Facsímil 12

Lógica e incertidumbre

Capítulo 02

Razonamiento bajo incertidumbre: Bayes y redes bayesianas

686f6c61 · 2026

Capítulo 02: Razonamiento bajo incertidumbre: Bayes y redes bayesianas

Entrando en el tema

En el capítulo anterior, una máquina razonaba con verdades y falsedades. Una proposición era verdadera o falsa; una regla se disparaba o no se disparaba; una contradicción cerraba el caso. Esa disciplina es preciosa cuando el mundo se deja describir con afirmaciones tajantes. El problema es que casi nunca lo hace.

Piensa en una frase tan corriente como “la prueba ha dado positiva”. La lógica del capítulo 1 quiere convertirla en “el paciente está enfermo: verdadero”. Pero esa traducción es falsa, y de una forma peligrosa. Un positivo no implica enfermedad: la sugiere, la hace más probable, mueve la creencia en una dirección. Cuánto la mueve depende de cosas que la lógica binaria no sabe representar: cómo de frecuente es la enfermedad, cómo de fiable es la prueba, cuántas falsas alarmas produce. Si fuerzas el mundo a entrar en verdadero/falso, pierdes justo la información que necesitas para decidir bien.

Este capítulo cambia la herramienta. En lugar de tratar cada hecho como verdadero o falso, le asignamos un número entre 0 y 1 que mide nuestro **grado de creencia**: 0 es “imposible”, 1 es “seguro”, y todo lo interesante vive en medio. Esa idea, la probabilidad entendida como creencia razonada y no solo como frecuencia de un experimento, es la que permite a una máquina razonar cuando los hechos son inciertos.[1] La lógica no se tira a la basura: se generaliza. La verdad y la falsedad pasan a ser los dos extremos de una escala continua.

Y no es un parche improvisado. La probabilidad tiene reglas tan estrictas como la lógica, y con dos herramientas (el teorema de Bayes para actualizar creencias y las redes bayesianas para organizarlas) se convierte en una maquinaria de razonamiento completa. Es la misma maquinaria que, por debajo, sostiene buena parte del aprendizaje automático del Facsímil 1 y la calibración de modelos del Facsímil 7. Empecemos por las reglas.

El grado de creencia tiene reglas

Si vamos a representar la incertidumbre con números, esos números no pueden ser arbitrarios. La teoría de la probabilidad descansa sobre tres axiomas, propuestos por Kolmogórov, que cualquier asignación de creencias coherente debe cumplir. Son pocos y sencillos:

AXIOMA	ENUNCIADO	LECTURA COTIDIANA
No negatividad	$P(A) \geq 0$	Ninguna creencia puede ser “menos que imposible”.
Normalización	$P(\Omega) = 1$	Algo del conjunto de resultados posibles ocurre con certeza.
Aditividad	Si A y B son incompatibles, $P(A \vee B) = P(A) + P(B)$	Las probabilidades de sucesos que no pueden coincidir se suman.

De ahí sale todo lo demás. Por ejemplo, como un suceso A y su negación $\neg A$ son incompatibles y juntos cubren todo lo posible, se deduce $P(A) + P(\neg A) = 1$. Esa pequeña igualdad ya es más de lo que la lógica binaria te daba: te dice que creer al 70 % en algo es exactamente creer al 30 % en lo contrario.

La pieza que de verdad enciende el razonamiento es la **probabilidad condicional**. Mide cuánto crees en A una vez que sabes que B ha ocurrido:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}, \qquad P(B) > 0$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(A \mid B)$	Creencia en A sabiendo B.	Probabilidad de estar enfermo dado un positivo.
$P(A \wedge B)$	Probabilidad de que ocurran ambos.	Estar enfermo y dar positivo a la vez.
$P(B)$	Probabilidad de la condición observada.	Probabilidad de dar positivo, sin más datos.

En palabras: la probabilidad de A sabiendo B es la fracción de los casos en que ocurre B que, además, cumplen A. Es lo mismo que preguntar «de la gente que ha dado positivo, ¿qué parte está enferma?».

La fórmula dice algo muy intuitivo: condicionar es **reducir el universo**. Antes mirabas todos los casos posibles; al saber que B ocurre, te quedas solo con la franja en la que B es cierto y vuelves a medir dentro de ella qué proporción cumple también A. Dividir por $P(B)$ es justamente reescalar esa franja para que vuelva a sumar 1.

Reordenando la definición aparece la **regla del producto**, que descompone la probabilidad conjunta en pasos encadenados:

$$P(A \wedge B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(A \wedge B)$	Probabilidad de que ocurran ambos sucesos.	Que llueva y que hayas cogido el paraguas.
$P(A \mid B)\,P(B)$	Descomposición fijando B primero.	Probabilidad de paraguas si llueve, por la de que llueva.
$P(B \mid A)\,P(A)$	La misma cantidad fijando A primero.	Probabilidad de lluvia si hay paraguas, por la de paraguas.

En palabras: la probabilidad de que dos cosas pasen a la vez es la de una de ellas multiplicada por la de la otra una vez supuesta la primera; y da igual cuál tomes como punto de partida, el resultado es el mismo.

Y de la regla del producto, aplicada a muchas variables, sale la **regla de la cadena**, que escribiremos en un momento porque es la columna vertebral de las redes bayesianas:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$X_1, \ldots, X_n$	Las variables, en un orden cualquiera que fijas tú.	Las palabras de una frase, de izquierda a derecha.
$\prod_{i=1}^{n}$	Producto de los n factores.	Multiplicar los términos uno tras otro.
$P(X_i \mid X_1, \ldots, X_{i-1})$	Probabilidad de cada variable dadas todas las anteriores.	La de cada palabra dado lo ya escrito.

En palabras: la probabilidad conjunta de muchas variables se monta en cadena: eliges un orden y vas multiplicando la probabilidad de cada una condicionada a todas las que ya han salido.

Falta una operación más, la **marginalización**, que hace lo contrario de condicionar: en lugar de fijar una variable, la elimina sumando sobre todos sus valores. Si conoces la conjunta de A y B pero solo te interesa A, barres B:

$$P(A) = \sum_{b} P(A \wedge B{=}b) = \sum_{b} P(A \mid B{=}b)\,P(B{=}b)$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(A)$	Probabilidad de A sola, sin condicionar a nada.	Probabilidad de dar positivo, mires a quien mires.
$\sum_b$	Suma sobre todos los valores posibles de B.	Sumar el caso «enfermo» y el caso «sano».
$P(A \mid B{=}b)\,P(B{=}b)$	Lo que aporta cada valor de B a A.	Positivos de los enfermos más positivos de los sanos.

En palabras: para quitarte de encima una variable que no te interesa, recorres todos sus valores posibles y sumas la probabilidad de A en cada caso, pesando cada uno por lo probable que era ese valor.

Esta suma se llama también **regla de la probabilidad total**, y es la que usaremos enseguida para calcular la probabilidad de "dar positivo" sumando las dos formas de llegar a un positivo: estando enfermo y no estándolo. Con condicional, producto y marginalización ya tenemos toda el álgebra que necesita el teorema de Bayes.

El teorema de Bayes: dar la vuelta a la pregunta

A menudo sabemos calcular una dirección de la probabilidad condicional, pero nos interesa la contraria. Un médico sabe, por estudios clínicos, con qué frecuencia una persona enferma da positivo en una prueba: eso es $P(\text{positivo} \mid \text{enfermo})$. Pero el paciente que tiene el positivo en la mano no quiere saber eso. Quiere saber $P(\text{enfermo} \mid \text{positivo})$: la pregunta al revés. El teorema de Bayes es exactamente el puente entre ambas direcciones.[2]

Se deduce en una línea igualando las dos formas de la regla del producto, $P(H \wedge E) = P(E \mid H)P(H) = P(H \mid E)P(E)$, y despejando:

$$P(H \mid E) = \frac{P(E \mid H)\,P(H)}{P(E)}$$

Cada pieza tiene un nombre y un papel, y conviene memorizarlos porque organizan toda la forma de pensar:

PIEZA	NOMBRE	QUÉ REPRESENTA
$P(H)$	Probabilidad previa	Lo que creías sobre la hipótesis H antes de ver nada.
$P(E \mid H)$	Verosimilitud	Cómo de esperable es la evidencia E si H fuese cierta.
$P(E)$	Evidencia	Cómo de esperable es E en total, sea cual sea la causa.
$P(H \mid E)$	Probabilidad posterior	Lo que crees sobre H después de incorporar E.

