

FUNDAMENTOS DE MACHINE LEARNING

Estadística Descriptiva III

Teoría de Probabilidades
Distribución de Probabilidades

DuocUC



ESCUELA DE
INFORMÁTICA Y
TELECOMUNICACIONES





¿Qué es la probabilidad?

La **probabilidad** es simplemente una manera de expresar qué tan probable es que ocurra un evento.

Puede ir desde 0 (imposible) hasta 1 (seguro).

Por ejemplo, lanzar una moneda y obtener cara tiene una probabilidad de 0.5.

Evento: Resultado posible de un experimento (ej: “obtener sello”).

Experimento: Actividad que genera un resultado (ej: lanzar una moneda).

Espacio muestral: Todos los resultados posibles (ej: {cara, sello}).

Podemos establecer la respuesta de antemano (a priori) sin necesidad de lanzar una moneda. No tenemos que efectuar experimentos para poder llegar a conclusiones sobre las monedas.

I. Probabilidad Clásica (A Priori)

Es la probabilidad que se asigna **antes de observar datos**, bajo el supuesto de que **todos los resultados posibles son igualmente probables**.

Probabilidad de un evento
$\text{Probabilidad de un evento} = \frac{\text{número de resultados en los que se presenta el evento}}{\text{número total de resultados posibles}}$

“Al lanzar un dado no cargado, hay 6 resultados posibles. ¿Cuál es la probabilidad de obtener un 4?”

$$P(4) = \frac{1}{6}$$

Podemos establecer la respuesta de antemano (a priori) sin necesidad de lanzar una moneda. No tenemos que efectuar experimentos para poder llegar a conclusiones sobre las monedas.

I. Probabilidad Clásica (A Priori)

Es la p
todos

En Machine Learning:

- Se usa para definir **priors** en modelos bayesianos cuando **aún no tenemos datos**.

Ejemplo: En clasificación, podríamos suponer que todas las clases tienen igual probabilidad si no hay información previa.

que

l es la

$$P(4) = \frac{1}{6}$$

Podemos establecer la respuesta de antemano (a priori) sin necesidad de lanzar una moneda. No tenemos que efectuar experimentos para poder llegar a conclusiones sobre las monedas.

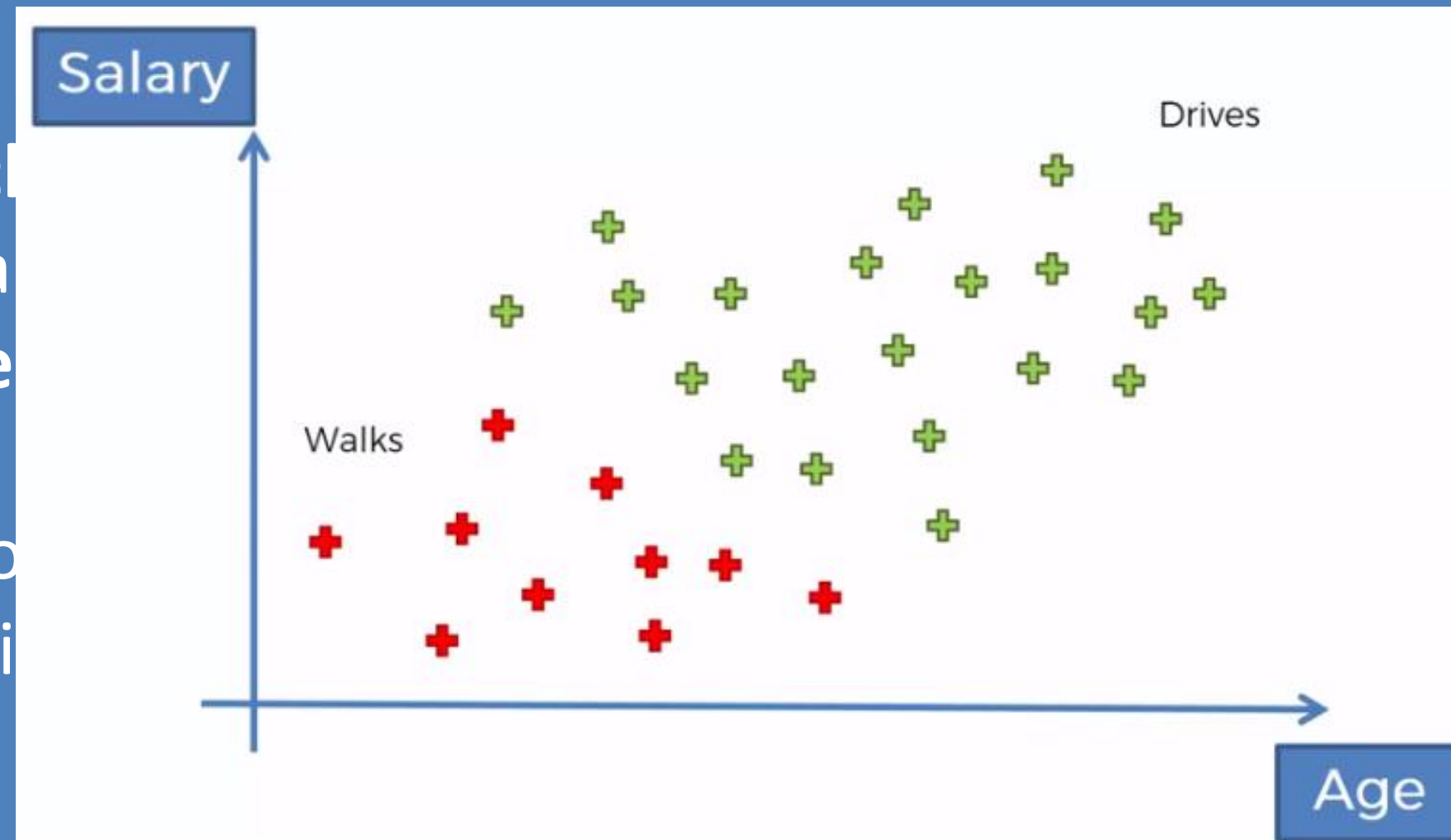
I. Probabilidad Clásica (A Priori)

