

# Razonamiento Probabilista

# Aprendizaje en redes de Bayes

Verónica E. Arriola-Rios

Facultad de Ciencias, UNAM

10 de agosto de 2021



# Introducción

- 1 Introducción
- 2 Máxima Verosimilitud
- 3 Estimación Bayesiana
- 4 Aprendizaje de la estructura
- 5 Datos incompletos

# Temas

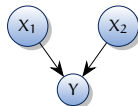
- 1 Introducción
  - Planteamiento
  - Tipos de problemas de aprendizaje

# Aprendizaje

- Se considera que existe una distribución de probabilidad verdadera  $P^*$ .

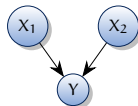
# Aprendizaje

- Se considera que existe una distribución de probabilidad verdadera  $P^*$ .
- Se aprenderá un Modelo Gráfico Probabilístico (MGP)  $\mathcal{M}^*$ , que podría modelar la distribución.



# Aprendizaje

- Se considera que existe una distribución de probabilidad verdadera  $P^*$ .
- Se aprenderá un Modelo Gráfico Probabilístico (MGP)  $\mathcal{M}^*$ , que podría modelar la distribución.



- Para ello se utilizan:
  - 1 Un conjunto de muestras  $D = \{d[1], \dots, d[M]\}$  obtenidas de  $P^*$ .
  - 2 El conocimiento de algún experto.

# ¿Por qué usar aprendizaje sobre MGP?

- Realiza predicciones sobre *objetos estructurados* (secuencias, gráficas, árboles).  
Permite explotar correlaciones entre varias variables predichas.
- Permite incorporar *conocimiento previo* en el modelo.
- Se puede aprender *un solo modelo* para varias tareas.
- Es un marco de trabajo para el *descubrimiento de conocimiento*.

# Conjuntos para el entrenamiento

El conjunto de muestras  $D = \{d[1], \dots, d[M]\}$  para el entrenamiento debe ser dividido en subconjuntos:

- Entrenamiento.
- Pruebas.
- (Opcionalmente) Validación.

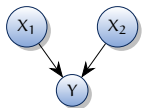


# Temas

- 1 Introducción
  - Planteamiento
  - Tipos de problemas de aprendizaje

# Tipos de problemas de aprendizaje

- Estructura conocida, datos completos.

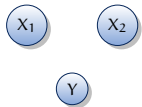


Muestras			
	$X_1$	$X_2$	$Y$
$d[0]$	$x_1^0$	$x_2^0$	$y^0$
$d[1]$	$x_1^1$	$x_2^1$	$y^1$
...			
$d[M]$	$x_1^M$	$x_2^M$	$y^M$



$X_1$	$P(X_1)$	$X_2$	$P(X_2)$	$X_1$	$X_2$	$Y$	$P(Y X_1, X_2)$
$x_1$		$x_2$		$x_1$	$x_2$	$y$	
$x'_1$		$x'_2$		$x_1$	$x_2$	$y'$	
...		...		...			
$x_1^n$		$x_2^n$		$x'_1$	$x'_2$	$y'$	

- Estructura desconocida, datos completos.



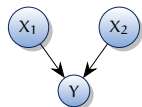
Muestras			
	$X_1$	$X_2$	$Y$
$d[0]$	$x_1^0$	$x_2^0$	$y^0$
$d[1]$	$x_1^1$	$x_2^1$	$y^1$
...			
$d[M]$	$x_1^M$	$x_2^M$	$y^M$



$X_1$	$P(X_1)$	$X_2$	$P(X_2)$	$X_1$	$X_2$	$Y$	$P(Y X_1, X_2)$
$x_1$		$x_2$		$x_1$	$x_2$	$y$	
$x'_1$		$x'_2$		$x_1$	$x_2$	$y'$	
...		...		...			
$x_1^n$		$x_2^n$		$x'_1$	$x'_2$	$y'$	

# Tipos de problemas de aprendizaje

- Estructura conocida, datos incompletos.

