Universidad ICESI

45 años

Aprendizaje Automático I

Milton Sarria-Paja, Ph.D.

2025-1

Universidad icesi

2025-1

Conceptos fundamentales

Milton Sarria-Paja, Ph.D.

Dependencia de Variables y Correlación en Machine Learning

Definición de dependencia de variables y correlación.

Importancia de estos conceptos en el análisis de datos y en la construcción de modelos de machine learning.

Dependencia de Variables

- Explicación de la dependencia entre variables:
  - Una variable depende de otra si su valor está influenciado por la otra.
  - Ejemplo: La altura y el peso de una persona.

- Diferencia entre dependencia funcional y estadística.

Dependencia de Variables

Dependencia Funcional: Se da cuando una variable determina completamente a la otra.

•**Ejemplo**: La temperatura en Fahrenheit (F) vs la temperatura en Celsius (C):

$$F=1.8C+32$$

Aquí, dado un valor de (C), el valor de (F) es completamente determinado.

Dependencia Estadística: Se da cuando una variable influye en otra, pero con cierta variabilidad.

•**Ejemplo**: A mayor nivel educativo, mayor suele ser el salario, pero no siempre ocurre exactamente así debido a otros factores como la experiencia laboral y la industria en la que trabaja una persona.

Covarianza vs. Correlación

- Definición de covarianza:
  - Indica si dos variables varían en la misma dirección (positiva) o en direcciones opuestas (negativa).

- Diferencias clave:
  - La covarianza proporciona la dirección de la relación, pero no la fuerza.
  - La correlación estandariza la covarianza, proporcionando tanto dirección como fuerza de la relación.

Covarianza

La **covarianza** mide la tendencia de dos variables a aumentar o disminuir juntas.

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

Donde:

- X_i, Y_i son los valores de las variables.
- $\bar{X}, \bar{Y}$ son las medias de X y Y.
- n es el número de observaciones.

La covarianza depende de la escala de las variables, lo que dificulta su interpretación.

Correlación

Medida estadística que indica la fuerza y dirección de una relación lineal entre dos variables.

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

La correlación **normaliza** la covarianza dividiéndola por las desviaciones estándar de las variables. Esto permite que el coeficiente de correlación esté **siempre entre -1 y 1**, facilitando su **interpretación**

Importancia en Machine Learning

- Selección de características:
  - Identificación de variables relevantes para mejorar la precisión del modelo.

- Detección de multicolinealidad:
  - Problemas que surgen cuando las variables predictoras están altamente correlacionadas entre sí.

  - Si en un dataset tenemos las variables **peso en kilogramos** y **peso en libras**, una de ellas puede eliminarse ya que aportan la misma información.

Métodos para Medir Dependencia y Correlación

Correlación de Pearson

Útil para medir relaciones **lineales** entre variables continuas.

Correlación de Spearman

Se basa en **rangos** en lugar de valores absolutos.

Se usa cuando la relación entre variables no es lineal.

Razón de Correlación

Se usa para evaluar la relación entre variables **categóricas** y **numéricas**.

Prueba de Chi-Cuadrado

Se usa para evaluar la relación entre variables **categóricas**.

Limitaciones de la Correlación

Correlación no implica causalidad
Ejemplo: Aumento de ventas de helados y número de ahogamientos en verano. No es que los helados causen los ahogamientos, sino que ambos están relacionados con el calor.

Sensibilidad a valores atípicos: Un valor extremo puede distorsionar la correlación.

Gracias!!