Universitat Oberta de Catalunya

Aula

M2.854 - Estadística avanzada aula 1

A1 - Preprocesado de datos

Inicio:	Entrega:	Solución:	Calificación:	Dedicación:
07/03/18	**03/04/18**	**11/04/18**	**11/04/18**	**20 %**

Descripció i enunciat

El objetivo de esta actividad es conocer y saber aplicar los métodos de preprocesado de los datos. El preprocesado de los datos es la fase del proceso de análisis de datos que sigue a la recolección de datos y es previa al modelado posterior, que constituirá el núcleo de la extracción de información de los datos. A menudo, se atribuye poca importancia a la fase de preproceso y en muchos cursos de analítica de datos se pasa por alto o se trata sólo superficialmente. Sin embargo, en la práctica real es una fase muy necesaria, sin la cual los datos no están preparados para el análisis posterior. Además, en la práctica real la fase de preproceso ocupa un elevado esfuerzo, tiempo y recursos.

El objetivo concreto de esta actividad es preparar el fichero de datos para su posterior análisis. Habitualmente, el analista de datos no tiene orientaciones precisas sobre qué tipo de preproceso debe aplicar, ya que ello surge de la inspección del fichero que el propio analista realiza. A continuación, se enumeran algunas de las situaciones que es necesario tratar:

- Verificación de la transmisión de la información
  - Tipos de variables estadísticas
  - Defectos de formato
  - Ingruencias de la información
- Limpieza de datos:
  - Errores sintácticos
  - Normalizar/estandarizar variables
  - Valores atípicos (outliers)
  - Valores perdidos (missing)

Enunciado

En el fichero adjunto se describe el enunciado especifico de la actividad junto con los pasos a seguir. Este enunciado estará disponible el día de inicio de la actividad.

FORMATO DE ENTREGA

Se debe enviar un **único fichero ZIP** que contenga los análisis realizados y sus resultados. El fichero ZIP debe contener:

1. El fichero Rmd (fichero fuente con el código y el texto)
2. El fichero de salida del Rmarkdown, es decir, el informe que genera, en formato html o en PDF.
3. El fichero de datos después del preproceso.

Para facilitar la corrección incluid vuestro apellido en los nombres de los ficheros, de la siguiente forma:

- **apellido_preproceso.zip** , el cual incluye:
  1. **apellido_preproceso.Rmd**
  2. **apellido_preproceso.html** o **apellido_preproceso.pdf**
  3. **apellido_immuebles_clean.csv**

Cuando realicéis el informe, debéis incluir para cada paso del enunciado:

- La pregunta particular a la que se quiere dar respuesta
- El código R utilizado
- La salida del código R
- Una explicación (si es necesaria) que complemente el resultado del código R.

Para mejor estructuración del código, se aconseja desarrollar cada una de las preguntas en un fragmento distinto (code chunk) de código R.

Es imprescindible que el informe se realice mediante la **herramienta RMarkdown** que permite generar informes dinámicos. En la actividad 0 tenéis instrucciones y ejemplos de cómo realizar un informe dinámico.

Es muy importante que recordéis que no se corrigirán entregas que contengan listados de datos de más de 1 página , puesto que genera informes muy largos y difíciles de seguir. En su lugar, podéis mostrar la cabecera del fichero (usando la función head de R) o sólo los cambios realizados.

No se corregirán entregas que sólo contengan el fichero .Rmd. El motivo es que para poder corregirlos, estos ficheros se deben ejecutar y al ejecutarlos pueden contener errores o puede que sea necesaria la instalación de librerías. A la hora de corregir, el profesor no puede subsanar estos posibles errores. Es por esto que debéis entregar el fichero de salida PDF o HTML donde se muestre la salida de la ejecución del código pertinente. El fichero RMD que entregáis sólo se usa para comprobaciones complementarias. Pero la revisión se realiza principalment sobre el fichero de salida PDF o HTML.

2016_raw.csv
A1_pre-proceso_Enunciado_CAST.pdf

Objectius i competències

Los objetivos a desarrollar en la actividad 1 son:

- Conocer en qué consiste el preprocesado de los datos y su importancia en el proceso de análisis de datos.
- Clasificar los distintos tipos de preprocesado y el objetivo de los mismos.
- Para cada tipo de preprocesado, enumerar qué tipos de métodos se pueden aplicar y para qué, así como conocer sus ventajas y limitaciones.
- Enumerar los tipos de errores/inconsistencias que pueden existir en los datos.
- Saber identificar en un fichero de datos qué tipos de preprocesado son necesarios, en función de los datos y según el contexto del análisis.
- Saber aplicar los métodos de preprocesado usando la herramienta R.
- Adquirir un buen manejo de las herramientas de R de manipulación de datos y de preprocesado de datos.
- Desarrollar un espíritu curioso para poder explorar en los datos y a la vez, analítico y crítico para elaborar conclusiones sobre los análisis, identificar limitaciones y alternativas de mejora.

Continguts i recursos

Para la realización de esta actividad y para alcanzar los objetivos de aprendizaje y competencias, es necesaria la lectura del módulo:

- Preprocesamiento de los datos

Está adjunto en este apartado. Asimismo, los materiales proporcionados en la actividad 0 son necesarios, especialmente los que se refieren al uso del lenguaje R (módulos R, ejercicios R y RMarkdown). Podéis consultar la actividad 0 para descargar estos materiales si no lo habéis hecho anteriormente.

Solució

La solución estará disponible el día que se establece en la actividad.