SISTEMAS INTELIGENTES

2025

PROYECTO DE AGRUPAMIENTO

Tema: Implementación y evaluación de clustering con el algoritmo K-means en el dataset *Adult* (UCI Census Income).

Fecha de entrega: 20 de octubre

En muchas aplicaciones del mundo real, los conjuntos de datos contienen variables de diferentes tipos: numéricas continuas, categóricas ordinales y nominales. Los algoritmos de clustering tradicionales como K-means están diseñados principalmente para datos numéricos, lo que limita su aplicabilidad en datasets mixtos.

La distancia de Gower (Gower, 1971) es una métrica de disimilitud que permite manejar eficientemente variables de diferentes tipos en un mismo análisis, asignando pesos apropiados según el tipo de variable y normalizando las contribuciones de cada atributo.

Conjunto de datos utilizado para este proyecto "Adult" (Censo de 1994)

El conjunto de datos **"Adult"**, también conocido como **"Census Income"**, es un repositorio de información demográfica ampliamente utilizado en el campo del **aprendizaje automático** (machine learning), particularmente para tareas de **clasificación binaria**. Origen y propósito :

- **Origen:** Fue extraído por Barry Becker de la base de datos del **Censo de los Estados Unidos de 1994**.
- **Propósito Principal (Tarea de Predicción):** Determinar si el **ingreso anual** de una persona es **superior a $50,000** o igual/inferior a esta cifra (una tarea de clasificación binaria: >50K o <=50K).

Para garantizar un conjunto de registros de calidad razonable, se aplicaron las siguientes condiciones durante la extracción, enfocándose en la población adulta con empleo e ingresos registrados:

- **Edad (AAGE)** mayor a 16 años.
- **Ingreso Ajustado Bruto (AGI)** mayor a 100.
- **Peso Final de la Muestra (AFNLWGT)** mayor a 1 (una variable que indica cuántas personas representa el registro censal).
- **Horas Trabajadas por Semana (HRSWK)** mayor a 0.

El conjunto de datos contiene una variedad de características que describen a cada individuo, incluyendo variables **categóricas** (como la clase de trabajo, educación, estado civil, ocupación, raza y país de origen) y **continuas** (como la edad, la ganancia/pérdida de capital y las horas trabajadas por semana).

- **Ejemplos de Atributos:**
  - age (Edad)
  - workclass (Clase de trabajo)
  - education (Nivel educativo)
  - marital-status (Estado civil)
  - occupation (Ocupación)
  - hours-per-week (Horas trabajadas por semana)
  - capital-gain y capital-loss (Ganancia/Pérdida de capital)
  - **Clase/Etiqueta:** income (Ingreso: >50K o <=50K).

Consideraciones claves

- **Valores Faltantes:** el dataset **sí contiene valores faltantes** (a menudo representados como ? o valores nulos), lo que requiere una fase de preprocesamiento de datos para su manejo antes de entrenar modelos de machine learning.
- **Desactualización y representatividad:** Dado que los datos son de 1994, no son representativos de la población actual de EE. UU. y deben interpretarse con precaución. No obstante, sigue siendo una referencia clásica para la evaluación de algoritmos de clasificación.

Entregables

1. Código fuente documentado.
2. Notebook con análisis paso a paso.
3. Informe técnico (máx. 10 páginas).

SISTEMAS INTELIGENTES

2025

Consignas y preguntas orientadoras

1. El script debe contener una sección de preparación de los datos

1. Cargá el dataset *Adult*. Consultar en https://archive.ics.uci.edu/dataset/2/adult
2. Identificá y separá las variables **numéricas** y **categóricas**.
3. Realizá un análisis descriptivo.
4. Decidir que tratamiento darle a valores faltantes.

Preguntas guía:

- ¿Qué variables numéricas y categóricas consideraste para el análisis?
- ¿Qué impacto tienen los valores faltantes en K-means?
- ¿Qué limitación supone ignorar las variables categóricas?

2. El script debe contener una implementación de la Distancia de Gower

1. Implementar la fórmula de Gower.
2. Validar con un subconjunto pequeño (ej. 5 filas).
3. Probar el cálculo de la matriz de distancias en una muestra pequeña del dataset.