Es la p
todos

En Mac

- Se usa
no tene

Ejemplo
clases ti



que

uando aún

as las
previa.

l es la

$$P(4) = \frac{1}{6}$$

II. Probabilidad Marginal (o Incondicional)

Es la probabilidad de que ocurra un evento específico, sin tener en cuenta ninguna otra información.

$$P(A) = \frac{\text{frecuencia de } A}{\text{total de observaciones}}$$

“En una rifa con 50 boletos y un solo ganador, si tú tienes 1 boleto:”

$$P(\text{ganar}) = \frac{1}{50}$$

↓



Concepto	¿Qué significa?	¿Cómo la saco?	¿Cuándo se usa?
Probabilidad marginal	Es la probabilidad de que pase algo, sin fijarte en otras cosas	Miro todos los casos y cuento cuántas veces pasa eso	Cuando ya tengo información y quiero ver la probabilidad total
Probabilidad a priori	Es lo que creo que va a pasar antes de ver datos nuevos	Me baso en lo que ya sabía o en experiencia previa	Antes de ver evidencia nueva, como en el Teorema de Bayes



Probabilidad marginal:

- Imagínate que ya revisaste 100 personas.
- Si 30 tienen fiebre, la probabilidad marginal de fiebre es $30/100 = 0,3$

Porque estás viendo el total de personas y contando las que tienen fiebre.

Probabilidad a priori:

Antes de revisar a nadie, tú ya sabes que más o menos el 10% de la gente en esta época del año tiene gripe.

Es lo que creías antes de mirar los datos nuevos.

III. Eventos Mutuamente Excluyentes

Dos eventos son mutuamente excluyentes si no pueden ocurrir al mismo tiempo.

$$P(A \cup B) = P(A) + P(B) \quad \text{si} \quad A \cap B = \emptyset$$

“Al lanzar una moneda:”

$A = \text{cara}, B = \text{sello}$

$$P(A \cup B) = P(\text{cara}) + P(\text{sello}) = 0.5 + 0.5 = 1$$

III. Eventos Mutuamente Excluyentes

Dos eventos

En Machine Learning:

- Las clases en clasificación binaria o multiclase suelen ser mutuamente excluyentes: un correo no puede ser simultáneamente “spam” y “no spam”.

Los algoritmos como **softmax** en redes neuronales también asumen exclusividad entre clases.

$A =$

$$P(A \cup B) = P(\text{cara}) + P(\text{sello}) = 0.5 + 0.5 = 1$$

IV. Eventos No Mutuamente Excluyentes

Son eventos que pueden ocurrir al mismo tiempo. Es decir, tienen intersección.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

“En un mazo de cartas, ¿cuál es la probabilidad de sacar un as o un corazón?”

Hay 4 ases, 13 corazones, y 1 as de corazones.

$$P(\text{As}) = \frac{4}{52}, P(\text{Corazón}) = \frac{13}{52}, P(\text{As y Corazón}) = \frac{1}{52}$$

$$P(\text{As o Corazón}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

IV. Eventos No Mutuamente Excluyentes

Son eventos

En Machine Learning:

En problemas de clasificación multilabel, una instancia puede tener múltiples clases a la vez: por ejemplo, una imagen puede ser “animal” y “mamífero”.

“En un

¿Por qué?”

Se usa en algoritmos de clasificación como **sigmoid** por salida, no **softmax**.

Hay 4

$$P(\text{As}) = \frac{4}{52}, P(\text{Corazon}) = \frac{13}{52}, P(\text{As y Corazon}) = \frac{1}{52}$$

$$P(\text{As o Corazón}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

V. Independencia Estadística

Dos eventos son independientes si la ocurrencia de uno no afecta la del otro.

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{si son independientes}$$

“Lanzar una moneda dos veces: la probabilidad de obtener dos caras es:”

$$P(\text{Cara 1 y Cara 2}) = 0.5 \cdot 0.5 = 0.25$$

V. Independencia Estadística

Dos ev


En Machine Learning:

Supuesto fundamental de Naive Bayes: que las características X_1, X_2, \dots, X_n son condicionalmente independientes dado Y .

“Lanz

Aunque esta suposición rara vez es 100% cierta, hace el modelo muy eficiente y rápido para grandes volúmenes de datos.

es:”



Imagina que estás tratando de adivinar si un correo es spam (Y) o no spam, y usas estas características:


X1: ¿Tiene la palabra “gratis”?

X2: ¿Tiene muchos signos de exclamación?

X3: ¿Viene de un remitente desconocido?

Naive Bayes asume que **una vez que sabes si es spam o no**, estas cosas no están relacionadas entre sí.

O sea, si ya sabes que es spam, **no importa si hay más signos de exclamación o si viene de alguien desconocido, cada cosa se analiza por separado.**



En probabilidades, todo gira en torno a saber si una cosa afecta a otra o no.

- Independencia: saber una cosa no cambia la probabilidad de otra.
- Dependencia: saber una cosa sí cambia la probabilidad de otra.
- Probabilidad incondicional: es la probabilidad "a secas", sin saber nada más.
- Probabilidad condicional: es la probabilidad sabiendo que otra cosa pasó.

