Fundamentos de Análisis Estadístico

Taller de Data Science - Intermedio

Contenido

Introducción y repaso

Análisis estadístico

Pruebas de hipótesis y correlación

Regresión

Introducción y repaso

Análisis Estadístico

- Estudio de los datos para comprender y resumir la información.
- Identificación de patrones, relaciones y tendencias.
- Cálculo de medidas de tendencia central y dispersión.
- Pruebas de hipótesis para la toma de decisiones basadas en evidencia estadística.

Aprendizaje Automático

- Rama de la inteligencia artificial.
- Construcción de algoritmos y modelos que aprenden de los datos.
- Reconocimiento de patrones, predicciones y toma de decisiones.
- Ejemplos de técnicas: clasificación, regresión y agrupamiento.

Conceptos básicos

- Variables y tipos de datos: variables numéricas, cadenas de texto, listas, diccionarios, etc.
- Estructuras de control: condicionales (if/else), bucles (for/while), y operadores lógicos.
- Funciones y módulos: definición y llamada de funciones, importación y uso de módulos predefinidos y personalizados.
- Manipulación de datos: lectura y escritura de archivos CSV o Excel, limpieza de datos, filtrado y transformación de datos.

Análisis estadístico

Medidas de tendencia central, posición y dispersión

Las medidas de tendencia central, de posición y de dispersión son herramientas estadísticas que nos permiten resumir y caracterizar un conjunto de datos.

Estas medidas proporcionan información importante sobre la distribución y la variabilidad de los valores en un conjunto de datos.

Medidas de tendencia central, posición y dispersión

Medidas de Tendencia Central:

- Media: Promedio de un conjunto de valores.
- Mediana: Valor central de un conjunto de valores ordenados.
- Moda: Valor(es) más frecuente(s) en un conjunto de valores.

Medidas de Posición:

- Cuartiles
- Deciles
- Percentiles

Medidas de Dispersión:

- Varianza: Mide la dispersión de los datos respecto a la media.
- Desviación Estándar: Medida de dispersión típica de los datos.
- Rango Intercuartílico: Diferencia entre el tercer y el primer cuartil.

Medidas de tendencia central, posición y dispersión

Interpretación e Importancia:

- Estas medidas resumen y caracterizan un conjunto de datos.
- Ayudan a entender la distribución y dispersión de los valores.
- Detectan valores atípicos y proporcionan información valiosa.
- Permiten tomar decisiones basadas en la variabilidad de los datos.
- Son fundamentales en el análisis, modelado y toma de decisiones.

Pruebas de hipótesis y correlación

Pruebas de hipótesis

Las pruebas de hipótesis son herramientas estadísticas utilizadas para tomar decisiones basadas en evidencia. En el análisis estadístico, se plantean dos hipótesis: la hipótesis nula (H0) y la hipótesis alternativa (H1). La hipótesis nula representa la afirmación que se intenta refutar o rechazar, mientras que la hipótesis alternativa representa la afirmación opuesta.

Prueba de hipótesis para la estatura media

- Hipótesis nula (H0): La estatura media de la población es igual a un valor específico (por ejemplo, 170 cm).
- Hipótesis alternativa (H1): La estatura media de la población es diferente de un valor específico (no igual a 170 cm).
- Nivel de significancia: Se selecciona un nivel de significancia previamente, como $\alpha = 0.05$.

Para realizar esta prueba, se puede utilizar la prueba t de Student. Se calcularía la media de estatura del grupo completo y se compararía con el valor específico (por ejemplo, 170 cm) utilizando la prueba t. Si el valor p obtenido es menor que α, se rechazaría la hipótesis nula y se concluiría que la estatura media es significativamente diferente del valor específico.

Prueba de hipótesis para la para la diferencia de medias entre hombres y mujeres

- Hipótesis nula (H0): La diferencia de medias entre la estatura de hombres y mujeres es igual a cero (es decir, no hay diferencia).
- Hipótesis alternativa (H1): La diferencia de medias entre la estatura de hombres y mujeres es diferente de cero (hay una diferencia significativa).
- Nivel de significancia: Se selecciona un nivel de significancia previamente, como $\alpha = 0.05$.

Para realizar esta prueba, se puede utilizar la prueba t de Student independiente. Se calcularían las medias de estatura para hombres y mujeres por separado y se compararían utilizando la prueba t independiente. Si el valor p obtenido es menor que α, se rechazaría la hipótesis nula y se concluiría que hay una diferencia significativa en la estatura media entre hombres y mujeres.

Regresión

Regresión simple y múltiple

La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente y una o más variables independientes. En el caso de la regresión lineal simple, se busca establecer una relación lineal entre una variable dependiente y una única variable independiente. Por otro lado, en la regresión lineal múltiple, se busca establecer una relación lineal entre una variable dependiente y múltiples variables independientes.

Modelo matemático

En ambos casos, el objetivo es encontrar la mejor línea recta que se ajuste a los datos observados, de manera que se minimice la diferencia entre los valores predichos por el modelo y los valores reales. La línea recta se define por una ecuación de la forma:

- Regresión lineal **simple**:

  $y = b0 + b1*x$

- Regresión lineal **múltiple**:

  $y = b0 + b1x1 + b2x2 + ... + bn*xn$

Donde:

- **y** es la variable dependiente que queremos predecir.
- **x, x1, x2, ..., xn** son las variables independientes que se utilizan para hacer la predicción.
- **b0, b1, b2, ..., bn** son los coeficientes de la ecuación de regresión, que representan la pendiente y el término de intercepción de la línea.

Modelo predictivo del peso

En el contexto de los datos de ejemplo (estatura, edad, peso y sexo), se puede aplicar una regresión lineal simple para predecir el peso en función de la estatura, o una regresión lineal múltiple para predecir el peso en función de la estatura y la edad. El análisis de los coeficientes de la regresión nos permitiría entender cómo la estatura y la edad influyen en el peso, y realizar predicciones para nuevas observaciones en base a los valores de estas variables.