Aprendizaje probabilista – clasificación mediante el método naïve Bayes

Cuando un meteorólogo proporciona un pronóstico del tiempo, las precipitaciones se describen típicamente con frases como “70 por ciento de probabilidad de lluvia”. Estos pronósticos se conocen como informes de probabilidad de precipitación.
¿Alguna vez has considerado cómo se calculan? Es una pregunta desconcertante porque, en realidad, lloverá o no con absoluta certeza.

Las estimaciones meteorológicas se basan en métodos probabilísticos, que son aquellos que se ocupan de describir la incertidumbre. Utilizan datos sobre eventos pasados para extrapolar eventos futuros. En el caso del clima, la probabilidad de lluvia describe la proporción de días anteriores con condiciones atmosféricas similares en los que se produjo precipitación. Un 70 por ciento de probabilidad de lluvia implica que, en 7 de cada 10 casos pasados con condiciones similares, la precipitación ocurrió en algún lugar del área.

Este documento cubre el algoritmo Naïve Bayes, que utiliza probabilidades de la misma manera que un pronóstico del tiempo. Mientras estudias este método, aprenderás sobre:

- Principios básicos de probabilidad
- Los métodos especializados y las estructuras de datos necesarias para analizar datos de texto con R
- Cómo emplear Naïve Bayes para crear un filtro de mensajes basura de Short Message Service (SMS)

Si has tomado una clase de estadística antes, parte del material de este documento puede ser un repaso.
Aun así, puede ser útil refrescar tus conocimientos de probabilidad. Descubrirás que estos principios son la base de cómo Naïve Bayes obtuvo un nombre tan extraño.

Entendiendo Naïve Bayes

Las ideas estadísticas básicas necesarias para entender el algoritmo Naïve Bayes han existido durante siglos. La técnica desciende del trabajo del matemático del siglo XVIII Thomas Bayes, quien desarrolló principios fundamentales para describir la probabilidad de eventos y cómo estas probabilidades deben revisarse a la luz de información adicional. Estos principios formaron la base de lo que ahora se conoce como métodos bayesianos.

Trataremos estos métodos con mayor detalle más adelante. Por ahora, basta con decir que una probabilidad es un número entre cero y uno (o de 0 a 100 por ciento) que captura la posibilidad de que ocurra un evento a la luz de la evidencia disponible. Cuanto menor sea la probabilidad, menos probable es que ocurra el evento. Una probabilidad de cero indica que el evento definitivamente no ocurrirá, mientras que una probabilidad de uno indica que el evento ocurrirá con absoluta certeza.

Los eventos más interesantes de la vida tienden a ser aquellos con probabilidad incierta; estimar la probabilidad de que ocurran nos ayuda a tomar mejores decisiones al revelar los resultados más probables.

Los clasificadores basados en métodos bayesianos utilizan datos de entrenamiento para calcular la probabilidad de cada resultado en función de la evidencia proporcionada por los valores de las características. Cuando el clasificador se aplica más tarde a datos no etiquetados, utiliza estas probabilidades calculadas para predecir la clase más probable para el nuevo ejemplo. Es una idea simple, pero da como resultado un método que puede tener resultados a la par de algoritmos más sofisticados. De hecho, los clasificadores bayesianos se han utilizado para:

- Clasificación de texto, como el filtrado de correo basura (spam)
- Detección de intrusiones o anomalías en redes computacionales
- Diagnóstico de afecciones médicas a partir de un conjunto de síntomas observados

Normalmente, los clasificadores bayesianos se aplican mejor a problemas para los que se debe considerar simultáneamente la información de numerosos atributos para estimar la probabilidad general de un resultado. Si bien muchos algoritmos de aprendizaje automático ignoran las características que tienen efectos débiles, los métodos bayesianos utilizan toda la evidencia disponible para cambiar sutilmente las predicciones. Esto implica que incluso si una gran parte de las características tienen efectos relativamente menores, su impacto combinado en un modelo bayesiano podría ser bastante grande.

Conceptos básicos de los métodos bayesianos

Antes de adentrarnos en el algoritmo Naïve Bayes, vale la pena dedicar un tiempo a definir los conceptos que se utilizan en los métodos bayesianos. Resumida en una sola oración, la teoría de probabilidad bayesiana se basa en la idea de que la probabilidad estimada de un **evento**, o resultado potencial, debe basarse en la evidencia disponible a partir de múltiples **ensayos** u oportunidades para que el evento ocurra.

