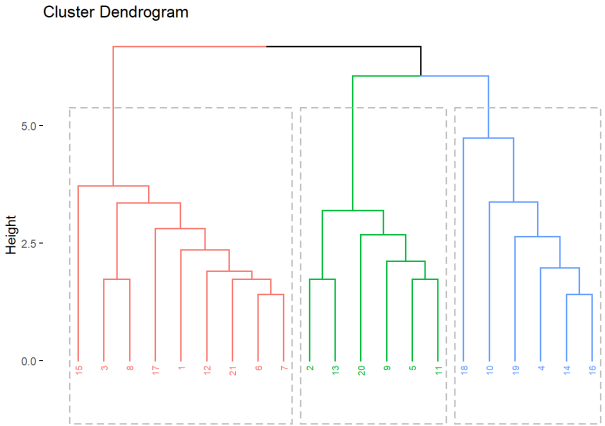
Métodos de agrupamiento jerárquico

Las estrategias para el agrupamiento jerárquico generalmente se dividen en dos categorías:

- **Aglomerativo**: este es un enfoque "de abajo hacia arriba", cada observación comienza en su propio grupo, y los pares de grupos se fusionan a medida que uno asciende en la jerarquía.
- **Divisivo**: este es un enfoque "de arriba hacia abajo", todas las observaciones comienzan en un grupo y las divisiones se realizan de forma recursiva a medida que se desciende en la jerarquía.

Dendrograma

Esta jerarquía usualmente es representada mediante un dendrograma.



Dendrograma

- Del dendrograma proviene el nombre "agrupamiento jerárquico"
- Estos algoritmos no solo proporcionan una partición del conjunto de datos, sino proporcionan una jerarquía de grupos
- Los grupos se fusionan unos con otros a ciertas distancias.
- En un dendrograma, el eje y marca la distancia en qué los grupos se fusionan, mientras que los objetos están colocados a lo largo del eje x, tal que los grupos se mezclan.

Algoritmos de agrupamiento jerárquico

- Para decidir como deben combinarse (por aglomeración) o dividirse (por división) un grupo, se requiere una medida de disimilitud.
- Esto se logra mediante el uso de una distancia d adecuada y un criterio de vinculación o enlace.
- La distancia determina qué individuos son más similares, mientras que el criterio de vinculación influye en la forma de las agrupaciones.
- Por ejemplo, el enlace completo tiende a producir más grupos esféricos que el enlace simple.

Criterios de enlace

Algunos criterios de enlace comúnmente utilizados entre dos conjuntos de observaciones A y B y una distancia d son:

- Enlace completo o agrupamiento máximo: $\max_{a \in A,\, b \in B} d(a, b)$
- Enlace simple o agrupamiento mínimo: $\min_{a \in A,\, b \in B} d(a, b)$
- Enlace promedio no ponderado (UPGMA): $\frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
- Enlace promedio ponderado (WPGMA): $d(i \cup j, k) = \frac{d(i,k) + d(j,k)}{2}$
- Enlace de Ward o incremento mínimo de la suma de los cuadrados (MISSQ): $\frac{|A| \cdot |B|}{|A \cup B|} \|\mu_A - \mu_B\|^2 = \sum_{x \in A \cup B} \|x - \mu_{A \cup B}\|^2 - \sum_{x \in A} \|x - \mu_A\|^2 - \sum_{x \in B} \|x - \mu_B\|^2$

Áreas de aplicación

- Este método ha sido utilizado para predecir el rendimiento de a partir de la información de 12 mil pozos con datos geológicos y geoquímicos en el sistema petrolero de Bekken, agrupando instancias según producción, localización y formación geológica (Chakhmakhchev et al., 2021).
- Otra área de aplicación es en la geoquímica del petróleo y su formación, ya que en el trabajo de Kwilosz (2022), presenta un método para encontrar similitudes entre estructuras geológicas que se diferencian en términos de sus propiedades de generación de hidrocarburos y cantidad de recursos.