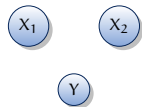


Muestras			
	$X_1$	$X_2$	$Y$
d[0]	?	$x_2^0$	$y^0$
d[1]	$x_1^1$	?	$y^1$
d[2]	?	$x_2^2$	$y^2$
...			
d[M]	$x_1^M$	$x_2^M$	?

 $\Rightarrow$ 

$X_1$	$P(X_1)$	$X_2$	$P(X_2)$	$X_1$	$X_2$	$Y$	$P(Y X_1, X_2)$
$x_1$		$x_2$		$x_1$	$x_2$	$y$	
$x_1'$		$x_2'$		$x_1$	$x_2$	$y'$	
...		...		...			
$x_1^n$		$x_2^n$		$x_1'$	$x_2'$	$y'$	

- Estructura desconocida, datos incompletos.



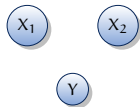
Muestras			
	$X_1$	$X_2$	$Y$
d[0]	?	$x_2^0$	$y^0$
d[1]	$x_1^1$	?	$y^1$
d[2]	?	$x_2^2$	$y^2$
...			
d[M]	$x_1^M$	$x_2^M$	?

 $\Rightarrow$ 

$X_1$	$P(X_1)$	$X_2$	$P(X_2)$	$X_1$	$X_2$	$Y$	$P(Y X_1, X_2)$
$x_1$		$x_2$		$x_1$	$x_2$	$y$	
$x_1'$		$x_2'$		$x_1$	$x_2$	$y'$	
...		...		...			
$x_1^n$		$x_2^n$		$x_1'$	$x_2'$	$y'$	

# Tipos de problemas de aprendizaje

- Variables latentes, datos incompletos.



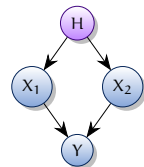
Muestras			
	X <sub>1</sub>	X <sub>2</sub>	Y
d[0]	?	x <sub>2</sub> <sup>0</sup>	y <sup>0</sup>
d[1]	x <sub>1</sub> <sup>1</sup>	?	y <sup>1</sup>
d[2]	?	x <sub>2</sub> <sup>2</sup>	y <sup>2</sup>
...			
d[M]	x <sub>1</sub> <sup>M</sup>	x <sub>2</sub> <sup>M</sup>	?

⇒

X <sub>1</sub>	P(X <sub>1</sub> )	X <sub>2</sub>	P(X <sub>2</sub> )
x <sub>1</sub>		x <sub>2</sub>	
x' <sub>1</sub>		x' <sub>2</sub>	
...		...	
x <sub>1</sub> <sup>n</sup>		x <sub>2</sub> <sup>n</sup>	

X <sub>1</sub>	X <sub>2</sub>	Y	P(Y X <sub>1</sub> , X <sub>2</sub> )
x <sub>1</sub>	x <sub>2</sub>	y	
x <sub>1</sub>	x <sub>2</sub>	y'	
...			
x' <sub>1</sub>	x' <sub>2</sub>	y'	



# Máxima Verosimilitud

- 1 Introducción
- 2 Máxima Verosimilitud
- 3 Estimación Bayesiana
- 4 Aprendizaje de la estructura
- 5 Datos incompletos

# Temas

- 2 Máxima Verosimilitud
  - Verosimilitud
  - Distribución de Bernoulli
  - Distribución Gaussiana
  - En redes Bayesianas

# Verosimilitud

## Definición

La *verosimilitud* es la probabilidad de haber observado un conjunto de datos, dado un modelo.

$$L(\mathcal{M} : D) = P(D|\mathcal{M}) = \prod_M P(d[m] : \mathcal{M}) \quad (1)$$

- La forma de aprendizaje que busca encontrar los parámetros para el modelo que maximicen esta cantidad se conoce como *máxima verosimilitud*.

