Apuntes - IA

Bioing. Ignacio Bosch

31 de agosto de 2023

I. Análisis de normalidad

Realizar tests de bondad de ajuste, como son los tests de normalidad, tiene por objetivos:

- Validar supuestos estadísticos: algunos métodos estadísticos, como las pruebas de hipótesis paramétricas, asumen que los datos provienen de una distribución específica, como puede ser la distribución normal. Al realizar tests de normalidad u otros tests de bondad de ajuste, es posible evaluar si estos supuestos se cumplen antes de aplicar métodos que los requieren.
- Seleccionar el modelo adecuado: el análisis de datos implica ajustar un modelo estadístico a los datos. Los tests de bondad de ajuste ayudan a evaluar qué modelo se ajusta mejor a los datos observados. Por ejemplo, si los datos se ajustan mejor a una distribución exponencial que a una normal, es mejor elegir un modelo basado en la distribución exponencial para realizar predicciones o inferencias.
- Detectar desviaciones significativas: al realizar un test de normalidad, es posible identificar si los datos se desvían significativamente de la normalidad. Esto sirve para detectar patrones inusuales o anomalías en los datos. A veces, las desviaciones de la normalidad pueden indicar la presencia de datos atípicos o errores de medición de los datos.
- Tomar decisiones: Comprender la distribución de los datos brinda información valiosa para la toma de decisiones. Por ejemplo, si los datos no siguen una distribución normal, es posible que se utilize métodos no paramétricos en lugar de métodos paramétricos para el análisis.

i. Shapiro-Wilk

El test de Shapiro-Wilk es uno de los tests más utilizados para verificar la normalidad de los datos. Es adecuado para muestras de tamaño pequeño a moderado (hasta aproximadamente 50 observaciones). El test proporciona un estadístico de prueba "W" y un valor "p" asociado que indica la probabilidad de obtener los datos observados si la muestra proviene de una distribución normal.

ii. Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov evalúa si una muestra de datos se ajusta a una distribución específica, no solo a la normal. El test compara la función de distribución acumulada "FDC" empírica con la función de distribución acumulada teórica. Es más adecuado para muestras de tamaño moderado a grande (mayor a 50 observaciones). Proporciona un estadístico de prueba "D" y un valor "p" asociado que indica la probabilidad de obtener los datos observados si la muestra proviene de la distribución teórica.

iii. Lilliefors

El test de Lilliefors es una variante del test de Kolmogorov-Smirnov que se utiliza para probar la normalidad de los datos. A diferencia del anterior, el test de Lilliefors utiliza una tabla de valores críticos ajustados específicamente para la distribución normal. Es adecuado para muestras de tamaño pequeño a moderado. Proporciona un estadístico de prueba "L" y un valor "p" asociado.

IV. Anderson-Darling

El test de Anderson-Darling es otro test de bondad de ajuste utilizado para evaluar si una muestra de datos sigue una distribución específica. A diferencia de los tests de normalidad como Shapiro-Wilk o Kolmogorov-Smirnov, el test de Anderson-Darling es más general y puede utilizarse para probar una variedad de distribuciones, no solo la normalidad. El test asigna una puntuación al ajuste de los datos a una distribución teórica y proporciona un valor crítico y un valor "p" asociado para determinar si se puede rechazar o no la hipótesis nula de que los datos siguen la distribución especificada.

V. Chi-cuadrado

Test de Chi-cuadrado se utiliza para evaluar si una muestra de datos se ajusta a una distribución teórica específica. Este test compara las frecuencias observadas y las esperadas en diferentes categorías y calcula un estadístico Chi-cuadrado que se compara con los valores críticos de la distribución Chi-cuadrado. Se utiliza para comparar la distribución que siguen los datos o bien se aplica para determinar si existe una asociación significativa entre dos variables categóricas (no numéricas).

II. Coeficientes de correlación

Los coeficientes de correlación evalúan la relación entre variables, mientras que los tests de bondad de ajuste (sección anterior) evalúan cómo los datos se ajustan a una distribución teórica (normal, uniforme, etc).