Ventajas

Ventajas:

- Si no sabe de antemano qué número de grupos se busca, el gráfico del dendrograma puede ayudarle a elegir.
- El dendrograma también puede dar una visión de la estructura de los datos y ayudar a identificar los valores atípicos.
- Se puede utilizar cualquier medida válida de distancia.
- Las observaciones en sí no son necesarias: todo lo que se utiliza es una matriz de distancias.

Desventajas

Desventajas:

- Excepto por el caso especial de enlace simple, ninguno de los algoritmos excepto la búsqueda exhaustiva puede garantizar encontrar la solución óptima.
- Puede ser altamente afectado por valores atípicos.
- Dependiendo de como se mida la similitud entre los datos el algoritmo puede ser sensible al ruido y a los valores atípicos.
- Para grandes conjuntos de datos puede ser costoso computacionalmente.

Software

Implementaciones de código abierto:

- `R`: tiene funciones integradas y paquetes que proporcionan funciones para el agrupamiento jerárquico.
- `SciPy`: implementa el agrupamiento jerárquico en Python, incluido el eficiente algoritmo SLINK.
- `scikit-learn`: también implementa el agrupamiento jerárquico en Python.
- `Weka`: incluye análisis de conglomerados jerárquicos.

Métodos de agrupamiento de particionamiento

Métodos de Agrupamiento de Particionamiento

Métodos de agrupamiento de particionamiento

- Son métodos de agrupamiento basados en el centroide.
- Cada grupo está representado por un vector central, que no es necesariamente un miembro del conjunto de datos.
- El número k de conglomerados se fija a priori.
- Se formula como un problema de optimización: encontrar los k centros de los conglomerados.
- Se asignan los individuos al centro del conglomerado más cercano.

Algoritmos de agrupamiento de particionamiento

- **k-medias**
- **k-medoides**

k-medias

- Tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.
- El objetivo consiste en optimizar dos métricas de ajuste: se minimizan las diferencias entre instancias de un mismo grupo y, al mismo tiempo se maximizan las diferencias entre grupos.
- La asignación inicial de grupos se lleva a cabo de forma aleatoria.
- Se actualizan las asignaciones de instancias a grupos hasta que ya no es posible mejorar las métricas de ajuste o se alcanza un número máximo de iteraciones.
- En cada iteración, se calculan k puntos (o centroides) que reemplazan a los puntos existentes y que se emplean para calcular las diferencias.

k-medoides

- La mecánica de funcionamiento es similar al método de k-medias.
- Las principales diferencias con respecto a k-medias, estriban en que los centroides se toman de instancias concretas del conjunto de datos y se conocen como medoides.
- En contraste a k-medias, donde los centroides corresponden a valores promedio de cada grupo de acuerdo con las asignaciones en la iteración actual.
- Se emplean métricas arbitrarias para el cálculo de distancias (o similitud), en contraste a k-medias, donde típicamente, la distancia por paradigma es la Euclidiana.

Áreas de aplicación

Algunos ejemplos de aplicación consisten en:

- Una de las aplicación de este algoritmo es el incorporar incertidumbre espacial en las superficies del yacimiento cuando este se modela (Hardy et al., 2019).
- Este algoritmo también se ha aplicado en la geoquímica, con el fin de determinar las mejores zonas de perforación y esto a partir de determinar áreas donde los datos lito geoquímicos muestren zonas de alta concentración (Shirazi, 2018).

Ventajas/desventajas

- **k-medias**
  - Buen desempeño para conjuntos de datos reales.
  - No hay garantía de agrupamientos idóneos,
  - Depende de una estimación inicial del número de grupos.
- **k-medoides**
  - Tiene ventaja cuando es complicado definir un valor promedio.
  - Mayor tolerancia a datos ruidosos y presencia de datos atípicos.
  - Su principal desventaja es determinar el número óptimo de grupos.
  - El desempeño tiende a la ineficiencia para conjuntos grandes de datos.

