Informática y comunicaciones
ies azarquiel

CURSO ESPECIALIZACIÓN INTELIGENCIA ARTIFICIAL Y BIG DATA

Tema 6: Naïve-Bayes

Sebastián Rubio

Noviembre 2025

1 Teorema de Bayes

El Teorema de Bayes es uno de los resultados más conocidos y útiles en el área de la probabilidad y estadística, y en particular en el estudio de la probabilidad condicional. Básicamente, el Teorema de Bayes nos dice cómo calcular la probabilidad de un suceso teniendo información a priori sobre dicho suceso.

Este teorema es una herramienta altamente usada por su simpleza y su rápida aplicación en distintas áreas del conocimiento, por ejemplo en medicina, biología, tecnología, negocios, o en cualquier área en la que se necesite tener una certeza sobre algún suceso dada información de antemano. Además, es común utilizar dicha herramienta consecutivamente para obtener una mayor certeza si el problema así lo requiere.

El teorema de Bayes es una fórmula matemática fundamental en la teoría de la probabilidad que describe cómo actualizar la probabilidad de una hipótesis cuando se tiene un nueva evidencia

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$
$$P(A|B) = \frac{P(B|A)P(A)}{P(B}$$

Veamos algunos ejemplos de como aplicar este teorema

Ejemplo 1

Este es un uso clásico de Bayes, donde se ajusta la probabilidad de tener una enfermedad después de obtener un resultado positivo en un prueba analítica.

Escenario:

- $P(E)$:Sólo el 1% de la población tiene una enfermedad rara (E). Esta es la probabilidad a priori
- Verdaderos positivos $P(p|E)$: La prueba es muy buena; detecta la enfermedad(p,positivo) en el 95% de los casos si la persona realmente la tiene.
- Falso positivos $P(p|NE)$: La prueba da positivo (P) en el 5% de los casos si la persona no tiene la enfermedad (NE)

Si una persona elegida al azar da positivo (p) en la prueba, ¿cuál es la probabilidad de que realmente tenga la enfermedad $P(E|p)$

Aplicamo Bayes

$$P(E|P) = \frac{P(P|E)\cdot P(E)}{P(P|E)\cdot P(E)+P(P|\text{NE})\cdot P(\text{NE})}$$

Sustituyendo valores:

- $P(E) = 0.01$
- $P(NE) = 1 - 0.01 = 0.99$
- $P(p|E) = 0.95$
- $P(p|NE) = 0.05$

$$P(E|p) = \frac{0.95.0.01}{0.95.0.01+0.05.0.99}$$
$$P(E|p) \approx 0.161$$

A pesar de dar positivo en una prueba con "95% de precisión", la probabilidad de que la persona realmente tenga la enfermedad es solo del 16.1%. La nueva evidencia (el resultado positivo) actualizó la probabilidad inicial (1%) a una probabilidad posterior (16.1%), pero la baja prevalencia de la enfermedad sigue siendo el factor dominante.

Ejemplo 2

Se utiliza para determinar la fuente más probable de un defecto, basándose en la tasa de producción y la tas de fallo de cada máquina.

Escenario:
Una fábrica tiene tres máquina (M_1, M_2, M_3) que produce el total de piezas

- Producción $P(M_i)$:M_1 produce el 50%, M_2 el 30
- Tasa de defectos $P(D|M_i)$: Cada máquina tiene una tasa de producción defectuosa $M_1 : 1\%$, $M_2 : 2\%$, $M_3 : 3\%$

Si se selecciona una pieza defectuosa (D) al azar, ¿cuál es la probabilidad de que haya sido producido por la máquie M_3 $P(M_3|D)$

$$P(D) = P(D|M_1)P(M_1) + P(D|M_2)P(M_2) + P(D|M_3)P(M_3)$$

$$P(D) = (0.01 \cdot 0.50) + (0.02 \cdot 0.30) + (0.03 \cdot 0.20)$$
$$P(D) = 0.005 + 0.006 + 0.006 = \mathbf{0.017}$$

