Métodos de clasificación Bayesianos [8.3]

Data Scientist
Jhoan Esteban Ruiz Borja Msc

"¿Qué son los clasificadores bayesianos?" Clasificadores bayesianos son clasificadores estadísticos. Pueden predecir las probabilidades de pertenencia de una clase como la probabilidad de que una tupla[1] determinada pertenece a una clase determinada.

Clasificación bayesiana se basa en el teorema de Bayes, que se describe a continuación. Estudios de comparación de algoritmos de clasificación han encontrado en un clasificador Bayesiano sencillo conocido como el **Clasificador Naïve Bayesiano** es comparable en rendimiento ha clasificadores cómo el árbol de decisión y las redes neuronales. Los clasificadores bayesianos también han exhibido una alta exactitud y velocidad cuando se aplica a grandes bases de datos.

El **Clasificador Naïve Bayesiano** asume que el efecto de un valor de atributo en una clase dada es independiente de los valores de otros atributos. Esta suposición se llama independencia condicional de clase. Se hace para simplificar los cálculos implicados y, en este sentido, se considera "ingenuo" (Naïve).

Teorema de Bayes [8.3.1]

Teorema de Bayes se nombra después Thomas Bayes, un clérigo inglés inconformista que hizo temprano trabajo en teoría de probabilidad y decisión durante el siglo XVIII. Sea $\boldsymbol{X}$[2] ser una tupla de datos. En términos bayesianos, $\boldsymbol{X}$ es considerado "evidencia". Como de costumbre, es descrito por las mediciones realizadas en un conjunto de $\boldsymbol{n}$ atributos. Sea $\boldsymbol{H}$ ser algunas hipótesis tales como que la tupla datos $\boldsymbol{X}$ pertenece a una clase especificada $\boldsymbol{C}$. Para problemas de clasificación, queremos determinar $\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{X})$, la probabilidad de que la hipótesis $\boldsymbol{H}$ se lleva a cabo teniendo en cuenta las "evidencias" u observaciones de la tupla de datos $\boldsymbol{X}$. En otras palabras, buscamos la probabilidad que esa tupla $\boldsymbol{X}$ pertenece a la clase $\boldsymbol{C}$, dado que sabemos la descripción de los atributos de $\boldsymbol{X}$.

$\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{X})$ es la **probabilidad posterior**, o la probabilidad a posteriori, de $\boldsymbol{H}$ condicionada en $\boldsymbol{X}$. En cambio, $\boldsymbol{P}(\boldsymbol{H})$ es la **probabilidad a prior** o probabilidad a priori, de $\boldsymbol{H}$. La probabilidad posterior $\boldsymbol{P}(\boldsymbol{H}|\boldsymbol{X})$, se basa en obtener más información que la probabilidad a priori, $\boldsymbol{P}(\boldsymbol{H})$, que es independiente de $\boldsymbol{X}$.

De manera similar, $\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{H})$ es la probabilidad posterior de $\boldsymbol{X}$ condicionada en $\boldsymbol{H}$. $\boldsymbol{P}(\boldsymbol{X})$ es la probabilidad a priori de $\boldsymbol{X}$.

---

[1] **Tupla:** En informática, o concretamente en el contexto de una base de datos relacional, un **registro** (también llamado **fila** o **tupla**) representa un objeto único de datos implícitamente estructurado en una tabla.
[2] Sea $\boldsymbol{X}$ el conjunto de $\boldsymbol{n}$ atributo como es el siguiente $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_n})$ donde $\boldsymbol{x_1}$ es un atributo.

"¿Cómo son estas probabilidades estimadas?" $\boldsymbol{P(H)}$, $\boldsymbol{P(X|H)}$ y $\boldsymbol{P(X)}$ pueden estimarse de los datos, como veremos a continuación. Teorema de Bayes es útil ya que es una manera de calcular la probabilidad posterior $\boldsymbol{P(H|X)}$, de $\boldsymbol{P(H)}$, $\boldsymbol{P(X|H)}$ y $\boldsymbol{P(X)}$. Teorema de Bayes es

$$\boldsymbol{P(H|X) = \frac{P(X|H)P(H)}{P(X)}} \qquad \boldsymbol{(8.10)}$$

Ahora, veremos cómo el teorema de Bayes se utiliza en clasificador naïve Bayesiano.