En palabras: tu nueva creencia en la hipótesis es la que ya tenías, multiplicada por cuánto encaja la evidencia con ella y rebajada por cuánto era de esperar esa evidencia en cualquier caso. Si la evidencia era esperable hasta sin la hipótesis, mueve poco; si solo la hipótesis la explicaba, mueve mucho.

El término del denominador, la evidencia, casi nunca se da directamente: se calcula marginalizando sobre las hipótesis posibles. Si H y $\neg H$ agotan los casos:

$$P(E) = P(E \mid H)\,P(H) + P(E \mid \neg H)\,P(\neg H)$$

En palabras: la probabilidad total de ver la evidencia es la suma de las dos formas de llegar a ella: que la hipótesis sea cierta y produzca la evidencia, o que sea falsa y aun así la evidencia aparezca (la falsa alarma).

Leído despacio, el teorema cuenta una historia: partes de una creencia (previa), la multiplicas por cuánto encaja la evidencia con esa creencia (verosimilitud) y la divides por cuánto era de esperar la evidencia en general (evidencia) para obtener tu creencia actualizada (posterior). La probabilidad posterior de hoy es la previa de mañana: cada nuevo dato vuelve a entrar en la misma maquinaria. Esa es la esencia del razonamiento bajo incertidumbre.

El ejemplo que casi todo el mundo falla

Vamos con el caso clásico, porque su resultado es tan contraintuitivo que conviene calcularlo a mano y no fiarse del instinto. Una enfermedad afecta al **1 %** de la población. Existe una prueba con estas características:

- Si una persona está enferma, la prueba da positivo el **99 %** de las veces (sensibilidad alta).
- Si una persona está sana, la prueba da positivo igualmente el **5 %** de las veces (falsos positivos).

Una persona se hace la prueba y da positivo. Pregunta directa: ¿qué probabilidad tiene de estar realmente enferma? El instinto de casi todo el mundo, incluidos muchos médicos, dispara una cifra alta, en torno al 95 %, confundiendo la fiabilidad de la prueba con la probabilidad de enfermedad. Hagamos las cuentas de verdad.

Primero, traduzcamos los datos al lenguaje de Bayes. Sea H = “estar enfermo” y E = “dar positivo”:

DATO DEL ENUNCIADO	NOTACIÓN	VALOR
Prevalencia (previa)	$P(H)$	$0{,}01$
Sano (probabilidad previa complementaria)	$P(\neg H)$	$0{,}99$
Sensibilidad (verosimilitud)	$P(E \mid H)$	$0{,}99$
Falsos positivos	$P(E \mid \neg H)$	$0{,}05$

Calculamos la evidencia, es decir, la probabilidad total de dar positivo, sumando las dos maneras de obtener un positivo:

$$P(E) = P(E \mid H)\,P(H) + P(E \mid \neg H)\,P(\neg H)$$

$$P(E) = (0{,}99)(0{,}01) + (0{,}05)(0{,}99) = 0{,}0099 + 0{,}0495 = 0{,}0594$$

Ahora aplicamos Bayes:

$$P(H \mid E) = \frac{P(E \mid H)\,P(H)}{P(E)} = \frac{0{,}0099}{0{,}0594} = 0{,}1667$$

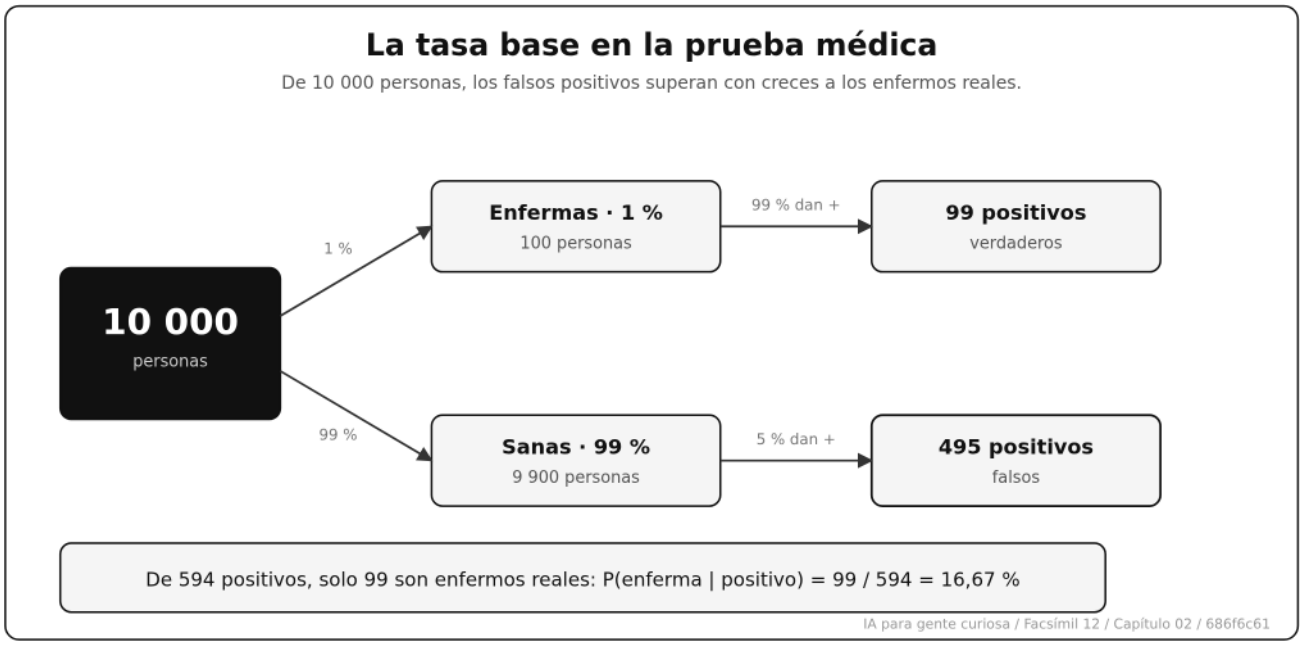
El resultado es **16,67 %**, no el 95 % que sugería el instinto. Una persona con un positivo en esta prueba tiene, todavía, una probabilidad de cinco sextos de estar sana. ¿Por qué? Porque la enfermedad es tan rara que los falsos positivos, ese 5 % aplicado al enorme 99 % de personas sanas, generan muchísimos más positivos que los verdaderos enfermos. Mira las cifras crudas sobre 10 000 personas:

GRUPO	PERSONAS	DAN POSITIVO	CÓMO SE OBTIENE
Enfermas	100	99	$10\,000 \times 0{,}01 \times 0{,}99$
Sanas	9 900	495	$10\,000 \times 0{,}99 \times 0{,}05$
Total positivos		**594**	99 verdaderos + 495 falsos

