

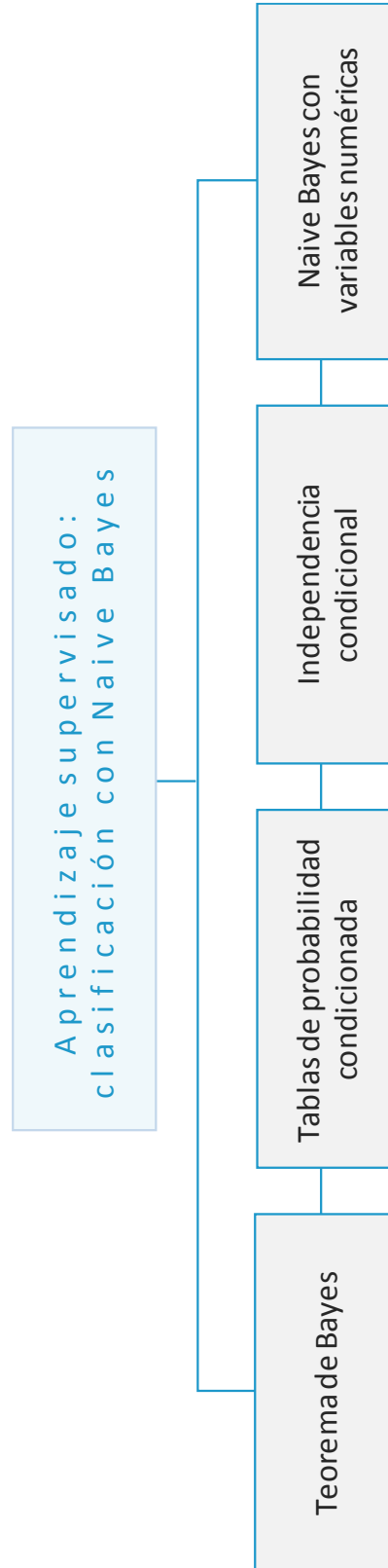
Aprendizaje Automático

Aprendizaje supervisado: clasificación con Naive Bayes

Índice

Esquema	3
Ideas clave	4
4.1. ¿Cómo estudiar este tema?	4
4.2. Teorema de Bayes	4
4.3. Tablas de probabilidad condicionada	7
4.4. Independencia condicional en el clasificador	
Naive Bayes	9
4.5. Clasificador Naive Bayes	11
4.6. Clasificador Naive Bayes con variables numéricas	13
4.7. Referencias bibliográficas	14
Lo + recomendado	15
+ Información	18
Actividades	19
Test	21

Esquema



4.1. ¿Cómo estudiar este tema?

Estudia este tema a través de las **Ideas clave** disponibles a continuación.

En este tema se va a introducir el clasificador Naive Bayes, el cual está basado en el teorema de Bayes para calcular las probabilidades a posteriori de los eventos a predecir.

Vamos a ver los siguientes puntos:

- ▶ En primer lugar, veremos el teorema de Bayes.
- ▶ A continuación, se describe la forma de calcular las tablas de probabilidad condicionada.
- ▶ Posteriormente se verá porque es importante asumir independencia condicional en el clasificador Naive Bayes.
- ▶ El clasificador Naive Bayes y finalmente como se puede utilizar con variables numéricas.

4.2. Teorema de Bayes

El Teorema de Bayes es una proposición planteada por el filósofo inglés Thomas Bayes (1702-1761) en el año 1763 en su artículo «An Essay towards solving a Problem in the Doctrine of Chances» publicado en la revista *Philosophical Transactions of the Royal Society of London*. Este teorema expresa la **probabilidad condicional** de un evento aleatorio A dado B en términos de la

distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de solo A.



Foto: Reverendo Thomas Bayes.

Recuperado de: https://upload.wikimedia.org/wikipedia/commons/d/d4/Thomas_Bayes.gif

El teorema de Bayes es bastante relevante porque **relaciona la probabilidad de dos eventos A y B utilizando la dependencia condicional de uno de ellos**. Es decir, relaciona la probabilidad de que ocurre el evento A y sabemos de antemano que ha ocurrido B, utilizando la probabilidad de que ocurra el evento B sabiendo que ha ocurrido A.