Concepto	¿Qué es?	Ejemplo simple	¿Cómo se escribe?
Probabilidad incondicional	Probabilidad de algo sin saber ninguna otra info	Probabilidad de que llueva	$P(\text{Lluvia})$
Probabilidad condicional	Probabilidad de algo sabiendo que otra cosa pasó	Probabilidad de que llueva si hay nubes	$P(\text{Lluvia} \mid \text{Nubes})$
Independencia	Dos cosas no se afectan entre sí	Lluvia y que te guste el helado	$P(A \mid B) = P(A)$
Dependencia	Una cosa sí cambia la probabilidad de la otra	Lluvia y que haya nubes	$P(\text{Lluvia} \mid \text{Nubes}) \neq P(\text{Lluvia})$
Naive Bayes (en ML)	Supone que las características son independientes si conocés la clase Y	Si ya sé que es spam, analizo cada palabra por separado	$P(X_1, X_2 \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y)$

VI. Dependencia Estadística y Probabilidad Condicional

Cuando la ocurrencia de un evento sí afecta la probabilidad del otro.

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

“Tenemos 10 bolas. 4 son de color: 3 con puntos y 1 con franjas. Si sabemos que se sacó una bola de color, ¿cuál es la probabilidad de que tenga puntos?”

$$P(\text{Puntos} | \text{Color}) = \frac{3}{4} = 0.75$$

VI. Dependencia Estadística v Probabilidad Condicional

Cuando

En Machine Learning:

Modelos como las Redes Bayesianas representan dependencias explícitas entre variables.

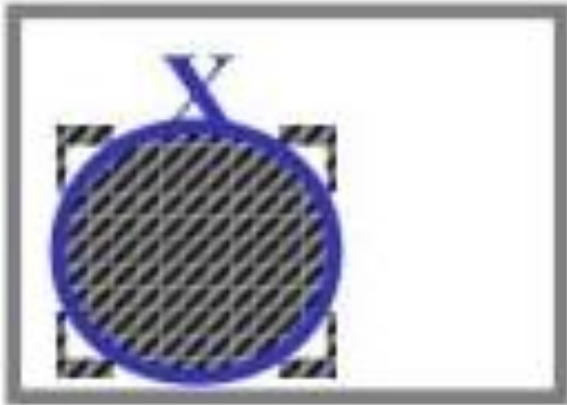

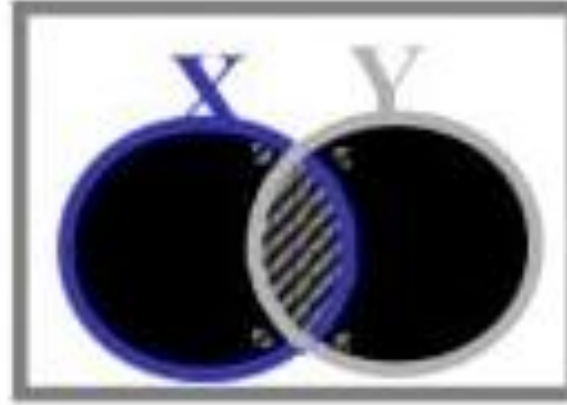

“Tenemos
se sabe

En algoritmos supervisados, las probabilidades condicionales como $P(Y|X)$ son justamente lo que aprendemos para predecir clases.

mos que
tos?”

$$P(\text{Puntos} \mid \text{Color}) = \frac{3}{4} = 0.75$$

tipos de probabilidad

Marginal	Unión	Conjunta	Condicional
$P(X)$ La probabilidad de que ocurra X	$P(X \cup Y)$ La probabilidad de que ocurra X o Y	$P(X \cap Y)$ La probabilidad de que ocurra X e Y	$P(X Y)$ La probabilidad de que ocurra X sabiendo que ha ocurrido Y
			

VII. Teorema de Bayes

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

Lo que viste	¿Qué aporta?
Probabilidad condicional	Te dice cómo cambia una probabilidad si sabés algo
Dependencia estadística	Justifica por qué importa saber una cosa para predecir otra
Bayes	Te permite calcular una probabilidad "al revés", usando lo que sabés y actualizando con datos nuevos

VII. Teorema de Bayes

Permite **actualizar una creencia (A)** cuando observamos nueva **evidencia (B)**. Este principio es la base del aprendizaje probabilístico.

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

“Dado que un correo contiene ‘descuento’, ¿cuál es la probabilidad de que sea spam?”

- $P(\text{spam}) = 0.2$
- $P(\text{'descuento'}|\text{spam}) = 0.6$
- $P(\text{'descuento'}) = 0.1$

$$P(\text{spam}|\text{'descuento'}) = \frac{0.6 \cdot 0.2}{0.1} = 1.2 \quad (\rightarrow \text{revisa datos, no puede ser } > 1)$$

Permite **actualizar una creencia (A)** cuando observamos nueva **evidencia (B)**. Este principio es la base del aprendizaje probabilístico.

