ENUNCIADO – PEC 1

Análisis Multivariante de Datos (Curso 2015-16)

Descripción de datos multivariantes, análisis gráfico y detección de valores atípicos

Introducción

Durante las primeras semanas de curso nos hemos iniciado en el análisis multivariante. Tras una introducción al tema y un repaso de conceptos de álgebra matricial, hemos estudiado diversas formas de describir los datos multivariantes y representarlos gráficamente. También estamos comenzado el estudio de los valores atípicos (*outliers*). En esta PEC se trabaja la aplicación de los conocimientos adquiridos durante estas primeras semanas de curso y se propone la ampliación del estudio de los valores atípicos.

Conjunto de datos

Se proporciona a los estudiantes un conjunto de datos multivariante conocido con el nombre de CENSUS. 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

[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

Objetivos

A partir del conjunto de datos CENSUS se realizar las siguientes tareas:

1. Describir el conjunto de datos
    a. Usar medidas de centralización y analizarlas
    b. Usar medidas de variabilidad y analizarlas
    c. Estudiar las dependencias lineales entre las variables
2. Representar gráficamente los datos
3. Estudiar la existencia de valores atípicos *(R puede ser de mucha ayuda)*
    a. Identificar y aplicar diversas (al menos dos) técnicas
    b. Comparar los resultados de las técnicas estudiadas
    c. Analizar de forma crítica los resultados

Evaluación

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

- 50% Aplicación de conocimientos adquiridos en los cap. 1 - 4 del texto base
- 20% Capacidad de usar R para resolver el problema planteado
- 30% Originalidad (especialmente el Objetivo 3), 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 las primeras semanas de curso, es decir, si el estudiante demuestra saber aplicar los conocimientos relativos a la descripción de datos multivariantes, su análisis gráfico y la detección de valores atípicos habrá superado la PEC. (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 (u otro software convenido, a priori con el profesor) para resolver el problema planteado. No es necesario tener un nivel elevado de programación en R sino demostrar que se saben usar algunas de sus librerías para tratar datos multivariantes y describirlos, analizarlos gráficamente y detectar valores atípicos.

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. Se recomienda usar LaTeX para la documentación (aunque no es obligatorio).

Fechas importantes

- Publicación del Enunciado: 30 de Marzo de 2016
- Entrega (Fecha límite): 20 de Abril de 2016
- Publicación de la Solución: 4 de Mayo de 2016
- Publicación de las calificaciones: 6 de Mayo de 2016