Software

Entre las implementaciones disponibles se han identificado:

- Statistics and Machine Learning Toolbox, Mathworks Matlab R2014b.
- Cluster-R, lenguaje de programación R.
- Scikit-learn, lenguaje de programación Python.

Flujo general de trabajo

Flujo general de trabajo:

1. Análisis exploratorio de los datos
2. Detección de valores atípicos
3. Selección de variables: ANOVA, PCA, AF, etc.
4. Selección del número óptimo de grupos o clases.
5. Aplicación
6. Evaluación de resultados
7. Interpretación de resultados

Análisis exploratorio de los datos

Por la estructura de datos:

- Global. Considera todas las instancias de datos.
- Intervalos (o unidades). Distingue subconjuntos de instancias según la unidad o intervalo geológico al cual pertenecen.
- Secciones. Distingue subconjuntos de instancias de acuerdo con la definición de secciones, grupos de pozos relacionados por la orientación espacial.
- Clasificaciones a priori. Distingue subconjuntos de instancias de acuerdo con clasificaciones presentes desde la etapa de Generación de la Base de Datos.
- Particular combinado. Considera dos o más criterios simultáneos.

Detección de valores atípicos

1. **Mediante la distancia de Mahalanobis (DM)**
   A la distancia de Mahalanobis previamente generada, se le calcula el primer cuartil (1Q), el tercer cuartil (3Q) y el rango intercuartil (RIQ), considerando los siguientes rangos:
   - Es atípico si DM $<$ 1Q – 1.5*RIQ.
   - Es atípico si DM $>$ 3Q + 1.5*RIQ.
2. **Mediante la Prueba Chi cuadrada (χ^2)**
   A la DM obtenida, se le calcula la Prueba Chi cuadrada (χ^2) con n-1 grados de libertad (n = no. de variables), y posteriormente se selecciona alguno de los criterios siguientes:
   - Etiquetar como atípicos las instancias con (1-χ^2) $<$ 0.001, ó
   - Calcular el cuantil 97.5 % (Q97_5) de χ^2 y etiquetar como atípicos las instancias donde DM $>$ DM en Q97_5.

Selección de variables

- **ANOVA/MANOVA**
- **Análisis Discriminante Lineal**
- **Análisis de Componentes Principales**
- **Análisis Factorial**

Selección del número óptimo de grupos o clases

1. **Método del Codo**: Se basa en la gráfica de la distancia media de las instancias a su centroide, con respecto al número de grupos.
2. **Índice Calinski-Harabasz**: Medida de qué tan similares son las instancias hacia dentro de su propio grupo (cohesión) en comparación con los demás grupos (separación).
3. **Método de la Silueta**: Este método mide la similitud de los datos hacia el grupo propio, contra el resto de los grupos.
4. **Índice Davies Bouldin**: Este es un esquema de evaluación interna, donde la validación de qué tan bien se ha realizado la agrupación se realiza utilizando cantidades y características inherentes al conjunto de datos.

Evaluación de resultados

- **Dendrogramas**: Gráfico que muestra las relaciones entre las instancias del conjunto de datos en forma de árbol.

- **Diagramas de Voronoi**: Divide un plano en regiones con base en un conjunto de puntos de referencia, de tal forma que cada punto del plano se encuentra lo más cerca al elemento de referencia al cual queda asignado.

Interpretación de resultados

- Ver que la cantidad del número óptimo de clases sea representativo y consistente con la información disponible.

Referencias

[1] DellÁversana, P. , "Cross-disciplinary Machine Learning - Part 1: Applications to rock classification, well log analysis and integrated geophysics", 182 pags. (2020).

[2] Leijnen, S. and Veen, F. V., *The neural network zoo*. Proceedings, **47(1)** (2020).

[3] Rosenblatt, F., "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain", *Psychological Review* **65 (6)** 386-408 (1958).

[4] Samuel, A. L. . "Some studies in machine learning using the game of checkers", *IBM Journal of Research and Development*, **3:3** 210–229 (1959).

Siguiente tema:

Métodos

de aprendizaje

supervisado