VII. Teorema de Bayes

En Machine Learning:
Naive Bayes Classifier
Aplicación del Teorema de Bayes

“Dado

que sea

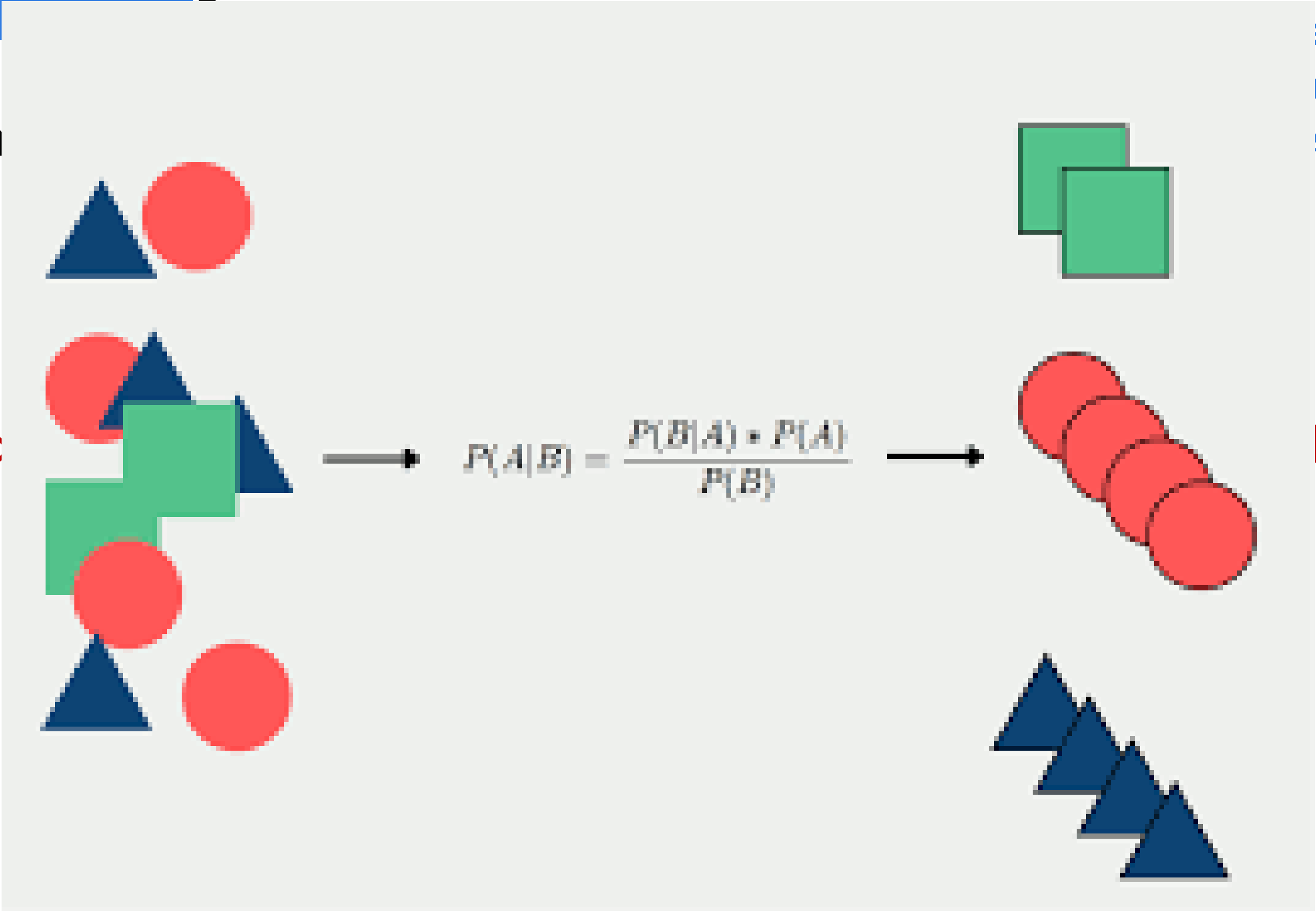
Filtros de spam, Clasificadores de texto, Análisis de sentimientos, Sistemas de diagnóstico médico, Reconocimiento de patrones, Inferencia bayesiana en redes neuronales

- $P(\text{spam}) = 0.6$
- $P(\text{'descuento'}) = 0.2$
- $P(\text{'descuento' | spam}) = 0.1$

$$P(\text{spam} | \text{'descuento'}) = \frac{0.6 \cdot 0.2}{0.1} = 1.2 \quad (\rightarrow \text{revisa datos, no puede ser } > 1)$$

VII. Teorema

“Dado c



encia (A) cuando
(B). Este principio
probabilístico.

lad de que sea



RESUMEN

- Conceptos de Probabilidad
- Probabilidad Condicional
- Teorema de Bayes

Distribución de Probabilidades

Distribución de Probabilidades

Una **distribución de probabilidad** es un modelo teórico que nos dice cómo esperamos que se distribuyan los posibles valores de una variable aleatoria.

Por ejemplo, si lanzamos una moneda dos veces, los posibles resultados (CC, CS, SC, SS) pueden representarse con una tabla de probabilidades.

Este es un ejemplo de **distribución discreta**.

Distribución Discreta: Toma valores finitos o contables.

Ejemplo clásico: Número de votos a un candidato, número de clics en un anuncio, número de errores en una red neuronal durante una época.

Distribución Continua: Toma cualquier valor en un intervalo.

Ejemplo: Tiempo que tarda en converger un modelo, precisión de un clasificador (0 a 1).

Distribución de Probabilidades

Representación de tipos de variables aleatorias

Solo puede tomar ciertos valores específicos (como contar cosas).

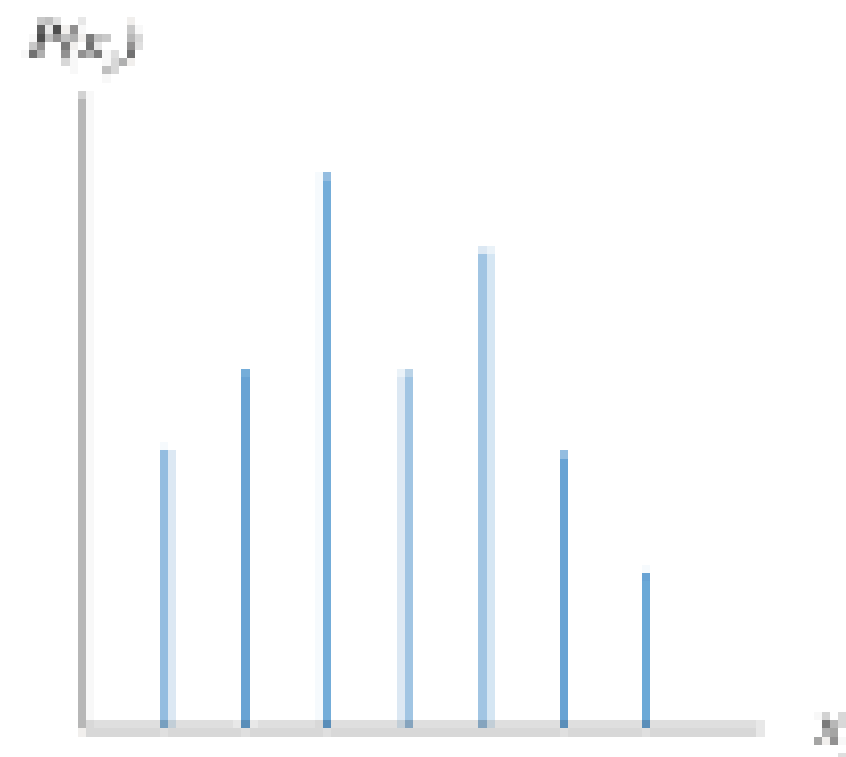
Barras separadas, cada una representa la probabilidad de un valor.

