Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Tarea 3: Minería de Datos IIC2433

Fecha de entrega : 10 de diciembre de 2019, 23:59 hrs.

Introducción

Clustering es una de las técnicas de análisis de datos exploratorios más comunes que se utilizan para obtener una intuición sobre la estructura de los datos. Se puede definir como la tarea de identificar subgrupos en los datos de modo que los puntos de datos en el mismo subgrupo (grupo) sean muy similares. Esta técnica se considera como un método de aprendizaje no supervisado ya que no tenemos la verdad fundamental para comparar la salida del algoritmo de agrupación con las etiquetas verdaderas para evaluar su rendimiento. Solo queremos intentar investigar la estructura de los datos agrupando los puntos de datos en subgrupos distintos

En esta tarea usted deberá implementar el algoritmo k-means.

k-means

El algoritmo Kmeans [Teknomo, 2006] es un algoritmo iterativo que intenta dividir el conjunto de datos en subgrupos (grupos) no superpuestos distintos donde cada punto de datos pertenece a un solo grupo. Trata de hacer que los puntos de datos entre clústeres sean lo más similares posible. Asigna puntos de datos a un grupo de modo que la suma de la distancia al cuadrado entre los puntos de datos y el centroide del grupo (media aritmética de todos los puntos de datos que pertenecen a ese grupo) es mínima. Cuanto menos variación tengamos dentro de los grupos, más homogéneos (similares) serán los puntos de datos dentro del mismo grupo.

En general:

- Especificar el número de grupos K.
- Inicializar los centroides para luego seleccionar aleatoriamente K puntos de datos para los centroides.
- Seguir iterando hasta que no haya cambios en los centroides, es decir, la asignación de puntos de datos a grupos no cambia.

Pasos a seguir

Para completar la tarea ustedes deben seguir los siguientes pasos:

1. **Cargar Datos**: deben preprocesar la base de datos *1-Base-de-Datos-Electoral.xlsx* [aud]
   Este archivo contiene una base de datos electoral, con datos de elecciones en Chile desde 1989 hasta 2013.

2. **Implementar** el algoritmo k-means desde cero. Para la implementación sólo podrá utilizar las librerías Numpy y Pandas.

3. **Aplicar y analizar**: el algoritmo programado deben aplicarlo a la base de datos y encontrar clusters, para las elecciones presidenciales, parlamentarias y municipales, de candidatos electos por partidos según las comunas. Considerar desde 1989 hasta 2013.
   Aquí, para facilitar el análisis pueden incorporar a independientes y partidos más pequeños dentro de los partidos más grandes. Para esto deben considerar los pactos y subpactos. En este punto es a libre elección cómo integrar los partidos, respetando siempre el espectro político.

4. **Visualizar**: Presentar una visualización de los clusters para las elecciones presidenciales, parlamentarias y municipales. Dar una breve intuición sobre lo obtenido.

1 Entrega

- La entrega debe ser realizada en un .zip con todos los archivos necesarios.
- El archivo de entrega debe ser subido al cuestionario abierto en el sistema SIDING específicamente para esta tarea hasta el día y hora señalada con el nombre $[numero\_alumno]\_T2$, tanto el .zip como el .ipynb.
- La tarea es estrictamente individual y el algoritmo debe ser implementado 100% (no usar funciones previamente implementadas o re-utilizar código).
- El documento principal debe ser un jupyter notebook con el código.
- Cualquier instrucción adicional y necesaria para la revisión debe ser escrita en un archivo README.txt contenido en el .zip

References

Kardi Teknomo. K-means clustering tutorial. *Medicine*, 100(4):3, 2006.

Auditoria a la democracia. `http://auditoriaalademocracia.org/`. Accessed: 2019-11-24.