I. Pearson

Evalúa la relación lineal entre dos variables continuas. Mide la fuerza y la dirección de la relación lineal y puede variar entre [-1,1]. Resulta más adecuado cuando se busca medir una relación lineal y se presume que las variables siguen una distribución normal.

II. Spearman

Evalúa la relación de clasificación o de orden entre dos variables. No asume una relación lineal, sino cualquier tipo de relación monótona, ya sea creciente o decreciente. También varía entre [-1,1]. Resulta más robusto frente a relaciones no lineales y no se presume distribuciones específicas.

III. Kendall

Mide la relación de orden entre dos variables. Es útil cuando los datos tienen empates (empates en los valores de las variables) y se enfoca en la concordancia o discordancia entre los órdenes de los pares de observaciones. También mide la relación monotónica entre dos variables y varía entre [-1,1].

III. Selección de características

El objetivo de la selección de características es elegir un subconjunto de variables de la entrada que pueda representar eficientemente los datos de entrada al tiempo que limita los impactos del ruido o las variables irrelevantes y al mismo tiempo proporciona una buena predicción resultados. Si continuamos utilizando nuestro modelo con estas variables irrelevantes, nuestro modelo tendrá una pobre generalización.

Al momento de elegir variables, siempre se va a cumplir el "Principio de Parsimonia": un modelo parsimonioso (menor número de variables) proporciona mejores predicciones que un modelo completo. Esto está relacionado con el problema del sobreajuste (overfitting). Los modelos con muchas variables pueden memorizar los datos, perdiendo poder de generalización frente a datos nuevos. A la vez que siempre se van a necesitar muchas variables para realizar una predicción.

Por lo que ser cuidadosos y criteriosos al momento de agregar variables es importante, para esto existen diversos métodos.

Existen tres formas de selección de características: Filtro, Envoltura y Embebido.

I. Filtro

Este método usa técnicas para rankear variables, como criterio primario para elegir variables en función de algún umbral para eliminar otras variables por debajo de ese umbral.

Si bien una característica puede considerarse irrelevante si es condicionalmente independiente de las etiquetas de clase. Básicamente afirma que si una característica es útil, puede ser independiente de los datos de entrada pero no de las etiquetas de clase, es decir. Se puede eliminar la característica que no tiene impacto en las etiquetas de clase. En otras palabras, la correlación entre características es crítica para determinar características únicas.

Se presentan 4 métodos de Filtro para selección de características: *correlación*, *umbral de varianza*, *chi cuadrado*, y *información mutua*.

II. Envoltura

Los enfoques envolventes (wrapper) evalúan el subconjunto de variables utilizando el algoritmo como una caja negra y el rendimiento del algoritmo como la función objetivo. Para seleccionar un subconjunto de características que maximice la función objetivo, se puede utilizar una variedad de estrategias de búsqueda. Sin embargo, la búsqueda crecería exponencialmente a medida que aumentara la cantidad de funciones. Para conjuntos de datos más grandes, los métodos de búsqueda exhaustivos pueden volverse computacionalmente intensivos. Por suerte existen librerías como scikit-learn que poseen métodos de búsqueda de features con **feature_selection**, el cual es un método que utiliza un algoritmo de clasificación o regresión y lo que hace es entrenar con diferentes combinaciones de variables, mientras monitorea los resultados de cada prueba con alguna métrica.

III. Embebido

Se refiere al método de selección de características que tienen algunos modelos como Random Forest dentro de su algoritmo. Random Forest es un algoritmo de aprendizaje automático que se utiliza tanto para la clasificación como para la regresión. La selección de características en Random Forest se realiza mediante un proceso de ïmportancia de características"que se calcula durante la construcción del bosque. Para esto se utilizan dos criterios (se debe elegir uno de estos dos): **Gini** y **Entropía**, se utilizan para medir la homogeneidad de un nodo en un árbol de decisión y ayudan a determinar cómo se debe realizar la división en ese nodo. Ambos criterios se utilizan para evaluar qué tan mezcladas están las clases en un nodo, y la división se elige para minimizar la mezcla.