PEC – 2

Análisis de Datos Multivariantes (Curso 2019-20)

Análisis de Conglomerados: Un caso práctico de protección de datos

Introducción

Durante la segunda parte del curso hemos discutido conceptos como el Análisis de Componentes Principales, Escalado multidimensional, Análisis de correspondencias y conglomerados. En esta prueba de evaluación se pide aplicar alguna de las técnicas estudiadas. Específicamente, nos centramos en el análisis de conglomerados y lo aplicamos a un problema de protección de datos.

Requisitos de la documentación: Máximo 15 páginas A4 con un tamaño de letra 11p. Por cada página extra se restará un punto a la nota final. La bibliografía y los apéndices con código R se pueden añadir sin limitación.

Conjunto de datos

Ponemos a disposición de los estudiantes el conjunto de datos AM20 – el mismo que en la PEC1. Éste es el conjunto de datos a analizar.

Protección de datos

Imaginemos que el conjunto de datos AM20 contiene datos de individuos. Cada registro (fila) pertenece a una persona. Cada variable (columna) de un registro hace referencia a una determinada característica de la persona (p.e., peso, altura, sueldo, etc.). En esta situación, el conjunto de datos no se podría hacer público puesto que contendría información privada. Sin embargo, si lo modificamos de algún modo podríamos distribuir el fichero sin vulnerar el derecho a la privacidad de los respondientes.

Hay muchos métodos que tienen como objetivo proteger los datos personales para que puedan ser divulgados a terceros para su posterior estudio. Uno de los métodos más comunes para hacerlo se basa en la agregación de registros similares. La agrupación (clústering) es un enfoque clásico para determinar qué registros agregar juntos. Primero, los registros se agrupan en subconjuntos (mediante el uso de técnicas de agrupación y análisis de conglomerados) y luego, cada registro (del conjunto de datos original) se reemplaza por el vector promedio del grupo en el que se ha agrupado, para crear un nuevo conjunto de datos modificado. Al hacerlo, los registros en este nuevo conjunto de datos no representan individuos únicos sino el promedio de varios individuos, por lo tanto, se protege su privacidad.

En un caso extremo, podríamos crear tantos clústeres como elementos en el conjunto de datos (clústeres de un solo elemento), en los que no hay protección en absoluto. En el otro extremo, podríamos crear un solo clúster (todo el conjunto de datos) y reemplazar todo el conjunto de datos por un solo promedio. Aunque este último enfoque garantiza la privacidad de las personas, la pérdida de información es tan alta que hace que el nuevo conjunto de datos sea inútil.

Nota: Para determinar la pérdida de información (IL), calculamos la diferencia valor a valor entre los elementos del conjunto de datos original y los del nuevo conjunto de datos modificado, luego los elevamos al cuadrado (para eliminar el signo) y los sumamos todos, como se muestra en la siguiente ecuación:

$$IL = \sum_{i=0}^{n} \sum_{j=0}^{p} (d_{ij} - \bar{\bar{d}}_{ij})^2$$

, donde n es el número de registros, p es el número de variables, d_{ij} es el elemento *ij* del conjunto de datos original, i $\bar{\bar{d}}_{ij}$ es el elemento *ij* del conjunto de datos modificado.

Por lo tanto, es necesario encontrar un equilibrio entre la protección y la pérdida de información. En general, para encontrar este equilibrio, los oficiales de protección de datos de las agencias de protección utilizan la cardinalidad de los grupos/clústeres creados. Dado un objetivo de cardinalidad (p.e., clústeres entre 10 y 15 elementos), se dice que el mejor algoritmo de agrupamiento es el que garantiza el objetivo de cardinalidad y obtiene la IL más baja.

Objetivos de la PEC

A partir del conjunto AM20 pedimos a los estudiantes que realicen las siguientes tareas:

1. Aplicar técnicas de clústering (al menos dos): Los estudiantes pueden usar métodos estudiados durante el curso, combinaciones de métodos e incluso sus propios métodos inventados con el objetivo de:
    a. Crear clústeres que cumplan con lo siguiente:
        i. CASO 1: Cardinalidad entre 10 y 15
        ii. CASO 2: Cardinalidad entre 20 y 30
2. Estudiar la pérdida de información asociada con los métodos de clústering usados, aplicados al CASO1 y al CASO2. Esto es, los estudiantes deberán computar el IL para cada caso.
3. Analizar los resultados obtenidos de forma crítica: Por ejemplo, analizar la relación entre la pérdida de información y la cardinalidad.

Evaluación

Para evaluar esta PEC consideramos los siguientes criterios:

- 50% Uso del conocimiento adquirido durante el curso, especialmente aquel relacionado con el análisis de conglomerados.
- 20% Capacidad de usar R para resolver el problema planteado.
- 30% Originalidad, corrección y calidad de la solución y la documentación.

Fechas importantes

- Enunciado: 06 mayo, 2020
- Entrega: 03 junio, 2020
- Solución: 10 junio, 2020
- Evaluación: 10 junio, 2020