La siguiente tabla ilustra eventos y ensayos para varios resultados del mundo real:

Evento	Ensayo
Resultado cara	Lanzamiento de una moneda
Tiempo lluvioso	Un solo día (u otro período de tiempo)
Mensaje spam	Un mensaje de correo electrónico entrante
Candidato se convierte en presidente	Una elección presidencial
Mortalidad	Un paciente de hospital
Ganar la lotería	Un billete de lotería

Los métodos bayesianos brindan información sobre cómo se puede estimar la probabilidad de estos eventos a partir de datos observados. Para ver cómo, necesitaremos formalizar nuestra comprensión de la probabilidad.

Comprensión de la probabilidad

La probabilidad de un evento se estima a partir de datos observados dividiendo el número de ensayos en los que ocurrió el evento por el número total de ensayos. Por ejemplo, si llovió 3 de 10 días con condiciones similares a las de hoy, la probabilidad de lluvia hoy se puede estimar como 3/10 = 0.30 o 30 por ciento. De manera similar, si 10 de los 50 mensajes de correo electrónico anteriores fueron spam, entonces la probabilidad de que cualquier mensaje entrante sea spam puede estimarse como 10/50 = 0.20 o 20 por ciento.

Para denotar estas probabilidades, utilizamos la notación en la forma $P(A)$, que significa la probabilidad del evento A. Por ejemplo, P(lluvia) = 0.30 para indicar una probabilidad del 30 por ciento de lluvia o P(spam) = 0.20 para describir una probabilidad del 20 por ciento de que un mensaje entrante sea spam.

Debido a que un ensayo siempre da como resultado algún resultado, la probabilidad de todos los resultados posibles de un ensayo siempre debe sumar uno. Por lo tanto, si el ensayo tiene exactamente dos resultados y los resultados no pueden ocurrir simultáneamente, entonces conocer la probabilidad de uno de los resultados revela la probabilidad del otro. Este es el caso de muchos resultados, como cara o cruz al lanzar una moneda, o mensajes de correo electrónico spam versus legítimos, también conocidos como “jamón (ham)”.

Usando este principio, sabiendo que P(spam) = 0.20 podemos calcular P(ham) = 1 – 0.20 = 0.80. Esto sólo funciona porque spam y ham son **eventos mutuamente excluyentes y exhaustivos**, lo que implica que no pueden ocurrir al mismo tiempo y son los únicos resultados posibles.

Un solo evento no puede ocurrir y no ocurrir simultáneamente. Esto significa que un evento siempre es mutuamente excluyente y exhaustivo con su complemento, o el evento que comprende todos los demás resultados en los que el evento de interés no ocurre. El complemento del evento *A* se denota típicamente *A*‘ o *A*’.

Además, la notación abreviada $P(A^c)$ o P(¬A) se puede utilizar para denotar la probabilidad de que el evento *A* no ocurra. Por ejemplo, la notación *P*(¬spam) = 0.80 sugiere que la probabilidad de que un mensaje no sea spam es del 80%.

Para ilustrar eventos y sus complementos, a menudo es útil imaginar un espacio bidimensional que se divide en probabilidades para cada evento. En el siguiente diagrama, el rectángulo representa los posibles resultados de un mensaje de correo electrónico. El círculo representa la probabilidad del 20 por ciento de que el mensaje sea spam. El 80 por ciento restante representa el complemento P(¬spam), o los mensajes que no son spam:

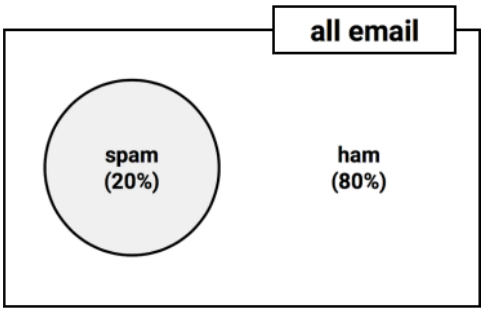

Figura 4.1: El espacio de probabilidad para todos los correos electrónicos se puede visualizar como particiones de spam y ham.

Comprender la probabilidad conjunta

A menudo, nos interesa monitorear varios eventos no mutuamente excluyentes en el mismo ensayo. Si ciertos eventos ocurren simultáneamente con el evento de interés, podemos usarlos para hacer predicciones.

Considera, por ejemplo, un segundo evento basado en el resultado de que un mensaje de correo electrónico contiene la palabra *Viagra*. El diagrama anterior, actualizado para este segundo evento, podría aparecer como se muestra en el siguiente diagrama (figura 4.2):

Observa en el diagrama que el círculo de Viagra se superpone con las áreas de spam y ham del diagrama y el círculo de spam incluye un área no cubierta por el círculo de Viagra. Esto implica que no todos los mensajes de spam contienen el término Viagra y algunos mensajes con el término Viagra son ham.