Preguntas guía:

- ¿Cómo se normalizan las variables numéricas antes de calcular la distancia?
- ¿Por qué Gower es más adecuado que la distancia euclídeana en datos mixtos?
- ¿Qué problemas aparecen al calcular matrices de distancias muy grandes (de conjuntos de datos muy grandes)?

3: El script debe dar tratamiento conjuntos de datos grandes

Consignas:

1. Tomar muestras de 5.000 y 10.000 registros.
2. Calcular la matriz de distancias Gower para cada muestra.
3. Medir tiempo de ejecución y uso de memoria.
4. Comparar estabilidad de clusters en muestras de distinto tamaño.

Preguntas guía:

- ¿Qué diferencias observaste en tiempos de ejecución al aumentar el tamaño de la muestra?
- ¿Qué trade-off o compromiso hay entre usar todo el dataset y usar una muestra más pequeña?
- ¿Qué técnicas se podrían aplicar en problemas reales para escalar este análisis a millones de registros?

4: El script debe realizar un agrupamiento o clustering con K-means

Consignas:

1. Seleccionar las variables numéricas del dataset.
2. Escalar si es necesario (normalizar)
3. Aplicar K-means con distintos valores de k.
4. Determinar el número óptimo de clusters (Silhouette y/o Davies-Bouldin).

Preguntas guía:

- ¿Qué criterios usaste para definir el número de clusters?
- ¿Qué patrones observaste en los clusters formados?
- ¿Qué limitaciones tiene K-means al ignorar las variables categóricas?

5: El script debe realizar un agrupamiento o Clustering con AgglomerativeClustering y distancia de Gower

Consignas:

1. Usar la matriz de distancias de Gower calculada en Fase 3.
2. Aplicar AgglomerativeClustering[1] con `affinity="precomputed"` y `linkage="average"`.
3. Determinar un número adecuado de clusters (Silhouette y/o Davies-Bouldin).

Preguntas guía:

- ¿Qué diferencia hay entre un dendrograma jerárquico y los centroides de K-means?
- ¿Qué criterio usaste para elegir el número de clusters finales?
- ¿Qué ventajas observaste al usar variables categóricas con Gower?

[1] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

6: Se debe realizar una evaluación de Clusters

Consignas:

1. Calcular métricas de calidad: Silhouette y Davies-Bouldin.
2. Comparar resultados entre K-means (numérico) y Agglomerative (Gower).
3. Documentar similitudes y diferencias.

Preguntas guía:

- ¿Qué algoritmo obtuvo mejores métricas y por qué?
- ¿Qué se pierde al no considerar las variables categóricas en K-means?
- ¿Qué diferencias notaste en la interpretación de los clusters entre ambos métodos?

7: Se debe realizar una interpretación y reporte de comunicación

Consignas:

1. Describir características comunes de cada cluster.
2. Generar visualizaciones (gráficos comparativos, mapas de calor, etc.).
3. Redactar informe con metodología, resultados y conclusiones.

Preguntas guía:

- ¿Qué patrones socioeconómicos identificaste en los clusters?
- ¿Cómo comunicarías los resultados a un público no técnico?
- ¿Qué limitaciones encontraste en tu análisis y qué mejoras propondrías?

Referencias Iniciales

- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 857-871.

SISTEMAS INTELIGENTES

2025

Reflexión y Feedback (No evaluable)

> ⚠️ Esta sección **no será evaluada** en la nota del proyecto. Su único objetivo es que reflexionen sobre el proceso de aprendizaje y brinden sugerencias para mejorar la actividad en próximas ediciones de la materia.

Preguntas orientadoras:

1. **Aprendizaje personal:**
    - ¿Qué conceptos nuevos aprendieron al trabajar con clustering y la distancia de Gower?
    - ¿Qué parte del proyecto les resultó más desafiante?
    - ¿Qué estrategias usaron para superar esas dificultades?
2. **Utilidad percibida:**
    - ¿Consideran que este proyecto les aportó herramientas prácticas aplicables a otros problemas?
    - ¿En qué situaciones reales creen que podrían aplicar lo trabajado en este proyecto?
    - ¿Qué parte del trabajo consideran que fue más valiosa para su formación?
3. **Feedback para la cátedra:**
    - ¿Qué aspectos de la consigna estuvieron claros y cuáles podrían mejorarse?
    - ¿El tiempo asignado fue suficiente?
    - ¿Qué sugerencias tienen para hacer este proyecto más interesante o útil el próximo año?