Este teorema **permite relacionar, entre otras cosas, síntomas y enfermedades**. Por ejemplo, sabiendo la probabilidad de tener dolor de garganta dado que se conoce que se tiene gripe, se puede obtener la probabilidad de tener gripe dado que se tiene dolor de garganta.

El teorema de Bayes **relaciona la comprensión de la probabilidad de aspectos causa-efecto dados los eventos dependientes observados**. Un evento dependiente es aquel cuyo resultado se ve afectado por el resultado de otro evento o serie de eventos. Los **eventos dependientes** son la base del modelado predictivo puesto que se busca obtener la probabilidad de que ocurra un suceso teniendo en cuenta la existencia de una serie de eventos dependientes.

En el caso de dos eventos dependientes A y B, el teorema de Bayes describe su relación de la siguiente manera:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Es decir, la probabilidad de A dado que ocurre B, es igual a, la probabilidad de que ocurra B dado que ha ocurrido A, multiplicado por la probabilidad de que ocurra A, dividido por la probabilidad de que ocurra B.

Este teorema se puede **generalizar para más de dos eventos** de la siguiente forma: en primer lugar, se entiende por evento mutuamente excluyente cuando dos resultados diferentes de un evento no pueden ocurrir al mismo tiempo. En el caso además de que los eventos sean exhaustivos, por lo menos uno de ellos debe de ocurrir.

De forma matemática, se puede definir el teorema de Bayes para n eventos: Sea $\{A_1, A_2, \dots, A_i, \dots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{P(B)}$$

Donde:

- ▶ $P(A_i)$ son las probabilidades *a priori*.
- ▶ $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i .
- ▶ $P(A_i|B)$ son las probabilidades *a posteriori*.
- ▶ $P(B)$ es la verosimilitud marginal.

Ejemplo:

Supongamos que deseamos estimar la probabilidad de que un mensaje de correo electrónico sea *spam*. Sin tener evidencias adicionales y, únicamente, por los mensajes previos obtenidos la probabilidad *a priori* de que un mensaje sea *spam* es 0,2. Ahora bien, si tenemos evidencia de que el mensaje contiene la palabra Viagra; por medio de conocer la probabilidad de que la palabra Viagra haya sido utilizada en mensajes de *spam* previos, lo cual se conoce como verosimilitud o *likelihood*; y por medio de la probabilidad de que la palabra Viagra aparezca en cualquier mensaje, lo que se conoce como verosimilitud marginal. Aplicando el teorema de Bayes se puede calcular la probabilidad *a posteriori* y si esta es mayor que 0,5 es más probable que el mensaje sea *spam*. En la siguiente formula se muestra el cálculo anteriormente descrito.

$$P(\text{Spam}_i | \text{Viagra}) = \frac{P(\text{Viagra} | \text{Spam}) * P(\text{Spam})}{P(\text{Viagra})}$$

El **teorema de Bayes** relaciona la probabilidad de dos o más eventos utilizando la dependencia condicional de cada uno de ellos. Los eventos deben ser dependientes y mutuamente excluyentes.

4.3. Tablas de probabilidad condicionada

Para obtener cada uno de los componentes de las formulas anteriores es necesario construir una **tabla de frecuencias** que indica el número de veces que el evento aparece en cada una de las situaciones. En nuestro ejemplo de *spam*, es necesario calcular el número de veces que la palabra Viagra ha aparecido en los mensajes de *spam*. Esta tabla de frecuencias se utiliza posteriormente para calcular las tablas de verosimilitud o de probabilidad condicionada.

Siguiendo con el ejemplo de los mensajes de *spam*, en el caso hipotético que tuviéramos la siguiente distribución histórica de 100 mensajes para la palabra Viagra.

Frecuencia	Si	No	Total
<i>Spam</i>	4	16	20
<i>Ham</i>	1	79	80
Total	5	95	100

Tabla 1. Ejemplo de distribución histórica de mensajes *spam* y *ham* para la palabra Viagra.

Obtendríamos la correspondiente tabla de verosimilitud:

Verosimilitud	Si	No	Total
<i>Spam</i>	4/20	16/20	20
<i>Ham</i>	1/80	79/80	80
Total	5/100	95/100	100

Tabla 2. Ejemplo de tabla de verosimilitud para mensajes *spam* y *ham* en función de la palabra Viagra.

