Tema 2: Preprocesamiento

El preprocesamiento de datos es un conjunto de tareas cruciales para preparar los datos antes de realizar análisis estadísticos o de aprendizaje automático. El objetivo principal es garantizar que los datos sean de alta calidad, consistentes y adecuados para el análisis posterior. A continuación, se presenta un resumen del preprocesamiento dividido en capítulos.

Capítulo 1: Estadística Descriptiva

Este capítulo se centra en comprender las características básicas de los datos mediante medidas estadísticas descriptivas.

- **Medidas de centralidad:** Estas medidas, como la media, la mediana y la moda, **resumen la tendencia central de los datos** . La media es el promedio de todos los valores, la mediana es el valor que divide los datos ordenados por la mitad, y la moda es el valor que aparece con mayor frecuencia. La elección de la medida de centralidad depende del tipo de datos y la presencia de valores atípicos. Por ejemplo, **la mediana es más robusta a los valores atípicos que la media**.
- **Medidas de dispersión:** Estas medidas, como la varianza, la desviación estándar y el rango, **cuantifican la dispersión de los datos alrededor de la medida de centralidad** . La varianza mide el promedio de las diferencias cuadradas entre cada valor y la media, la desviación estándar es la raíz cuadrada de la varianza, y el rango es la diferencia entre el valor máximo y el valor mínimo. Estas medidas ayudan a comprender **cuán dispersos están los datos**. Un rango intercuartílico alto indica una mayor dispersión de los datos.

Capítulo 2: Tratamiento de valores perdidos

Los valores perdidos (missing values) son un problema común en los conjuntos de datos. Este capítulo aborda las estrategias para manejarlos:

- **Identificación:** El primer paso es identificar los valores perdidos en el conjunto de datos.
- **Análisis:** Es importante analizar la naturaleza de los valores perdidos. ¿Son aleatorios o hay un patrón en su aparición?
- **Imputación:** Si es apropiado, los valores perdidos se pueden reemplazar con valores estimados. Existen varios métodos de imputación, como la imputación por la media, la mediana o la moda, o mediante métodos más sofisticados como la regresión o la imputación múltiple.
- **Eliminación:** En algunos casos, puede ser necesario eliminar los registros con valores perdidos si la cantidad de valores perdidos es significativa y la imputación no es factible.

Capítulo 3: Detección y tratamiento de valores atípicos (outliers)

Los valores atípicos (outliers) son observaciones que se desvían significativamente del patrón general de los datos. Este capítulo se centra en identificar y tratar estos valores:

- **Detección:** Se pueden utilizar métodos como el análisis de diagramas de caja (boxplots) y la identificación de valores que se encuentren fuera de un rango determinado (por ejemplo, más de 3 desviaciones estándar de la media) para detectar outliers. El rango intercuartílico también se utiliza para detectar outliers leves y extremos.
- **Tratamiento:** Las opciones para tratar los valores atípicos incluyen la eliminación, la transformación (por ejemplo, aplicar una transformación logarítmica) o la sustitución por valores más plausibles. La elección del método depende de la naturaleza del outlier y el objetivo del análisis.

Capítulo 4: Escalamiento de datos (Estandarización)

El escalamiento de datos es una técnica que se utiliza para ajustar el rango de las variables a una escala común. Esto es especialmente importante cuando las variables se miden en unidades diferentes o tienen rangos muy diferentes. Este capítulo explora diferentes métodos de escalamiento:

- **Estandarización por rangos:** Esta técnica escala los datos para que estén en un rango de 0 a 1.
- **Estandarización Z-score:** Esta técnica transforma los datos para que tengan una media de 0 y una desviación estándar de 1.
- **Escalamiento decimal:** Implica dividir las variables por potencias de 10.

Capítulo 5: Selección de variables

La selección de variables consiste en elegir las variables más relevantes para el análisis, lo que puede mejorar la precisión del modelo y reducir el costo computacional. Este capítulo analiza diferentes enfoques para la selección de variables:

- **Análisis de correlación:** Se utilizan medidas de correlación, como el coeficiente de correlación de Pearson, para identificar las variables que están más relacionadas con la variable objetivo.
- **Técnicas de reducción de dimensionalidad:** Métodos como el análisis de componentes principales (PCA) se utilizan para reducir la dimensionalidad de los datos al tiempo que se conserva la mayor parte de la información original.

Capítulo 6: Ponderación de variables

La ponderación de variables consiste en asignar diferentes pesos a las variables en función de su importancia para el análisis. Este capítulo describe diferentes métodos de ponderación que se pueden utilizar en diferentes contextos, como el análisis de conglomerados y los problemas de regresión/clasificación.

Tema 3: Aprendizaje Supervisado

El **aprendizaje supervisado** es un tipo de aprendizaje automático que utiliza datos etiquetados para entrenar un modelo que pueda predecir resultados futuros. El modelo aprende de las relaciones entre las variables de entrada y las variables de salida, utilizando conocimientos previos (ejemplos) para obtener resultados futuros.

Capítulo 1: Clasificación vs Regresión

Dentro del aprendizaje supervisado, se pueden distinguir dos tareas principales: **clasificación** y **regresión**. La elección entre una u otra depende de la naturaleza de la variable que se quiere predecir.

1.1. Clasificación

En la clasificación, el campo objetivo es **categórico**, es decir, la etiqueta que acompaña a cada ejemplo es **discreta** y compuesta por valores finitos. El objetivo es predecir a qué categoría pertenece una nueva instancia basándose en los ejemplos de entrenamiento. Algunos ejemplos de problemas de clasificación son:

- Filtrado de spam: clasificar correos electrónicos como spam o no spam.
- Diagnóstico médico: predecir si un paciente tiene una enfermedad cardiovascular o no.
- Etiquetado de noticias: asignar etiquetas de contenido a una noticia, como deportes o política.

