Lic. en Sistemas de Información
TP. 2 - Elementos de Inteligencia Artificial
Preprocesamiento y análisis de datos

La calidad de los datos y la cantidad de información útil que contienen son factores clave que determinan lo bien que puede aprender un algoritmo de aprendizaje automático. Por lo tanto, es absolutamente imprescindible que nos aseguremos de examinar y preprocesar un conjunto de datos antes de lanzarlo a un algoritmo de aprendizaje. (Sebastian Raschka - Vahd Mirjalili, 2019).
La performance de un sistema de machine learning es altamente dependiente de la calidad del conjunto de datos de entrenamiento. (Abhiskey Mishra. 2019)

EJERCICIO 1 - ESTANDARIZACIÓN DEL DATO.

En IA se suelen estandarizar los datos antes de ingresarlos a los algoritmos. En forma esquemática, se reduce la escala de los datos según se ilustra en la Figura 1.

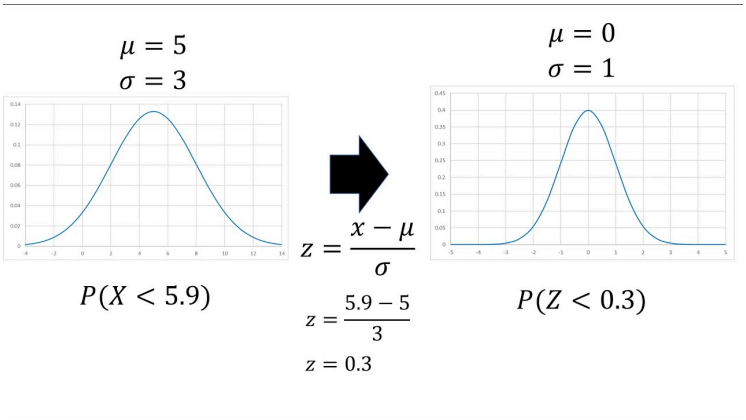

Fig 1. Estandarización del dato.

El desvío estándar de una muestra proveniente de una población determinada, puede calcularse con la siguiente fórmula:

$\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)}$ donde x_i es el i-ésimo elemento de la muestra, n es el tamaño muestral y $\bar{x}$ es la media muestral.

1. Desarrolle una función en Python para estandarizar los datos un arreglo numérico genérico. Por ejemplo, entradas para el programa podrían ser:

$entrada1 = [2, 3, 4, 5, 6, 1, 1, 1, 1, 2, 3, 4, 6, 9]$.
$entrada2 = [2.3, 2.3, 2.3]$
$entrada3 = [2.11111, 4, 4, 4, 4, 4, 4, 4, 4, 99, 99, 1, -5, -99]$
Sugerencia: utilizar **len()** para calcular la cantidad de elementos de una lista.
Recordar que la fórmula de estandarización es $z = \frac{x-\bar{x}}{\sigma}$

2. Para el caso de la $entrada3$, ¿cómo la media es afectada por los valores atípicos? ¿Es una buena medida para representar la tendencia central de los datos frente a un conjunto de datos que presenta múltiples valores atípicos ?
3. graficar una distribución normal, investigue cómo realizarlo en python. Compare con las distribuciones de entrada 1, entrada 2 y entrada 3 y saque conclusiones.

EJERCICIO 2. TIPOS DE VARIABLES Y CODIFICACION

Por cada grupo, visite el siguiente enlace: https://www.kaggle.com/datasets y elija un dataset, cada dataset debe ser diferente por grupo.

1. Las variables de tipo cualitativas pueden ser ordinales y nominales. Investigar el significado de esto y plantear dos ejemplos de cada una. ¿Qué diferencia tiene una y otra?
2. Seleccione un dataset de su preferencia y codifique al menos dos variables ordinales y nominales en su correspondiente codificación numérica.

- Para el tipo ordinal, dar un orden en formato de ranking
- Para el tipo nominal, descompone la variable en dummy variables.

EJERCICIO 3. MANEJO DE VALOR NULO.

Por cada grupo, visite el siguiente enlace: https://www.kaggle.com/datasets y elija un dataset, cada dataset debe ser diferente por grupo. Estudiar el capítulo 4 del libro sugerido por la cátedra de procesos de datos para manejar los valores nulos. Debe realizar las siguientes actividades en el dataset. Tome al menos 2 columnas, una categórica y otra numérica que contengan al menos un valor nulo y realice las siguientes operaciones utilizando pandas.

3. 1.Exploración Inicial:

- Carga el conjunto de datos.
- Realiza un análisis exploratorio para identificar la cantidad de valores nulos por columna.
- Visualiza la distribución de los datos para cada columna, incluyendo los valores nulos.

3.2. Estrategias de Manejo de Datos Nulos:

- Estudiar cuándo utilizar una técnica de imputación u otra. La defensa se basa en que puedan entender como trabajar el valor nulo.
- Revisar si corresponde:
    - Eliminación: Elimina las filas que tienen cualquier valor nulo y analiza el impacto de esta eliminación en el tamaño del conjunto de datos y en las distribuciones de las variables.
    - Imputación: Decide una estrategia de imputación (por ejemplo, usar la media o mediana para numéricas, utilizar más frecuente para categóricas)

EJERCICIO 4 - INTRODUCCIÓN A LA SELECCIÓN DE VARIABLES IMPORTANTES

El método más simple de selección de variables en estadística y aprendizaje automático es probablemente el "Filtro de Variables" (Feature Filtering). Este método implica evaluar la importancia de cada variable independientemente del modelo que se va a utilizar. La selección se realiza en función de medidas estadísticas que capturan la relevancia de las variables con respecto a la variable objetivo.
Ejemplo de técnicas comunes son: Anova, Chi-Cuadrado, Correlación de Spearman / Pearson
Por cada grupo, visite el siguiente enlace: https://www.kaggle.com/datasets y elija un dataset, cada dataset debe ser diferente por grupo. Estudie como las variables estan correlacionadas entre sí, focalizándose sólo en la correlación de Spearman y Pearson. Estudie cómo crear matrices o heatmaps con python para evaluar variables de forma bi-variada y cruzada, y analizar si ellas son o no independientes.

Bibliografía consultada

Python Machine Learning - Aprendizaje automático y aprendizaje profundo con Python - scikit-learn y TensorFlow. Sebastian Raschka - Vahd Mirjalili Segunda Edición, año 2019.

Machine Learning in the AWS Cloud. Add INtelligence to Applications with Amazon SageMaker and Amazon Rekognition. Abhiskey Mishra. 2019.