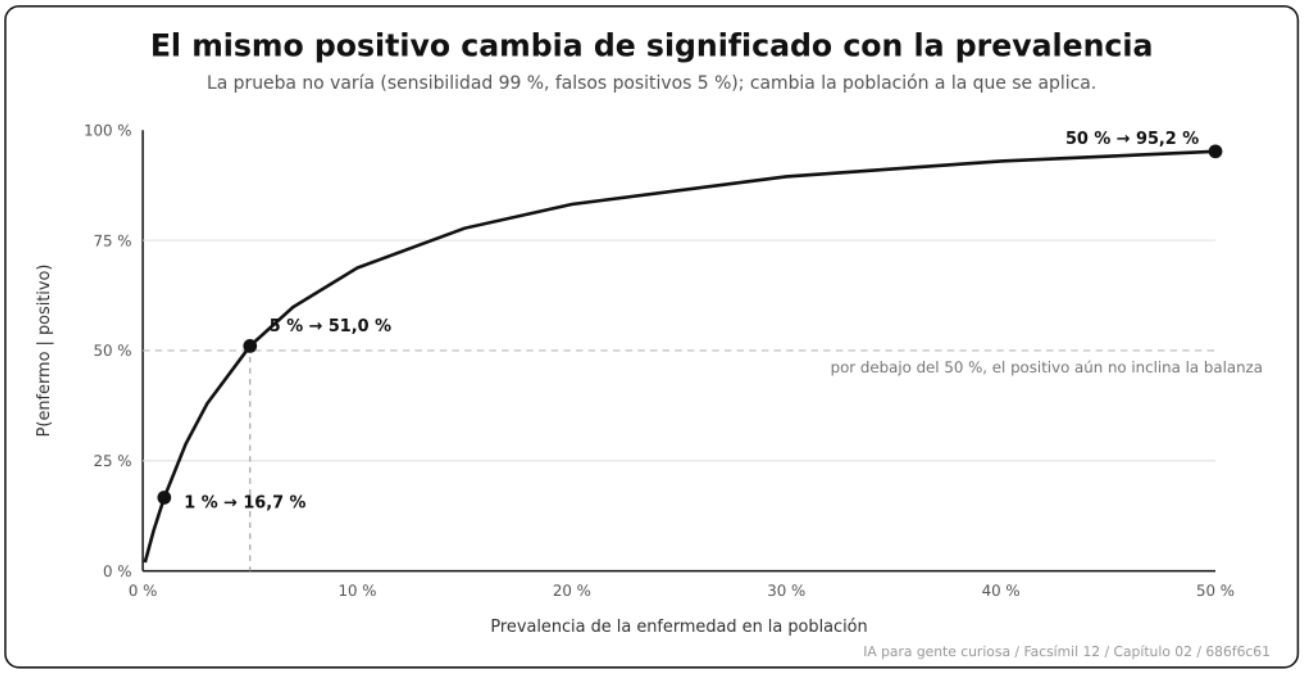
De los 594 positivos, solo 99 son personas realmente enfermas: $99/594 = 0{,}1667$, el mismo 16,67 %. Esa es la lección de la **tasa base**: ignorar lo rara que es una hipótesis previa lleva a sobreinterpretar la evidencia. Es el mismo error que cometería un detector de fraude que

confunde su precision con la probabilidad real de fraude, el problema de clases desbalanceadas que vimos en el Facsímil 1, capítulo 11. La probabilidad previa no es un adorno: es la mitad del cálculo.

El dibujo deja a la vista por qué el instinto se equivoca: el grupo de sanos es tan grande que su pequeño 5 % de falsos positivos arroja muchísimas más alarmas que el 99 % de aciertos sobre el minúsculo grupo de enfermos.



Lo más revelador es que **el mismo positivo significa cosas distintas según la prevalencia**. La prueba no cambia (sigue acertando el 99 % en enfermos y fallando el 5 % en sanos), pero su veredicto se reinterpreta según lo frecuente que sea la enfermedad en la población a la que se aplica. Si barremos la prevalencia de 0 a 50 % y recalculamos Bayes en cada punto, sale esta curva:



La curva sube deprisa al principio y luego se aplana: con la enfermedad rara (1 %) el positivo apenas alcanza el 16,7 %, no cruza el umbral del 50 % hasta una prevalencia cercana al 5 % y solo se acerca a la certeza cuando la enfermedad es muy común. Por eso una misma prueba sirve para confirmar en una población de riesgo y produce sobre todo falsas alarmas en un cribado masivo: el instrumento es el mismo, pero la tasa base lo recoloca. Es exactamente lo que un médico hace al decidir a quién conviene hacerle la prueba.

Naive Bayes: clasificar con una hipótesis “ingenua”

El teorema de Bayes con una sola evidencia ya es útil. Pero los problemas reales tienen muchas evidencias a la vez. Un correo no trae un solo indicio de spam: trae decenas de palabras, cada una un pequeño dato. Un ticket de soporte trae un texto entero. ¿Cómo aplicamos Bayes cuando la evidencia es un vector $E = (w_1, w_2, \ldots, w_n)$ de muchos rasgos?

El teorema sigue valiendo, con la clase c como hipótesis:

$$P(c \mid w_1, \ldots, w_n) = \frac{P(w_1, \ldots, w_n \mid c)\, P(c)}{P(w_1, \ldots, w_n)}$$

El problema está en la verosimilitud conjunta $P(w_1, \ldots, w_n \mid c)$. Estimarla bien exigiría conocer cómo se combinan todas las palabras entre sí dentro de cada clase, una tabla astronómica que jamás tendríamos datos para llenar. Aquí entra la **hipótesis ingenua** (de ahí el nombre): suponer que, **dada la clase**, los rasgos son condicionalmente independientes entre sí. Es decir, que conocer la clase “spam” hace que la presencia de “gratis” no aporte información extra sobre la presencia de “premio”. Con esa suposición, la verosimilitud conjunta se descompone en un producto:

$$P(w_1, \ldots, w_n \mid c) \approx \prod_{i=1}^{n} P(w_i \mid c)$$

En palabras: si suponemos que, una vez fijada la clase, las palabras no se influyen entre sí, la probabilidad de verlas todas juntas se vuelve el simple producto de la probabilidad de cada una por separado. Eso convierte una tabla imposible de llenar en una multiplicación de números que sí sabemos contar.

Y el clasificador elige la clase con mayor probabilidad posterior. Como el denominador $P(w_1, \ldots, w_n)$ es el mismo para todas las clases, se puede ignorar para decidir, y basta comparar el numerador:

$$\hat{c} = \arg\max_{c} \; P(c) \prod_{i=1}^{n} P(w_i \mid c)$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$\hat{c}$	Clase elegida por el clasificador.	spam o legítimo.
$P(c)$	Probabilidad previa de cada clase.	Fracción de correos que son spam.
$P(w_i \mid c)$	Verosimilitud de cada palabra en la clase.	Frecuencia de “gratis” entre los spam.
$\arg\max_c$	La clase que maximiza el producto.	La hipótesis más creíble dadas las palabras.

En palabras: el clasificador prueba cada clase posible, multiplica lo frecuente que es esa clase por lo bien que sus palabras típicas explican el mensaje, y se queda con la clase que da el número mayor. No necesita la probabilidad exacta, solo saber cuál gana.

Un clasificador de spam, calculado a mano

Veámoslo con un corpus diminuto para que se vean todos los números. Tenemos seis mensajes de entrenamiento, tres spam y tres legítimos:

CLASE	MENSAJES DE ENTRENAMIENTO
spam	“gana dinero gratis” · “dinero gratis ya” · “gratis premio dinero”
legítimo	“reunión proyecto mañana” · “envío informe proyecto” · “reunión informe mañana”

El vocabulario completo tiene 10 palabras distintas: gana, dinero, gratis, ya, premio, reunión, proyecto, mañana, envío, informe. Cada clase aporta 9 palabras contadas con repetición. Los recuentos en la clase spam son: dinero 3, gratis 3, gana 1, ya 1, premio 1. Como las clases

tienen el mismo número de mensajes, las probabilidades previas son $P(\text{spam}) = P(\text{legítimo}) = 0{,}5$.

Para estimar $P(w \mid c)$ usamos el modelo multinomial con **suavizado de Laplace**: sumamos 1 a cada recuento para que ninguna palabra tenga probabilidad cero (si no, una sola palabra nunca vista en una clase anularía todo el producto). La fórmula es:

$$P(w \mid c) = \frac{\text{recuento}(w, c) + 1}{N_c + |V|}$$

donde $N_c = 9$ es el total de palabras de la clase y $|V| = 10$ el tamaño del vocabulario, así que el denominador es $9 + 10 = 19$.