Ejemplo: número de hijos en una familia (0, 1, 2, 3...)

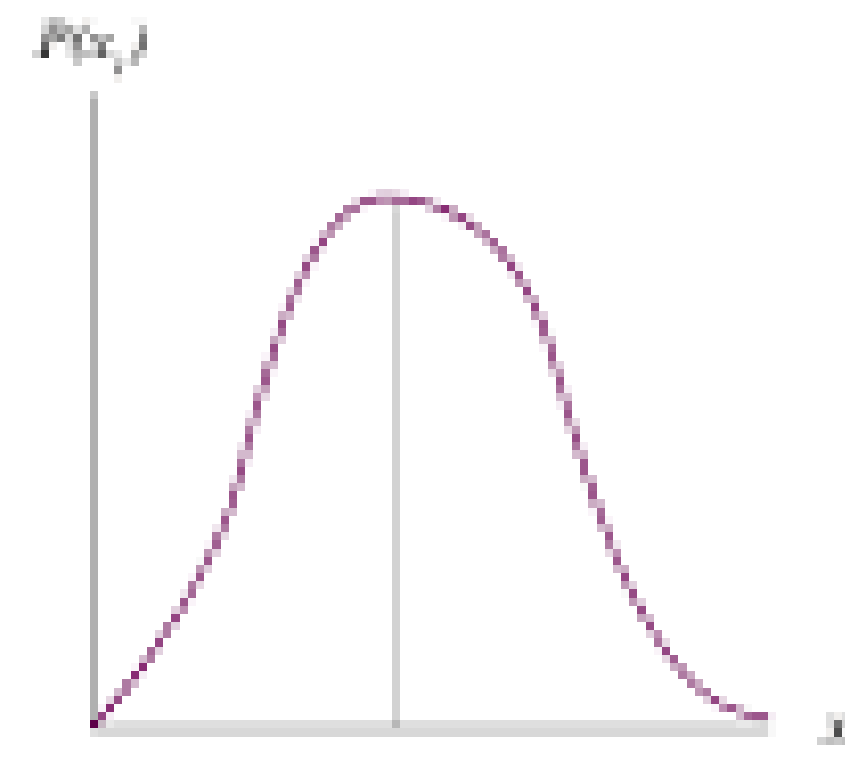
La probabilidad de cada valor se muestra como una barra.

Solo se pueden tener ciertos valores enteros
→ no podés tener 2,7 hijos.

Discreta



Continua



Puede tomar infinitos valores dentro de un rango (como medir cosas).

Una curva suave, que representa densidad de probabilidad.

Ejemplo: altura de personas, temperatura, tiempo.

Puede tomar cualquier valor en un intervalo, incluso con decimales.

No se usan barras, sino una curva que muestra la densidad (no la probabilidad exacta de un punto).

Distribución de Probabilidades

Característica	Variable Discreta	Variable Continua
Tipo de valores	Específicos (enteros)	Infinitos (decimales posibles)
Cómo se representa	Barras	Curva
Probabilidad exacta	Tiene valor	Casi cero (se usa intervalo)
Ejemplos	Número de hijos, dados	Estatura, peso, tiempo

Variables Aleatorias: Discretas vs Continuas

Acá viene un concepto clave. Hay dos tipos de variables con las que trabajamos:

Característica	Variable Discreta	Variable Continua
Tipo de valores	Específicos (enteros)	Infinitos (decimales posibles)
Cómo se representa	Barras	Curva
Probabilidad exacta	Tiene valor	Casi cero (se usa intervalo)
Ejemplos	Número de hijos, dados	Estatura, peso, tiempo

Discretas

Toman valores que se pueden contar.

Ejemplo: cantidad de emails marcados como spam por un modelo.

Continuas

Pueden tomar cualquier valor dentro de un rango.

Ejemplo: el tiempo (en segundos) que tarda tu modelo en clasificar una imagen.

Variables Aleatorias: Discretas vs Continuas

Acá viene un concepto clave. Hay dos tipos de variables con las que trabajamos:

Variable discreta:

Número de objetos detectados por imagen

- Solo puede ser: 0, 1, 2, 3, etc.
- No se puede detectar "2.5 personas" → tiene que ser un número entero.
- Eso lo convierte en una variable discreta.

En una imagen detectas 3 autos → ese "3" es un valor discreto.

Variable continua:

Tiempo que tarda el modelo en analizar una imagen (inferir)

- Puede ser 0.8 segundos, 1.24 segundos, 2.015 segundos, etc.
- Son valores que pueden tener decimales y cambiar muy poco.
- Por eso es una variable continua.

El modelo tarda 1.37 segundos en procesar una imagen → eso es un valor continuo.

El valor esperado (el promedio “esperado”)

El valor esperado (el promedio “esperado”)

¿Te suena lógico que si algo pasa mucho, en promedio, esperes que pase?

