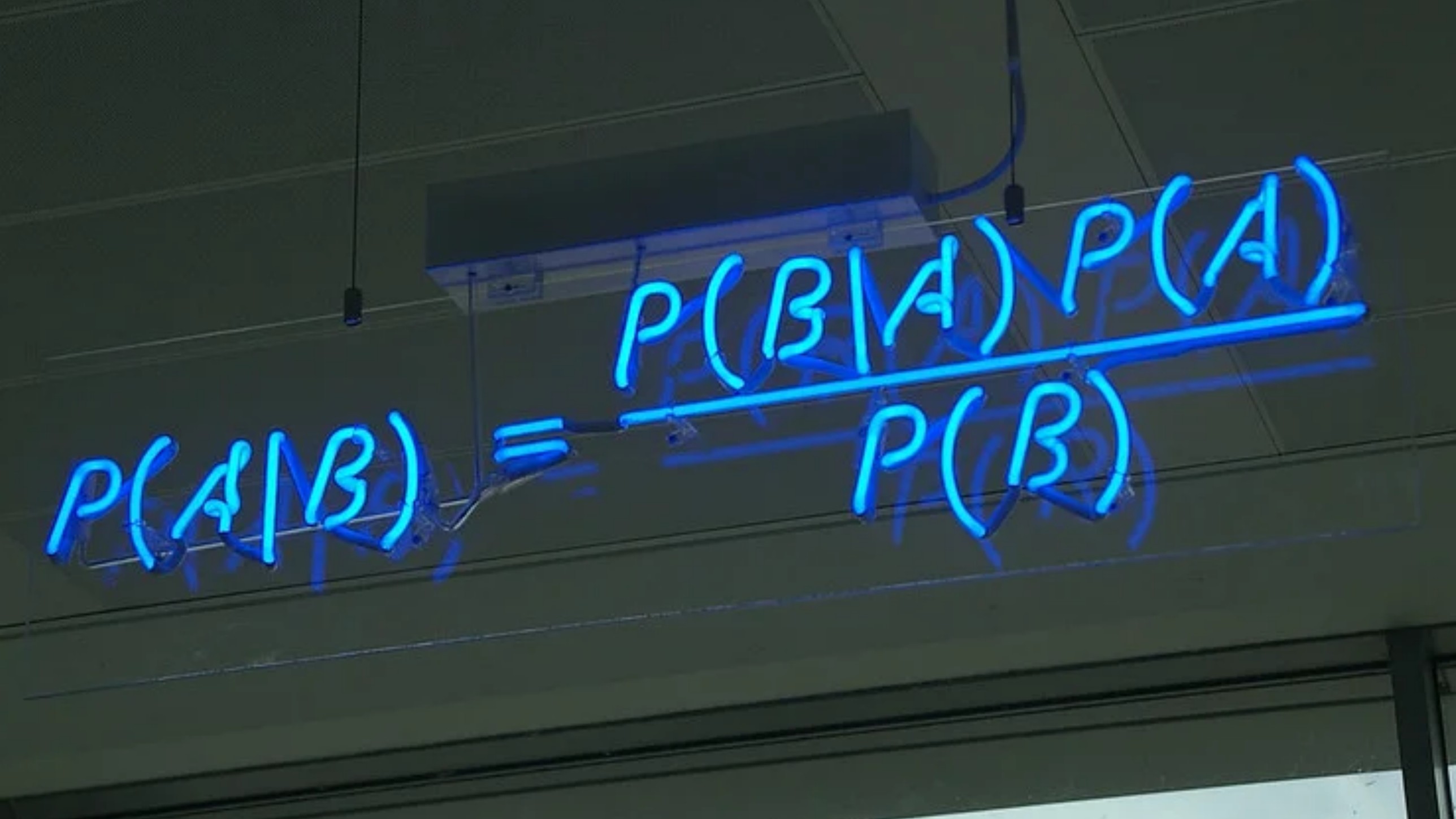
Naive Bayes

Maestría en Ciencia de Datos, CUCEA, Universidad de Guadalajara.