SÍMBOLO	SIGNIFICADO	EJEMPLO
$\text{recuento}(w, c)$	Veces que la palabra w aparece en la clase c.	«gratis» sale 3 veces en spam.
N_c	Total de palabras de la clase c.	9 en cada clase del ejemplo.
$	V	$

En palabras: en lugar de dividir recuento entre total, sumamos un 1 a cada palabra antes de dividir. Así ninguna palabra recibe probabilidad cero, y una sola palabra nunca vista en una clase no puede tumbar todo el producto a cero. El precio es repartir un poquito de probabilidad a lo que aún no hemos visto.

Clasifiquemos el mensaje nuevo **“dinero gratis reunión”**. Necesitamos seis verosimilitudes:

PALABRA	EN SPAM	$P(W \mid \text{SPAM})$	EN LEGÍTIMO	$P(W \mid \text{LEGITIMO})$
dinero	3	$(3{+}1)/19 = 4/19$	0	$(0{+}1)/19 = 1/19$
gratis	3	$4/19$	0	$1/19$
reunión	0	$(0{+}1)/19 = 1/19$	2	$(2{+}1)/19 = 3/19$

Multiplicamos cada probabilidad previa por sus tres verosimilitudes:

$$P(\text{spam}) \prod P(w \mid \text{spam}) = 0{,}5 \cdot \frac{4}{19} \cdot \frac{4}{19} \cdot \frac{1}{19} = 0{,}5 \cdot \frac{16}{6859} = \frac{8}{6859}$$

$$P(\text{legítimo}) \prod P(w \mid \text{legítimo}) = 0{,}5 \cdot \frac{1}{19} \cdot \frac{1}{19} \cdot \frac{3}{19} = 0{,}5 \cdot \frac{3}{6859} = \frac{1{,}5}{6859}$$

Para convertir estos dos números en probabilidades que sumen 1, normalizamos dividiendo por su suma, $9{,}5/6859$:

$$P(\text{spam} \mid \text{mensaje}) = \frac{8}{9{,}5} = 0{,}842, \qquad P(\text{legítimo} \mid \text{mensaje}) = \frac{1{,}5}{9{,}5} = 0{,}158$$

El clasificador marca el mensaje como **spam con un 84,2 % de probabilidad**. Tiene sentido: dos de sus tres palabras son típicas de spam y solo una es neutra.

Por qué funciona pese a ser "ingenuo"

La hipótesis de independencia condicional es, casi siempre, **falsa**. En un correo de spam, "dinero" y "gratis" aparecen juntas mucho más de lo que predeciría tratarlas como independientes; no son sucesos que se ignoren entre sí. Entonces, ¿por qué Naive Bayes clasifica bien tan a menudo?

La clave es que para **decidir** una clase no necesitamos que las probabilidades sean exactas, solo que el orden entre clases sea correcto. Naive Bayes estima mal las magnitudes (sus probabilidades posterior suelen ser demasiado extremas, cercanas a 0 o a 1, justo el problema de calibración del Facsímil 7), pero acierta el ganador con frecuencia sorprendente.[3] Es rápido, necesita pocos datos, se entrena en una sola pasada contando frecuencias y es difícil de batir como punto de partida (esa *baseline* del Facsímil 1, capítulo 11, que cualquier modelo más complejo debe superar para justificarse). Por eso sigue vivo en filtros de spam, clasificación de tickets y análisis de sentimiento, décadas después de inventarse.

Redes bayesianas: organizar muchas creencias

Naive Bayes hace una suposición drástica: que **todos** los rasgos son independientes dada la clase. Es eficaz, pero burdo. El mundo real tiene una estructura de dependencias más rica: unas variables influyen en otras de formas concretas, y muchas son independientes solo en ciertas condiciones. Necesitamos un lenguaje para representar exactamente esa estructura. Ese lenguaje son las **redes bayesianas**, formalizadas por Judea Pearl.[4]

Una red bayesiana es un **grafo dirigido acíclico** (un DAG, como los del capítulo de SAT y CSP, pero con otra lectura): cada nodo es una variable aleatoria, y cada flecha de X a Y expresa que X influye directamente en Y, que X es un "padre" de Y. El que sea acíclico importa: ninguna variable puede ser, dando la vuelta, causa de sí misma.

La idea central es que el grafo codifica una afirmación de independencia muy concreta: **cada variable es independiente de sus no-descendientes una vez conocidos sus padres**. Dicho de otro modo, los padres de un nodo resumen toda la influencia que el resto de la red ejerce sobre él. Eso permite reescribir la regla de la cadena de forma mucho más barata. La conjunta general necesitaría condicionar cada variable sobre todas las anteriores; con la red, basta condicionar cada variable sobre sus padres:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{padres}(X_i))$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
$P(X_1, \ldots, X_n)$	Probabilidad conjunta de todas las variables.	Que coincidan robo, alarma y las dos llamadas.
$\prod_{i=1}^{n}$	Producto sobre todos los nodos de la red.	Una entrada de CPT por nodo.
$\text{padres}(X_i)$	Las variables con flecha directa hacia X_i.	Robo y Terremoto son los padres de Alarma.
$P(X_i \mid \text{padres}(X_i))$	La tabla (CPT) del nodo X_i.	Probabilidad de que suene la alarma según robo y terremoto.

En palabras: la probabilidad de un escenario completo es el producto de la probabilidad de cada variable mirando solo a sus padres, sin arrastrar el resto de la red. El grafo es justamente el mapa de qué se puede ignorar al calcular cada factor.

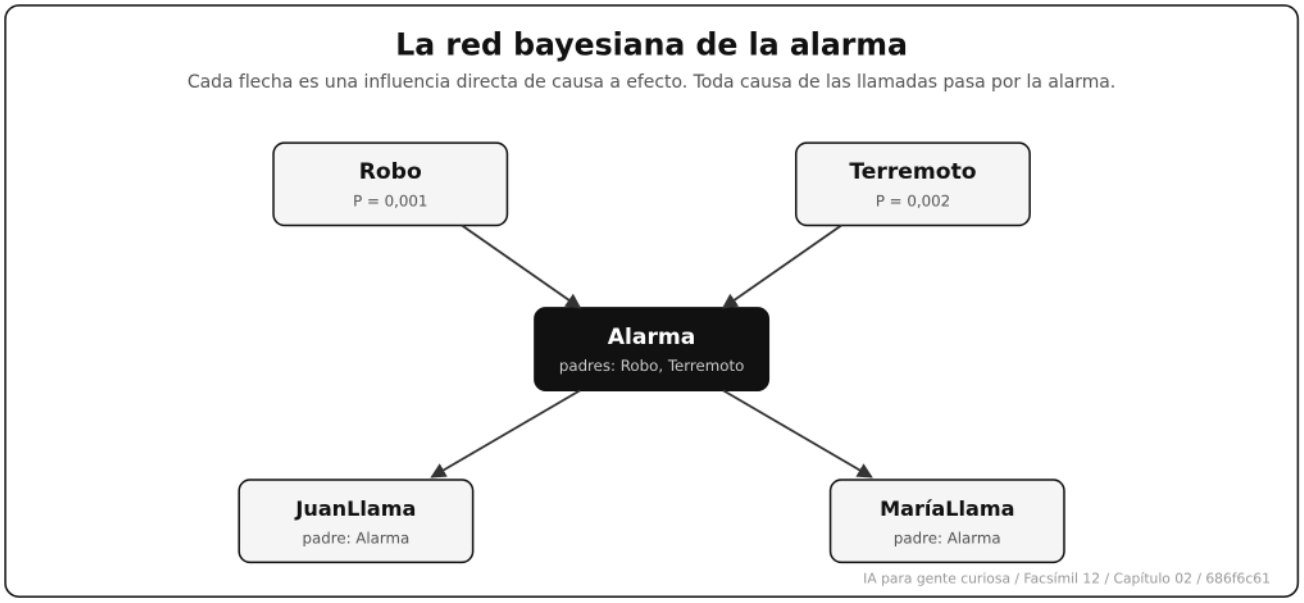
Esta **factorización** es la razón de ser de las redes bayesianas. La distribución conjunta de n variables booleanas tiene $2^n - 1$ números independientes, una cifra que explota enseguida. La red la sustituye por una tabla pequeña por nodo, la **tabla de probabilidad condicional** o CPT, que da la probabilidad de la variable para cada combinación de valores de sus padres. Si los padres son pocos, las tablas son pequeñas, y el producto reconstruye la conjunta entera sin necesidad de almacenarla.

