

Inteligencia Artificial

Probabilidad



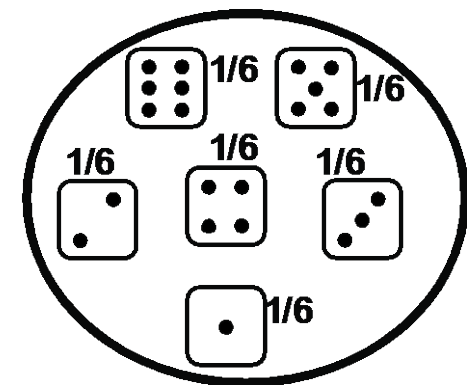
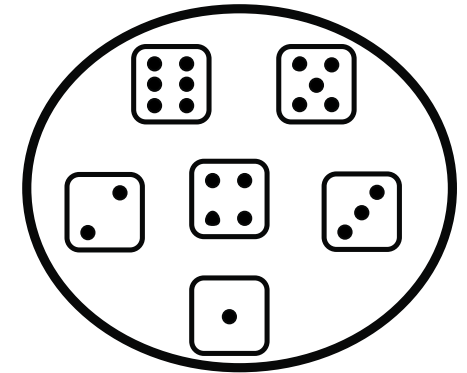
[Transparencias adaptadas de Dan Klein and Pieter Abbeel: CS188 Intro to AI, UC Berkeley (ai.berkeley.edu)]

Incertidumbre

- El mundo real está lleno de incertidumbre
 - P. ej., si salgo hacia Barajas 90 minutos antes de mi vuelo, ¿llegaré en hora?
- Problemas:
 - observabilidad parcial (estado de la carretera, planes de otros coches, etc.)
 - sensores con ruido (informes de tráfico por radio, mapas de Google)
 - inmensa complejidad de modelización y predicción del tráfico, la cola de seguridad, etc.
 - desconocimiento de la dinámica del mundo (¿reventará el neumático?)
- Las afirmaciones probabilísticas resumen los efectos de la **ignorancia y la pereza**
- La teoría de la probabilidad es necesaria para tomar decisiones racionales
 - **Maximizar la utilidad esperada** : $a^* = \operatorname{argmax}_a \sum_s P(s \mid a) U(s)$

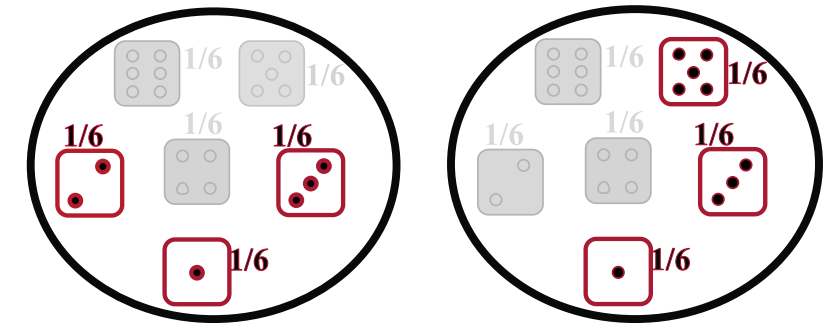
Leyes básicas de la probabilidad (discreta)

- Empezando con un conjunto Ω de mundos posibles
 - P.ej., 6 posibles resultados de la tirada de un dado, $\{1, 2, 3, 4, 5, 6\}$
- Un **modelo de probabilidad** asigna un número $P(\omega)$ a cada mundo ω
 - P.ej., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.
- Estos números deben satisfacer:
 - $0 \leq P(\omega) \leq 1$
 - $\sum_{\omega \in \Omega} P(\omega) = 1$



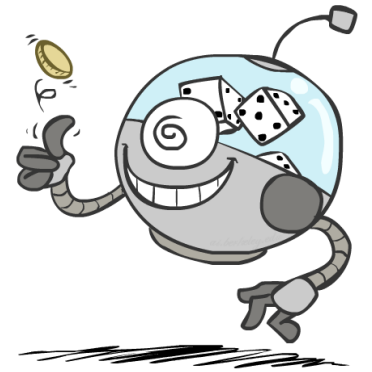
Leyes básicas de la probabilidad (continúa)

- Un **suceso** (o evento) es cualquier subconjunto de Ω
 - P. ej., “tirada < 4” es el conjunto {1,2,3}
 - P. ej., “la tirada es impar” es el conjunto {1,3,5}
- La probabilidad de un suceso es la **suma** de probabilidades sobre sus mundos
 - $P(A) = \sum_{\omega \in A} P(\omega)$
 - P.ej., $P(\text{tirada} < 4) = P(1) + P(2) + P(3) = 1/2$
- De Finetti (1931): cualquiera que apueste según probabilidades que violen estas leyes puede verse obligado a perder dinero en cada serie de apuestas



Variables aleatorias

- Una variable aleatoria (normalmente en mayúscula) es un aspecto del mundo sobre el que no tenemos certeza.
- Formalmente, una **función determinista** de ω
- El **rango** de una variable aleatoria es el conjunto de posibles valores:
 - $Impar = \text{¿La tirada de dados es un número impar?} \rightarrow \{\text{verdadero, falso}\}$
 - P.ej. $Impar(1) = \text{verdadero}$, $Impar(6) = \text{falso}$
 - A menudo, se escribe el suceso $Impar=\text{verdadero}$ como $impar$, $Impar=\text{falso}$ como $\neg impar$
 - $T = \text{¿Hace calor o frío?} \rightarrow \{\text{calor, frío}\}$
 - $D = \text{¿Cuánto se tarda en llegar al aeropuerto?} \rightarrow [0, \infty)$
 - $L_{Ghost} = \text{¿Dónde está el fantasma?} \rightarrow \{(0,0), (0,1), \dots\}$
- La **distribución de probabilidad** de una variable aleatoria X da la probabilidad para cada valor x en su rango (probabilidad del suceso $X=x$)
 - $P(X=x) = \sum_{\{\omega: X(\omega)=x\}} P(\omega)$
 - $P(x)$ para abreviar (cuando no es ambiguo)
 - $P(X)$ se refiere a toda la distribución (considérala un vector o una tabla)

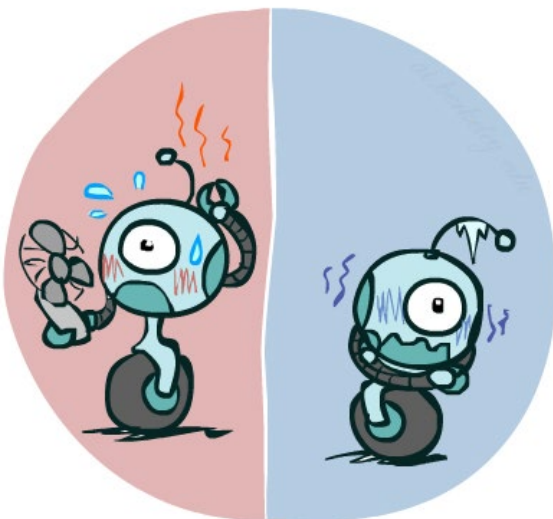


Distribuciones de probabilidad

- Asociar una probabilidad con cada valor; la suma tiene que ser 1
 - Temperatura:
 - Tiempo:
 - Distribución conjunta***