Guadalajara, Jal., agosto de 2025


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

Inferencia Estadística

Introducción

Un análisis estadístico incluye:

- Parámetros
- Datos
- Modelos

Inferencia Estadística

Parámetros

- **¿Qué son?** - cantidades (desconocidas) que caracterizan el fenómeno y al modelo.
- **Ejemplos generales**
  - Regresión lineal: $\beta_0, \beta_1, \ldots$
  - Normal: media y varianza, μ, σ^2
  - Poisson: tasa λ

Inferencia Estadística

Datos

- **¿Qué son?** - observaciones disponibles para aprender e inferir.
- **Estructura básica**
  - Matriz de características: $\mathbf{X} \in \mathbb{R}^{n \times d}$
  - Etiquetas (en clasificación): $\mathbf{y} \in 1, \ldots, K^n$

Inferencia Estadística

Modelos

- **¿Qué son?** - supuestos probabilísticos que conectan parámetros y datos para explicar/generar observaciones.
- **Variantes del modelo Naive Bayes**
  - Multinomial NB: texto con conteos (bolsa de palabras).
  - Bernoulli NB: texto binarizado (presencia/ausencia) o rasgos 0/1
  - Gaussian NB: rasgos continuos (asume normalidad por clase y rasgos)

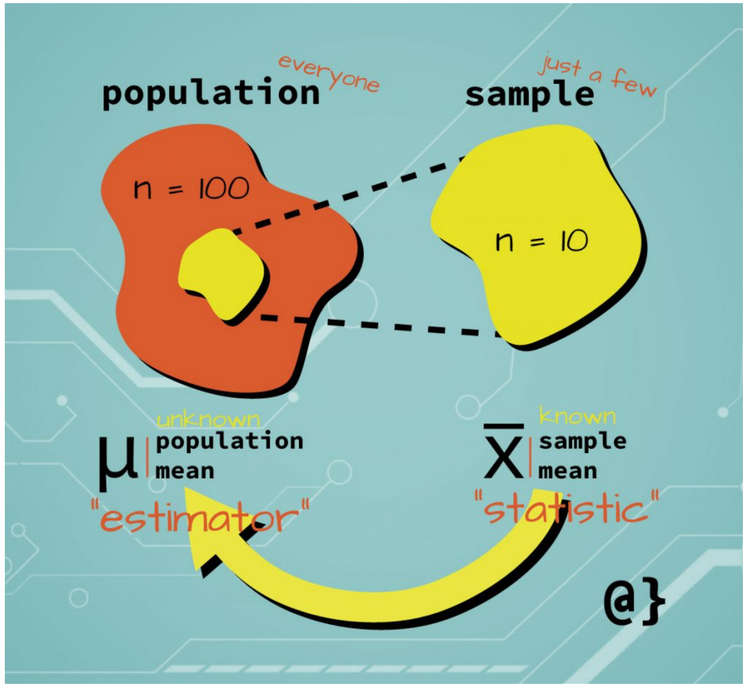
Frecuentista vs Bayesiano

Visión Rápida

- **Frecuentista**
  - **Datos**: aleatorios
  - **Parámetros**: fijos (desconocidos)
  - La probabilidad está sobre los datos (muestras repetidas)
- **Bayesiano**
  - **Datos**: aleatorios
  - **Parámetros**: aleatorios
  - Posterior: $p(\theta \mid \mathscr{D}) \propto p(\mathscr{D} \mid \theta)p(\theta)$

Inferencia Estadística

¿Qué es un estimador?



- Un **estimador** es una función de la muestra completa

$$\hat{\theta} = g\big((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\big) \, .$$

- Ejemplos: media muestral ($\bar{x}$), MLE (Estimación de máxima verosimilitud), MAP (Estimación máxima a posteriori), promedios por clase, etc.

Inferencia Estadística

Ejemplo de estimador (punto de vista frecuentista)

- Se ha recopilado información sobre la prevalencia de TDAH en una población definida. Nuestra muestra incluye n niños/niñas, de los cuales tenemos etiquetas $y_i \in 0,1$ para un diagnóstico positivo/negativo, respectivamente. Sea $S = \sum_i y_i$, podríamos construir un estimador para la proporción.

$$\hat{p} = \frac{S}{n}.$$

- Con ciertas propiedades que vienen de la distribución muestral de una proporción, bajo el teorema del límite central,

$$\mathbb{E}[\hat{p}] = p$$

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$$\widehat{\text{SE}}\,(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Modelo de Naive Bayes

Idea del modelo

- Un modelo de Naive Bayes es un clasificador probabilístico, fundamentado en el teorema de Bayes, y es uno de los métodos supervisados de clasificación más simples.

- Uno tiene una colección de N pares entrenados $\{(\vec{x}_i, y_i)\}_{i=1}^{N}$, donde el vector $\vec{x}_i \in \mathbb{R}^D$ son las D variables de entrada (o características) de las i muestras o registros, siendo $y_i \in \{1,\ldots,C\}$ es la variable dependiente/objetivo (categórica) que denota a cuál clase C pertenece cada registro i.