El ejemplo canónico: la alarma de Pearl

El ejemplo de referencia, debido a Pearl, es una alarma antirrobo. Vives en una casa con una alarma que puede dispararse por un robo, pero también por un pequeño terremoto. Tienes dos vecinos, Juan y María, que te llaman al móvil si oyen la alarma, aunque no siempre: Juan a veces la confunde con el teléfono y llama sin motivo, y María a veces no la oye porque pone música alta. Cinco variables booleanas, cada una verdadera o falsa:

VARIABLE	SIGNIFICADO	PADRES
Robo	Hay un robo en la casa.	(ninguno)
Terremoto	Ha habido un terremoto.	(ninguno)
Alarma	La alarma se dispara.	Robo, Terremoto
JuanLlama	Juan te llama.	Alarma
MaríaLlama	María te llama.	Alarma

La estructura del grafo dice cosas precisas. Robo y Terremoto son causas independientes que confluyen en Alarma. Y, crucialmente, JuanLlama y MaríaLlama no dependen directamente del robo ni del terremoto: dependen solo de la alarma. Si supieras con certeza que la alarma sonó, saber además que hubo un robo no cambiaría tu predicción sobre si Juan llama. Toda la influencia del robo sobre Juan pasa, obligatoriamente, por la alarma. Ese es el contenido de las flechas:



Cada nodo lleva su CPT. Robo y Terremoto, al no tener padres, solo necesitan su probabilidad previa. La alarma necesita una fila por cada combinación de sus dos padres. Las llamadas, una fila por cada valor de la alarma. Estas son las tablas, las mismas del ejemplo original de Pearl:

CPT de Robo y de Terremoto (sin padres):

VARIABLE	$P(\text{VERDADERO})$
Robo	0,001
Terremoto	0,002

CPT de Alarma, condicionada a sus dos padres:

ROBO	**TERREMOTO**	$P(\text{ALARMA} = =\text{SI})$
sí	sí	0,95
sí	no	0,94
no	sí	0,29
no	no	0,001

CPT de JuanLlama y MaríaLlama, condicionadas a la alarma:

ALARMA	$P(\text{JUAN}=\text{SI})$	$P(\text{MARIA}=\text{SI})$
sí	0,90	0,70
no	0,05	0,01

Con estas cuatro tablas, la red codifica la distribución conjunta completa de las cinco variables. Por ejemplo, la probabilidad de que no haya robo ni terremoto, suene la alarma y llamen los dos vecinos se calcula multiplicando una entrada de cada CPT:

$$P(\neg r, \neg t, a, j, m) = P(\neg r)\, P(\neg t)\, P(a \mid \neg r, \neg t)\, P(j \mid a)\, P(m \mid a)$$

$$= 0{,}999 \cdot 0{,}998 \cdot 0{,}001 \cdot 0{,}90 \cdot 0{,}70 = 0{,}000628$$

En palabras: para saber lo probable que es un escenario concreto, basta tomar de cada tabla la casilla que le corresponde —no robo, no terremoto, alarma sí, Juan sí, María sí— y multiplicarlas todas. Cada factor es una lectura directa de una CPT.

Es un evento poco probable (una falsa alarma que moviliza a los dos vecinos), y la red lo cuantifica sin esfuerzo. Treinta y un números harían falta para la conjunta a pelo; aquí han bastado las diez probabilidades de las CPT.

Cuándo dos variables son independientes: la d-separación

La estructura del grafo no solo factoriza la conjunta: permite **leer las independencias de un vistazo**, sin tocar un solo número, mediante un criterio gráfico llamado **d-separación**. La idea es seguir los caminos entre dos variables y ver si la información puede "fluir" por ellos o queda bloqueada. Hay tres patrones de conexión, y se comportan de forma distinta:

PATRÓN	FORMA	SIN OBSERVAR EL NODO CENTRAL	OBSERVANDO EL NODO CENTRAL
Cadena	$X \to Z \to Y$	la información fluye	el camino se bloquea
Bifurcación	$X \leftarrow Z \to Y$	la información fluye	el camino se bloquea
Colisión	$X \to Z \leftarrow Y$	el camino está bloqueado	el camino se abre

Las dos primeras son intuitivas: un padre común o un eslabón intermedio transmiten dependencia, y observarlo la corta. En nuestra red, JuanLlama y MaríaLlama están conectadas por la bifurcación $\text{Juan} \leftarrow \text{Alarma} \to \text{María}$. Sin saber nada de la alarma, las dos llamadas están correlacionadas: que Juan llame te hace creer más que la alarma sonó, y eso te hace esperar también la llamada de María. Pero **dada** la alarma, se vuelven independientes: si ya sabes que la alarma sonó, la llamada de Juan no añade nada sobre la de María. Eso es exactamente la independencia condicional que asumía Naive Bayes, aquí leída en el grafo.

El tercer patrón, la **colisión**, es el más sutil y el más interesante. Robo y Terremoto chocan en Alarma: $\text{Robo} \to \text{Alarma} \leftarrow \text{Terremoto}$. De partida, sin observar nada, son independientes; los robos no causan terremotos. Pero si **observas la alarma** (sabes que sonó), Robo y Terremoto pasan a estar correlacionados de forma negativa. Es el fenómeno llamado *explaining away*, descartar por explicación: si la alarma sonó y te enteras de que ha habido un terremoto, baja tu sospecha de robo, porque el terremoto ya explica la alarma. Una causa, al confirmarse, resta credibilidad a la otra. La lógica binaria no sabría representar esto; la red lo hace de forma natural.

Responder preguntas: inferencia por eliminación de variables

Tener la conjunta factorizada está muy bien, pero lo que queremos es **hacer preguntas**: dado que Juan y María han llamado, ¿qué probabilidad hay de que haya un robo? Esto es $P(\text{Robo} \mid \text{Juan}{=}\text{sí}, \text{María}{=}\text{sí})$. La consulta tiene una variable de interés (Robo), unas variables de evidencia (las dos llamadas, fijadas a "sí") y unas variables ocultas que no aparecen ni en una ni en otra (Terremoto y Alarma) y que hay que marginalizar.

El método exacto es la **eliminación de variables**: usamos Bayes para pasar a la conjunta, sumamos sobre las ocultas y normalizamos. Por Bayes,

$$P(r \mid j, m) = \frac{P(r, j, m)}{P(j, m)} \propto P(r, j, m) = \sum_{t} \sum_{a} P(r)\, P(t)\, P(a \mid r, t)\, P(j \mid a)\, P(m \mid a)$$

SÍMBOLO	SIGNIFICADO	EJEMPLO
r	La variable que preguntamos (Robo), fijada a un valor.	¿Hubo robo, sí o no?
j, m	Las variables de evidencia, ya observadas.	Juan y María han llamado.
$\sum_t \sum_a$	Suma sobre las variables ocultas.	Recorrer terremoto y alarma, sí y no.
$\propto$	«Proporcional a»: falta dividir por $P(j, m)$.	Normalizar al final para que sume 1.

En palabras: para responder la consulta reconstruimos la conjunta, sumamos sobre las variables que no nos importan (terremoto y alarma) recorriendo todos sus valores, y al final normalizamos para que las dos opciones de robo vuelvan a sumar 1. El símbolo $\propto$ avisa de que, hasta ese último paso, manejamos números sin normalizar.