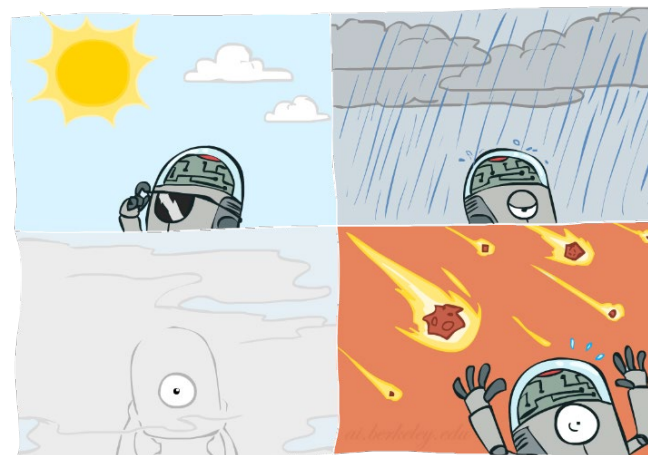
$P(T)$

T	P
calor	0,5
frío	0,5



$P(W)$

W	P
soleado	0,6
lluvioso	0,1
niebla	0,3
meteoritos	0,0



$P(T,W)$

		Temperatura	
		calor	frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

Distribuciones marginales

Distribuciones de probabilidad (continúa)

- Normalmente, de las distribuciones marginales no puede sacarse la distribución conjunta
 - En ésta, se recoge información adicional: la correlación entre variables.
 - Ej.: si hace calor, es más probable que esté soleado
- De la distribución conjunta, sí pueden sacarse las marginales.
 - Sólo hay que sumar (lo vamos a ir viendo)

Creando mundos posibles

- En muchos casos:
 - empezamos con variables aleatorias y sus dominios
 - construimos mundos posibles como asignaciones de valores a todas las variables
- P. ej., dos tiradas del dado *Tirada₁* y *Tirada₂*
 - ¿Cuántos mundos posibles?
 - ¿Cuáles son sus probabilidades?
- ¿Tamaño de distribución para *n* variables con rango de tamaño *d*? d^n
- Ninguna distribución, salvo las más pequeñas, **se puede escribir a mano**

Probabilidades de sucesos

- Recuerda que la probabilidad de un suceso es la suma de las probabilidades de sus mundos:

- $P(A) = \sum_{\omega \in A} P(\omega)$

- Así, dada una distribución conjunta sobre todas las variables, **se puede calcular la probabilidad de cualquier suceso**

- ¿Probabilidad de que haga calor Y haga sol?
 - ¿Probabilidad de que haga calor?
 - ¿Probabilidad de que haga calor O no haya niebla?

- *Distribución conjunta*

$P(T,W)$

		Temperatura	
		calor	Frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

Distribuciones marginales

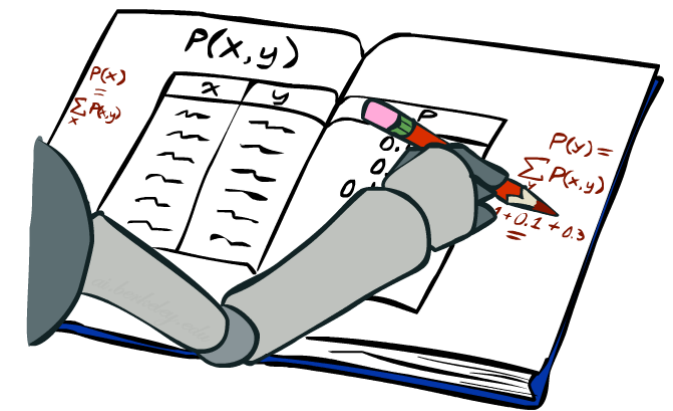
- Las distribuciones marginales son subtablas que eliminan variables
- Marginalización:** Colapsar una dimensión sumando

$$P(X=x) = \sum_y P(X=x, Y=y)$$

		Temperatura		
		calor	frío	
Tiempo	soleado	0,45	0,15	0,60
	lluvioso	0,02	0,08	0,10
	niebla	0,03	0,27	0,30
	meteor.	0,00	0,00	0,00
		0,50	0,50	

$P(T)$

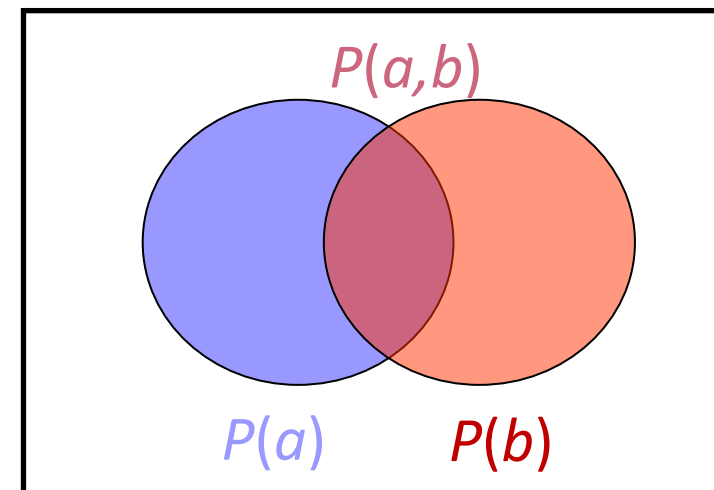
$P(W)$



Probabilidades condicionales

- “Probabilidad de **a** dado **b**”

$$P(a \mid b) = \frac{P(a, b)}{P(b)}$$



$P(T, W)$

		Temperatura	
		calor	frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

$$P(W=s \mid T=f) = \frac{P(W=s, T=f)}{P(T=f)} = 0.15 / 0.50 = 0.3$$

$$= P(W=s, T=f) + P(W=l, T=f) + P(W=n, T=f) + P(W=m, T=f) \\ = 0,15 + 0,08 + 0,27 + 0,00 = 0,50$$

Distribuciones condicionales

- Distribuciones para un conjunto de variables dado otro conjunto

$$P(T,W)$$

		Temperatura	
		calor	frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

$$P(W \mid T=c)$$

calor

0,90
0,04
0,06
0,00

$$P(W \mid T=f)$$

frío

0,30
0,16
0,54
0,00

$$P(W \mid T)$$

calor

frío

0,90	0,30
0,04	0,16
0,06	0,54
0,00	0,00

Normalizando una distribución

- Hacer que las probabilidades sumen 1
- Procedimiento:
 - Multiplicar cada entrada por $\alpha = 1/(\text{suma sobre todas las entradas})$