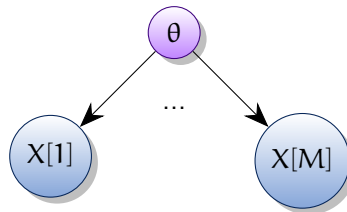
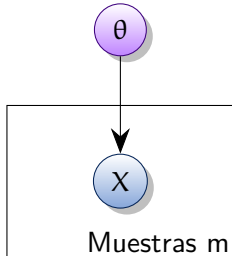
# Temas

- 2 Máxima Verosimilitud
  - Verosimilitud
  - Distribución de Bernoulli
  - Distribución Gaussiana
  - En redes Bayesianas



# Ejemplo: Distribución de Bernoulli

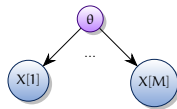
Para una moneda cargada:



$$P(X[m]|\theta) = \begin{cases} \theta & \text{si } X[m] = \text{Sol}^1 \\ 1 - \theta & \text{si } X[m] = \text{Águila}^0 \end{cases}$$

Meta: encontrar  $\theta$  tal que *prediga bien* los resultados de futuros experimentos D.

# Ejemplo: Distribución de Bernouli (Solución)



Se busca maximizar la verosimilitud de las muestras obtenidas:

$$L(\theta : D) = P(D|\theta) = \prod_{m=1}^M P(X[m] : \theta) \quad (2)$$

Por ejemplo:

$$\begin{aligned} L(\theta : < S, A, A, S, S >) &= P(S|\theta)P(A|\theta)P(A|\theta)P(S|\theta)P(S|\theta) \\ &= \theta(1 - \theta)(1 - \theta)\theta\theta \\ &= \theta^3(1 - \theta)^2 \end{aligned} \quad (3)$$

La máxima verosimilitud está dada por:

$$\max(L(\theta : < S, A, A, S, S >)) = \theta^3(1 - \theta)^2 \quad (4)$$

En general:

$$\max(L(\theta : D)) = \theta^{NS}(1 - \theta)^{NA} \quad (5)$$

El máximo de la verosimilitud es también el máximo del *logaritmo de la verosimilitud* llamada, por comodidad, **log-verosimilitud**:

$$\max(\log(L(\theta : D))) = NS \log \theta + NA \log(1 - \theta) \quad (6)$$

Utilizando cálculo se obtiene:

$$\theta = \frac{NS}{NS + NA} = \frac{NS}{N} \quad (7)$$

# Máxima verosimilitud para una distribución multivaluada

- Si la variable  $X$  puede tomar  $k$  valores distintos y  $M_i$  representa el número de muestras donde  $X = x_i$ :

$$L(\theta : D) = \prod_{i=1}^k \theta_i^{M_i} \quad (8)$$

con  $\theta_i$  la probabilidad de obtener el  $i$ -ésimo valor y  $\sum \theta_i = 1$ .

- Entonces la probabilidad aprendida para cada valor posible de  $X$  es:

$$\theta_i = \frac{M_i}{M} \quad (9)$$

con  $M$  el número total de muestras.

# Temas

- 2 Máxima Verosimilitud
  - Verosimilitud
  - Distribución de Bernoulli
  - Distribución Gaussiana
  - En redes Bayesianas

# Máxima verosimilitud para una distribución Gaussiana

Si la distribución de probabilidad se describe con la función normal:

$$P(x) \sim N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} \quad (10)$$

Entonces la verosimilitud viene dada por:

$$L(\mu, \sigma^2; x_1, \dots, x_n) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2} \quad (11)$$

El máximo de la verosimilitud es también el máximo del logaritmo de la verosimilitud:

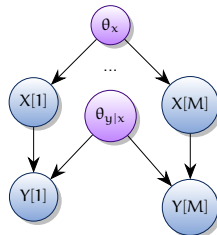
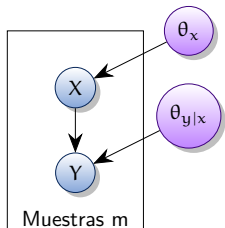
$$l(\mu, \sigma^2; x_1, \dots, x_n) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n (x_j - \mu)^2 \quad (12)$$