La clave para que esto sea barato es **empujar las sumas hacia dentro**: cada variable solo se suma sobre los factores que la contienen, no sobre todo el producto. Empecemos definiendo el factor que combina las dos llamadas en función de la alarma, $f(a) = P(j \mid a)\, P(m \mid a)$:

ALARMA	$P(J \mid A)$	$P(M \mid A)$	$F(A)$
sí	0,90	0,70	0,630
no	0,05	0,01	0,0005

Ahora, para cada combinación de Robo y Terremoto, sumamos sobre la alarma el producto $P(a \mid r, t)\, f(a)$. Llamemos a ese resultado $g(r, t)$:

$$g(r, t) = P(a{=}\text{sí} \mid r, t) \cdot 0{,}630 + P(a{=}\text{no} \mid r, t) \cdot 0{,}0005$$

ROBO	TERREMOTO	$P(A{=}\text{SI} \mid R, T)$	$G(R, T)$
sí	sí	0,95	$0{,}95 \cdot 0{,}630 + 0{,}05 \cdot 0{,}0005 = 0{,}598525$
sí	no	0,94	$0{,}94 \cdot 0{,}630 + 0{,}06 \cdot 0{,}0005 = 0{,}592230$
no	sí	0,29	$0{,}29 \cdot 0{,}630 + 0{,}71 \cdot 0{,}0005 = 0{,}183055$
no	no	0,001	$0{,}001 \cdot 0{,}630 + 0{,}999 \cdot 0{,}0005 = 0{,}001130$

A continuación sumamos sobre Terremoto, ponderando por su probabilidad previa ($P(t{=}\text{sí}) = 0{,}002$, $P(t{=}\text{no}) = 0{,}998$):

$$\text{Robo sí:} \quad 0{,}002 \cdot 0{,}598525 + 0{,}998 \cdot 0{,}592230 = 0{,}592243$$

$$\text{Robo no:} \quad 0{,}002 \cdot 0{,}183055 + 0{,}998 \cdot 0{,}001130 = 0{,}001493$$

Finalmente multiplicamos por la probabilidad previa de Robo ($P(r{=}\text{sí}) = 0{,}001$, $P(r{=}\text{no}) = 0{,}999$) para obtener los dos números sin normalizar:

$$P(r{=}\text{sí}, j, m) = 0{,}001 \cdot 0{,}592243 = 0{,}000592$$

$$P(r{=}\text{no}, j, m) = 0{,}999 \cdot 0{,}001493 = 0{,}001492$$

Normalizamos dividiendo por la suma, que es la evidencia $P(j, m) = 0{,}000592 + 0{,}001492 = 0{,}002084$:

$$P(\text{Robo} \mid \text{Juan}, \text{María}) = \frac{0{,}000592}{0{,}002084} = 0{,}284$$

El resultado es **28,4 %**. Vuelve a ser una lección sobre la tasa base: aunque las dos llamadas son indicios fuertes de robo, el robo era tan improbable de partida (una entre mil) que, incluso con ambos vecinos llamando, lo más probable sigue siendo que **no** haya robo. La evidencia ha multiplicado tu creencia por más de 280 (de 0,001 a 0,284), pero partía de tan abajo que no llega ni a la mitad. La red ha hecho automáticamente lo que a mano nos costaría: combinar muchas creencias coherentemente.

Cuando lo exacto no basta: inferencia aproximada

La eliminación de variables es exacta, pero su coste crece con la estructura de la red. En grafos grandes y densamente conectados, las sumas intermedias se vuelven intratables: el problema de inferencia exacta es, en general, NP-difícil, el mismo muro de complejidad que

vimos con SAT.[5] Cuando eso ocurre, se renuncia a la respuesta exacta y se estima mediante **muestreo**: en lugar de sumar sobre todas las combinaciones, se simulan miles de mundos posibles según las CPT y se cuenta en qué fracción ocurre lo que preguntamos.

La forma más simple es el **muestreo directo**: recorrer la red de padres a hijos, sorteando cada variable según su CPT dado lo ya sorteado, hasta generar un "mundo" completo. Para responder a una consulta con evidencia se descartan los mundos que no encajan con lo observado (muestreo por rechazo) o se ponderan (muestreo por verosimilitud); técnicas más sofisticadas, como las cadenas de Markov de Montecarlo, exploran el espacio de forma más eficiente. La idea de fondo es la misma que reaparece, a otra escala, en los métodos modernos: cuando no puedes calcular una probabilidad, la aproximas con muchas simulaciones.[6]

En el día a día

Aunque parezcan formalismos de manual, estas ideas están funcionando constantemente a tu alrededor. El filtro de spam de tu correo es, en su versión clásica, un Naive Bayes contando palabras; los filtros modernos son más sofisticados, pero la intuición de "esta palabra hace más probable el spam" sigue ahí. Los sistemas de diagnóstico médico asistido combinan síntomas, pruebas y prevalencias exactamente con la mecánica del ejemplo de la enfermedad rara, y su mayor valor es precisamente recordar la tasa base que el instinto humano olvida.

Las redes bayesianas aparecen en diagnóstico de averías (un coche o una máquina industrial cuyos sensores apuntan a una causa raíz), en evaluación de riesgo crediticio, en sistemas de recomendación y en modelos de fiabilidad de equipos. Y cada vez que una aplicación combina varias señales débiles para tomar una decisión (un antifraude que junta ubicación, importe y hora; un detector que pondera varios indicios), por debajo hay una actualización de creencias que es Bayes, se llame así o no.

Hay además un sitio donde estas ideas viven sin que casi nadie lo mencione: dentro de los modelos de lenguaje. Un LLM es, en su núcleo, un modelo probabilístico que estima $P(\text{siguiente palabra} \mid \text{texto anterior})$. No razona con redes bayesianas explícitas, pero su salida es una distribución de probabilidad sobre el vocabulario, y todo lo que sabemos de probabilidad se aplica a ella.

Por qué debería importarte

Porque la incertidumbre no es un defecto que la IA vaya a eliminar; es la condición normal de cualquier sistema que actúa en el mundo real. La lógica del capítulo anterior te enseña a razonar cuando los hechos son seguros. Bayes te enseña a razonar cuando no lo son, que es casi siempre. Y te da algo que el instinto no tiene: un método para combinar la frecuencia de fondo de algo (la tasa base) con la fuerza de una evidencia concreta, sin sobreinterpretar ninguna de las dos.

Esto conecta directamente con cómo se debe usar un modelo de IA hoy. Cuando un LLM o un clasificador te da una probabilidad, esa cifra solo vale si está bien **calibrada**: si entre todos los casos en que dice “0,8” aproximadamente el 80 % resultan ciertos. Esa es la pregunta del Facsímil 7, capítulo 5, y es heredera directa de este: una probabilidad es una promesa sobre frecuencias, y Bayes es la gramática que la hace coherente. Entender el grado de creencia como un número con reglas es lo que te permite no tragarte una confianza inflada ni descartar una alerta solo porque “suena improbable”.

Dónde solía tropezar yo

ERROR	POR QUÉ ES UN ERROR	ANTÍDOTO
Olvidar la tasa base	Confundir $P(E \mid H)$ (cómo de fiable es la prueba) con $P(H \mid E)$ (la probabilidad real de la hipótesis) ignora lo rara que era la hipótesis previa.	Calcula siempre la evidencia $P(E)$ completa. Si la hipótesis es rara, hasta una prueba buena deja una probabilidad posterior baja.
Confundir $P(A \mid B)$ **con** $P(B \mid A)$	Son cantidades distintas; intercambiarlas es el error que produce el famoso 95 % falso en la prueba médica.	Escribe explícitamente qué condicionas sobre qué antes de poner números. La barra no es simétrica.
Tomar las probabilidades de Naive Bayes al pie de la letra	La hipótesis ingenua infla las cifras hacia 0 y 1; el orden suele ser correcto, pero la magnitud no es fiable.	Usa Naive Bayes para decidir la clase, no como probabilidad calibrada. Si necesitas la cifra, calíbrala.
Dibujar flechas como si fueran causalidad garantizada	En una red bayesiana la flecha codifica dependencia e influencia directa, pero la dirección la eliges tú; mal orientada, la factorización deja de ser económica.	Orienta las flechas de causa a efecto cuando puedas, y comprueba que cada nodo depende de pocos padres.
Creer que observar más siempre acerca dos variables	En una colisión ocurre lo contrario: observar el efecto común correlaciona causas antes independientes (*explaining away*).	Identifica el patrón (cadena, bifurcación, colisión) antes de afirmar una independencia.