Con estos datos, para calcular la probabilidad *a posteriori* de que un mensaje sea *spam* dado que nos ha llegado la palabra Viagra, tendríamos que hacer el siguiente calculo:

$$P(\text{Spam}|\text{Viagra}) = [(4/20) * (20/100)] / (5/100) = 0.8$$

Es decir, con los datos anteriores, la probabilidad de que un correo electrónico que contenga la palabra Viagra sea *spam* es del 0,8.

Ahora supongamos que deseamos añadir a este cálculo otros términos más comunes que aparecen en los mensajes *spam*, como pueden ser: *money*, *groceries* y *unsubscribe*.

En este caso, tendríamos la siguiente tabla de verosimilitud:

	Viagra (W1)		Money (W2)		Groceries (W3)		Unsubscribe (W4)		
Verosimilitud	Si	No	Si	No	Si	No	Si	No	Total
<i>Spam</i>	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
<i>Ham</i>	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
Total	5/10	95/10	24/10	76/10	8/100	91/100	35/100	65/100	100

Tabla3. Ejemplo de tabla de verosimilitud para los mensajes *spam* y *ham* en función de las palabras *Viagra*, *money*, *groceries* y *unsubscribe*.

En este caso, si llega un nuevo mensaje que contiene las palabras *Viagra* y *unsubscribe*, pero no *money* ni *groceries*; utilizando el teorema de Bayes habría que calcular la siguiente formula:

$$P(\text{Spam}|\text{Viagra}) = \frac{P(\text{Viagra}|\text{Spam}) * P(\text{Spam})}{P(\text{Viagra})}$$

El cálculo de la fórmula anterior es **computacionalmente costoso**, puesto que a medida que se añaden nuevos términos son necesarias grandes cantidades de memoria para almacenar todas las combinaciones.

4.4. Independencia condicional en el clasificador

Naive Bayes

Debido a que el cálculo riguroso de la fórmula del teorema de Bayes, como en el ejemplo anterior, es computacionalmente costoso, el clasificador Naive Bayes se basa en una **modificación sencilla**. Básicamente asume **independencia condicional** entre los eventos, a pesar de que si todos los eventos fueran independientes sería imposible predecir ningún evento con los datos

observados por otro. Formalmente, **dos eventos son independientes** si el resultado del segundo evento no es afectado por el resultado del primer evento. Si A y B son eventos independientes, la probabilidad de que ambos eventos ocurran es el producto de las probabilidades de los eventos individuales.

Por otro lado, los **eventos dependientes** son la base del modelado predictivo, puesto que permiten predecir la presencia de un evento en función de otro. Por ejemplo, la presencia de nubes suele ser un evento predictivo de un día lluvioso, o la presencia de la palabra *viagra* en un correo electrónico suele ser un evento predictivo de *spam*.

No obstante, al no poder asumir dependencia condicional por el alto coste computacional, el clasificador Naive Bayes asume **independencia condicional** entre los eventos condicionados al mismo valor de la clase. Este hecho es el que le ha dado el adjetivo de *naive* al clasificador.

En nuestro ejemplo anterior, asumiendo independencia condicional de las palabras para obtener la probabilidad de *spam*, tendríamos la siguiente formula:

$$\begin{aligned}
 P(\text{Spam}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \\
 &= \frac{P(W_1|\text{Spam})P(\neg W_2|\text{Spam})P(\neg W_3|\text{Spam})P(W_4|\text{Spam}) * P(\text{Spam})}{P(W_1)P(\neg W_2)P(\neg W_3)P(W_4)} \\
 &= (4/20) * (10/20) * (20/20) * (12/20) * (20/100) = 0.012
 \end{aligned}$$

Por otro lado, para obtener la probabilidad de *ham*, tendríamos:

$$\begin{aligned}
 P(\text{Ham}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \\
 &= \frac{P(W_1|\text{Ham})P(\neg W_2|\text{Ham})P(\neg W_3|\text{Ham})P(W_4|\text{Ham}) * P(\text{Ham})}{P(W_1)P(\neg W_2)P(\neg W_3)P(W_4)} \\
 &= (1/80) * (66/80) * (71/80) * (23/80) * (80/100) = 0.002
 \end{aligned}$$

Como $0,012/0,002 = 6$, se puede afirmar que es 6 veces más probable que el mensaje sea *spam* que *ham*.