# Temas

- 2 Máxima Verosimilitud
  - Verosimilitud
  - Distribución de Bernoulli
  - Distribución Gaussiana
  - En redes Bayesianas

# Máxima verosimilitud para Redes Bayesianas

## Con dos variables



$$\begin{aligned} L(\Theta : D) &= \prod_{m=1}^M P(x[m], y[m] : \Theta) \\ &= \left( \prod_{m=1}^M P(x[m] : \theta_x) \right) \left( \prod_{m=1}^M P(y[m] | x[m] : \theta_{y|x}) \right) \end{aligned}$$



## En general

$$\begin{aligned} L(\Theta : D) &= \prod_m P(X[m] : \Theta) \\ &= \prod_m \prod_i P(x_i[m] | \text{Padres}_{x_i}[m] : \theta_{x_i | \text{Padres}_{x_i}}) \\ &= \prod_i \underbrace{\prod_m P(x_i[m] | \text{Padres}_{x_i}[m] : \theta_{x_i | \text{Padres}_{x_i}})}_{L_i(\theta_{x_i | \text{Padres}_{x_i}} : D)} \\ &= \prod_i L_i(\theta_{x_i | \text{Padres}_{x_i}} : D) \end{aligned}$$

- La máxima verosimilitud de la red se puede maximizar considerando cada probabilidad condicional por separado.

## Con variables discretas

$$\theta_{x_i | \text{padres}_{x_i}} = \frac{N(x, \text{padres}_x)}{N(\text{padres}_x)} \quad (13)$$

# Estimación Bayesiana

- 1 Introducción
- 2 Máxima Verosimilitud
- 3 Estimación Bayesiana**
- 4 Aprendizaje de la estructura
- 5 Datos incompletos

# Estimación Bayesiana

*Cualquier valor sobre el que haya incertidumbre debería ser una variable aleatoria cuya distribución de probabilidad es actualizada conforme reunimos datos.*

# Inferencia con estimación Bayesiana

- Para realizar inferencia se trata a los parámetros  $\Theta$  como a cualquier otra variable.
- Para realizar una consulta, se marginaliza la variable aleatoria correspondiente al(los) parámetro(s) desconocido(s).
- Para realizar la marginalización se suma la distribución de probabilidad sobre todos los posibles valores (se integra cuando ésta es continua).

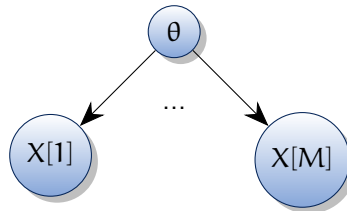
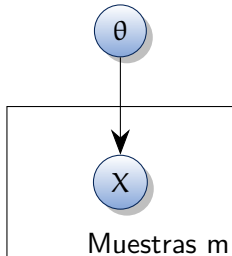
# Temas

## 3 Estimación Bayesiana

- Un parámetro
- Consultas
- Distribución a priori del parámetro

# Estimación de parámetros como un MGP

- Para una moneda cargada sea:  
 $\theta$  una variable aleatoria continua en  $[0, 1]$ .
- Dado que el valor de  $\theta$  es desconocido, fluye información entre los resultados de cada volado  $X[i]$ .



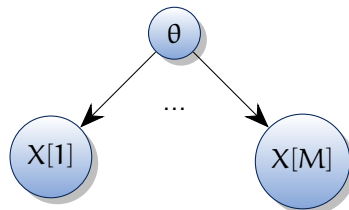
# Temas

## 3 Estimación Bayesiana

- Un parámetro
- Consultas
- Distribución a priori del parámetro



# Ejemplo



- Distribución de probabilidad conjunta:

$$P(x[1], \dots, x[m], \theta) = P(x[1], \dots, x[M]|\theta)P(\theta) \quad (14)$$

$$= P(\theta) \prod_{i=1}^M P(X[i]|\theta) \quad (15)$$

$$= P(\theta)\theta^{NS}(1-\theta)^{NA} \quad (16)$$

# Consultas

- A partir del planteamiento:

$$P(X[1], \dots, X[m], \theta) = P(\theta)\theta^{Ns}(1 - \theta)^{Na} \quad (17)$$

