La ingeniería de características

La ingeniería de características es un proceso fundamental para mejorar la representación de los datos para los algoritmos de aprendizaje que se emplearán en el modelado. Algunas definiciones que podemos encontrar son:

- "…proceso de transformar los datos en características que representan mejor el problema subyacente, lo que permite mejorar el aprendizaje automático…" (Ozdemir y Susarla, 2018).
- "… proceso de crear representaciones de los datos para incrementar la efectividad de un modelo…" (Kuhn y Johnson, 2020).

Se trata entonces de aplicar un conjunto de técnicas que nos permitan construir una mejor representación de los datos, tomando en cuenta las características del problema bajo estudio y, a su vez, los requerimientos de entrada de los algoritmos de aprendizaje. Es importante que siempre tengas en cuenta que la forma de representar los datos puede tener un gran efecto en el rendimiento de los modelos que se generan. ¿Qué podemos hacer para construir esta nueva representación? Es decir, ¿qué conlleva la preparación de los datos? La Fig. 1 nos muestra:

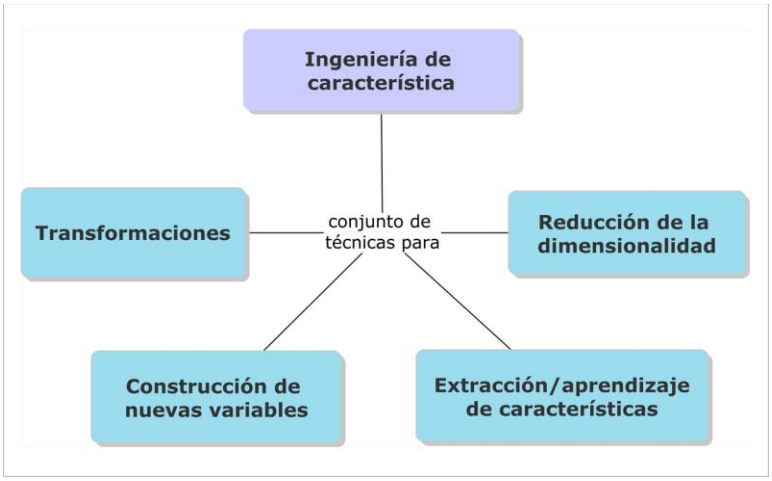

Fig. 1. Técnicas de la ingeniería de características.

Transformaciones. Permiten cambiar el formato de las variables originales mediante mapeos funcionales. Esto da lugar a nuevas variables, que sustituyen a las primeras en el conjunto de datos y, de alguna forma, el resultado puede ser interpretado como una manera de "ver" la variable bajo otra perspectiva. Algunos tipos de transformaciones son (ver Fig. 2):

Funcionales. Se aplica una función a cada valor de la variable. Por ejemplo, para atributos que tienen distribuciones asimétricas positivas se pueden usar las transformaciones $\sqrt{x}$ y log(x), que comprimen los valores altos y expanden los pequeños. El objetivo es producir una variable con una distribución simétrica y más cercana a la distribución normal. También es posible aplicar transformaciones a todo el conjunto de datos. Un ejemplo es la transformación polinomial, la cual genera nuevas variables cuyo número dependerá del grado del polinomio