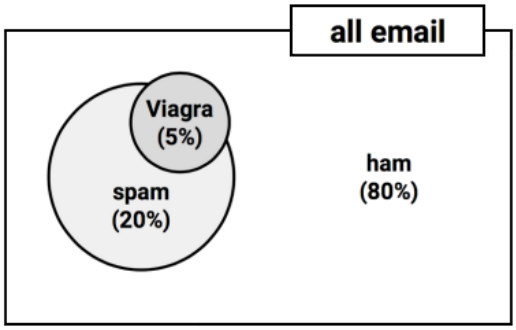

Figura 4.2: Los eventos no mutuamente excluyentes se representan como particiones superpuestas.

Sin embargo, debido a que esta palabra aparece muy raramente fuera de spam, su presencia en un nuevo mensaje entrante sería una fuerte evidencia de que el mensaje es spam.

Para ampliar la imagen y ver mejor la superposición entre estos círculos, utilizaremos una visualización conocida como diagrama de Venn. Utilizado por primera vez a finales del siglo XIX por el matemático John Venn, el diagrama utiliza círculos para ilustrar la superposición entre conjuntos de elementos. Como en la mayoría de los diagramas de Venn, el tamaño y el grado de superposición de los círculos en la representación no son significativos. En cambio, se utilizan como recordatorio para asignar probabilidad a todas las combinaciones de eventos. Un diagrama de Venn para spam y Viagra podría representarse de la siguiente manera:

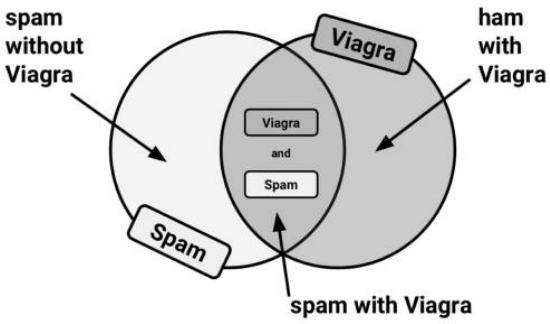

Figura 4.3: Un diagrama de Venn ilustra la superposición de los eventos spam y Viagra.

Sabemos que el 20 por ciento de todos los mensajes eran spam (el círculo de la izquierda) y el 5 por ciento de todos los mensajes contenían la palabra Viagra (el círculo de la derecha). Nos gustaría cuantificar el grado de superposición entre estas dos proporciones. En otras palabras, esperamos estimar la probabilidad de que ocurran tanto *P*(spam) como *P*(Viagra), que se puede escribir como *P*(spam ∩ Viagra). El símbolo ∩ significa la intersección de los dos eventos; la notación *A* ∩ *B* se refiere al evento en el que ocurren tanto *A* como *B*.

El cálculo de *P*(spam ∩ Viagra) depende de la **probabilidad conjunta** de los dos eventos, o de cómo la probabilidad de un evento se relaciona con la probabilidad del otro. Si los dos

eventos no están relacionados en absoluto, se denominan eventos independientes. Esto no quiere decir que los **eventos independientes** no puedan ocurrir al mismo tiempo; la independencia de los eventos simplemente implica que conocer el resultado de un evento no proporciona ninguna información sobre el resultado del otro. Por ejemplo, el resultado de un lanzamiento de moneda con cara es independiente de si el clima es lluvioso o soleado en un día determinado.

Si todos los eventos fueran independientes, sería imposible predecir un evento observando otro.
En otras palabras, los **eventos dependientes** son la base del modelado predictivo. Así como la presencia de nubes predice un día lluvioso, la aparición de la palabra Viagra predice un correo electrónico no deseado.

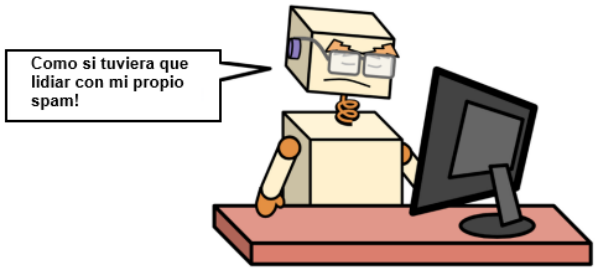

Figura 4.4: Se requieren eventos dependientes para que las máquinas aprendan a identificar patrones útiles.

Calcular la probabilidad de eventos dependientes es un poco más complejo que calcular la probabilidad de eventos independientes.

Si *P*(spam) y *P*(Viagra) fueran independientes, podríamos calcular fácilmente *P*(spam ∩ Viagra), la probabilidad de que ambos eventos ocurran al mismo tiempo. Como el 20 por ciento de todos los mensajes son spam y el 5 por ciento de todos los correos electrónicos contienen la palabra Viagra, podríamos suponer que el 1 por ciento de todos los mensajes con el término Viagra son spam.