El 1,7% de las piezas son defectuosas Ahora aplicamos el Teorema de Bayes par M_3

$$P(M_3|D) = \frac{P(D|M_3)P(M_3}{P(D)}$$
$$P(M_3|D) = \frac{0.03.0.20}{0.017}$$
$$P(M_3|D) \approx 0.353$$

Si encontramos una pieza defectuosa, la probabilidad de que provenga de la Máquina 3 es del 35.3%. A pesar de que M_3 produce la menor cantidad total de piezas (20%), su alta tasa de defectos (3%) la convierte en la fuente más probable de un fallo.

2 Clasificación por Naïve-Bayes

Es un algoritmo supervisado, que se fundamenta en la probabilidad. Es rápido, y funciona bien con los textos (NLP).En este algoritmo se asume que las características son independientes, conocido un valor,es decir no hay ninguna correlación. Pero esta propiedad puede no cumplirse en algunos problemas de clasificación, por lo que es conveniente comprobar si es propiedad del conjunto de datos se cumple. No requiere grandes cantidades de ejemplos de entrenamiento.

El uso de recursos de computación para obtener un modelo es significativamente más pequeño que los requeridos por otros algoritmos de aprendizaje automático.

Es un algoritmo de referencia que siempre debe probarse a la hora de resolver cualquier problema de clasificación. Puede trabajar con atributos continuos,discretos y categóricos.

Los ejemplos de entrenamiento viene dados por valores discretos o continuos. Se basa en estimar la probabilidad de pertenencia a una clase con las probabilidades condicionadas, mediante el teorema de Bayes. La principal característica de este tipo de clasificadores es que el modelo obtenido es capaz de predecir con máxima probabilidad los ejemplos de entrenamiento. Cada clase de entrenamiento debe pertenecer a una clase C_k,etiqueta. La forma de proceder con ese algoritmo es:

- Calcular la probabilidad a priori de que cada clase $P(C_k)$.
- Para cada instancia de entrenamiento,que pertenece a una clase, hay que dis-

tinguir si el atributo es discreto ó continuo.

- Si el atributo es discreto, para estimar las probabilidades, hay que calcular la frecuencia de aparición de un valor del atributo en el conjunto de entrenamiento.
- Si el atributo es continuo,para calcular las probabilidades se realiza usando la media μ y la desviación típica σ del valor del atributo en el conjunto de ejemplos de entrenamiento.
- Para hacer el cálculo de probabilidades vamos a suponer que estamos una distribución normal (gaussiana)

La media

$$\mu = \frac{\sum x_i}{n} \tag{1}$$

donde x_i es el valor del atributo del ejemplo de entrenamiento i-ésimo y n es es el total de ejemplos que tienen el ese mismo valor en el correspondiente atributo. Para calcular la desviación típica σ, vamos a seguir la siguiente formula

$$\sigma = f(\mu) = \sqrt{\frac{\sum (x_i - \mu)^2}{n-1}} \tag{2}$$

Una vez que ya hemos realizado todos estos cálculos estamos en disposición de hallar la probabilidad condicionada, es decir aplicar Bayes. Ante un ejemplo nuevo, el modelo lo clasificará en una de las clases.

2.1 Cálculos a realizar

- Calcular la probabilidad a priori, $P(C_k), k = 1, 2$
- Calcular la probabilidad condicionada de los atributos x_j
  - Discretos: Calcular la probabilidad $P(x_j|C_k)$
  - Contínuos:
    - Calcular la media,μ
    - Calcular la desviación típica, σ
    - Calcular la distribución normal, $f(\mu, \sigma)$
- Aplicar el teorema de Bayes, $P(C_k|X_i)$

3 Ejemplo

EJEMPLO:Naïve-Bayes. Los datos de entrenamiento hace referencia a clientes de un banco. El objetivo que se persigue es obtener un modelo clasificador, para decidir si a un nuevo cliente se le concede o no un préstamo. (Este ejemplo es recogido de los apuntes de la UNED)