Cómo encaja todo

Este capítulo es la bisagra del facsímil, y conviene verlo así: a un lado queda la certeza, al otro la creencia con grados. Venimos de la lógica del capítulo 1, donde cada hecho era verdadero o falso sin matices, y lo que hemos hecho aquí no ha sido tirarla a la basura, sino estirarla. La verdad y la falsedad se han convertido en los dos extremos, 1 y 0, de una regla continua, y entre medias cabe todo lo que de verdad nos pasa por la cabeza cuando decidimos

con información incompleta. Por eso el capítulo siguiente, el 3 de este mismo Facsímil 12, viene a continuación de forma natural: explora qué hacer cuando ni siquiera podemos poner un número limpio a la creencia, cuando la incertidumbre no es probabilística sino imprecisa o vaga, y para eso recurre a los factores de certeza y a la lógica difusa. Si Bayes te exige una probabilidad previa exacta («el robo ocurre una vez entre mil»), el capítulo 3 acepta que muchas veces lo único honesto que puedes decir es «poco probable, pero no sabría cuantificarlo», y construye una herramienta para esa penumbra. Son dos respuestas distintas a la misma pregunta incómoda de cuánto fiarte.

Hacia atrás, el capítulo cierra un círculo muy concreto con el Facsímil 1. El Naive Bayes que aquí hemos calculado a mano es exactamente uno de los clasificadores clásicos del capítulo 11 de aquel facsímil, y no es casualidad que reaparezca: allí lo presentamos como la baseline, la línea de salida que cualquier modelo más complejo tiene que superar para justificar su coste. La razón de que sea tan buena referencia es justo lo que hemos visto: se entrena contando frecuencias en una sola pasada y casi nunca falla el orden de las clases. Y la lección de la tasa base reaparece allí con otro nombre, el de las clases desbalanceadas. Un detector de fraude que ve un caso de cada diez mil cae en el mismo espejismo que el positivo de la enfermedad rara: confundir lo fiable que es la señal con lo probable que es el suceso. La matriz de confusión y el coste de los errores de aquel capítulo son, vistos desde aquí, maneras prácticas de no dejarse engañar por una probabilidad previa minúscula.

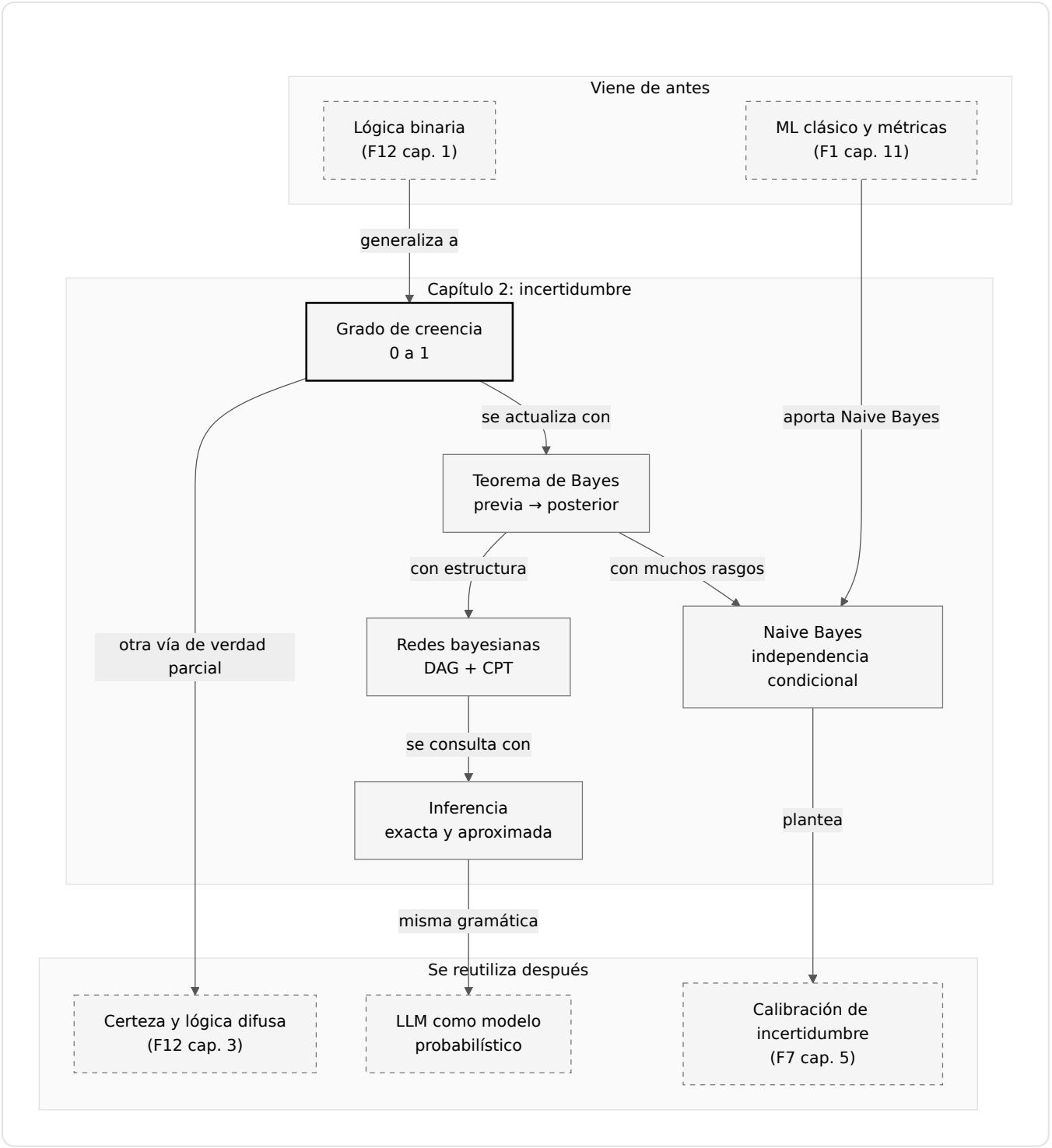
Hacia delante, el hilo más directo lleva al Facsímil 7, capítulo 5, sobre calibración e incertidumbre. Cuando un modelo —el Naive Bayes de aquí o un sistema mucho mayor— te dice «0,8», esa cifra es una promesa: que de cada cien veces que diga 0,8, alrededor de ochenta deberían acertar. Calibrar es comprobar si la promesa se cumple, y es la versión moderna de la pregunta que recorre todo este capítulo, «¿cuánto debería fiarme de esto?», pero formulada ya no para un cálculo a mano sino para un modelo que escupe miles de probabilidades al día. De hecho aquí ya avisamos del problema: Naive Bayes acierta el ganador pero miente en la magnitud, porque sus probabilidades se pegan a 0 y a 1. Esa patología tiene nombre y remedio, y son precisamente los del Facsímil 7. Es la diferencia entre un sistema que dice «99 % seguro» y se equivoca una de cada tres veces, y otro cuyo «80 %» puedes tomarte al pie de la letra.

Conecta también con lo que hoy llamamos IA moderna por una idea más sencilla de lo que parece. Un modelo de lenguaje, de los del Facsímil 3, no es en su núcleo más que una máquina probabilística: ante un texto, estima la probabilidad de cada posible palabra siguiente y elige según esa distribución. Cuando escribes «el cielo está» y el modelo continúa con «despejado», por dentro ha hecho lo mismo que nosotros con el spam: comparar verosimilitudes y quedarse con la más creíble dado el contexto. Toda la gramática de este capítulo —probabilidad condicional, marginalización, probabilidad posterior— se aplica a esa salida. Por eso entender Bayes no es un capricho de matemático: es lo que te permite leer con sentido crítico lo que afirma un chatbot, sabiendo que detrás hay una distribución de probabilidad, no una verdad revelada.

