Resumen corto de lecturas sesión asincrónica.

Lectura procesamiento de datos

Fuente: <https://towardsdatascience.com/data-preprocessing-in-python-b52b652e37d5>

El preproceso de datos se compone primero por selección de los datos y atributos relacionados, y segundo por creación y cambio de atributos. Para iniciar el proceso se usa lenguaje Python y el navegador de Anaconda. Se inicia por importar diferentes librerias, se destacan las vistas en clase como numpy y pandas. Segundo se importa el dataset o información en “csv”, en artículo también definen “x” y “y”, y se genera una tabla resumen.

Se continua con la revisión de los valores faltantes, se remplazan con nan y por el promedio de los valores de la columna. Luego se da el manejo de variables de texto dividiéndolas en nuevas columnas de tipo binario. Posterior se indica la opción de separar los datos en entrenamiento y prueba con porcentajes diferentes, no se profundiza el tema. Finalmente se realiza una transformación de variables para que en las diferentes columnas se trabajen las mismas escalas.

Lectura de preprocesamiento de datos con Python Pandas: Binning

Fuente: <https://towardsdatascience.com/data-preprocessing-with-python-pandas-part-5-binning-c5bd5fd1b950>

Se explica corresponde a la agrupación de datos para reemplazar valores por uno representativo del grupo para mejorar eficiencia de predicción. Se usa para transformar datos en números a caracteres y/o reducir cantidad de valores. Se presenta la función de indicar valor mínimo y máximo para luego crear 4 límites para el caso observado, se utiliza la función linspace() para el cálculo equilibrado de los mismos. Luego se usa función cut con las categorías creadas definiendo “bins”.

Se presentan otras formas de definir los límites en categorias como manualmente o por frecuencia calculada. Se explica el sampling donde disminuyen la cantidad de datos, unificando por funciones estadísticas.

Lectura de normalización de datos

Fuente: no funciona.