Conjunto de entrenamiento						
Ejemplo	Atributos Continuos		Atributos Discretos			Clase
D	Edad (E)	Sueldo (S)	Casa Propia(CP)	Segunda Vivienda (SV)	E. Civil (EC)	Crédito
d_1	24	50	No	No	Soltero	Denegar
d_2	34	90	No	No	Casado	Denegar
d_3	45	125	Sí	No	Casado	Denegar
d_4	60	60	No	No	Casado	Denegar
d_5	31	200	No	No	Casado	Denegar
d_6	29	140	No	No	Soltero	Denegar
d_7	50	250	Si	Si	Soltero	Conceder
d_8	41	320	Si	No	Casado	Conceder
d_9	30	670	No	No	Casado	Conceder
d_{10}	35	400	Si	No	Soltero	Conceder
d_{11}	47	360	Si	No	Casado	Conceder
d_{12}	29	140	Si	No	Soltero	Conceder

Table 1: Datos de entrenamiento con atributos continuos y discretos

En primer lugar vamos a calcular las probabilidades a priori. Al analizar la tabla vemos que tenemos dos clases K=2:{Denegar, Conceder}, por lo tanto la probabilidad de cada clase es: Ahora vamos a calcular las probabilidades de los atributos discretos.

Clase	$P(C_k), k = 1, 2$
$C_1(Denegar)$	$6/12 = 0.50$
$C_2(Conceder)$	$6/12 = 0.5$

Table 2: Probabilidades de clase

Este cálculo se realiza con la frecuencia de aparición del valor del atributo en el conjunto de ejemplos de entrenamiento . Hay que hacerlo para cada atributo discreto

Casa en Propiedad (CP)	$P(CP = NO\|C_k), k = 1, 2$	$P(CP = SI\|C_k), k = 1, 2$
$C_1(Denegar)$	$5/6 = 0.83$	$1/6 = 0.17$
$C_2(Conceder)$	$1/6 = 0.17$	$5/6 = 0.83$

Table 3: Notas matemáticas y lengua

Segunda vivienda (SV)	$P(SV = NO\|C_k), k = 1, 2$	$P(SV = SI\|C_k), k = 1, 2$
$C_1(Denegar)$	$6/6 = 0.95$	$0/6 = 0.05$
$C_2(Conceder)$	$1/6 = 0.17$	$5/6 = 0.83$

Table 4: Notas matemáticas y lengua

Estado Civil (EC)	$P(EC = Soltero\|C_k), k = 1, 2$	$P(EC = Casado\|C_k), k = 1, 2$
$C_1(Denegar)$	$2/6 = 0.33$	$4/6 = 0.67$
$C_2(Conceder)$	$3/6 = 0.0.5$	$5/6 = 0.50$

Table 5: Notas matemáticas y lengua

El cálculo de las probabilidades implicará la asunción de un tipo de distribución asociada a el atributo (normalmente, se considerará una distribución normal $f(\mu, \sigma)$

El cálculo de la probabilidades condicionadas al atributo continuo de la edad (E) y el atributo continuo denegar (C_1),ser realiza como vemos a continuación

Calculamos la media de edad $\mu = (24 + 34 + 45 + 60 + 32 + 29)/6 = 223/6 = 37.17$
A continuación calculamos la desviación típica
$\sigma = \sqrt{\frac{(24-37.17)^2+(34-37.17)^2+(45-37.17)^2+(60-37.17)^2+(31-37.17)^2+(29-37.17)^2}{6-1}} =$
$= \sqrt{174.13} = 13.20$

El cálculo de la probabilidades condicionadas al atributo continuo de la edad (E) y el atributo continuo conceder (C_2),ser realiza como vemos a continuación

Calculamos la media de edad $\mu = (50 + 41 + 30 + 35 + 47 + 42)/6 = 245/6 = 40.83$