Esto se debe a que 0.05 * 0.20 = 0.01. De manera más general, para los eventos independientes *A* y *B*, la probabilidad de que ambos ocurran se puede calcular como $P(A \cap B) = P(A) * P(B)$.

Dicho esto, sabemos que es probable que *P*(spam) y *P*(Viagra) sean altamente dependientes, lo que significa que este cálculo es incorrecto. Para obtener una estimación más razonable, necesitamos utilizar una formulación más cuidadosa de la relación entre estos dos eventos, que se basa en métodos bayesianos más avanzados.

Cálculo de la probabilidad condicional con el teorema de Bayes

Las relaciones entre eventos dependientes se pueden describir utilizando el **teorema de Bayes**, que proporciona una forma de pensar sobre cómo revisar una estimación de la probabilidad de un evento a la luz de la evidencia proporcionada por otro. Una formulación es la siguiente:

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

La notación $P(A|B)$ se lee como la probabilidad del evento A dado que ocurrió el evento B. Esto se conoce como **probabilidad condicional** ya que la probabilidad de A depende (es decir, es condicional) de lo que sucedió con el evento B.

El teorema de Bayes establece que la mejor estimación de $P(A|B)$ es la proporción de ensayos en los que A ocurrió con B, de todos los ensayos en los que ocurrió B. Esto implica que la probabilidad del evento A es mayor si A y B ocurren a menudo juntos cada vez que se observa B. Ten en cuenta que esta fórmula ajusta $P(A \cap B)$ para la probabilidad de que ocurra B. Si B es extremadamente raro, $P(B)$ y $P(A \cap B)$ siempre serán pequeñas; sin embargo, si A casi siempre ocurre junto con B, $P(A|B)$ seguirá siendo alto a pesar de la rareza de B.

Por definición, $P(A \cap B) = P(A|B) * P(B)$, un hecho que se puede derivar fácilmente aplicando un poco de álgebra a la fórmula anterior. Reordenando esta fórmula una vez más con el conocimiento de que $P(A \cap B) = P(B \cap A)$ se llega a la conclusión de que $P(A \cap B) = P(B|A) * P(A)$, que podemos utilizar en la siguiente formulación del teorema de Bayes:

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

De hecho, esta es la formulación tradicional del teorema de Bayes por razones que se aclararán a medida que lo apliquemos al aprendizaje automático. Primero, para entender mejor cómo funciona el teorema de Bayes en la práctica, revisemos nuestro filtro de spam hipotético.

Sin el conocimiento del contenido de un mensaje entrante, la mejor estimación de su estado de spam sería $P(\text{spam})$, la probabilidad de que cualquier mensaje anterior fuera spam. Esta estimación se conoce como **probabilidad a priori**. Anteriormente, habíamos descubierto que era del 20 por ciento.

Supongamos que obtuviste evidencia adicional al observar con más atención el conjunto de mensajes recibidos anteriormente y examinar la frecuencia con la que aparecía el término Viagra. La probabilidad de que la palabra Viagra se haya utilizado en mensajes de correo no deseado anteriores, o *P*(Viagra|spam), se denomina likelihood (probabilidad). La probabilidad de que Viagra apareciera en cualquier mensaje, o *P*(Viagra), se conoce como **probabilidad marginal**.

Al aplicar el teorema de Bayes a esta evidencia, podemos calcular una **probabilidad a posteriori** que mide la probabilidad de que un mensaje sea correo no deseado. Si la probabilidad posterior es mayor del 50 por ciento, es más probable que el mensaje sea correo no deseado que correo basura, y tal vez deba filtrarse. La siguiente fórmula muestra cómo se aplica el teorema de Bayes a la evidencia proporcionada por mensajes de correo electrónico anteriores:

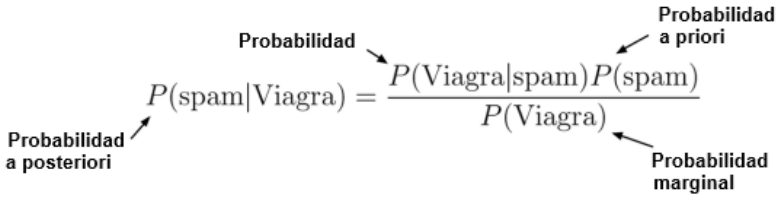

Figura 4.5: El teorema de Bayes actúa sobre los correos electrónicos recibidos anteriormente.

Para calcular los componentes del teorema de Bayes, resulta útil construir una **tabla de frecuencias** (que se muestra a la izquierda en las tablas que siguen) que registre la cantidad de veces que apareció Viagra en mensajes de spam y de ham. Al igual que en una tabulación cruzada de dos vías, una dimensión de la tabla indica los niveles de la variable de clase (spam o ham), mientras que la otra dimensión indica los niveles de las características (Viagra: sí o no). Las celdas indican entonces la cantidad de instancias que tienen la combinación especificada del valor de clase y el valor de característica.