Errores	Probabilidad
0	20%
1	50%
2	30%

$$(0 \times 0.2) + (1 \times 0.5) + (2 \times 0.3) = 1.1 \text{ errores por lote}$$

La distribución binomial (cuando algo pasa o no pasa)

La distribución binomial (cuando algo pasa o no pasa)

La distribución binomial (cuando algo pasa o no pasa)

¿Cuándo se usa?

- Tienes muchos intentos.
- En cada uno solo puede pasar A o B (por ejemplo, acierto o error).
- Todo tiene la misma probabilidad.
- Los intentos son independientes.

Fórmula binomial

$$\text{Probabilidad de } r \text{ éxitos en } n \text{ intentos} = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

Tienes un modelo de clasificación binaria que acierta el 80% de las veces. Lo pruebas 10 veces.

¿Qué probabilidad hay de que acierte exactamente 8 veces?

$$P(8 \text{ aciertos}) = \text{"combinaciones"} \times (0.8)^8 \times (0.2)^2$$

Te da alrededor del 30% de probabilidad. O sea, no es tan raro que tu modelo acierte 8 veces si es 80% preciso.

Distribución de Poisson (cuando contás eventos raros)

Esto sirve cuando quieres contar cuántas veces pasa algo, pero no sabes cuándo exactamente.

¿Cuántas veces llegan alertas de fraude en una hora?

¿Cuántas veces falla un sensor por día?

Supongamos que, en promedio, tienes 5 alertas de seguridad por día.

¿cuál es la chance de que un día haya 0 o 1 alertas?

Se usa esta fórmula con una letra griega (λ , que es la media), y te da una probabilidad concreta (en este caso, alrededor de 4,5% de que pase eso).

Fórmula de Poisson

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

Distribución normal (la famosa “campana”)

Esta sí que es una celebridad. La curva en forma de campana que aparece por todos lados: estatura, peso, tiempos de espera... y sí, también en Machine Learning.

¿Qué tiene de especial?

- Es simétrica.
- La mayoría de los datos se acumulan alrededor del centro (la media).
- A medida que te alejas, las probabilidades bajan.

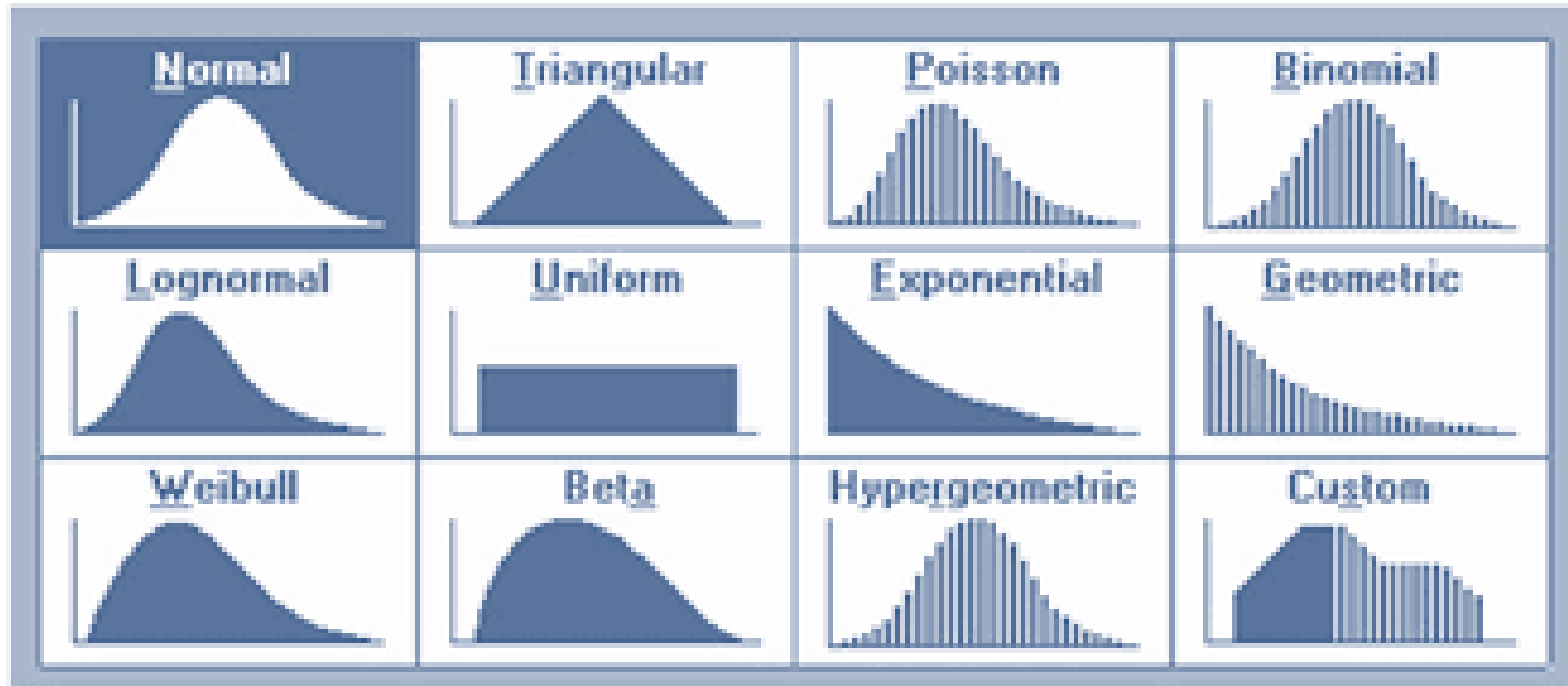
Tienes un sistema de entrenamiento que tarda, en promedio, **500 horas**, y la mayoría de los casos están dentro de **± 100 horas**.

Si quieres saber qué tan probable es que tarde **entre 500 y 650 horas**, puedes usar la distribución normal y te dice: **43% de probabilidad**.

¿Cómo se usa esto en ML?