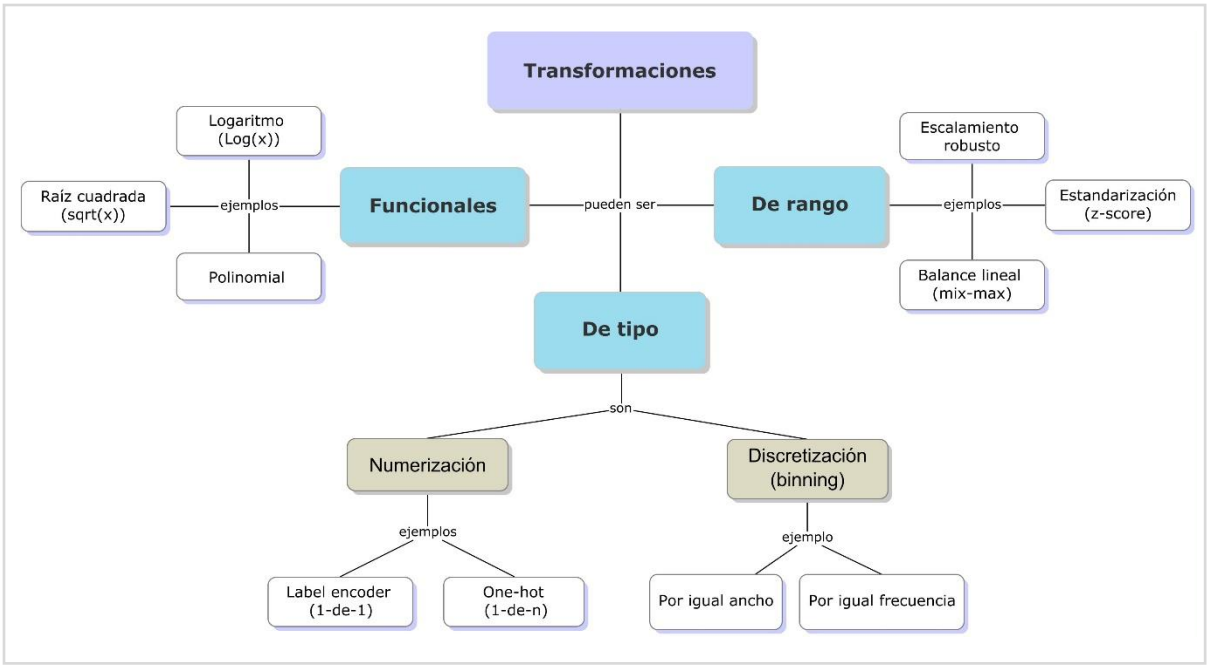

Fig.2. Algunos tipos de transformaciones.

De tipo. Estas transformaciones actúan en dos sentidos: cambiando un atributo categórico en una representación numérica, proceso que se conoce como numerización, o cambiando un atributo numérico en uno categórico, lo que se conoce como discretización o *binning*.

De rango. Estas transformaciones cambian el rango de valores de una variable numérica. También se conoce como escalado de los datos. Es importante aplicarla para los algoritmos de aprendizaje basados en métodos de distancias y los que utilizan procedimientos de optimización con el fin de mejorar la convergencia.

Construcción de nuevos atributos. Es posible generar nuevas variables a partir de las originales mediante el conocimiento del dominio. La idea es que estos atributos, que se añaden al conjunto de datos, capturen mejor la información más pertinente para la tarea. Por ejemplo, en un contexto médico relacionado con la determinación de los factores que más inciden en el riesgo de sufrir una enfermedad cardiovascular, podría resultar más informativo contar con el índice de obesidad calculado a partir de las variables peso y estatura como:

$$\text{Índice de obesidad} = \text{peso} / (\text{estatura})^2$$

Este nuevo atributo se añade al conjunto de datos. Las variables estatura y peso podrían entonces ser descartadas, lo cual implica, además, una forma de reducción de la dimensionalidad.

Extracción/aprendizaje de características. Se construyen nuevas características aplicando algoritmos desarrollados para este fin. Por ejemplo, en un problema relacionado con la clasificación de imágenes, se podrían extraer descriptores que resuman información para discriminar entre las clases, como la media y la varianza de los tres canales (en un formato RGB). También se pueden aplicar transformaciones a todo el conjunto de datos para generar otro espacio de características, que es utilizado entonces como entrada para el algoritmo de aprendizaje. Es el caso del análisis de componentes principales (PCA), el cual, además, se utiliza para la reducción de la dimensionalidad. Por último, el algoritmo podría aprender este nuevo espacio. Por ejemplo, las redes neuronales son capaces de aprender nuevas características a través de sus capas ocultas. Esto se conoce como aprendizaje de representación.

Selección de características. El conjunto de datos puede contener variables que no son informativas para el problema a resolver, así como características redundantes, irrelevantes (por ejemplo, números de identificación) o correlacionados. Utilizando diferentes criterios, los métodos de selección nos permitirán determinar las variables más informativas del conjunto de datos o descartar aquellas que no serán útiles para resolver el problema. De esta forma, podríamos obtener

un espacio de entrada de baja dimensionalidad, pero con la máxima información. Un aspecto para resaltar es que, al hacer la reducción de la dimensionalidad con estos métodos, se mantiene el espacio original de representación. Para realizar esta tarea podemos utilizar técnicas del análisis exploratorio, aplicar ciertos test estadísticos o utilizar métodos que, de manera automática, nos indican las variables más importantes para el problema.

Por último, es importante que tengas en cuenta que cada conjunto de datos tendrá diferentes caminos de preparación, por lo que será necesario determinar cuál es el adecuado para la tarea. Además, si los datos son complejos, como textos, habrá que cumplir más pasos de preparación para llevarlos a la representación vectorial que requieren, en general, los algoritmos de aprendizaje.

Bibliografía

Ozdemir, S., Susarla, D. (2018). Feature Engineering Made Easy. Packt Publishing Ltd.

Kuhn, M., Johnson, K. (2020). Feature Engineering and Selection. A Practical Approach for Predictive Models. CRC press.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.