La tabla de frecuencias se puede utilizar entonces para construir una **tabla de probabilidad**, como se muestra a la derecha en las tablas siguientes. Las filas de la tabla de probabilidad indican las probabilidades condicionales de Viagra (sí/no), dado que un correo electrónico fue spam o ham.

	Viagra		
Frequency	**Yes**	**No**	**Total**
spam	4	16	20
ham	1	79	80
Total	5	95	100

	Viagra		
Likelihood	**Yes**	**No**	**Total**
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

Figura 4.6: Las tablas de frecuencias y de probabilidad son la base para calcular la probabilidad posterior de spam.

La tabla de probabilidad revela que *P*(Viagra=Sí|spam) = 4/20 = 0.20, lo que indica que hay un 20 por ciento de probabilidad de que un mensaje contenga el término Viagra dado que el mensaje es spam.

Además, dado que $P(A \cap B) = P(B|A) * P(A)$, podemos calcular $P(\text{spam} \cap \text{Viagra})$ como $P(\text{Viagra}|\text{spam}) * P(\text{spam}) = (4/20) * (20/100) = 0.04$. El mismo resultado se puede encontrar en la tabla de frecuencia, que indica que 4 de cada 100 mensajes eran spam y contenían el término Viagra. De cualquier manera, esto es cuatro veces mayor que la estimación anterior de 0.01 que calculamos como $P(A \cap B) = P(A)*P(B)$ bajo el supuesto falso de independencia. Esto, por supuesto, ilustra la importancia del teorema de Bayes para estimar la probabilidad conjunta.

Para calcular la probabilidad posterior, $P(\text{spam}|\text{Viagra})$, simplemente tomamos $P(\text{Viagra}|\text{spam}) * P(\text{spam}) / P(\text{Viagra})$, o $(4/20) * (20/100) / (5/100) = 0.80$. Por lo tanto, la probabilidad de que un mensaje sea spam es del 80 por ciento dado que contiene la palabra Viagra. A la luz de este hallazgo, cualquier mensaje que contenga este término probablemente debería filtrarse.

Así es como funcionan los filtros de spam comerciales, aunque consideran una cantidad mucho mayor de palabras simultáneamente al calcular las tablas de frecuencia y probabilidad. En la siguiente sección, veremos cómo se puede adaptar este método para dar cabida a casos en los que se involucran características adicionales.

El algoritmo Naïve Bayes

El algoritmo Naïve Bayes define un método simple para aplicar el teorema de Bayes a los problemas de clasificación. Aunque no es el único método de aprendizaje automático que utiliza métodos bayesianos, es el más común. Su popularidad aumentó gracias a sus éxitos en la clasificación de textos, donde en su día fue el estándar de facto. Las fortalezas y debilidades de este algoritmo son las siguientes:

Fortalezas	Debilidades
• Sencillo, rápido y muy eficaz • Funciona bien con datos ruidosos y faltantes y con una gran cantidad de características • Requiere relativamente pocos ejemplos para el entrenamiento • Es fácil obtener la probabilidad estimada de una predicción	• Se basa en una suposición, a menudo errónea, de características igualmente importantes e independientes • No es ideal para conjuntos de datos con muchas características numéricas • Las probabilidades estimadas son menos fiables que las clases predichas

El algoritmo Naïve Bayes se denomina así porque hace algunas suposiciones denominadas "ingenuas" sobre los datos. En particular, Naïve Bayes supone que todas las características del conjunto de datos son **igualmente importantes e independientes**. Estas suposiciones rara vez son ciertas en la mayoría de las aplicaciones del mundo real.

Por ejemplo, al intentar identificar el correo no deseado mediante el control de los mensajes de correo electrónico, es casi seguro que algunas características serán más importantes que otras. Por ejemplo, el remitente del correo electrónico puede ser un indicador más importante de correo no deseado que el texto del mensaje.

Además, las palabras del cuerpo del mensaje no son independientes entre sí, ya que la aparición de algunas palabras es un buen indicio de que es probable que aparezcan también otras palabras. Un mensaje con la palabra Viagra probablemente también contenga la palabra *prescripción* o *medicamentos*.

Sin embargo, en la mayoría de los casos, incluso cuando se violan estos supuestos, Naïve Bayes sigue funcionando sorprendentemente bien. Esto es así incluso en circunstancias en las que se encuentran fuertes dependencias entre las características.