Si queremos calcular la probabilidad de que el mensaje sea *spam*, sería igual a la verosimilitud de que el mensaje sea *spam* dividido por la verosimilitud de que sea *spam* o *ham*: $0,012 / (0,012 + 0,002) = 0,857$

Análogamente, la probabilidad de ser *ham* es: $0,002 / (0,002 + 0,012) = 0,143$

Por tanto, podemos estimar que, dadas las palabras del mensaje, hay una probabilidad de 0,857 de que sea *spam* y de 0,143 de que sea *ham* y como son eventos mutuamente excluyentes suman 1.

4.5. Clasificador Naive Bayes

Como hemos comentado previamente, el teorema de Bayes es la **base** para el clasificador Naive Bayes. Este clasificador utiliza las probabilidades *a priori* de los eventos para estimar la **probabilidad de eventos futuros** por medio del teorema de Bayes.

Este clasificador **utiliza datos históricos o de entrenamiento** para calcular la probabilidad observada de cada evento en función de su vector de características. Para realizar una predicción, el clasificador es utilizado con datos que tienen la clase desconocida y se utilizan las probabilidades observadas para estimar la clase más probable.

La fórmula general del clasificador Naive Bayes se puede definir de la siguiente manera: la probabilidad del nivel L de la clase C , dada la evidencia proporcionada por las variables de F_1, F_n es igual al producto de las probabilidades de cada evidencia condicionada al nivel de la clase, la probabilidad *a priori* del nivel de la clase y un factor de escala $1/Z$ que convierte el resultado en probabilidad:

$$P(C_L|F_1, \dots, F_n) = \frac{1}{Z} p(C_L) \prod_{i=1}^n p(F_i|C_L)$$

Este clasificador se utiliza principalmente para **clasificar texto, para detección de intrusos en redes de computadores, diagnósticos médicos**, etc. Por ejemplo, se puede utilizar la frecuencia de las palabras de los correos electrónicos para identificar nuevos correos *spam* en el futuro.

Combinaciones Desconocidas

Supongamos que ahora recibimos un mensaje que contiene las palabras *Viagra*, *groceries*, *money* y *unsubscribe*, y queremos estimar la probabilidad de que el mensaje sea *Viagra*. En este caso la verosimilitud de *spam* es:

$$(4/20) * (10/20) * (0/20) * (12/20) * (20/100) = 0$$

Por otro lado, la verosimilitud de *ham* es:

$$(1/80) * (14/80) * (8/80) * (23/80) * (80/100) = 0.00005$$

La probabilidad de *spam* es:

$$0/(0 + 0.0099) = 0$$

Y la probabilidad de *ham* es:

$$0.00005 / (0 + 0.00005) = 1$$

Este problema sucede cuando un evento nunca ha ocurrido para una o más categorías de las clases. Por ejemplo, si nunca se ha visto el termino *groceries* en un mensaje spam $P(\text{Spam} | \text{groceries}) = 0$.

La solución es añadir un pequeño número a todas las clases en la tabla, para asegurarse que no existe ninguna combinación con probabilidad de ocurrir igual a 0, esto se conoce con el nombre de **estimador de Laplace**.

Por ejemplo, si usamos un valor de 1, la verosimilitud de *spam* y *ham* quedaría:

$$(5/24) * (11/24) * (1/24) * (13/24) * (20/100) = 0.0004$$
$$(2/84) * (15/84) * (9/84) * (24/84) * (80/100) = 0.0001$$

Lo que indica que la probabilidad de que el mensaje sea spam es del 0,8 y de que sea *ham* del 0,2.

4.6. Clasificador Naive Bayes con variables numéricas

Debido a que el clasificador Naive Bayes utiliza tablas de frecuencias para calcular las probabilidades, cada una de las variables utilizada debe de ser **categorica** y no se pueden utilizar de forma directa variables numéricas.

Una solución sencilla es **discretizar las variables numéricas en N conjuntos, agrupamientos o bins**. Este método es ideal cuando hay grandes cantidades de datos. Una cuestión importante aquí es considerar el punto de corte óptimo para hacer cada

uno de los agrupamientos. Una buena solución suele ser explorar los datos para observar los puntos de corte en la distribución de los datos.