$P(T,W)$

		Temperatura	
		calor	frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

$P(W, T=f)$

0,15
0,08
0,27
0,00

Normalizar
 $\alpha = 1/0,50 = 2$

0,30
0,16
0,54
0,00

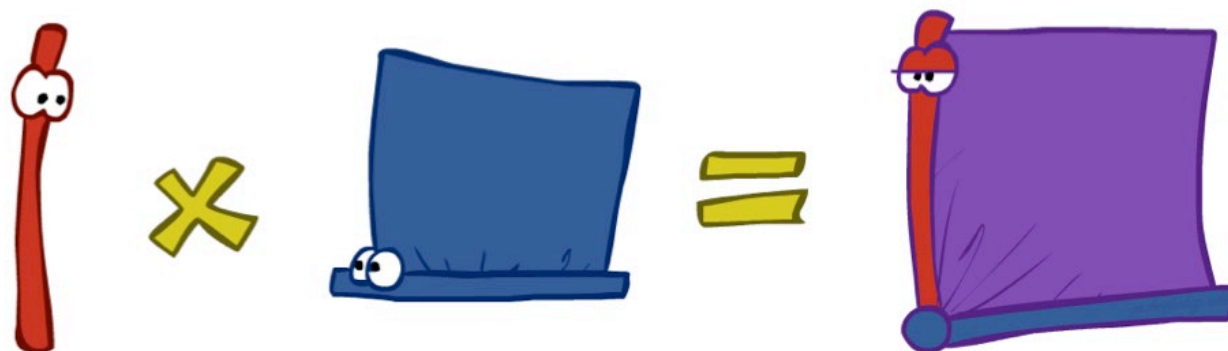
$P(W | T=f) = P(W, T=f) / P(T=f)$
 $= \alpha P(W, T=f)$

- Es otra notación (más concisa) para el denominador de las probabilidades condicionales

La regla del producto

- A veces tenemos distribuciones condicionales, pero queremos la conjunta

$$P(a \mid b) P(b) = P(a, b) \quad \longleftrightarrow \quad P(a \mid b) = \frac{P(a, b)}{P(b)}$$



La regla del producto: ejemplo

$$P(W \mid T) P(T) = P(W, T)$$

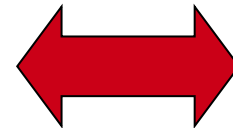
$P(W \mid T)$

calor frío

0,90	0,30
0,04	0,16
0,06	0,54
0,00	0,00

$P(T)$

T	P
calor	0,5
frío	0,5



$P(W, T)$

		Temperatura	
		calor	frío
Tiempo	soleado	0,45	0,15
	lluvioso	0,02	0,08
	niebla	0,03	0,27
	meteor.	0,00	0,00

La regla de la cadena

- Una distribución conjunta puede escribirse como un producto de distribuciones condicionales mediante la aplicación repetida de la regla del producto:
- $P(x_1, x_2, x_3) = P(x_3 \mid x_1, x_2) P(x_1, x_2) = P(x_3 \mid x_1, x_2) P(x_2 \mid x_1) P(x_1)$
- $P(x_1, x_2, \dots, x_n) = \prod_i P(x_i \mid x_1, \dots, x_{i-1})$

Inferencia probabilística

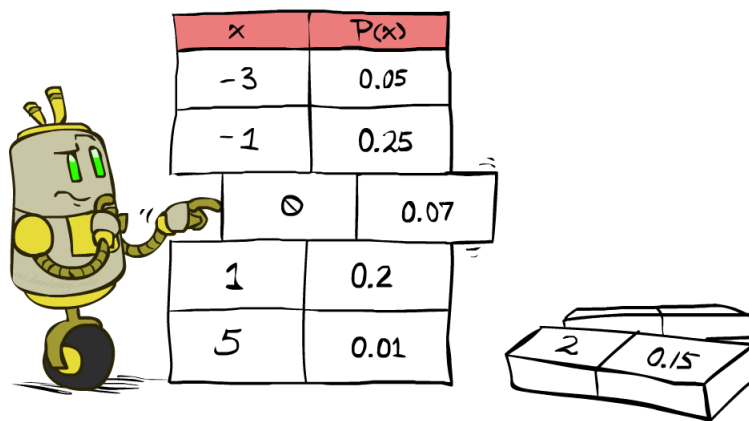
- Inferencia probabilística: calcular una probabilidad deseada a partir de un modelo de probabilidad
 - Típicamente para una **variable de consulta** dada la **evidencia**
 - P.ej., $P(\text{aeropuerto a tiempo} \mid \text{no accidentes}) = 0.90$
 - Representan las creencias del agente dadas las pruebas
- Las probabilidades cambian con nuevas evidencias:
 - $P(\text{aeropuerto a tiempo} \mid \text{no accidentes}, 5 \text{ a.m.}) = 0.95$
 - $P(\text{aeropuerto a tiempo} \mid \text{no accidentes}, 5 \text{ a.m., lloviendo}) = 0.80$
 - La observación de nueva evidencia hace que **se actualicen las creencias**



Inferencia por enumeración

- Dado un modelo de probabilidad $P(X_1, \dots, X_n)$
- Particiona las variables X_1, \dots, X_n en conjuntos como sigue:
 - Variables de evidencia: $E = e$
 - Variables de consulta: Q
 - Variables ocultas: H
 - Queremos: $P(Q \mid e)$

- Paso 1: Seleccionar las entradas coherentes con las evidencias

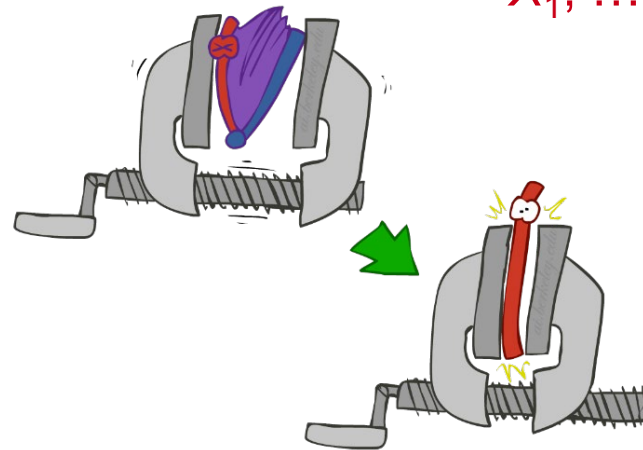


x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01

2 0.15

- Paso 2: Marginalizar H del modelo para obtener la unión de la consulta y la evidencia

$$P(Q, e) = \sum_h \underbrace{P(Q, h, e)}_{X_1, \dots, X_n}$$



- Paso 3: Normalizar

$$P(Q \mid e) = \alpha P(Q, e)$$

Inferencia por enumeración

- ¿ $P(W)$?