- Para generar datos sintéticos con ciertas distribuciones.
- Para evaluar modelos, especialmente con métricas que siguen alguna de estas distribuciones.
- Para hacer detección de anomalías: lo que se aleja mucho de la curva normal suele ser algo raro.

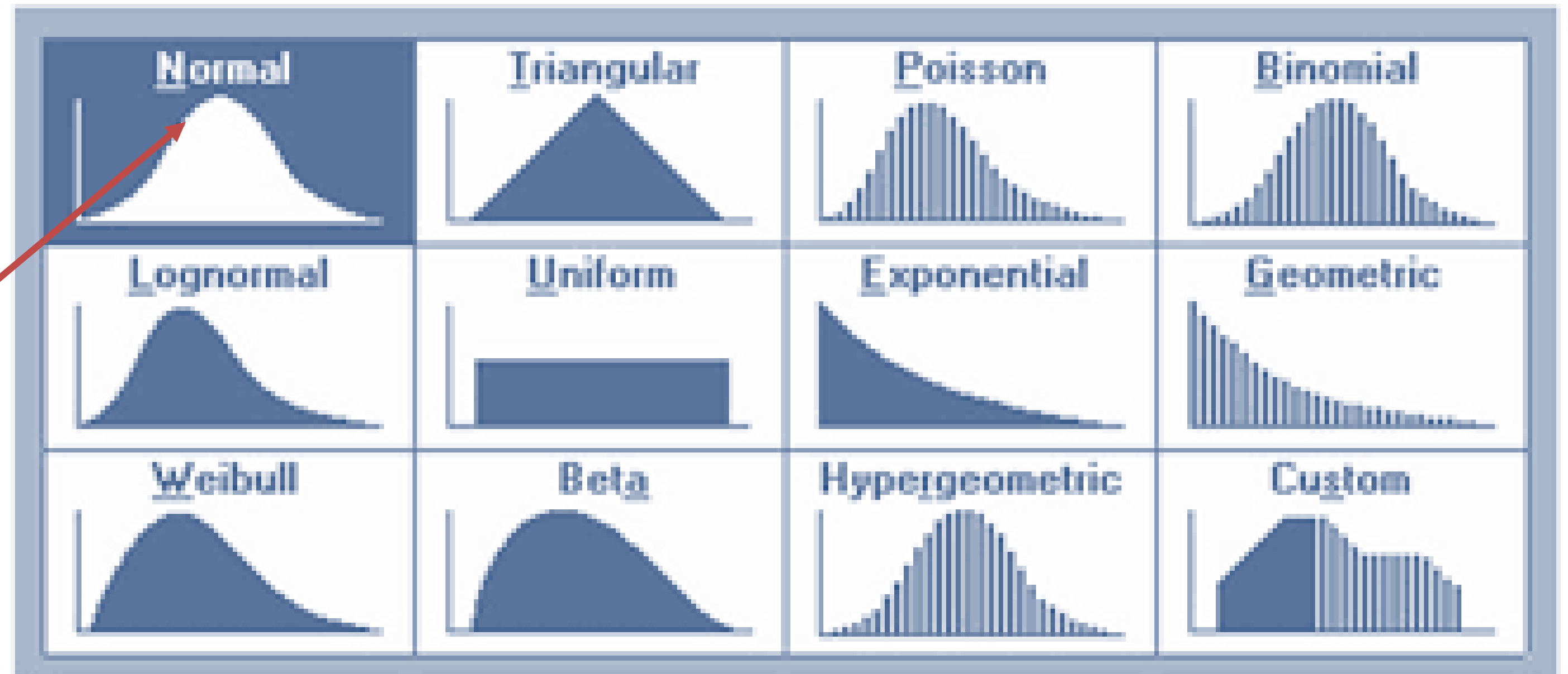
¿Cómo se usa esto en ML?



¿Cómo se usa esto en ML?

Se usa mucho en evaluación de modelos (por ejemplo, los errores residuales suelen seguir esta distribución).

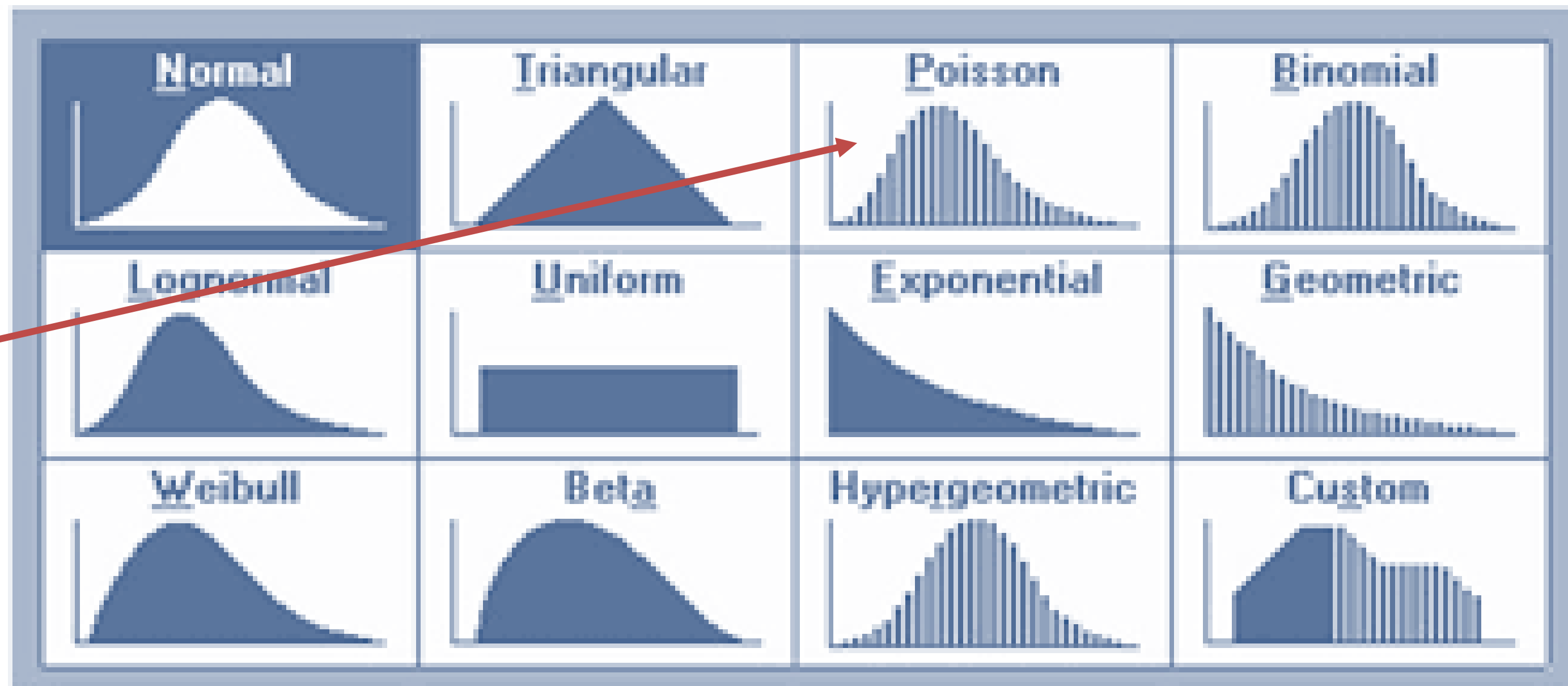
Es base para técnicas como regresión lineal y normalización de datos.