Debido a la versatilidad y precisión del algoritmo en muchos tipos de condiciones, en particular con conjuntos de datos de entrenamiento más pequeños, Naïve Bayes suele ser un candidato de referencia razonable para las tareas de aprendizaje de clasificación.

La razón exacta por la que Naïve Bayes funciona bien a pesar de sus supuestos erróneos ha sido objeto de mucha especulación. Una explicación es que no es importante obtener una estimación precisa de la probabilidad siempre que las predicciones sean precisas.
Por ejemplo, si un filtro de spam identifica correctamente el spam, ¿importa si la probabilidad prevista de spam fue del 51 por ciento o del 99 por ciento? Para una discusión de este tema, consulta el artículo On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Domingos, P. y Pazzani, M., Machine Learning, 1997, Vol. 29, pp. 103-130.

Clasificación con Naïve Bayes

Ampliemos nuestro filtro de spam agregando algunos términos adicionales para monitorear además del término Viagra: money (dinero), groceries (comestibles) y unsubscribe (cancelar suscripción). El aprendiz de Naïve Bayes se entrena construyendo una tabla de probabilidad para la aparición de estas cuatro palabras (etiquetadas W^1, W^2, W^3 y W^4), como se muestra en el siguiente diagrama para 100 correos electrónicos:

	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		
Likelihood	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Total**
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

Figura 4.7: Una tabla expandida agrega probabilidades para términos adicionales en mensajes de spam y ham.

A medida que se reciben nuevos mensajes, necesitamos calcular la probabilidad posterior para determinar si es más probable que sean spam o ham, dada la probabilidad de que las palabras se encuentren en el texto del mensaje. Por ejemplo, supongamos que un mensaje contiene los términos Viagra y cancelar suscripción, pero no contiene dinero ni alimentos.

Utilizando el teorema de Bayes, podemos definir el problema como se muestra en la siguiente fórmula. Esta calcula la probabilidad de que un mensaje sea spam dado que Viagra = Sí, Dinero = No, Comestibles = No y Cancelar suscripción = Sí:

$$P(\text{spam}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4|\text{spam})P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

Por dos razones, esta fórmula es difícil de resolver desde el punto de vista computacional. En primer lugar, a medida que se añaden características adicionales, se necesitan enormes cantidades de memoria para almacenar las probabilidades de todos los posibles eventos que se intersectan. Imagina la complejidad de un diagrama de Venn para los eventos de cuatro palabras, y mucho menos para cientos o más. En segundo lugar, muchas de estas posibles intersecciones nunca se habrán observado en datos anteriores, lo que llevaría a una probabilidad conjunta de cero y a problemas que se aclararán más adelante.

El cálculo se vuelve más razonable si aprovechamos el hecho de que el método Naïve Bayes hace la suposición ingenua de independencia entre los eventos. En concreto, supone la **independencia condicional de clase**, lo que significa que los eventos son independientes siempre que estén condicionados al mismo valor de clase. La suposición de independencia condicional nos permite utilizar la regla de probabilidad para eventos independientes, que establece que $P(A \cap B) = P(A) * P(B)$. Esto simplifica el numerador al permitirnos multiplicar las probabilidades condicionales individuales en lugar de calcular una probabilidad conjunta condicional compleja.

Por último, como el denominador no depende de la clase de destino (spam o ham), se trata como un valor constante y se puede ignorar por el momento. Esto significa que la probabilidad condicional de spam se puede expresar como:

$$P(\text{spam}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto$$

$$P(W_1|spam)P(\neg W_2|spam)P(\neg W_3|spam)P(W_4|spam)P(spam)$$

Y la probabilidad de que el mensaje sea ham se puede expresar como:

$$P(\text{ham}|W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto$$
$$P(W_1|ham)P(\neg W_2|ham)P(\neg W_3|ham)P(W_4|ham)P(ham)$$

Observa que el símbolo igual se ha reemplazado por el símbolo proporcional (similar a un "8" abierto y de lado) para indicar el hecho de que se ha omitido el denominador.

Usando los valores en la tabla de probabilidad, podemos comenzar a completar números en estas ecuaciones. La probabilidad general de spam es entonces:

(4 / 20) * (10 / 20) * (20 / 20) * (12 / 20) * (20 / 100) = 0.012

Mientras que la probabilidad de ham es:

(1 / 80) * (66 / 80) * (71 / 80) * (23 / 80) * (80 / 100) = 0.002

Como 0.012 / 0.002 = 6, podemos decir que este mensaje tiene 6 veces más probabilidades de ser spam que ham. Sin embargo, para convertir estos números en probabilidades, necesitamos un último paso para reintroducir el denominador que se ha excluido. Básicamente, debemos volver a escalar la probabilidad de cada resultado dividiéndola por la probabilidad total de todos los resultados posibles.