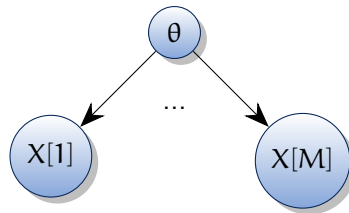
se pueden realizar las preguntas siguientes:

- Distribución de probabilidad a posteriori para el parámetro  $\theta$ :

$$P(\theta|X[1], \dots, X[M]) = \frac{P(X[1], \dots, X[M]|\theta)P(\theta)}{P(X[1], \dots, X[M])} = \frac{P(\theta) \prod_{i=1}^M P(X[i]|\theta)}{Z} \quad (18)$$

- Probabilidad de obtener un cierto valor en el volado siguiente, dados los volados anteriores:

$$P(X[M+1]|X[1], \dots, X[m]) = \frac{P(X[M+1], X[1], \dots, X[M])}{Z} \quad (19)$$



$$\begin{aligned}
 &P(X[M+1]|X[1], \dots, X[M]) \\
 &= \int_{\theta} P(X[M+1], \theta | X[1], \dots, X[M]) d\theta \\
 &= \int_{\theta} P(X[M+1] | X[1], \dots, X[M], \theta) P(\theta | X[1], \dots, X[M]) d\theta \\
 &= \int_{\theta} P(X[M+1] | \theta) P(\theta | X[1], \dots, X[M]) d\theta
 \end{aligned}$$

# Temas

## 3 Estimación Bayesiana

- Un parámetro
- Consultas
- Distribución a priori del parámetro

# Distribución a priori de $\theta$

- Sin embargo todas las consultas anteriores han dejado un pendiente:

$$P(\theta) =? \quad (20)$$

# Distribución de Dirichlet

- Aplica cuando el parámetro  $\theta$  describe una distribución multinomial sobre  $k$  valores posibles de la variable aleatoria  $x$ .
- La *Distribución de Dirichlet* es una distribución continua dada por:

$$P(\theta) = \text{Dirichlet}(\alpha_1, \dots, \alpha_k) = \frac{1}{Z} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

con la constante de normalización

$$Z = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma\left(\sum_{t=1}^k \alpha_t\right)}$$

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

donde  $\Gamma(x)$  es considerada la generalización de la función factorial para números reales.

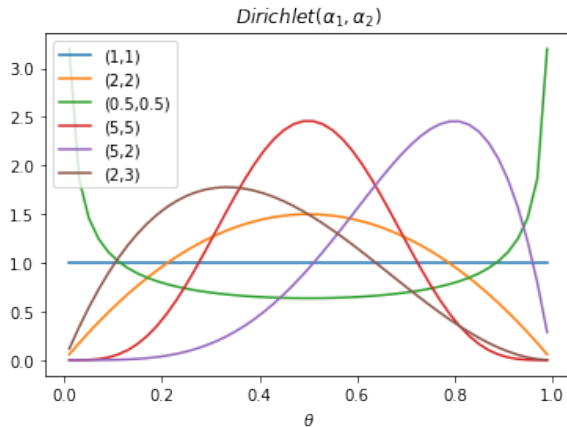


Figura: Caso con dos valores posibles para la variable  $x$ .

- Intuitivamente los hiperparámetros  $\alpha$  corresponden al número de muestras vistas.