Estación	Temp	Tiempo	P
verano	calor	sol	0,30
verano	calor	lluvia	0,04
verano	calor	niebla	0,01
verano	calor	meteor.	0,00
verano	frío	sol	0,10
verano	frío	lluvia	0,02
verano	frío	niebla	0,03
verano	frío	meteor.	0,00
invierno	calor	sol	0,10
invierno	calor	lluvia	0,02
invierno	calor	niebla	0,03
invierno	calor	meteor	0,00
invierno	frío	sol	0,15
invierno	frío	lluvia	0,12
invierno	frío	niebla	0,08
invierno	frío	meteor.	0,00

Inferencia por enumeración

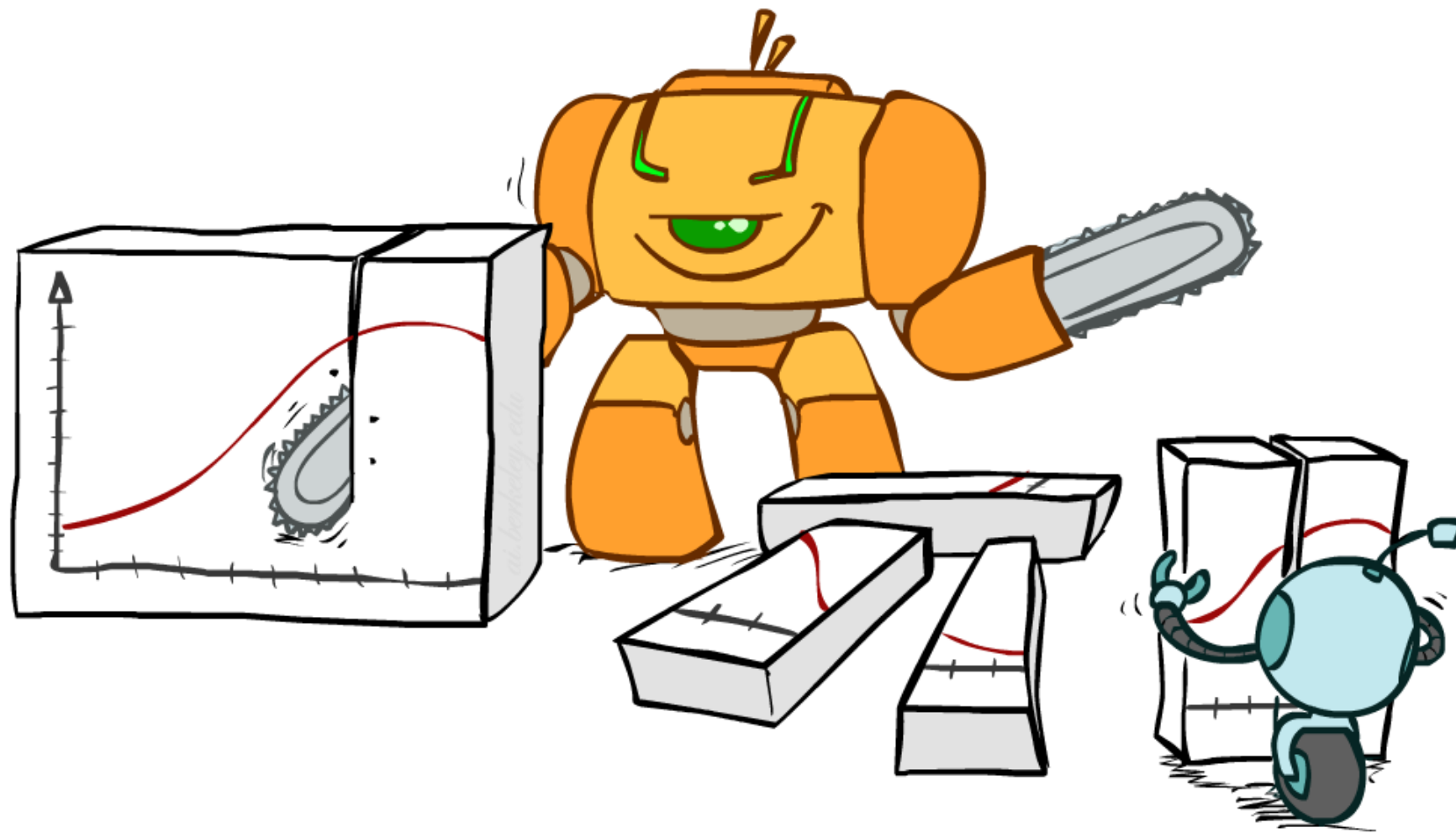
- ¿ $P(W)$?
- ¿ $P(W \mid \text{invierno})$?

Estación	Temp	Tiempo	P
verano	calor	sol	0,30
verano	calor	lluvia	0,04
verano	calor	niebla	0,01
verano	calor	meteor.	0,00
verano	frío	sol	0,10
verano	frío	lluvia	0,02
verano	frío	niebla	0,03
verano	frío	meteor.	0,00
invierno	calor	sol	0,10
invierno	calor	lluvia	0,02
invierno	calor	niebla	0,03
invierno	calor	meteor.	0,00
invierno	frío	sol	0,15
invierno	frío	lluvia	0,12
invierno	frío	niebla	0,08
invierno	frío	meteor.	0,00

Inferencia por enumeración

- Problemas a simple vista:
 - Complejidad temporal en el peor de los casos $O(d^n)$ (exponencial en el número de variables ocultas)
 - Complejidad espacial (memoria) $O(d^n)$ para almacenar la distribución conjunta
 - $O(d^n)$ puntos de datos para estimar las entradas de la distribución conjunta

Regla de Bayes



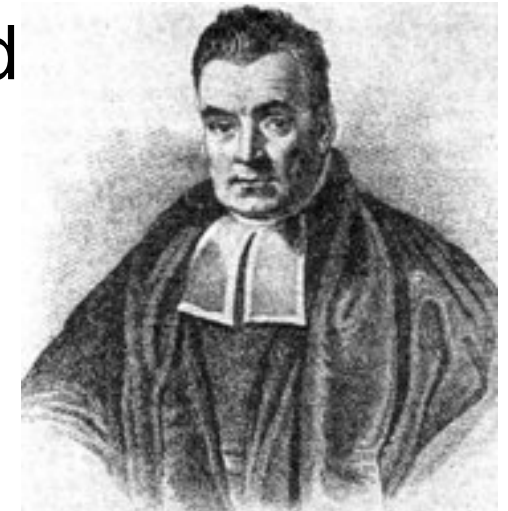
Regla de Bayes

- Si escribimos la regla del producto en ambos sentidos

$$P(a | b) P(b) = P(a, b) = P(b | a) P(a)$$

- Dividiendo las expresiones izquierda y derecha:

$$P(a | b) = \frac{P(b | a) P(a)}{P(b)}$$



- ¿Y esto sirve para algo?
 - Nos permite obtener una condicional a partir de su inversa
 - A menudo, una de las condiciones es complicada (causa dado el efecto), pero la otra es sencilla (efecto dada la causa)
 - Describe un paso de "actualización" de la probabilidad a priori o incondicional $P(a)$ a la probabilidad a posteriori o condicional $P(a | b)$
 - Por tanto, proporciona una teoría sencilla y formal del aprendizaje

Inferencia con la regla de Bayes

- Ejemplo: Probabilidad diagnóstica a partir de probabilidad causal:

$$P(\text{causa} \mid \text{efecto}) = \frac{P(\text{efecto} \mid \text{causa}) P(\text{causa})}{P(\text{efecto})}$$

- Ejemplo:

- M: meningitis, T: tortícolis

$$\left. \begin{array}{l} P(t \mid m) = 0.8 \\ P(m) = 0.0001 \\ P(t) = 0.01 \end{array} \right\} \text{Ejemplos dados}$$

$$P(m \mid t) = \frac{P(t \mid m) P(m)}{P(t)} = \frac{0.8 \times 0.0001}{0.01}$$

- Nota: la probabilidad a posteriori de meningitis sigue siendo muy pequeña: 0,008 (80 veces mayor, ¿por qué?)