De esta manera, la probabilidad de spam es igual a la probabilidad de que el mensaje sea spam dividida por la probabilidad de que el mensaje sea spam o ham:

0.012 / (0.012 + 0.002) = 0.857

De manera similar, la probabilidad de ham es igual a la probabilidad de que el mensaje sea ham dividida por la probabilidad de que el mensaje sea spam o ham:

0.002 / (0.012 + 0.002) = 0.143

Dado el patrón de palabras encontrado en este mensaje, esperamos que el mensaje sea spam con una probabilidad del 85.7 por ciento y ham con una probabilidad del 14.3 por ciento. Debido a que estos son eventos mutuamente excluyentes y exhaustivos, las probabilidades suman 1.

El algoritmo de clasificación Naïve Bayes utilizado en el ejemplo anterior se puede resumir con la siguiente fórmula. La probabilidad de nivel *L* para la clase *C*, dada la evidencia proporcionada por las características F_1 a F_n, es igual al producto de las probabilidades de cada evidencia condicionada al nivel de clase, la probabilidad previa del nivel de clase y un factor de escala 1/*Z*, que convierte los valores de probabilidad en probabilidades. Esto se formula como:

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

Aunque esta ecuación parece intimidante, como lo ilustra el ejemplo del filtrado de spam, la serie de pasos es bastante sencilla. Comienza por construir una tabla de frecuencias, utilízala para construir una tabla de probabilidad y multiplica las probabilidades condicionales con el supuesto ingenuo de independencia.
Finalmente, divida por la probabilidad total para transformar la probabilidad de cada clase en una probabilidad. Después de intentar este cálculo unas cuantas veces a mano, se convertirá en algo natural.

El estimador de Laplace

Antes de emplear el algoritmo Naïve Bayes en problemas más complejos, hay algunos matices que considerar.
Supongamos que recibimos otro mensaje, esta vez con los cuatro términos: Viagra, groceries, money y unsubscribe. Usando el algoritmo Naïve Bayes como antes, podemos calcular la probabilidad de spam como:

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

Y la probabilidad de ham es:

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

Por lo tanto, la probabilidad de spam es:

0 / (0 + 0.00005) = 0

Y la probabilidad de jamón es:

0.00005 / (0 + 0. 0.00005) = 1

Estos resultados sugieren que el mensaje es spam con una probabilidad del 0 por ciento y ham con una probabilidad del 100 por ciento. ¿Tiene sentido esta predicción? Probablemente no. El mensaje contiene varias palabras que suelen asociarse con el spam, incluida Viagra, que rara vez se utiliza en mensajes legítimos.
Por lo tanto, es muy probable que el mensaje se haya clasificado incorrectamente.

Este problema surge si un evento nunca ocurre para uno o más niveles de la clase y, por lo tanto, las probabilidades resultantes son cero. Por ejemplo, el término comestibles (*groceries*) nunca había aparecido anteriormente en un mensaje de spam. En consecuencia, *P*(groceries|spam) = 0%.

Ahora bien, como las probabilidades en la fórmula de Naïve Bayes se multiplican en una cadena, este valor de cero por ciento hace que la probabilidad posterior de spam sea cero, lo que le da a la palabra groceries (comestibles) la capacidad de anular y anular de manera efectiva todas las demás pruebas. Incluso si se esperaba abrumadoramente que el correo electrónico fuera spam, la ausencia de la palabra comestibles en el spam siempre vetará las demás pruebas y dará como resultado que la probabilidad de spam sea cero.

Una solución a este problema implica utilizar algo llamado **estimador de Laplace**, que lleva el nombre del matemático francés Pierre-Simon Laplace. El estimador de Laplace suma un número pequeño a cada uno de los recuentos en la tabla de frecuencias, lo que garantiza que cada característica tenga una probabilidad distinta de cero de ocurrir con cada clase. Normalmente, el estimador de Laplace se establece en uno, lo que garantiza que cada combinación de clase y característica se encuentre en los datos al menos una vez.

El estimador de Laplace se puede establecer en cualquier valor y no necesariamente tiene que ser el mismo para cada una de las características. Si fuera un bayesiano devoto, podría usar un estimador de Laplace para reflejar una probabilidad previa a priori de cómo una característica se relaciona con una clase. En la práctica, dado un conjunto de datos de entrenamiento lo suficientemente grande, esto es excesivo. En consecuencia, casi siempre se usa el valor de uno.