Modelo de Naive Bayes

Idea del modelo

- Dada una nueva muestra $\vec{x}$, el objetivo es determinar cual sería su clase predicha $\hat{y}$. Esta predicción de la clase se realiza escogiendo la clase con mayor probabilidad a *posteriori* bajo la suposición simple (ingenua, naive), de que cada una de estas variables de entrada $\vec{x}$ son independientes una de la otra.

$$\hat{y} = \arg\max_{y \in [C]} \mathscr{P}(y \mid \vec{x}) = \arg\max_{y \in [C]} \frac{\mathscr{P}(\vec{x} \mid y)\mathscr{P}(y)}{\mathscr{P}(\vec{x})} \propto \arg\max_{y \in [C]} \mathscr{P}(\vec{x} \mid y)\mathscr{P}(y)$$

$$= \arg\max_{y \in [C]} \mathscr{P}(y) \prod_{j=1}^{D} \mathscr{P}(x_j \mid y)$$

- Donde $\mathscr{P}$ es la probabilidad a posteriori que se busca maximizar.

Posterior Probability

Likelihood of Observations

Prior Probability

$$\Pr(\theta|y) = \frac{\Pr(y|\theta)\Pr(\theta)}{\Pr(y)}$$

Normalizing Constant

Modelo de Naive Bayes

Ejemplo

- Supongamos que nuestro objetivo es desarrollar un filtro de spam utilizando este modelo. Para esto, se nos proporciona un dataset que se muestra en la siguiente tabla.

Features		Muestras(emails)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	congratulations	1	1	1	1	1	0	0	0	0	1
2	you	1	1	1	0	0	0	1	1	0	0
3	won	0	1	1	1	1	1	0	0	0	1
4	free	1	1	1	1	1	1	1	0	0	0
5	gift	0	0	1	1	1	1	0	1	0	0
6	attached	0	0	1	0	0	0	1	1	1	0
7	sincerely	1	0	1	0	0	1	0	0	1	1
8	thanks	0	1	0	1	1	0	1	1	0	0
Categoría		**Spam**						**No Spam**			

Modelo de Naive Bayes

Ejemplo

	Features	Muestras(emails)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	congratulations	1	1	1	1	1	0	0	0	0	1
2	you	1	1	1	0	0	0	1	1	0	0
3	won	0	1	1	1	1	1	0	0	0	1
4	free	1	1	1	1	1	1	1	0	0	0
5	gift	0	0	1	1	1	1	0	1	0	0
6	attached	0	0	1	0	0	0	1	1	1	0
7	sincerely	1	0	1	0	0	1	0	0	1	1
8	thanks	0	1	0	1	1	0	1	1	0	0
	Categoría	**Spam**						**No Spam**			

- Dado un nuevo mail que contiene el mensaje “congratulations, you won free gift”, el objetivo es determinar si este mensaje es clasificado como spam o no.

- Para esto, el modelo de Naive Bayes estima primero la probabilidad total de que un correo sea spam $\hat{\mathscr{P}}(y = 1)$ (spam=1, nospam=0) (Maximum Likelihood Estimador). En teoría, uno podría tener un estimado de cuál sería la probabilidad global de que correo sea spam en toda la población, pero si no se tiene este dato, lo recomendable es utilizar la información del dataset mismo.

$$\hat{\mathscr{P}}(y = \text{spam}) = \frac{6}{10} = \frac{3}{5}$$

$$\hat{\mathscr{P}}(y = \text{no spam}) = \frac{4}{10} = \frac{2}{5}$$

Modelo de Naive Bayes

Ejemplo

	Features	Muestras(emails)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	congratulations	1	1	1	1	1	0	0	0	0	1
2	you	1	1	1	0	0	0	1	1	0	0
3	won	0	1	1	1	1	1	0	0	0	1
4	free	1	1	1	1	1	1	1	0	0	0
5	gift	0	0	1	1	1	1	0	1	0	0
6	attached	0	0	1	0	0	0	1	1	1	0
7	sincerely	1	0	1	0	0	1	0	0	1	1
8	thanks	0	1	0	1	1	0	1	1	0	0
	Categoría	**Spam**						**No Spam**			

- De manera similar, uno puede obtener para cada variable de entrada (cada feature, en este caso cada palabra) x_j una probabilidad condicional $\hat{\mathscr{P}}(x_j = 1 \,|\, y)$ como la fracción correspondiente por cada clase, utilizando la información del dataset.