Independencia

- Dos variables X e Y son (absolutamente) *independientes* si:

$$\forall x, y \quad P(x, y) = P(x) P(y)$$

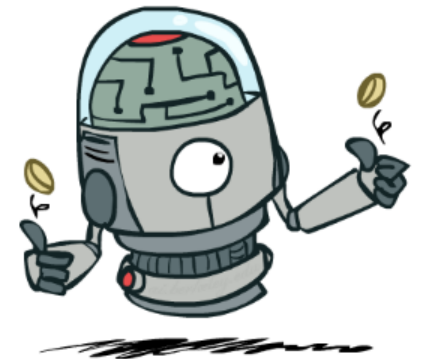
- Es decir, la distribución conjunta se factoriza en un producto de dos distribuciones más simples
- De forma equivalente, mediante la regla del producto

$$P(x, y) = P(x | y) P(y),$$

$$P(x | y) = P(x) \quad \text{o} \quad P(y | x) = P(y)$$

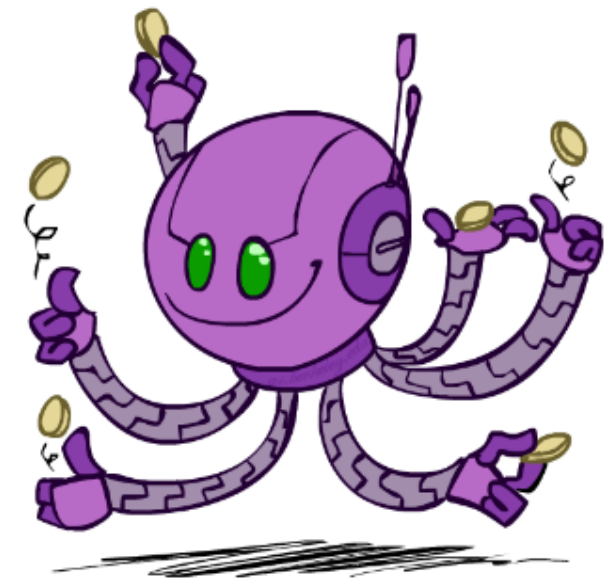
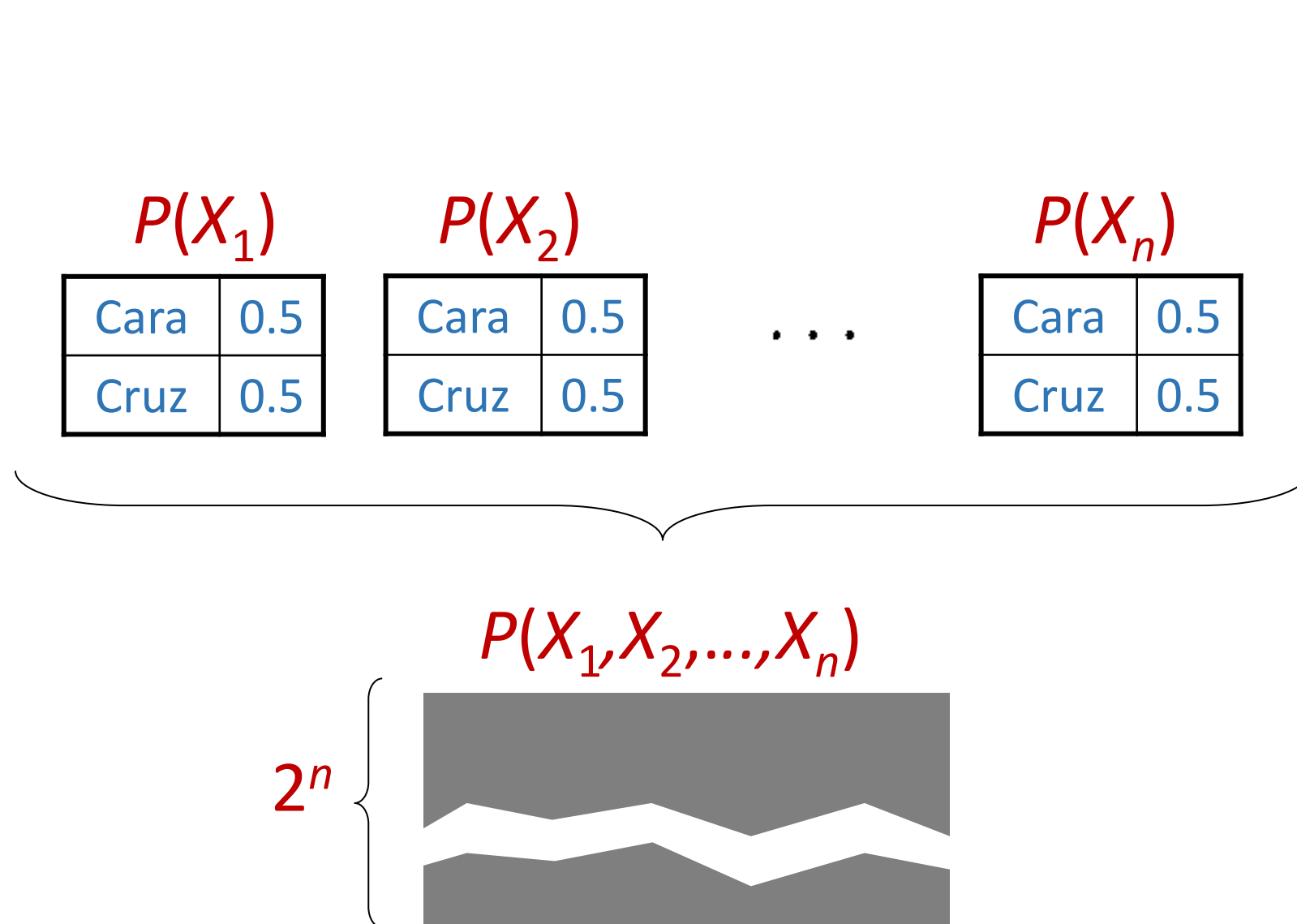
- Ejemplo: dos tiradas de dados *Tirada₁* y *Tirada₂*

- $P(\text{Tirada}_1=5, \text{Tirada}_2=3) = P(\text{Tirada}_1=5) P(\text{Tirada}_2=3) = 1/6 \times 1/6 = 1/36$
- $P(\text{Tirada}_2=3 | \text{Tirada}_1=5) = P(\text{Tirada}_2=3)$



