ENUNCIADO – PEC 2

Análisis Multivariante de Datos (Curso 2015-16)

Análisis de conglomerados: Un caso práctico aplicado a la protección de datos

Introducción

En la segunda parte del curso hemos trabajo conceptos de Análisis de Componentes Principales, Escalado Multidimensional, Análisis de Correspondencias y Análisis de Conglomerados. En esta PEC se pide la aplicación de algunos de los conocimientos adquiridos durante la segunda parte del curso. En especial nos centramos en el análisis de conglomerados, que aplicaremos a un problema de protección datos.

Conjunto de datos

Para facilitar el trabajo a los estudiantes, usaremos el conjunto de datos CENSUS, que es el que ya se ha usado en la PEC1. Este conjunto de datos ha sido usado en el proyecto CASC (http://neon.vb.cbs.nl/casc/index.htm) como conjunto de test para estudiar y proponer técnicas de protección de datos en el campo del control de la revelación estadística (en Inglés: Statistical Disclosure Control)[1]. CENSUS contiene 1080 registros (filas) y 13 atributos (columnas) con datos extraídos de la base de datos del *U. S. Bureau of the Cesus*. La información de CENSUS hace referencia a salarios, impuestos, beneficios y sueldos pagados por empresas y particulares Americanos durante 1995. Para una descripción más completa puede consultarse (http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf).

El conjunto de datos CENSUS se encuentra públicamente disponible en: http://neon.vb.cbs.nl/casc/CASCrefmicrodata.zip . La Tabla 1 contiene una muestra de los datos de CENSUS.

AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
270914	45554	4173	4621	45527	1428	30809	27	27	45500	3480	45500	45500
250802	57610	2639	6045	42008	1902	39234	1008	808	41000	3136	41000	41000
...	...	...	...	...	...	...	...	...	...	...	...	...

Tabla 1: Muestra de los datos contenidos en el conjunto CENSUS

Protección de datos

Existen muchos métodos que tienen por objetivo proteger datos de carácter personal provenientes de individuos o empresas concretas (i.e. microdatos) con el fin de poder cederlos a centros de investigación y universidades para su uso secundario.

[1] Si se desea más información sobre el control de la revelación estadística, se puede consultar el siguiente manual gratuito: http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

El conjunto CENSUS, es un conjunto de microdatos y el objetivo de esta PEC será protegerlo. Los dos métodos más comunes para la protección de datos para uso secundario son:

- Adición de ruido Gaussiano: El método de adición de ruido Gaussiano se basa en perturbar los datos reales con ruido resultado de muestrear una distribución Gaussiana (habitualmente con media nula). Al añadir este ruido a los datos se busca evitar que un atacante pueda re-identificar a los individuos o empresas cuyos datos están siendo publicados. Por ejemplo, si añadimos ruido a la cuenta de resultados de una empresa será más difícil identificar a dicha empresa a partir de esta información (puesto que los datos no coincidirán con lo esperado por el atacante).

- Microagregación: El método de microagregación persigue el mismo objetivo que el de ruido Gaussiano (i.e. proteger a los individuos y empresas de la re-identificación por parte de un atacante) pero sigue un proceso distinto basado en agrupación. El procedimiento de microagregación de un conjunto de datos puede resumirse en los siguientes pasos:
    1. Agrupación de los registros del conjunto original: Se aplican técnicas de análisis de conglomerados para agrupar los distintos registros (1080 en el caso de CENSUS) en subconjuntos disjuntos. Los subconjuntos acostumbran a tener una cardinalidad[2] constante (habitualmente $k = 3, 4, 5... 10$) pero también puede ser variable.
    2. Para cada subconjunto S_i de registros obtenido en el paso anterior, se calcula su vector de medias V_i.
    3. Se substituye cada registro del conjunto original por el vector de medias V_i del subconjunto en el que se ha clasificado el registro. Con esta substitución se consigue un nuevo conjunto de datos formado únicamente por vectores de medias (repetidos), con lo que las empresas o individuos pertenecientes al mismo grupo son indistinguibles.

Ambos métodos (i.e. adición de ruido y microagregación) persiguen el objetivo de modificar los datos de forma suficiente como para proteger a los individuos frente a la re-identificación pero a la vez se quiere mantener un alto grado de utilidad de los datos (por lo que no se pueden modificar demasiado). Para el caso de la microagregación, el criterio de protección se fija mediante la cardinalidad de los subconjuntos (como hemos dicho en el punto 1). Por regla general a mayor cardinalidad se obtiene mayor protección pero también mayor pérdida de información.

Para calcular la perdida de información se acostumbra a usar la suma del error cuadrado (SSE), es decir, se calcula la diferencia valor-a-valor entre el conjunto de datos original y el conjunto de datos modificado, se eleva al cuadrado y se suma. Fijada una cardinalidad, cuando menor sea el valor de SSE mejor se considera el método de microagregación.

---

[2] Número de elementos en el conjunto

Objetivos

A partir del conjunto de datos CENSUS se pide:

1. Aplicar análisis de conglomerados: Los estudiantes pueden usar los métodos, conjuntos de métodos y/o modificaciones propias de los métodos estudiados u otros con el objetivo de:
    a. Realizar subconjuntos de cardinalidad variable, es decir, cada subconjunto puede tener distinta cardinalidad.
    b. Opcionalmente (para mejorar la calificación de la PEC) realizar subconjuntos de cardinalidad constante *(para k=3, k=4, k=5, k=10)* que reduzcan la pérdida de información (SSE). Nótese que obtener conjuntos de cardinalidad constante no es inmediato (p.e. al aplicar k-means, el número de grupos es "k" pero la cardinalidad de cada grupo no es fija.
2. Estudiar la pérdida de información asociada a los métodos estudiados en el apartado anterior. Es decir calcular el SSE para cada método y cardinalidad estudiados.
3. Analizar críticamente los resultados obtenidos: Por ejemplo, se puede estudiar la relación entre la cardinalidad y la pérdida de información, la distancia entre los centros de masa de los grupos, etc..

Evaluación

Para evaluar la PEC se tendrán en cuenta los siguientes conceptos:

- 50% Aplicación de conocimientos adquiridos en los cap. 5 - 8 del texto base, con especial énfasis en los métodos de agrupación.
- 20% Capacidad de usar R para resolver el problema planteado
- 30% Originalidad, corrección y calidad de la solución y la documentación

¿Porqué evaluamos así?

En esta PEC se da un peso del 50% a la aplicación de conocimientos adquiridos durante la segunda parte del curso. Nos concentramos en análisis de conglomerados pero otras técnicas de reducción dimensional también pueden ser de utilidad. (Para obtener la máxima calificación en este apartado no es necesario usar R).

Dado que el planteamiento de la asignatura es aplicado, se valora hasta un 20% la capacidad de usar R para resolver el problema planteado.

Finalmente un 30% de la calificación recae en la originalidad, corrección y calidad de la solución propuesta y la documentación entregada por el estudiante. Dado que se está cursando un Máster, se espera que el estudiante demuestre capacidad de resolver problemas de forma original y sea crítico con la calidad y corrección de la solución y la documentación.

Fechas importantes

- Publicación del Enunciado: 12 de Mayo de 2016
- Entrega (Fecha límite): 3 de Junio de 2016
- Publicación de la Solución: 8 de Junio de 2016
- Publicación de las calificaciones: 8 de Junio de 2016