Por ejemplo, el siguiente histograma:

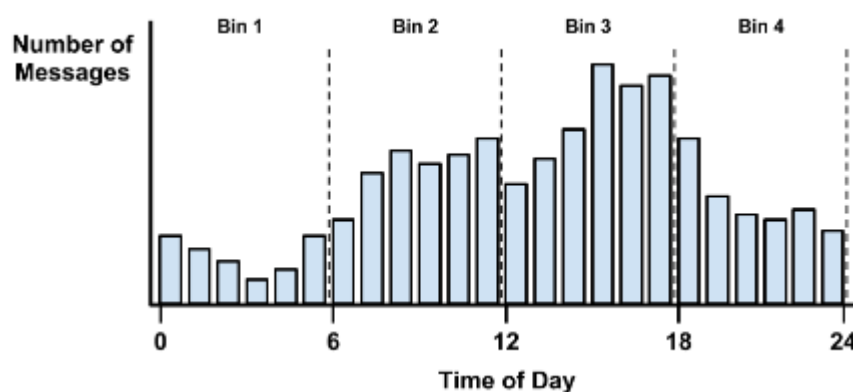


Gráfico 1. Fuente: Lantz, 2013.

Sugiere realizar una división en cuatro *bins*.

La **discretización** siempre se traduce en una reducción de la información, ya que la granularidad inicial se reduce. Por tanto, es importante mantener un balance entre el número de *bins*, puesto que con muy pocas se pierda mucha información y con muchas el proceso es muy costoso.

4.7. Referencias bibliográficas

Lantz, B. (2013) *Machine Learning with R*. Packt

Lo + recomendado

No dejes de leer

Naïve Bayes

Scikit learn. (s.f.) Naive Bayes.

Ejemplo de creación de un modelo Naive Bayes utilizando la librería scikit-learn de Python.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

http://scikit-learn.org/stable/modules/naive_bayes.html

How to Implement Naive Bayes From Scratch in Python

Brownlee, J. (diciembre, 2014) How to Implement Naive Bayes From Scratch in Python. *machinelearningmastery.com*.

Ejemplo de creación de un modelo de Naive Bayes desde cero utilizando el lenguaje de programación Python.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

<https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>

Naive Bayes en Python

Greg (11 de diciembre de 2014) Naive Bayes en Python [Mensaje en un blog]. *The Yat Blog*.

Explicación de cómo funciona el algoritmo Naive Bayes y ejemplo de implementación utilizando el lenguaje Python.

Accede al artículo a través del aula virtual o desde la siguiente dirección web:

<http://blog.yhat.com/posts/naive-bayes-in-python.html>

Doing Naive Bayes Classification: aplicado al libro «50 sombras de Grey»

Cherny, L. (octubre de 2015) Doing Naive Classification.

Ejemplo de un clasificador Naive Bayes sobre el texto de las páginas del libro 50 sombras de grey con el objetivo de clasificar el contenido de las páginas.

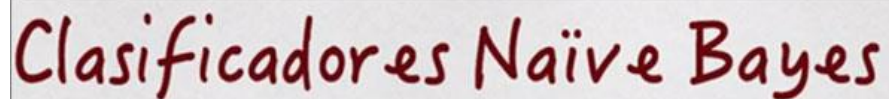
Accede al artículo a través del aula virtual o desde la siguiente dirección web:

http://nbviewer.jupyter.org/github/arnicas/NLP-in-Python/blob/master/4.%20Naive%20Bayes%20Classification.ipynb?imm_mid=0cd660&cmp=em-data-na-na-newsltr_20150225

No dejes de ver

Clasificadores Naive Bayes

Breve introducción al funcionamiento de los clasificadores Naive Bayes.



Clasificadores Naïve Bayes

Accede al vídeo a través del aula virtual o desde la siguiente dirección web:

<https://www.youtube.com/watch?v=oQ1OyqvL7dQ>

Bibliografía

Lantz, B. (2013) *Machine Learning with R*. Packt

Trabajo: Clasificación con Naive Bayes

Descripción del trabajo

Mediante este trabajo se pretende que pongas en práctica la creación de modelos basados en el clasificador Naive Bayes. El objetivo es que comprendas de forma práctica con un problema determinado los pasos que hay que realizar para construir un clasificador Naive Bayes que infiera la clase más probable de dos eventos.