La validación de modelos de clasificación se realiza mediante el porcentaje de elementos clasificados correctamente, utilizando herramientas como la **matriz de confusión** y la **curva ROC**.

1.2. Regresión

En la regresión, el objetivo es predecir un **valor numérico continuo**. Se busca inferir las relaciones entre las variables para ofrecer una predicción sobre la salida requerida. Algunos ejemplos son:

- Predicción del precio de una casa.
- Pronóstico de ventas de un producto.
- Estimación de ingresos potenciales de un cliente.

La validación de modelos de regresión se basa en **métricas de error**, como el error absoluto medio (MAE) y el coeficiente de determinación (R^2).

Capítulo 2: Proceso del Aprendizaje Supervisado

El proceso de aprendizaje supervisado generalmente sigue estos pasos:

1. **Fase de entrenamiento/construcción del modelo:** Se utiliza un conjunto de datos de entrenamiento con ejemplos previamente etiquetados para encontrar un modelo que explique el valor objetivo. Este modelo puede ser una función matemática que asigna cada ejemplo de entrada a una de las clases predefinidas.
2. **Fase de test/uso del modelo:** Se evalúa la capacidad del modelo para procesar nuevas instancias utilizando un conjunto de datos etiquetados que no se utilizaron en el entrenamiento. El objetivo es proporcionar una estimación imparcial del rendimiento del modelo en producción.

Es crucial dividir los datos en entrenamiento y test **antes** de realizar cualquier transformación para evitar la fuga de datos.

2.1. Ajuste y Sobreajuste

Un aspecto importante del aprendizaje supervisado es la **capacidad de generalización del modelo**, es decir, su flexibilidad ante nuevos casos. El modelo debe encontrar un equilibrio entre ajustarse a los datos de entrenamiento y ser lo suficientemente general para predecir con precisión nuevas instancias.

Para lograr un buen ajuste y evitar el sobreajuste, se utiliza la **validación cruzada**. Este método consiste en dividir el conjunto de entrenamiento en subconjuntos más pequeños y entrenar/validar el modelo en diferentes combinaciones de estos subconjuntos.

Capítulo 3: Clasificación Bayesiana

Los modelos Naive Bayes son algoritmos de clasificación rápidos y sencillos, adecuados para conjuntos de datos de alta dimensionalidad. Se basan en el **Teorema de Bayes**, que describe la relación entre las probabilidades condicionales de diferentes variables.

3.1. Clasificación Bayesiana

La clasificación bayesiana busca la probabilidad de una etiqueta dada en función de las características observadas, expresado como P(L | F) donde F son las *features*. El Teorema de Bayes permite calcular esto en términos de cantidades más directas:

$$P(L/F) = \frac{P(F/L) \times P(L)}{P(F)}$$

El objetivo es construir un **modelo generativo** que establezca el proceso aleatorio hipotético que genera los datos. Para simplificar el cálculo, se utiliza la **hipótesis de independencia condicional**, que asume que los valores de los atributos son independientes entre sí dada la clase.

3.2. Clasificadores Gaussianos

Estos clasificadores utilizan **modelos generativos basados en la distribución gaussiana** para cada clase. Se ajustan encontrando la media y la desviación estándar de los puntos dentro de cada clase. Aunque se basan en suposiciones que no siempre se cumplen, los clasificadores gaussianos pueden ser útiles como punto de referencia inicial.

3.3. Clasificadores Multinomiales

En los clasificadores multinomiales, se asume que las características se generan a partir de una **distribución multinomial**, adecuada para características que representan conteos o tasas de recuento. Son muy utilizados en la **clasificación de textos**, donde las características se basan en la frecuencia de palabras en los documentos.

3.4. Ventajas e Inconvenientes de la Clasificación Bayesiana

Ventajas:

- Rápidos para el entrenamiento y la predicción.
- Proporcionan una predicción probabilística directa.
- Fáciles de interpretar.
- Pocos parámetros ajustables.
- Robustos al ruido y a datos incompletos.

Desventajas:

- Suposiciones muy estrictas sobre los datos, lo que puede afectar su precisión.
- Requieren eliminar las funciones correlacionadas.
- Problemas con frecuencias cero, que se pueden abordar con técnicas de suavizado como la estimación de Laplace.

Capítulo 4: Vecinos más Cercanos (KNN)

El algoritmo de vecinos más cercanos (kNN) es un algoritmo de aprendizaje supervisado sencillo que funciona por **analogía**. Se basa en la idea de que **objetos similares se encuentran cerca en el espacio de características**.

4.1. El Algoritmo

El algoritmo kNN se basa en encontrar los **k vecinos más cercanos** a una nueva instancia en el espacio de características. La predicción se realiza en función de la clase o valor de estos vecinos.

- **Distancia:** La similitud entre instancias se mide utilizando una métrica de distancia, como la distancia euclidiana.
- **Votación/Promedio:** En clasificación, la clase se decide por votación mayoritaria entre los vecinos. En regresión, se promedia el valor de los vecinos.
- **Parametrización:** Es importante elegir un valor adecuado para k y una función de ponderación para los vecinos.

4.2. Ventajas e Inconvenientes

Ventajas:

- Coste de entrenamiento nulo.
- No requiere suposiciones previas sobre los datos.
- Tolerante al ruido.
- Se puede usar para clasificación y regresión.

Desventajas:

- Alto coste computacional en la predicción.
- Sensible a la maldición de la dimensionalidad.

4.3. Variantes y Mejoras

Existen estrategias para mejorar la eficiencia del algoritmo kNN, como la reducción del conjunto de referencia y la preindexación del conjunto de datos.