Calculamos también la desviación típica
$\sigma = \sqrt{\frac{(50-40.83)^2+(41-40.83)^2+(30-40.83)^2+(35-40.83)^2+(47-40.83)^2+(42-40.83)^2}{6-1}} = \sqrt{54.97} = 7.41$

Realizados los cálculos anteriores, podemos obtener la distribución Para realizar el

Edad (E)	μ	σ	$1/\sqrt{(2\pi\sigma^2)}$	$\frac{1}{2\sigma^2}$	$g(x,\mu,\sigma)$
$C_1(Denegar)$	37.17	13.20	0.0302	0.0029	$\mu.e^{[-0.0029.(x-37.17)^2]}$
$C_2(Conceder)$	40.83	7.41	o.0583	0.0091	$\mu.e^{[-0.0091.(x-40.83)^2]}$

Table 6: Distribuciones

cálculo de las probabilidades condicionadas del atributo sueldo. Procedemos de la misma forma que en el caso del atributo edad, pero ahora calculando la media del sueldo. Los resultados son los siguientes:

Para la variable Denegar

- $\mu = 110.83$
- $\sigma = 56.07$

Para la variable Conceder

- $\mu = 385$
- $\sigma = 148.42$

La tabla de distribuciones Una vez realizado todos los cálculos hemos (entrenado el

Sueldo (S)	μ	σ	$1/\sqrt{(2\pi\sigma^2)}$	$\frac{1}{2\sigma^2}$	$g(x,\mu,\sigma)$
$C_1(Denegar)$	110.83	56.07	0.0071	0.0002	$\mu.e^{[-0.0002.(x-110.83)^2]}$
$C_2(Conceder)$	385.00	148.42	0.0027	0.0001	$\mu.e^{[-0.0001.(x-385.00)^2]}$

Table 7: Distribuciones

modelo) vamos a predecir el valor correspondiente a un nuevo ejemplo de entrada que no se ha utilizado en el entrenamiento.

El **nuevo cliente** $X = [Edad(E) = 22, Sueldo(S) = 200, CasaPropia(CP) = SI, SegundaVivienda(SV) = SI, EstdoCivil(EC) = soltero]$
Tenemos que calcular:

- $A = P(X|C_{Denegar})$
- $B = P(X|C_{Conceder})$

- $Max(A, B) = Clase$

Por el teorema de Bayes tenemos que $P(X|C_{Denegar})$:

$P(X|C_{Denegar}) = P(E_{22}|C_{Denegar}).P(S_{200}|C_{Denegar}).P(CP_{Si}|C_{Denegar}).$
$P(SV_{Si}|C_{Denegar}).P(EC_{Soltero}|C_{Denegar}).P(C_{Denegar})=$
$=(0.0302.e^{[-0.0029.(22-37.17)^2]}).(0.0071.e^{[-0.0002.(200-110.83)^2]}).(0.17).(0.05).(0.33).(0.5) = (29.5).10^{-9}$

Para el calculo de $P(X|C_{Conceder})$

$P(X|C_{Conceder}) = P(E_{22}|C_{Conceder}).P(S_{200}|C_{Conceder}).P(CP_{Si}|C_{Conceder}).$
$P(SV_{Si}|C_{Conceder}).P(EC_{Soltero}|C_{Conceder}).P(C_{Conceder})=$
$=(0.0538.e^{[-0.0091.(22-40.83)^2]}).(0.0027.e^{[-0.0001.(200-385)^2]}).(0.83).(0.83).(0.5).(0.5) = (3.6).10^{-9}$

Para encontrar que clase pertenece

$Max(P(X|C_{Denegar}), P(X|C_{Conceder})) = Max((29.5).10^{-9}, (3.6).10^{-9}) = (29.5).10^{-9}$
Esto se correspode con $P(X|C_{Denegar})$, por lo que al nuevo cliente X se le denegará el crédito.

