Naive Bayes: Teoría y Aplicación con Ejemplo Práctico

NAIVE BAYES: TEORÍA Y APLICACIÓN CON EJEMPLO PRÁCTICO

🧐 ¿Qué es Naive Bayes?

Naive Bayes es un método de clasificación probabilístico basado en el Teorema de Bayes, con la suposición ingenua (naive) de que los atributos son independientes entre sí dado el resultado o clase. A pesar de esta suposición simplificadora, funciona sorprendentemente bien en muchos contextos reales.

🔹 Teorema de Bayes (base teórica)

El teorema de Bayes se expresa como: $P(A|B) = P(B|A) * P(A) / P(B)$

En clasificación:

- A es la clase (por ejemplo, "Se jugó = Sí" o "No")

- B representa los atributos observados (como cielo, temperatura, etc.)

Pero en Naive Bayes, no calculamos esa probabilidad de manera teórica, sino a partir de datos empíricos.

🤢 Cómo funciona Naive Bayes

Dado un conjunto de atributos $X_1, X_2, ..., X_n$, Naive Bayes predice la clase C que maximiza:

$P(C \mid X_1, ..., X_n) \propto P(C) * \prod P(X_i \mid C)$

Nota: usamos el símbolo $\propto$ (proporcional a), porque el denominador $P(X_1, ..., X_n)$ es el mismo para todas las clases y no afecta la comparación.

🔎 Estimación de probabilidades condicionales

En lugar de usar la fórmula teórica de probabilidad condicional:

$P(A \mid B) = P(A \cap B) / P(B)$

En la práctica usamos frecuencias observadas:

P(X = x | C = c) = Frecuencia de x en clase c / Total de casos en clase c

⚠️ Problema: Probabilidad cero

Si algún valor de atributo nunca aparece en una clase, su probabilidad es 0. Eso hace que todo el producto sea 0, anulando la predicción.

🔧 Solución: Suavizado de Laplace

Usamos el suavizado de Laplace para evitar ceros:

P(X = x | C = c) = (frecuencia(x, c) + 1) / (total en clase c + k)

Donde:

- 1: es el ajuste por suavizado
- k: cantidad de valores posibles para el atributo X

¿Qué significa k?

k representa la cantidad de valores distintos que puede tomar un atributo. Por ejemplo, si el atributo "Cielo" tiene tres posibles valores: Soleado, Nublado y Lluvia, entonces k = 3.

Este valor se suma al denominador para compensar el hecho de que estamos agregando 1 al numerador de cada categoría posible (aunque no haya aparecido en los datos), manteniendo así una distribución de probabilidad válida.

Esto garantiza que todas las probabilidades sean > 0.

📊 Ejemplo práctico: Predecir si se jugará

Vamos a usar esta tabla de datos (14 casos):

Día	Cielo	Temp	Humedad	Viento	Se Jugó
1	Soleado	Alta	Alta	Débil	No
2	Soleado	Alta	Alta	Fuerte	No
3	Nublado	Alta	Alta	Débil	Sí
4	Lluvia	Media	Alta	Débil	Sí
5	Lluvia	Baja	Normal	Débil	Sí
6	Lluvia	Baja	Normal	Fuerte	No
7	Nublado	Baja	Normal	Fuerte	Sí
8	Soleado	Media	Alta	Débil	No
9	Soleado	Baja	Normal	Débil	Sí
10	Lluvia	Media	Normal	Débil	Sí
11	Soleado	Media	Normal	Fuerte	Sí
12	Nublado	Media	Alta	Fuerte	Sí
13	Nublado	Alta	Normal	Débil	Sí
14	Lluvia	Media	Alta	Fuerte	No

🔢 Paso a paso para predecir: si se juega o no, con las siguientes condiciones.

1. Cielo = Soleado
2. Temperatura = Alta
3. Humedad = Normal
4. Viento = Fuerte

Paso 1: Probabilidades a priori

1. P(Sí) = 9/14
2. P(No) = 5/14

Paso 2: Probabilidades condicionales (con suavizado)

- Para clase "Sí":

- Cielo = Soleado: (2+1)/(9+3) = 3/12 = 0.25

- Temperatura = Alta: (3+1)/(9+3) = 4/12 = 0.33

- Humedad = Normal: (6+1)/(9+2) = 7/11 ≈ 0.636

- Viento = Fuerte: (3+1)/(9+2) = 4/11 ≈ 0.364

P(Sí | datos) $\propto$ (9/14) * 0.25 * 0.33 * 0.636 * 0.364 ≈ 0.0123

- Para clase "No":

- Cielo = Soleado: (3+1)/(5+3) = 4/8 = 0.5

- Temperatura = Alta: (2+1)/(5+3) = 3/8 = 0.375

- Humedad = Normal: (1+1)/(5+2) = 2/7 ≈ 0.286

- Viento = Fuerte: (2+1)/(5+2) = 3/7 ≈ 0.429

P(No | datos) $\propto$ (5/14) * 0.5 * 0.375 * 0.286 * 0.429 ≈ 0.0102

✅ Resultado:

Como 0.0123 > 0.0102, Naive Bayes predice que se jugará el partido.

🎓 Conclusión

1. En Naive Bayes no usamos la fórmula teórica de condicional.
2. Usamos frecuencias empíricas, y aplicamos suavizado de Laplace para evitar probabilidades cero.
3. La clasificación se hace comparando probabilidades proporcionales.
4. Es un método simple, pero poderoso para clasificación en muchos contextos reales.