¿Cómo se usa esto en ML?

Muy útil en *detección de anomalías*, cuando los eventos son infrecuentes.

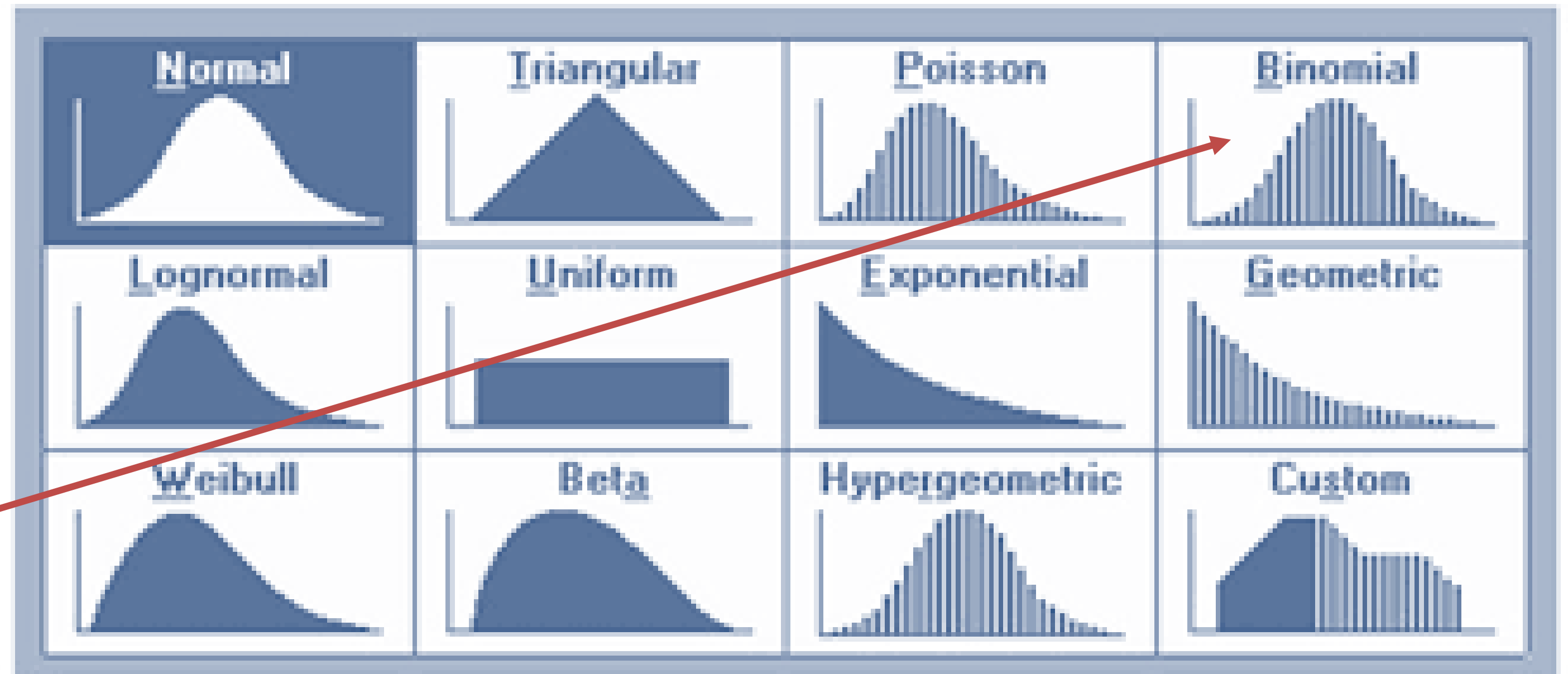
También en modelos que predicen "conteo de eventos" como cuántos usuarios harán clic o cuántos errores generará un sistema.



¿Cómo se usa esto en ML?

Sirve para problemas de clasificación binaria (sí/no, spam/no spam).

También se usa para analizar cuántas veces esperamos que un clasificador acierte en un conjunto de pruebas.



FUNDAMENTOS DE MACHINE LEARNING

Estadística Descriptiva III

Teoría de Probabilidades
Distribución de Probabilidades

DuocUC



ESCUELA DE
INFORMÁTICA Y
TELECOMUNICACIONES