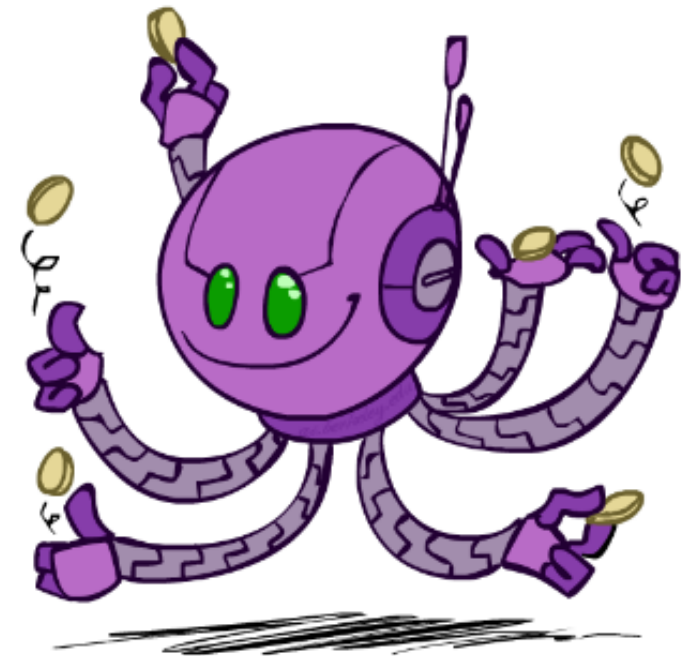
Ejemplo: independencia

- n tiradas de moneda al aire, sin truco e independientes

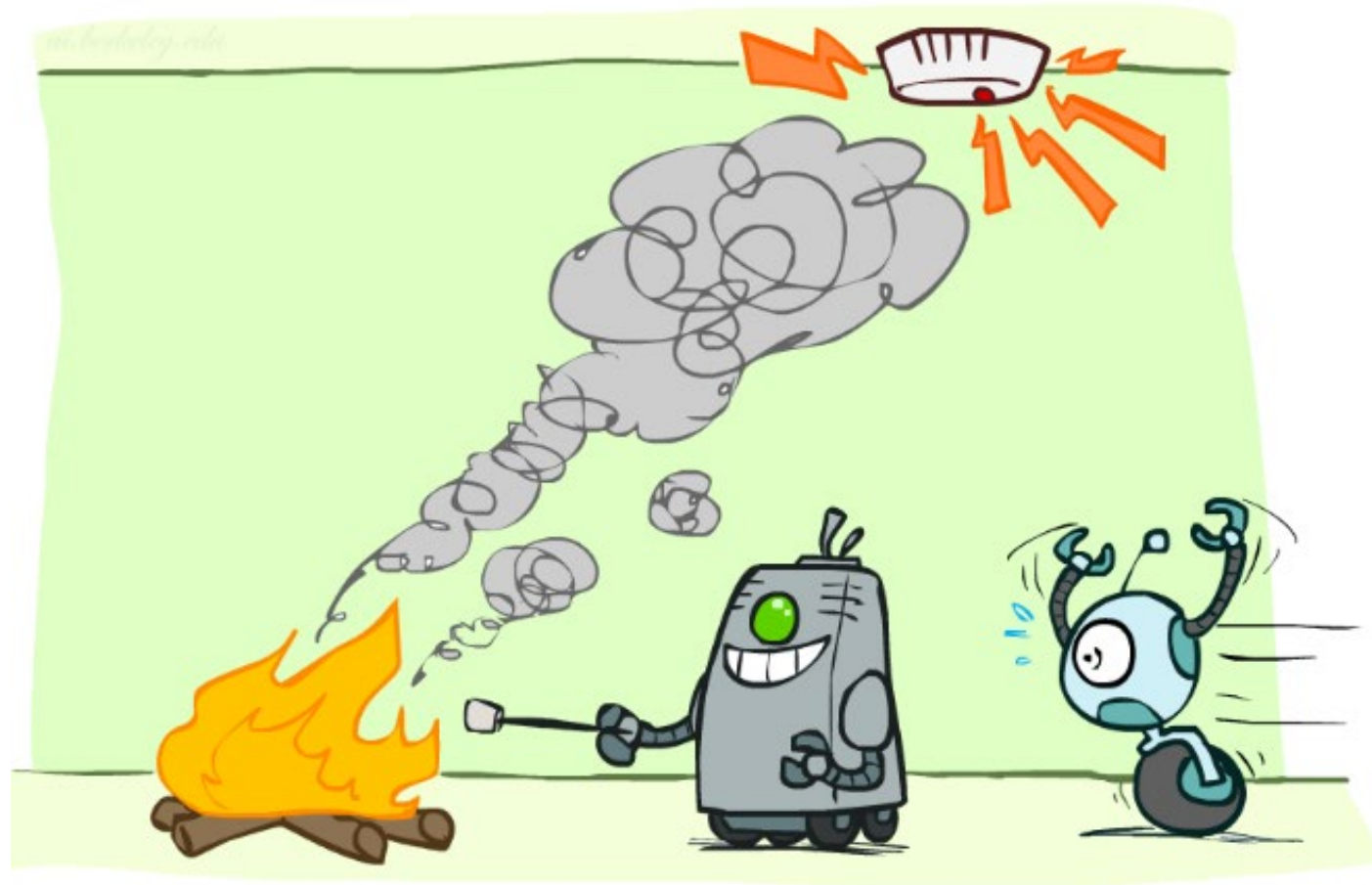


Independencia (continúa)

- La independencia es increíblemente poderosa
 - Reducción exponencial en el tamaño de representación
- La independencia también es extremadamente rara
- Pero la independencia **condicional** está en todas partes

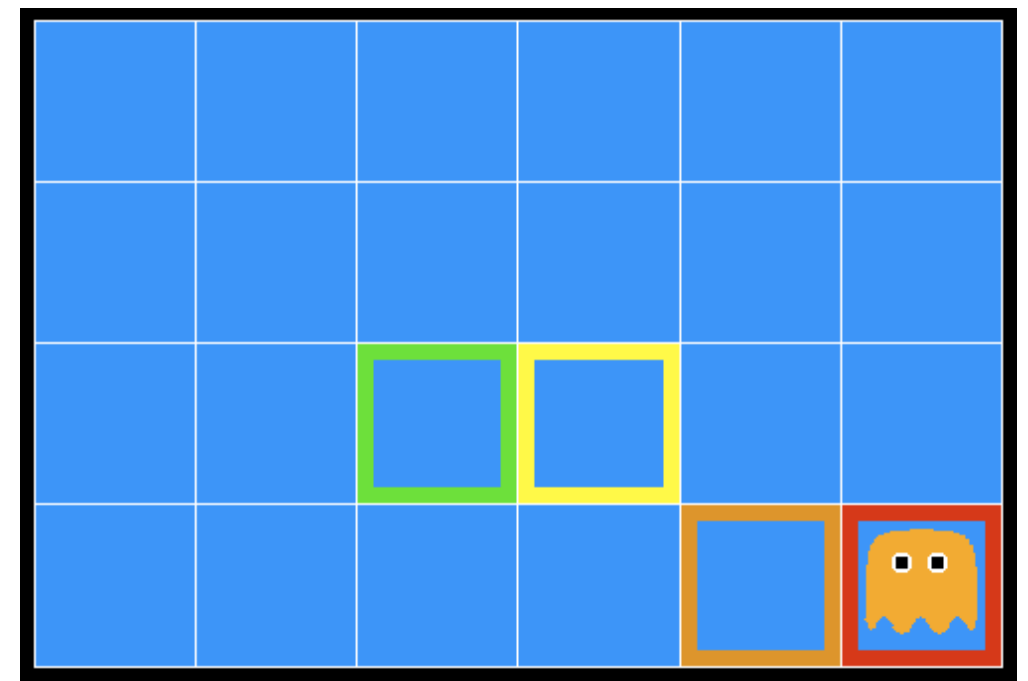


Independencia condicional



Cazafantasmas

- Un fantasma está en el tablero en alguna parte
- Las lecturas de los sensores indican lo cerca que está un cuadrado del fantasma
 - En el fantasma: normalmente rojo
 - 1 o 2 de distancia: principalmente naranja
 - 3 o 4 de distancia: normalmente amarillo
 - A más de 5 años: a menudo verde
- Ir midiendo el color de las casillas hasta estar seguros de la ubicación y, entonces, “disparar” a esa casilla



