

Procesamiento de datos

Pre-procesado de datos

Introducción

Tenemos una serie de datos, ya observados, la idea es suministrar estos datos al machine-learning y la máquina v **intentar definir una variable dependiente.**

Notas Python y R

- Cambiar formato decimales en spider por 0f, si no obtendremos los números en anotación científica.
- Python inicia a contar desde 0
- R inicia a contar desde 1

Variables Machine Learning

- Variables independientes → Son las variables que le daremos al algoritmo para intentar predecir.
- Variables dependientes → Es la variable que queremos predecir.

Datos Desconocidos

Cuando nos encontramos con ausencia de valores podemos optar por introducir la media o la moda de dicha columna, siempre será mejor esto que poner ese valor a 0.

Datos Categóricos

Se tratan de esos valores que su columna en vez de tener un número contiene un valor para catalogar o clasificar un usuario.

Variable dummy → traducir una variable a variable categorica sin orden. La variable dummy clasificada en activa

Set de entrenamiento y Set de Test

Over fitting → problema que hay que intentar evitar. El algoritmo no tiene comparaciones suficientes y aprende lo

- 70% o 80% para entrenamiento
- 20% o 30% para testing

Escalado de datos

Diferencias de rango de valores, ejemplo edad(27) y salarios(51000). El efecto de la edad pasaría inadvertido en nuestro algoritmo de machine learning.

Si tenemos una variable cuyo rango de valor es muy superior a las otras, podría ser un problema porque las varia

- Normalización de valores menor valor -1 mayor valor 1: Standarización o Normalización
- Standarización → Permite aglutinar valores en torno a la media
- Normalización → El más pequeño es 0 el mayor es 1

[Normalizacion.png](#) unknown

Revision #1

Created 28 November 2023 19:33:54 by adminROM

Updated 28 November 2023 19:34:04 by adminROM