Clasificación Naïve Bayesiano [8.3.2]

El clasificador **Naïve Bayesiano**, o **simple clasificador Bayesiano,** trabaja como sigue:

1. Sea $\boldsymbol{D}$ un conjunto de entrenamiento de tuplas, junto con sus etiquetas de la clase asociada. Como de costumbre, cada tupla está representada por un vector **$\boldsymbol{n}$-dimensional** de atributos $\boldsymbol{X = (x_1, x_2, .., x_n)}$, que representan $\boldsymbol{n}$ mediciones hechas en la tupla de $\boldsymbol{n}$ atributos, respectivamente $\boldsymbol{A_1, A_2, \dots, A_n}$.
2. Suponga que hay $\boldsymbol{m}$ clases $\boldsymbol{C_1, C_2, \dots, C_m}$. Dado una tupla $\boldsymbol{X}$, el clasificador predice que $\boldsymbol{X}$ pertenece a la clase con la más alta probabilidad a posteriori, condicionada sobre $\boldsymbol{X}$. Es decir el clasificador Naïve Bayesiano predice que la tupla $\boldsymbol{X}$ pertenece a la clase $\boldsymbol{C_i}$ si y sólo si:

   $$\boldsymbol{P(C_i|X) > P(C_j|X) \qquad para \quad 1 \le j \le m, \qquad j \ne i}$$

   Así, maximizar $\boldsymbol{P(C_i|X)}$. La clase $\boldsymbol{C_i}$ para que $\boldsymbol{P(C_i|X)}$ es maximizada es llamada la hipótesis de máxima a posteriori. Por el teorema de Bayes (Ecuación 8.10):

   $$\boldsymbol{P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}} \qquad \boldsymbol{(8.11)}$$

3. Como $\boldsymbol{P(X)}$ es constante para todas las clases, solo $\boldsymbol{P(X|C_i)P(C_i)}$ necesita ser maximizada. Si las probabilidades a priori de las clases no son conocidas, entonces es común asumir que las clases son igualmente probables, es decir $\boldsymbol{P(C_1) = P(C_2) = \cdots = P(C_m)}$, y por tanto se podría maximizar $\boldsymbol{P(X|C_i)}$. En otro caso, maximizamos $\boldsymbol{P(X|C_i)P(C_i)}$. Note que las probabilidades a priori de las clases pueden ser estimadas por $\boldsymbol{P(C_i) = \frac{|C_{i,D}|}{|D|}}$, donde $\boldsymbol{|C_{i,D}|}$ es el número de tuplas de entrenamiento de la clase $\boldsymbol{C_i}$ en $\boldsymbol{D}$.
4. Dados conjuntos de datos con muchos atributos, sería extremadamente costoso computacionalmente para calcular $\boldsymbol{P(X|C_i)}$. Para reducir el costo de la computación en la evaluación de $\boldsymbol{P(X|C_i)}$, se hace el ingenuo supuesto de la independencia condicional de clase. Esto asume que los valores de los atributos son condicionalmente independientes uno del otro, dado la etiqueta de clase de la tupla (es decir, que no hay ninguna relación de dependencia entre los atributos). Por lo tanto:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times, \dots, \times P(x_n|C_i) \quad (8.12)$$

Podemos estimar fácilmente las probabilidades $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ de la tupla de entrenamiento. Recordemos que aquí x_k se refiere a los valores del atributo A_k para la tupla X. Para cada atributo, veremos si el atributo es de valores categóricos o continuos. Por ejemplo, para calcular $P(X|C_i)$, tenemos en cuenta lo siguiente:

i) Si el atributo A_k es categórico, entonces $P(x_k|C_i)$ es el número de tuplas de la clase C_i en D que tiene el valor de x_k para A_k, dividido por $|C_{i,D}|$, el número de tuplas de la clase C_i en D.
ii) Si el atributo A_k es de valores continuos, entonces tenemos que hacer un poco más de trabajo pero el cálculo es bastante sencillo. Un atributo de valores continuos es típicamente asumido que tiene una distribución Gaussiana con una media μ y desviación estándar σ, definida por:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8.13)$$

Así que

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (8.14)$$

Estas ecuaciones pueden aparecer enormes proporciones, ¡pero espera! Necesitamos calcular μ_{C_i} y σ_{C_i}, que son la media (es decir, promedio) y la desviación estándar, respectivamente, de los valores de atributo A_k para las tuplas de entrenamiento de clase C_i. A continuación, enchufe estas dos cantidades en la ecuación (8.13), junto con x_k para estimar $P(x_k|C_i)$.

5. Para predecir la etiqueta de la clase de X, $P(X|C_i)P(C_i)$ se evalúa para cada clase C_i. El clasificador predice que la etiqueta de la clase de la tupla X es la clase C_i si y sólo si:

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad para\ 1 \le j \le m, j \ne i \quad (8.15)$$

En otras palabras, la etiqueta de la clase predicha es la clase C_i para los que $P(X|C_i)P(C_i)$ es el máximo.

"¿Qué tan efectivos son los clasificadores bayesianos?" Varios estudios empíricos de este clasificador en comparación con árbol de decisión y los clasificadores de redes neuronales han encontrado que es comparable en algunos dominios. En teoría, los clasificadores bayesianos tienen la tasa de error mínimo en comparación con todos los otros clasificadores. Sin embargo, en la práctica esto no siempre es el caso, debido a imprecisiones en las suposiciones hechas para su uso, como la independencia de clase condicional, y la falta de datos disponibles de probabilidad.

Los clasificadores bayesianos también son útiles ya que proporcionan una justificación teórica para otros clasificadores que no utilizan explícitamente el teorema de Bayes. Por ejemplo, en ciertos supuestos, se puede demostrar que la red neural y la salida de los algoritmos de ajuste de curva la hipótesis de máximo a posteriori, como lo hace el clasificador bayesiano ingenuo.