- Por ejemplo, de los datos, observamos que la probabilidad de encontrar la palabra "congratulations" en un mensaje de spam sería $\hat{\mathscr{P}}(x = \text{congratulations} \,|\, \text{y} = \text{spam}) = \frac{5}{6}$, pues según los datos, esta palabra se encuentra en 5 de cada 6 mensajes catalogados como spam.

Modelo de Naive Bayes

Ejemplo

Features		Muestras(emails)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	congratulations	1	1	1	1	1	0	0	0	0	1
2	you	1	1	1	0	0	0	1	1	0	0
3	won	0	1	1	1	1	1	0	0	0	1
4	free	1	1	1	1	1	1	1	0	0	0
5	gift	0	0	1	1	1	1	0	1	0	0
6	attached	0	0	1	0	0	0	1	1	1	0
7	sincerely	1	0	1	0	0	1	0	0	1	1
8	thanks	0	1	0	1	1	0	1	1	0	0
Categoría		**Spam**						**No Spam**			

		$P(x_j = 1\|y)$	
		Spam	**No Spam**
1	congratulations	5/6	1/4
2	you	1/2	1/2
3	won	5/6	1/4
4	free	1	1/4
5	gift	2/3	1/4
6	attached	1/6	3/4
7	sincerely	1/2	1/2
8	thanks	1/2	1/2

Modelo de Naive Bayes

Ejemplo

		$P(x_j = 1 \mid y)$	
		Spam	**No Spam**
1	congratulations	5/6	1/4
2	you	1/2	1/2
3	won	5/6	1/4
4	free	1	1/4
5	gift	2/3	1/4
6	attached	1/6	3/4
7	sincerely	1/2	1/2
8	thanks	1/2	1/2

- Utilizando el mismo formato del dataset, el mensaje “congratulations, you won free gift” estaría codificado como un vector $\vec{x} = \left[1,1,1,1,1,0,0,0\right]$.
- Utilizando la información de la tabla de probabilidades creada, podemos calcular la probabilidad condicional de que este mensaje sea spam (y no spam) dado este contenido de palabras,

$$\hat{\mathscr{P}}(y = \text{spam} \mid \vec{x}) \propto \hat{\mathscr{P}}(y = \text{spam}) \prod_{j=1}^{D} \hat{\mathscr{P}}(x_j \mid y = \text{spam})$$

$$\hat{\mathscr{P}}(y = \text{no spam} \mid \vec{x}) \propto \hat{\mathscr{P}}(y = \text{no spam}) \prod_{j=1}^{D} \hat{\mathscr{P}}(x_j \mid y = \text{no spam})$$

Modelo de Naive Bayes

Ejemplo

		$P(x_j = 1 \mid y)$	
		Spam	**No Spam**
1	congratulations	5/6	1/4
2	you	1/2	1/2
3	won	5/6	1/4
4	free	1	1/4
5	gift	2/3	1/4
6	attached	1/6	3/4
7	sincerely	1/2	1/2
8	thanks	1/2	1/2

- Por ejemplo, dado nuestro vector a probar $\vec{x} = [1,1,1,1,1,0,0,0]$, tenemos que,

$$\hat{\mathscr{P}}(y = \text{spam} \mid \vec{x}) \propto \hat{\mathscr{P}}(y = \text{spam}) \prod_{j=1}^{D} \hat{\mathscr{P}}(x_j \mid y = \text{spam})$$

$$= \frac{3}{5} \cdot \hat{\mathscr{P}}(x_1 = 1 \mid y = \text{spam}) \cdot \hat{\mathscr{P}}(\text{x}_2 = 1 \mid \text{y} = \text{spam}) \cdot \hat{\mathscr{P}}(\text{x}_3 = 1 \mid \text{y} = \text{spam})$$

$$\cdot \hat{\mathscr{P}}(x_4 = 1 \mid y = \text{spam}) \cdot \hat{\mathscr{P}}(\text{x}_5 = 1 \mid \text{y} = \text{spam}) \cdot \hat{\mathscr{P}}(\text{x}_6 = 0 \mid \text{y} = \text{spam})$$

$$\cdot \hat{\mathscr{P}}(x_7 = 0 \mid y = \text{spam}) \cdot \hat{\mathscr{P}}(\text{x}_8 = 0 \mid \text{y} = \text{spam})$$

$$= \frac{3}{5} \cdot \frac{5}{6} \cdot \frac{1}{2} \cdot \frac{5}{6} \cdot 1 \cdot \frac{2}{3} \cdot \frac{5}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.0289$$

- Lo mismo se puede hacer para la otra probabilidad condicional $\hat{\mathscr{P}}(y = \text{no spam} \mid \vec{\text{x}})$, y escoger la clase de aquella con mayor probabilidad posterior.