Veamos cómo afecta esto a nuestra predicción para este mensaje. Usando un valor de Laplace de 1, agregamos 1 a cada numerador en la función de probabilidad. Luego, necesitamos agregar 4 a cada denominador de probabilidad condicional para compensar los 4 valores adicionales agregados al numerador. Por lo tanto, la probabilidad de spam es:

$$(5 / 24) * (11 / 24) * (1 / 24) * (13 / 24) * (20 / 100) = 0.0004$$

Y la probabilidad de ham es:

$$(2 / 84) * (15 / 84) * (9 / 84) * (24 / 84) * (80 / 100) = 0.0001$$

Al calcular 0.0004 / (0.0004 + 0.0001), encontramos que la probabilidad de spam es del 80 por ciento y, por lo tanto, la probabilidad de ham es de aproximadamente el 20 por ciento.

Este es un resultado más plausible que el *P*(spam) = 0 calculado cuando el término comestibles (*groceries*) por sí solo determina el resultado.

Aunque el estimador de Laplace se agregó al numerador y al denominador de las probabilidades, no se agregó a las probabilidades anteriores (los valores de 20/100 y 80/100). Esto se debe a que nuestra mejor estimación de la probabilidad general de spam y ham sigue siendo del 20 % y el 80 % respectivamente, dado lo que se observó en los datos.

Uso de características numéricas con Naïve Bayes

Naïve Bayes utiliza tablas de frecuencia para aprender los datos, lo que significa que cada característica debe ser categórica para crear las combinaciones de valores de clase y característica que componen la matriz. Dado que las características numéricas no tienen categorías de valores, el algoritmo anterior no funciona directamente con datos numéricos. Sin embargo, existen formas de abordar este problema.

Una solución fácil y eficaz es **discretizar** las características numéricas, lo que simplemente significa que los números se colocan en categorías conocidas como contenedores (**bins**). Por este motivo, la discretización también se denomina a veces **binning**. Este método funciona mejor cuando hay grandes cantidades de datos de entrenamiento.

Existen varias formas diferentes de discretizar una característica numérica. Quizás la más común sea explorar los datos en busca de categorías naturales o **puntos de corte** en la distribución.
Por ejemplo, supongamos que agregó una característica al conjunto de datos de correo no deseado que registra la hora del día o de la noche en que se envió el correo electrónico, desde las 0 hasta las 24 horas después de la medianoche. Representados mediante un histograma, los datos de tiempo podrían verse como el siguiente diagrama:

En las primeras horas de la mañana, la frecuencia de los mensajes es baja. La actividad aumenta durante el horario comercial y disminuye por la noche. Esto crea cuatro compartimentos naturales de actividad, divididos por las líneas discontinuas. Estos indican

lugares donde los datos numéricos podrían dividirse en niveles para crear una nueva característica categórica, que luego podría usarse con Naive Bayes.

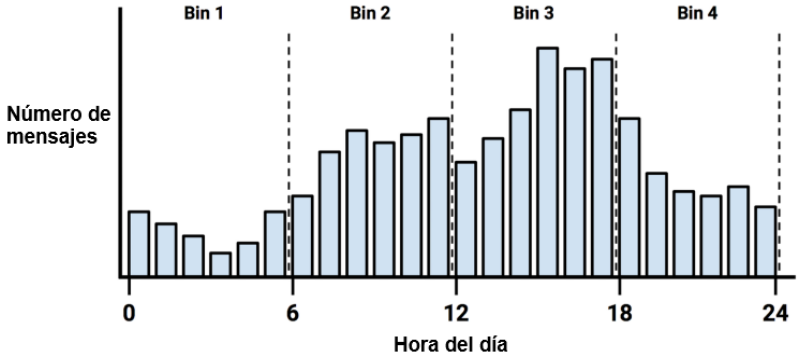

Figura 4.8: Un histograma que visualiza la distribución de la hora en que se recibieron los correos electrónicos.

La elección de cuatro compartimentos se basó en la distribución natural de los datos y en una intuición sobre cómo podría cambiar la proporción de spam a lo largo del día. Podríamos esperar que los spammers operen en las últimas horas de la noche, o pueden hacerlo durante el día, cuando es más probable que la gente revise su correo electrónico. Dicho esto, para capturar estas tendencias, podríamos haber usado fácilmente tres compartimentos o doce.

Si no hay puntos de corte obvios, una opción es discretizar la característica usando cuantiles. Puedes dividir los datos en tres compartimentos con terciles, cuatro compartimentos con cuartiles o cinco compartimentos con quintiles.

Una cosa que debes tener en cuenta es que la discretización de una característica numérica siempre da como resultado una reducción de la información, ya que la granularidad original de la característica se reduce a un número menor de categorías. Es importante lograr un equilibrio. Si hay muy pocos intervalos, pueden quedar ocultas tendencias importantes. Si hay demasiados intervalos, pueden aparecer recuentos pequeños en la tabla de frecuencias de Naïve Bayes, lo que puede aumentar la sensibilidad del algoritmo a los datos ruidosos.