Cuestiones y terminología de la Ciencia de Datos

Christopher González Delgado

1. Cuestiones y Términos II

1.1. ¿Qué es el pre-procesamiento de datos?

El pre-procesamiento de datos consiste en convertir o "mapear"los datos desde su primera versión en crudo a un formato que permita preparar los datos para un análisis posterior.

1.2. Objetivos y tareas del pre-procesamiento de datos

Los principales objetivos y tareas del pre-procesamiento son, principalmente :

- Identificar y tratar con datos desconocidos (missing values)
- Formateo edición del formato de los datos
- Normalización de datos
- *Binning* de los datos
- Conversión de variables categóricas en numéricas

1.3. Valores faltantes o desaparecidos (missing values). Posibles soluciones

Los valores faltantes son aquellos de los que no se tienen registro para una característica en una observación dada. Se puede representar como *?*, *N/A*... Las principales soluciones que se dan para los datos faltantes son:
Eliminar valores: en caso de no disponer de de un dato, podemos proceder a eliminar esa característica o la entrada. **Reemplazar el valor:** en caso de ser numérico podemos reemplazar el dato desaparecido por la media. En caso de ser categórico, podemos utilizar la categoría más frecuente (reemplazo por frecuencia).

1.4. Formateo o edición del formato de los datos

Es un proceso que se encarga de convertir los datos a un mismo formato. Cuando los datos son recolectados, por lo general se obtienen de diferentes

fuentes y se almacenan en distintos formatos, e.g. fechas, horas o abreviaturas. El formateo de datos consiste en convertir los datos en una forma estándar con el fin de trabajar con ellos de una forma más cómoda, más clara y que facilita las comparaciones o la visualización.

1.5. Normalización de datos

Este proceso es bastante importante a la hora de trabajar con datos. La normalización consiste en llevar los valores numéricos de nuestras características a un mismo rango de forma que podamos asegurar que todas las entradas tengan la misma importancia o el mismo impacto. Por ejemplo, si tenemos dos características *edad* y *cantidad de pasos diarios*, veremos que son datos que se encuentran en rangos numéricos realmente diferentes y que hará que sea muy difícil su comparación. El número de pasos, al ser mucho mayor en número que la edad, tendrá más importancia.

Tras la normalización de los datos tendremos todos los valores en rangos similares y además se logrará que la influencia de cada característica sea la misma.

1.6. Binning de datos

Binning consiste en la separación o distribución de los datos en grupos. En lugar de trabajar con todas las edades, se puede trabajar con niños, adultos y ancianos. Esto permite agrupar números en variables categóricas. En algunas ocasiones, realizar este procedimiento puede mejorar la precisión del modelo predictivo.

1.7. Conversión de variables categóricas en numéricas

Existe un problema y es que resulta que la mayoría de modelos estadísticos no permiten textos (strings) como variable de entrada. La solución a este problema consiste en asignar una variable muda a cada categoría y darle valores numéricos, e.g., 0 o 1. Por ejemplo, si tenemos una categoría çolor"que incluye azul o rojo, podemos sustituir la categoría color por dos categorías .ªzulz rojo". Se asignará el valor 0 o 1 a azul o rojo según el color.

1.8. ¿Qué es el Análisis Exploratorio de Datos?

El Análisis Exploratorio de Datos (en inglés EDA, Exploratory Data Analysis) son una serie de pasos preliminares en el análisis de datos que permiten resumir las principales características de los datos, obtener un entendimiento más profundo del set de datos, descubir relaciones entre las variables y extraer variables importantes que posiblemente pasáramos por alto en primera instancia.

1.9. Correlación entre variables

La correlación nos permite conocer en qué medida las diferentes variables son interdependientes. Es decir, si tenemos dos variables y modificamos una de ellas, ¿de qué forma varía la otra? Un ejemplo de dos variables correlacionadas son el uso de paraguas y si está lloviendo o no.

1.10. Correlación de Pearson

La correlación de Pearson mide cómo de correlacionadas están dos variables, i.e. la fuerza con la que dos variables están correlacionadas. Este método nos proporciona dos valores: el coeficiente de correlación y el valor P.
El coeficiente de correlación nos indica, de alguna forma, la cantidad de correlación entre dos variables. Toma valores entre [-1,1], pudiendo ser gran correlación positiva o negativa en los casos de extremo.
El valor P nos dice qué tan seguros estamos del resultado que hemos obtenido.