En el siguiente ejemplo vamos entrenar un modelo en el que una característica es numérica y la otra categórica

Table 8: Placeholder Caption

ID	Clase	X (numérica)	Y (categórica)
1	A	5.0	red
2	A	6.0	red
3	A	7.0	blue
4	B	3.0	blue
5	B	4.0	blue
6	B	2.0	green
7	C	8.0	green
8	C	9.0	green
9	C	7.0	red

Cada clase tiene tiene tres ejemplo de 9 totales, las probabilidades a priori son:

$$P(A) = P(B) = P(C) = 3/9 \approx 0.333333$$

Para los parámetros de la distribución usaremos la varianza

$$\sigma^2 = \frac{1}{n}\sum(x_i - \mu)^2$$

Los datos de $X_A = \{5, 6, 7\}$

- $\mu_A = (5 + 6 + 7)/3 = (18/3) = 6.0$
- Desviaciones cuadradas: $(5-6)^2 = 1, (6-6)^2 = 0, (7-6)^2$, la suma $= 2$
- $\sigma_A^2 = 2/3 \approx 0.6666667$
- $\sigma_A \approx 0.81649655809$

Para $X_B = \{3, 4, 2\}$

- $\mu = (3 + 4 + 2)/3 = 3$
- Las desviaciones cuadradas suman 2; $\sigma_B^2 = 2/3, \sigma_B \approx 0.81649655809$

En la clase $X_C = \{8, 9, 7\}$

- $\mu_C = (8 + 9 + 7)/3 = 24/3 = 8.0$
- Desviaciones cuadradas suman 2; $\sigma_C^2 = 2/3, \sigma_C \approx 0.81649655809$

El siguiente paso es calcular la probabilidades condicionadas para la característica categórica Y (Usamos Laplace suavizado). Para cada clase el denominador es $n_{clase} + k = 3+3$. Es la suma del número de clases y el número de categorías, tomaremos $\alpha = 1$

$$P(Y = y \mid Clase = y) = \frac{conteo(y,c)+\alpha}{N_c+\alpha K}$$

Los conteos son los siguientes:

- $conteo(red, A) = 2, conteo(red, B) = 0, conteo(red, C) = 1$
- $conteo(blue, A) = 1, conteo(blue, B) = 2, conteo(blue, C) = 0$
- $conteo(green, A) = 0, conteo(green, B) = 1, conteo(green, C) = 2$

Para la clase A (Tiene: red=2,blue=1,gree=0)

- $P(red \mid A) = (2 + 1)/6 = 3/6 = 0.5$
- $P(blue \mid A) = (1 + 1)/6 = 2/6 = 0.3333333.$
- $P(green \mid A) = (0 + 1)/6 = 1/6 = 0.1666667.$

Para la clase B (Tiene: red=0,blue=2,gree=1)

- $P(red \mid B) = 1/6 = 0.16666667$
- $P(blue \mid B) = 3/6 = 2 = 0.5$
- $P(green \mid B) = (2/6 = 0.3333333.$

Para la clase C (Tiene: red=1,blue=0,gree=2)

- $P(red \mid C) = 2/6 = 0.3333333$
- $P(blue \mid C) = 1/6 = 2 = 0.166667$
- $P(green \mid C) = (3/6 = 1/6 = 0.5.$

Calculamos los scores no normalizados según la siguiente ecuación

$$score(clase) = P(clase).P(6,5 \mid clase).P(red \mid clase)$$

Aplicando la fórmula obetenmos:

- $score(A) = P(A).P(6,5 \mid A).P(red \mid A) \approx 0.0675109516$
- $score(B) = P(B).P(6,5 \mid B).P(red \mid B) \approx 0.000002777171$
- $score(C) = P(C).P(6,5 \mid C).P(red \mid C) \approx 0.010042486294$

Ahora normalizamos $S = score(A) + score(B) + score(C) \approx 0.0775562150628$ Por último calculamos las probabilidades condicionadas de la observación, es decir la probabilidad de que se de cada una de las clases:

- $P(A \mid obs) = \frac{score(A)}{S} \approx 0.8704776470(87,5)\%$
- $P(B \mid obs) = \frac{score(B)}{S} \approx 0.0000358085(0.0036\%)$
- $P(C \mid obs) = \frac{score(C)}{S} \approx 0.1294865445(12.95\%)$

La observación tiene mayor probabilidad en la clase A, por tanto decimos que cons esos valores (X=6,5,Y=red) es de la clase A

4 Naive-Bayes scikit-learn

En la librería scikit-learn no tenemos ningún método para entrenar modelo Naive-Bayes mixtos El enfoque más común en la práctica para la clasificación con tipos de datos mixtos es:

1. Transformar todas las características categóricas en numéricas (usando One-Hot Encoding).Escalar todas las características numéricas (discretas y continuas)
2. .Aplicar el modelo GaussianNB al conjunto de datos totalmente transformado.Esto se debe a que las nuevas características binarias (provenientes de la categórica) y las características discretas escaladas pueden ser modeladas razonablemente bien por la distribución Gaussiana.

Esto funciona bien porque GaussianNB puede manejar tanto los valores continuos (si tu discreta se trata como continua) como las nuevas características binarias generadas por el One-Hot Encoding.

Ejercicio 1

```
# Paso 1: Codificar la característica categórica (e.g., Tipo-Cliente)
# Le aplicamos One-Hot Encoding
preprocesador-categorico = OneHotEncoder(handle-unknown='ignore')

# Paso 2: Escalar la característica discreta (e.g., Num-Compras-Mes)
# Le aplicamos un escalador para 'normalizarla' antes de usarla en GaussianNB
# Si usaramos MultinomialNB, podríamos omitir el escalado.
preprocesador-discreto = StandardScaler()

# ColumnTransformer:  Aplica el preprocesador correcto a la columna correcta
preprocesador = ColumnTransformer(
transformers=[
('cat', preprocesador-categorico, ['Tipo-Cliente']), # Aplicar OHE a la columna 1
('disc', preprocesador-discreto, ['Num-Compras-Mes']) # Aplicar escalado a la columna 2
],
remainder='passthrough' # Mantiene el resto de las columnas sin cambios
)
```

3. Crear el Pipeline Final. El Pipeline une el preprocesamiento y el modelo final en un solo objeto:Python # Ejemplo de Pipeline (enfoque más simple usando GaussianNB)
El modelo GaussianNB es tolerante con datos binarios (OHE) y escalados (discretos),
por lo que a menudo se utiliza como una solución de compromiso.
modelo-final = Pipeline(steps=[
('preprocesador', preprocesador),
('clasificador', GaussianNB()) # Usamos GaussianNB como solución "todo en uno"
])

Ahora puedes entrenar con una sola línea
modelo-final.fit(X-entrenamiento, y-entrenamiento)

Ejercicio 1

Construir un clasificador Naive-Bayes para discriminar si un paciente tiene una determinada patología a partir del resultado de dos pruebas clínicas (PC_1 y PC_2) y de la existencia o no de un determinado sintoma (S_1).Para aprender el clasificador, se utilizarán los datos de la tabla. Se pide:

1. Calcular las diferentes tablas de probabilidad y distribuciones gaussianas necesarias para construir el clasificador
2. Clasifica un paciente, con valores $[PC_1 = 120, PC_2 = Normal, S_1 = si]$, como patológico o normal, de acuerdo al clasificador construido

Paciente	PC_1	PC_2	S_1	**PATOLOGÍA**
1	100	Bajo	sí	**no**
2	125	Bajo	sí	**no**
3	145	Normal	no	**sí**
4	135	Bajo	no	**no**
5	100	Normal	no	**no**
6	140	Alto	no	**sí**
7	138	Alto	sí	**sí**
8	110	Normal	no	**no**
9	105	Bajo	no	**no**

Para realizar el ejercicio puedes codificar la característica PC_2 o bien crear tres columnas, esto te puede complicar los cálculos (un ordenador lo haría muy bien)