Demostración Cazafantasmas con probabilidad




Modelo Cazafantasmas

- Variables y rangos:
 - G (localización del fantasma) en $\{(1,1), \dots, (3,3)\}$
 - $C_{x,y}$ (color medido en la casilla x,y) en $\{\text{rojo}, \text{naranja}, \text{amarillo}, \text{verde}\}$
- Físicas del Cazafantasmas (*Ghostbuster physics*)
 - **Distribución uniforme de probabilidades a priori** sobre la localización del fantasma: $P(G)$
 - **Modelo del sensor.** $P(C_{x,y} \mid G)$ (depende solo de la distancia a G)
 - P. ej.: $P(C_{1,1} = \text{amarillo} \mid G = (1,1)) = 0.1$

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

Modelo Cazafantasmas (continúa)

- $P(G, C_{1,1}, \dots, C_{3,3})$ tiene $9 \times 4^9 = 2.359.296$ de entradas en la tabla
- Independencia en el Cazafantasmas:
 - ¿Son $C_{1,1}$ y $C_{1,2}$ independientes?
 - P.ej., ¿es $P(C_{1,1} = \text{amarillo}) = P(C_{1,1} = \text{amarillo} \mid C_{1,2} = \text{naranja})$?
- Física del Cazafantasmas otra vez:
 - $P(C_{x,y} \mid G)$ depende solo de la distancia a G
 - Así que $P(C_{1,1} = \text{amarillo} \mid \underline{G = (2,3)}) = P(C_{1,1} = \text{amarillo} \mid \underline{G = (2,3)}, C_{1,2} = \text{naranja})$
 - Es decir, $C_{1,1}$ es **condicionalmente independiente** de $C_{1,2}$ dada G

0.11		0.11
0.11	0.11	0.11
0.11	0.11	0.11

Modelo Cazafantasmas (continúa)

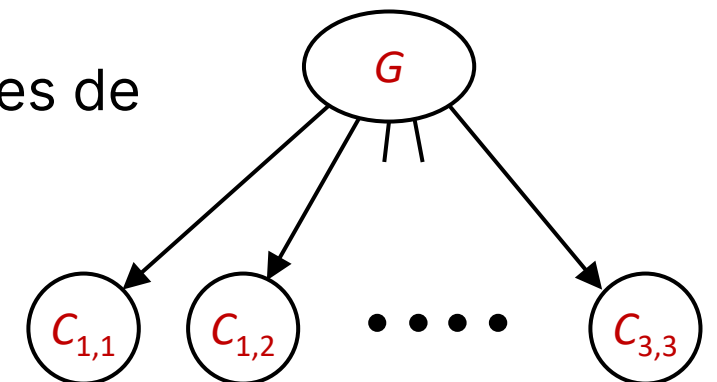
- Aplicar la regla de la cadena para descomponer el modelo de probabilidad conjunta:

$$P(G, C_{1,1}, \dots, C_{3,3}) = P(G) P(C_{1,1} \mid G) P(C_{1,2} \mid G, C_{1,1}) P(C_{1,3} \mid G, C_{1,1}, C_{1,2}) \dots P(C_{3,3} \mid G, C_{1,1}, \dots, C_{3,2})$$

- Ahora simplificar utilizando la independencia condicional:

$$P(G, C_{1,1}, \dots, C_{3,3}) = P(G) P(C_{1,1} \mid G) P(C_{1,2} \mid G) P(C_{1,3} \mid G) \dots P(C_{3,3} \mid G)$$

- Es decir, las propiedades de independencia condicional de la física de los cazafantasmas simplifican el modelo de probabilidad de **exponencial a cuadrático** en el número de casillas
- Esto se denomina modelo de Bayes ingenuo (**Naive Bayes**):
 - Una variable de consulta discreta (a menudo denominada variable de **clase** o **categoría**)
 - Todas las demás variables son (potencialmente) variables de evidencia
 - Las variables de evidencia son todas condicionalmente independientes dada la variable de consulta



Independencia condicional

- La **independencia condicional** es nuestra forma más básica y robusta de conocimiento sobre entornos inciertos
- X es condicionalmente independiente de Y dado Z si y solo si:

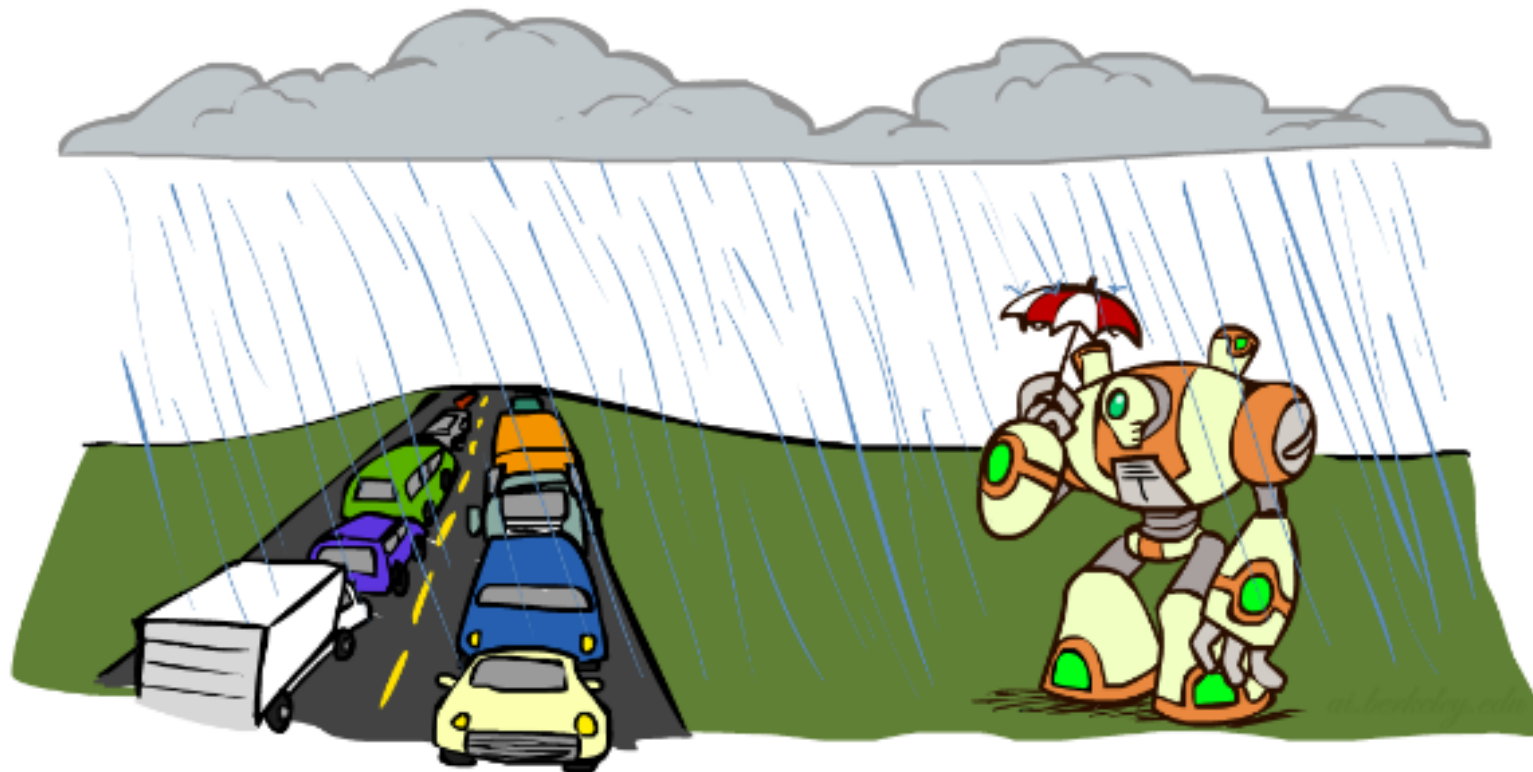
$$\forall x, y, z \quad P(x \mid y, z) = P(x \mid z)$$