Bernoulli Naive Bayes

For binary or boolean features.

EXAMPLE

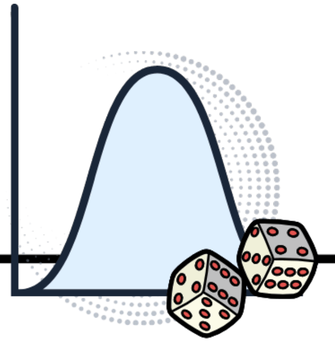
HAS 4 LEGS?	CAN FLY?	LAY EGGS?
1	0	1
0	1	1
0	1	1
0	1	0
0	0	1

Multinomial Naive Bayes

For discrete features (like word count)

EXAMPLE

WORD the	WORD but	WORD is
4	1	2
3	0	0
3	0	0
1	0	4
0	2	3



Gaussian Naive Bayes

For continuous, real-valued attributes.

EXAMPLE

AGE	WEIGHT	HEIGHT
17.4	56.5	145.2
25.4	71.2	170.4
18.8	70.3	164.5
21.1	51.2	140.5
20.9	81.5	182.2

Modelo de Naive Bayes (Gaussiano)

Ejemplo

- Supongamos que eres un detective a cargo de un homicidio, y pudiste extraer la siguiente evidencia del asesino:
  - Tamaño del calzado: 42 cm
  - Altura: 180 cm
  - Ritmo: 5.5 minutos/km
- Como comienzo, quieres ir reduciendo la lista de sospechosos que ya tienes, y te preguntas simplemente si el asesino es de sexo masculino o femenino, utilizando esta aproximación Naive Bayes. La información de tus sospechosos es la siguiente:

			Samples (People)					
			1	2	3	4	5	6
Features	1	Shoe size (cm)	41	43	44	45	37	39
	2	Height (cm)	170	175	185	180	160	170
	3	Max Speed (min/mile)	6	7	6.5	7.5	6.5	7
y		**Sex**	**Male**				**Female**	

Modelo de Naive Bayes (Gaussiano)

Ejemplo

- De lo anterior, el estimador de máxima verosimilitud para el prior sería la fracción de muestras en cada clase,

$$\hat{\mathscr{P}}(y = \text{male}) = \frac{2}{3}$$

$$\hat{\mathscr{P}}(y = \text{female}) = \frac{1}{3}$$

- A continuación, modelamos de manera normal estas características. Es decir, se tratan como variables aleatorias con los siguientes estimadores de máxima verosimilitud de sus medias y varianzas,

	Male		**Female**	
	Mean	Variance	Mean	Variance
Shoe size	43.25	2.9	38	2
Height	177.5	41.7	165	50
Max Speed	6.75	0.42	6.75	0.125

Modelo de Naive Bayes (Gaussiano)

Ejemplo

- Luego, podemos hacer la estimación en base a nuestro nuevo dato $\vec{x} = \left[42,180,5.5\right]$ de acuerdo a esta distribución normal,
- $$\hat{\mathscr{P}}(x_1 \,|\, y = \text{male}) = \frac{1}{\sqrt{2\pi(2.9)}} e^{-\frac{(42 - 43.25)^2}{2(2.9)}} = 0.1787$$
- $$\hat{\mathscr{P}}(x_2 \,|\, y = \text{male}) = \frac{1}{\sqrt{2\pi(41.7)}} e^{-\frac{(180 - 177.5)^2}{2(41.7)}} = 0.0573$$
- $$\hat{\mathscr{P}}(x_3 \,|\, y = \text{male}) = \frac{1}{\sqrt{2\pi(0.42)}} e^{-\frac{(5.5 - 6.75)^2}{2(0.42)}} = 0.0948$$
- Con esto podemos calcular la probabilidad posterior

$$\hat{\mathscr{P}}(y = \text{male} \,|\, \vec{\text{x}}) = \mathscr{P}(\text{y} = \text{male}) \prod_{\text{j}=1}^{\text{D}} \hat{\mathscr{P}}(\text{x}_\text{i} \,|\, \text{y} = \text{male}) = 6.4745 \times 10^4$$

- Y lo mismo podemos hacer para $\hat{\mathscr{P}}(y = \text{female} \,|\, \vec{\text{x}})$ y escoger aquella clase con la probabilidad posterior más grande.