## Ejemplo volados

- Se utiliza la distribución  $\text{Dirichlet}(\alpha_S, \alpha_A)^{[1]}$  para representar la creencia inicial  $P(\theta)$  que se tiene sobre la posibilidad de que la moneda esté cargada.
- Esa posibilidad se puede interpretar como los conteos simulados  $\alpha$  de unos *volados virtuales* lanzados antes de iniciar los experimentos reales.
- La distribución a posteriori toma la forma  $\text{Dirichlet}(\alpha_S + NS, \alpha_A + NA)$

$$P(x|D) = \frac{\alpha_x + M[x]}{\alpha + M} \quad (21)$$

En el ejemplo equivalente con **dados** se tendría un conteo simulado por cada cara del dado  $\text{Dirichlet}(\alpha_1, \dots, \alpha_6)$ , posteriormente  $\text{Dirichlet}(\alpha_1 + N_1, \dots, \alpha_6 + N_6)$ , contando cuántas veces  $N_i$  ha salido cada número.

---

<sup>[1]</sup>En el caso particular de dos variables, también se le llama función beta. 

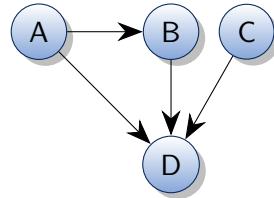
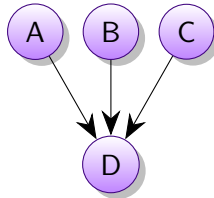
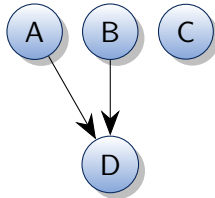


# Aprendizaje de la estructura

- 1 Introducción
- 2 Máxima Verosimilitud
- 3 Estimación Bayesiana
- 4 Aprendizaje de la estructura**
- 5 Datos incompletos

# Aprendizaje de la estructura

- Se asume una distribución verdadera  $P^*$ .



## Omitir una arista

- Las independencias son incorrectas.
- No se puede aprender  $P^*$ .
- Generaliza mejor.

## Añadir una arista

- Introduce dependencias falsas.
- Puede aprender  $P^*$ .
- Hay más parámetros que aprender.
- Peor para generalizar.

# Función de evaluación

- Se utiliza una función para decidir qué gráfica  $\mathcal{G}$  es mejor.
- Por ejemplo, se usa la verosimilitud.
- Sin embargo, la verosimilitud siempre aumenta con el número de aristas, por lo tanto:
  - Se añade un peso que castiga la inclusión de aristas:

$$\text{eva}_{\text{BIC}} = l(\hat{\Theta}_{\mathcal{G}} : D) - \frac{\log M}{2} \text{Dim}[\mathcal{G}] \quad (22)$$

donde  $M$  es el número de ejemplares de entrenamiento y  $\text{Dim}[\mathcal{G}]$  es el número de parámetros independientes en el modelo.

# Datos incompletos

- 1 Introducción
- 2 Máxima Verosimilitud
- 3 Estimación Bayesiana
- 4 Aprendizaje de la estructura
- 5 Datos incompletos**

# Datos incompletos

x	y
?	y=0
x=0	y=1
?	y=0

$$L(D : \Theta) = P(y = 0)P(x = 0, y = 1)P(y = 0) \quad (23)$$

$$= \left( \sum_{x=Val(X)} P(X, y = 0) \right)^2$$



$$P(x = 0, y = 1) \quad (24)$$

$$= (\theta_{x=0}\theta_{y=0|x=0} + \theta_{x=1}\theta_{y=0|x=0})^2$$

$$\theta_{x=0}\theta_{y=1|x=0} \quad (25)$$

No es posible descomponer esta ecuación ni por *variables*, *distribución de probabilidad condicional* y requiere realizar *inferencia*.

# Referencias I

-  Koller, Daphne (15 de jun. de 2016). *Programa especializado: Probabilistic Graphical Models*. Ed. por Coursera. URL: <https://www.coursera.org/specializations/probabilistic-graphical-models>.
-  Koller, Daphne y Nir Friedman (2009). *Probabilistic Graphical Models, Principles and Techniques*. MIT Press Cambridge.

# Licencia

Creative Commons  
Atribución-No Comercial-Compartir Igual