"¿Qué pasa si encuentro valores de probabilidad de cero?" Recuerde que en la ecuación (8.12), estimamos $\boldsymbol{P(X|C_i)}$ como el producto de las probabilidades $\boldsymbol{P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)}$, basado en la suposición de independencia condicional-condicional. Estas probabilidades se pueden estimar a partir de las tuplas de entrenamiento (paso 4). Necesitamos calcular $\boldsymbol{P(X|C_i)}$ para cada clase ($\boldsymbol{i = 1, 2, \dots, m}$) para encontrar la clase $\boldsymbol{C_i}$ para la cual $\boldsymbol{P(X|C_i)P(C_i)}$ es el máximo (paso 5). Consideremos este cálculo. Para cada par de atributo-valor (es decir, $\boldsymbol{A_k = x_k}$ para $\boldsymbol{k = 1, 2, \dots, n}$) en la tupla $\boldsymbol{X}$, necesitamos contar el número de tuplas que tienen ese par de atributo-valor, por clase (es decir, Por $\boldsymbol{C_i}$, para $\boldsymbol{i = 1, 2, \dots, m}$).

Ejemplo de clasificación

Las siguientes tablas se usan para los ejemplos de clasificación:

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta (80)	Alta (90)	Si	No
3	Nublado	Alta (83)	Alta (86)	No	Si
4	Lluvioso	Media (70)	Alta (96)	No	Si
5	Lluvioso	Baja (68)	Normal (80)	No	Si
6	Lluvioso	Baja (65)	Normal (70)	Si	No
7	Nublado	Baja (64)	Normal (65)	Si	Si
8	Soleado	Media (72)	Alta (95)	No	No
9	Soleado	Baja (69)	Normal (70)	No	Si
10	Lluvioso	Media (75)	Normal (80)	No	Si
11	Soleado	Media (75)	Normal (70)	Si	Si
12	Nublado	Media (75)	Alta (90)	Si	Si
13	Nublado	Alta (81)	Normal (75)	No	Si
14	Lluvioso	Media (71)	Alta (91)	Si	No

Tabla 1. Datos del ejemplo

El ejemplo empleado tiene dos atributos, temperatura y humedad, que pueden emplearse como simbólicos o numéricos. Entre paréntesis se presentan sus valores numéricos. Además, se mostrarán ejemplos que permiten observar el funcionamiento del algoritmo, para lo que se utilizará la tabla 1, que presenta un sencillo problema de clasificación consistente en, a partir de los atributos que modelan el tiempo (vista, temperatura, humedad y viento), determinar si se puede o no jugar al tenis.

Proceso de Aprendizaje

Vistas			Temperatura			Humedad			Viento			Jugar		
	Si	No		Si	No		Si	No		Si	No		Si	No
Soleado	2	3	Alta	2	2	Alta	3	4	Si	3	3		9	5
Nublado	4	0	Media	4	2	Normal	6	1	No	6	2			
Lluvioso	3	2	Baja	3	1									
Soleado	2/9	3/5	Alta	2/9	2/5	Alta	3/9	4/5	Si	3/9	3/5		9/14	5/14
Nublado	4/9	0/5	Media	4/9	2/5	Normal	6/9	1/5	No	6/9	2/5			
Lluvioso	3/9	2/5	Baja	3/9	1/5									

Clasificación de un ejemplo de test

Vistas	Temperatura	Humedad	Viento	Jugar
Soleado	Baja	Alta	Si	?

$$\boldsymbol{P(Si \mid D) = P(Si) \times \prod_i P(A_i \mid Si) = \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.0053}$$

$$\boldsymbol{P(No \mid D) = P(No) \times \prod_i P(A_i \mid No) = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = 0.0206}$$

$$\boldsymbol{Normalizado} \begin{cases} \boldsymbol{P(Si \mid D) = \dfrac{0.0053}{0.0053 + 0.0206} = 20.5\%} \\ \boldsymbol{P(No \mid D) = \dfrac{0.0206}{0.0053 + 0.0206} = 79.5\%} \end{cases}$$

Respuesta no se juega.

Vistas	Temperatura	Humedad	Viento	Jugar
Soleado	Baja	Normal	No	?

$$\boldsymbol{P(Si \mid D) = P(Si) \times \prod_i P(A_i \mid Si) = \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{6}{9} = 0.0211}$$

$$\boldsymbol{P(No \mid D) = P(No) \times \prod_i P(A_i \mid No) = \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} = 0.00342}$$

$$\boldsymbol{Normalizado} \begin{cases} \boldsymbol{P(Si \mid D) = \dfrac{0.0211}{0.0211 + 0.00342} = 86.1\%} \\ \boldsymbol{P(No \mid D) = \dfrac{0.00342}{0.0211 + 0.00342} = 13.9\%} \end{cases}$$

Respuesta si se juega.

Bibliografía

Data Mining Concepts and Techniques [Jiawei Han | Micheline Kamber | Jian Pei] Third Edition 2012