Para este trabajo puedes utilizar la herramienta **R** y el entorno de desarrollo **RStudio**, por tanto, debes tener correctamente descargado e instalado estos programas en tu ordenador.

- ▶ El intérprete del lenguaje R, lo puedes descargar desde aquí: <https://cran.r-project.org/mirrors.html>
- ▶ El programa RStudio lo puedes descargar desde aquí: <https://www.rstudio.com/products/rstudio/download/#download>

Además, puedes usar el paquete **e1017** de R para la implementación del Naive Bayes: <https://cran.r-project.org/web/packages/e1071/e1071.pdf> y el paquete **tm** para llevar a cabo operaciones básicas de preparación de texto: <https://cran.r-project.org/web/packages/tm/tm.pdf>

Elaboración del trabajo

Durante este trabajo utilizarás los paquetes anteriores para realizar un clasificador Naive Bayes. Para ello es necesario procesar los datos de entrada (entrenamiento)

que serán proporcionados para el trabajo y entrenar los modelos correspondientes para llevar a cabo las predicciones sobre otros conjuntos de test.

Entrega del trabajo

Tras la realización del trabajo deberás entregar por un lado un fichero con el código que has realizado para generar los modelos y las predicciones y un informe que contenga los resultados obtenidos y una explicación del modelo desarrollado.

El informe tendrá una **extensión máxima de 3 páginas** siendo la fuente utilizada Georgia 11 e interlineado, 1,5.

Evaluación

El criterio de evaluación de esta actividad será tanto la capacidad predictiva obtenida en el conjunto de test, como la creatividad y rigurosidad llevada a cabo para obtener la solución.

1. El teorema de Bayes:
 - A. Fue propuesto por el reverendo Thomas Bayes.
 - B. Relaciona la probabilidad de dos eventos A y B utilizando la dependencia condicional de uno de ellos.
 - C. Relaciona la probabilidad de dos eventos A y B utilizando la dependencia condicional de ambos de ellos.
2. En el teorema de Bayes
 - A. Los eventos deben ser dependientes y mutuamente excluyentes.
 - B. Los eventos deben ser independientes y mutuamente excluyentes.
 - C. Los eventos deben ser independientes.
3. Si dos eventos son exhaustivos:
 - A. Deben ocurrir los dos.
 - B. Al menos debe ocurrir uno de ellos.
 - C. Ninguna de las anteriores es correcta.
4. Un evento mutuamente excluyente:
 - A. Cuando siempre debe ocurrir el mismo evento.
 - B. Cuando dos resultados diferentes de un mismo evento no pueden ocurrir al mismo tiempo.
 - C. Ninguna de las anteriores es correcta.

5. Las tablas de frecuencias:
- A. Indican el número de veces que el evento aparece en cada una de las situaciones.
 - B. Sirven para medir el éxito del modelo.
 - C. Son la base para la construcción del modelo Naive Bayes.
6. Los eventos dependientes:
- A. Permiten estimar la presencia de un evento en función del otro.
 - B. Implica que siempre ocurren a la vez.
 - C. Implica que la existencia de uno puede conllevar la existencia del otro.
7. Cuáles de las siguientes afirmaciones son ciertas sobre el clasificador de Naive Bayes
- A. Utiliza datos históricos para obtener la probabilidad observada de cada evento en función de su vector de características.
 - B. Asume independencia condicional entre los eventos.
 - C. El cálculo riguroso del teorema de Bayes es computacionalmente costoso.
8. Cuando existen combinaciones desconocidas en los datos de entrada:
- A. Las probabilidades a posteriori obtenidas pueden no tener sentido.
 - B. El teorema de Bayes utiliza el estimador de Laplace.
 - C. Se eliminan estas combinaciones de los datos de entrada.
9. La discretización de variables:
- A. Es una técnica que se aplica para utilizar el clasificador Naive Bayes con variables numéricas.
 - B. Es ideal cuando hay grandes cantidades de datos.
 - C. Funciona mejor cuando hay pocos datos.

10. La discretización de variables:

- A. Siempre se traduce en reducción de información.
- B. Nunca se traduce en reducción de información.
- C. Ninguna de las anteriores es correcta.