Y hay una conexión fácil de pasar por alto que el Facsímil 8 desarrolla: de dónde salen las probabilidades previas. Aquí te las hemos dado hechas —la enfermedad afecta al 1 %, el robo ocurre una vez entre mil—, pero en la práctica esos números se estiman a partir de datos, y

los datos arrastran sus propios sesgos. Si tu muestra sobrerrepresenta a un grupo, la probabilidad previa que aprendes queda torcida, y como Bayes la multiplica por la verosimilitud, ese sesgo se propaga sin resistencia hasta la probabilidad posterior. Un sistema de diagnóstico entrenado sobre una población distinta de la tuya parte de una tasa base equivocada y se equivoca con precisión matemática. Por eso el Facsímil 8, sobre datos y sesgos, es el complemento incómodo de este: Bayes es escrupulosamente honesto con las creencias que le metes, pero no las pone en duda; esa vigilancia te toca a ti.

Visto en conjunto, este capítulo no es una parada aislada sino un cruce de caminos: recoge el aprendizaje automático clásico que vino antes, presta su gramática a la calibración y a los modelos de lenguaje que vienen después, y deja una puerta abierta hacia las incertidumbres que ni siquiera se dejan escribir como una probabilidad. El diagrama lo resume:


Viene de antes
Lógica binaria
(F12 cap. 1)
ML clásico y métricas
(F1 cap. 11)
generaliza a
Capítulo 2: incertidumbre
Grado de creencia
0 a 1
se actualiza con
aporta Naive Bayes
Teorema de Bayes
previa → posterior
con estructura
con muchos rasgos
otra vía de verdad
parcial
Redes bayesianas
DAG + CPT
Naive Bayes
independencia
condicional
se consulta con
Inferencia
exacta y aproximada
plantea
misma gramática
Se reutiliza después
Certeza y lógica difusa
(F12 cap. 3)
LLM como modelo
probabilístico
Calibración de
incertidumbre
(F7 cap. 5)

Vocabulario aprendido

TÉRMINO	DEFINICIÓN
Probabilidad condicional	Probabilidad de un suceso dado que otro ha ocurrido, $P(A \mid B)$. Condicionar reduce el universo a los casos donde B es cierto.
Teorema de Bayes	Regla que actualiza la creencia en una hipótesis al observar evidencia, invirtiendo la dirección de la condicional.
Probabilidad previa	Creencia en una hipótesis antes de ver la evidencia; incluye la tasa base.
Verosimilitud	Probabilidad de la evidencia suponiendo cierta la hipótesis, $P(E \mid H)$.
Probabilidad posterior	Creencia en la hipótesis después de incorporar la evidencia; la probabilidad posterior de hoy es la previa de mañana.
Marginalización	Eliminar una variable de una conjunta sumando sobre todos sus valores.
Naive Bayes	Clasificador que aplica Bayes asumiendo independencia condicional entre rasgos dada la clase.
Independencia condicional	Dos variables son independientes una vez conocida una tercera.
Red bayesiana	Grafo dirigido acíclico que factoriza una distribución conjunta mediante CPT.
Tabla de probabilidad condicional (CPT)	Tabla que da la probabilidad de una variable para cada combinación de valores de sus padres.
d-separación	Criterio gráfico para decidir independencias condicionales según los patrones de cadena, bifurcación y colisión.
Eliminación de variables	Algoritmo de inferencia exacta que suma las variables ocultas para responder una consulta.

Antes de pasar página

- [ ] ¿Sé explicar por qué la lógica binaria del capítulo 1 no basta cuando los hechos son inciertos? (Si no, vuelve a «Entrando en el tema».)
- [ ] ¿Puedo escribir el teorema de Bayes y nombrar sus cuatro piezas? (Si no, vuelve a «El teorema de Bayes: dar la vuelta a la pregunta».)
- [ ] ¿Entiendo por qué un positivo en una prueba fiable puede dejar solo un 16,67 % de probabilidad de enfermedad? (Si no, vuelve a «El ejemplo que casi todo el mundo falla».)

- [ ] ¿Sé qué supone la hipótesis «ingenua» de Naive Bayes y por qué funciona aun siendo falsa? (Si no, vuelve a «Naive Bayes».)
- [ ] ¿Puedo escribir la factorización de la conjunta de una red bayesiana a partir de su grafo? (Si no, vuelve a «Redes bayesianas».)
- [ ] ¿Distingo cadena, bifurcación y colisión, y entiendo el *explaining away*? (Si no, vuelve a «Cuándo dos variables son independientes: la d-separación».)
- [ ] ¿He seguido el cálculo de $P(\text{Robo} \mid \text{Juan}, \text{María}) = 0{,}284$ por eliminación de variables? (Si no, vuelve a «Responder preguntas».)
- [ ] ¿He ejecutado el cuaderno y comprobado que `scikit-learn` y `pgmpy` reproducen los números?

En resumen

IDEA FUERZA	DETALLE
La probabilidad es lógica con grados.	El grado de creencia entre 0 y 1 generaliza el verdadero/falso, con reglas tan estrictas como las de la lógica.
Bayes invierte la pregunta.	Convierte “cómo de fiable es la prueba” en “qué probabilidad tengo de estar enfermo”, combinando la previa, la verosimilitud y evidencia.
La tasa base decide.	Ignorar lo rara que es una hipótesis previa es el error que produce el falso 95 %; la probabilidad posterior real era 16,67 %.
Naive Bayes acierta sin ser realista.	La independencia condicional es falsa, pero basta para ordenar bien las clases; por eso sigue siendo una baseline difícil de batir.
Las redes bayesianas organizan muchas creencias.	Un DAG con CPT factoriza la conjunta, hace legibles las independencias y permite inferir consultas exactas o aproximadas.
Es la gramática de lo moderno.	Un LLM estima probabilidades condicionales; calibrar y razonar sobre ellas exige las reglas de este capítulo.

Para saber más

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370-418. https://doi.org/10.1098/rstl.1763.0053

Duda, R. O., Hart, P. E. y Stork, D. G. (2001). *Pattern classification* (2.ª ed.). Wiley.

Koller, D. y Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.

Nilsson, N. J. (1986). Probabilistic logic. *Artificial Intelligence*, 28(1), 71-87. https://doi.org/10.1016/0004-3702(86)90031-7

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Russell, S. y Norvig, P. (2021). *Artificial intelligence: a modern approach* (4.ª ed.). Pearson.

Notas

1. Nilsson, N. J. (1986). Probabilistic logic. *Artificial Intelligence*, 28(1), 71-87. https://doi.org/10.1016/0004-3702(86)90031-7. Nilsson propuso una lógica probabilística que extiende la lógica clásica asignando probabilidades a las proposiciones en lugar de valores de verdad rígidos, y muestra que la lógica binaria es el caso límite cuando las probabilidades son 0 o 1. ↩

2. Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370-418. https://doi.org/10.1098/rstl.1763.0053. El ensayo, publicado tras la muerte de Bayes por su amigo Richard Price, plantea por primera vez el problema de inferir la causa probable a partir de los efectos observados, que es la estructura de toda inferencia bayesiana posterior. ↩

3. Duda, R. O., Hart, P. E. y Stork, D. G. (2001). *Pattern classification* (2.ª ed.). Wiley. El texto analiza el clasificador bayesiano ingenuo y muestra por qué su frontera de decisión puede ser correcta aunque las estimaciones de densidad sean imprecisas, siempre que los errores no alteren cuál es la clase de mayor probabilidad. ↩

4. Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann. Pearl introdujo las redes bayesianas como representación gráfica de la independencia condicional y desarrolló los algoritmos de propagación de creencias que hicieron tratable la inferencia probabilística en sistemas con muchas variables. ↩

5. Russell, S. y Norvig, P. (2021). *Artificial intelligence: a modern approach* (4.ª ed.). Pearson. Los capítulos sobre razonamiento probabilístico presentan la eliminación de variables y los métodos de Montecarlo, y demuestran que la inferencia exacta en redes bayesianas generales es computacionalmente difícil, lo que motiva los métodos aproximados. ↩

6. Koller, D. y Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press. El libro es el tratado de referencia sobre modelos gráficos probabilísticos y cubre con detalle tanto la inferencia exacta como la familia de métodos de muestreo y de propagación aproximada para redes de gran tamaño. ↩