o, equivalentemente, si y solo si

$$\forall x, y, z \quad P(x, y \mid z) = P(x \mid z) P(y \mid z)$$

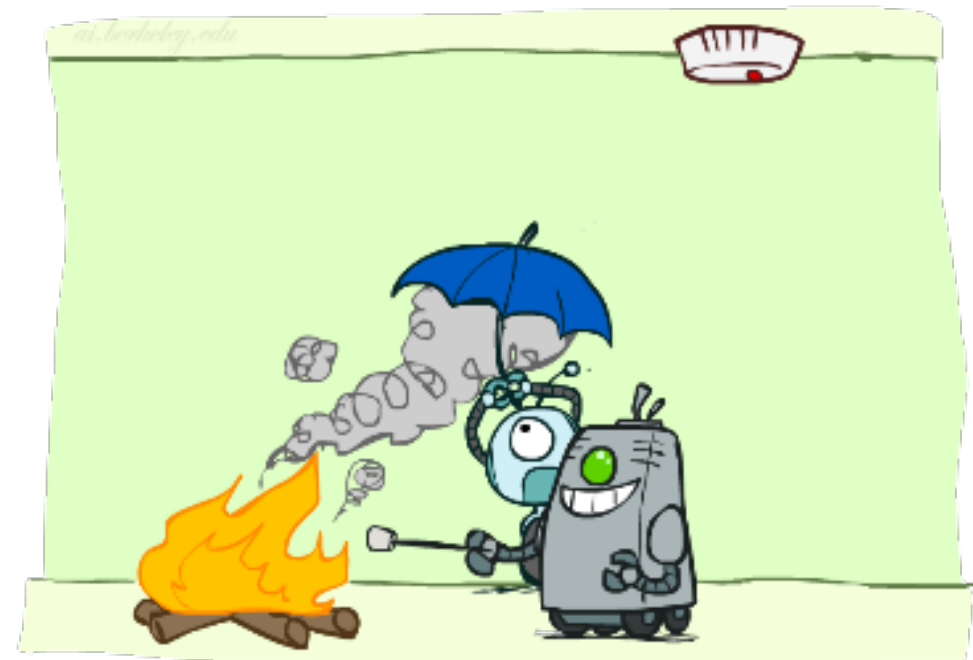
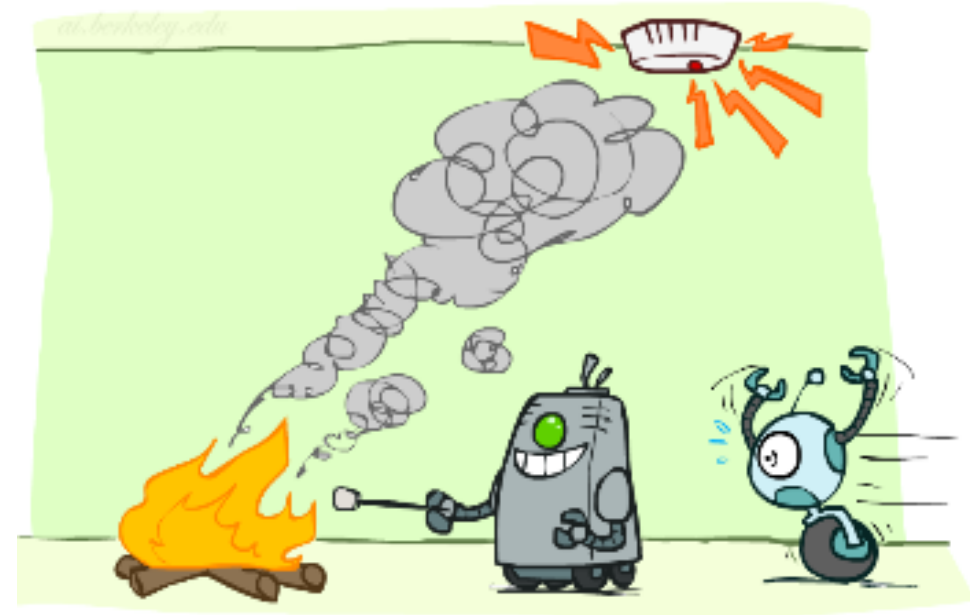
Independencia condicional

- ¿Qué pasa en este dominio?
 - Atasco
 - Paraguas
 - Lluvia



Independencia condicional

- ¿Y en este dominio?
 - Fuego
 - Humo
 - Alarma



Resumen: Probabilidad elemental

- Leyes básicas: $0 \leq P(\omega) \leq 1$ $\sum_{\omega \in \Omega} P(\omega) = 1$
- Sucesos: subconjuntos de Ω : $P(A) = \sum_{\omega \in A} P(\omega)$
- La variable aleatoria $X(\omega)$ tiene un valor en cada ω
 - La distribución $P(X)$ da la probabilidad para cada valor posible x
 - La distribución conjunta $P(X,Y)$ da la probabilidad total por cada combinación x,y
- Marginalización: $P(X=x) = \sum_y P(X=x,Y=y)$
- Probabilidad condicional: $P(X|Y) = P(X,Y)/P(Y)$
- Regla del producto: $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
 - Generalización a la regla de la cadena: $P(X_1, \dots, X_n) = \prod_i P(X_i | X_1, \dots, X_{i-1})$