

Explicant les dades probabilísticament

Coneixement, Raonament i Incertesa.

El contingut d'aquest document s'ha derivat de material provinent de Tom Mitchell, William Cohen, Andrew Moore, Aarti Singh, Eric Xing, Carlos Guestrin.

On som?

1. Necessitem 2^m files en la joint distribution per poder fer inferència (m és el número de variables)

Solució? No sempre podem assegurar independència

2. No sempre tenim informació de tots els casos

Solució? Buscar maneres alternatives a la 'joint distribution'

Simplifiquem el món: Naïve Bayes

No hi ha connexió entre les propietats (variables aleatòries) que defineixen els nostres objectes

Descrivim les propietats (variables aleatòries) per una funció de probabilitat segons el que observem

Tipus de variables aleatòries

- Discretes
binària, 'multivariades'

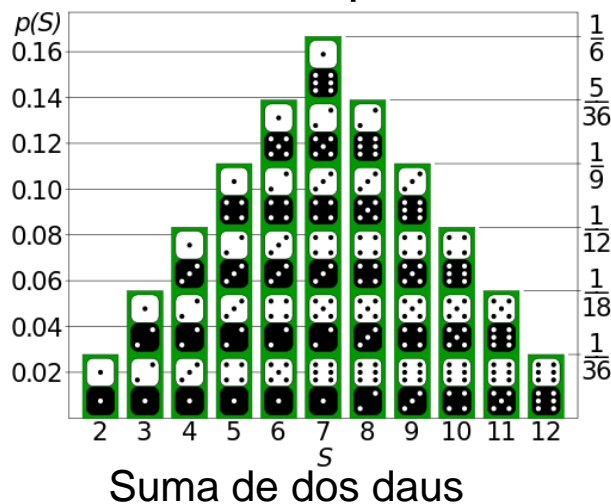
Moneda, dau, ...

- Continues

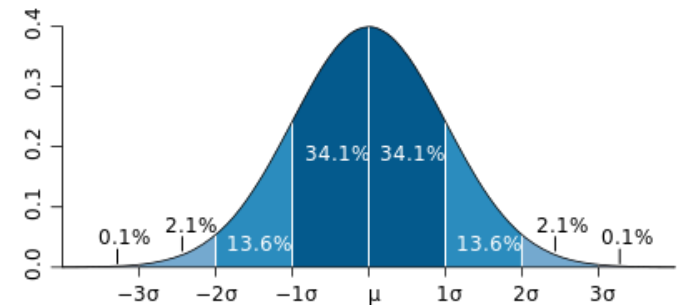
Alçada, ingressos, ...

Com les caracteritzem?

Distribució de probabilitat



Densitat de probabilitat



Alçada població, ...

Example :

The numeric weather data with summary statistics

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	mean	73	74.6	mean	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	std dev	6.2	7.9	std dev	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Tipus de variables aleatòries

- Discretes

Booleans, multiavaluades

$$\sum_u \Pr(X = u) = 1$$

$$P(X = x_i \cap X = x_j) = 0 \text{ if } i \neq j$$

$$P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j) \text{ if } i \neq j$$

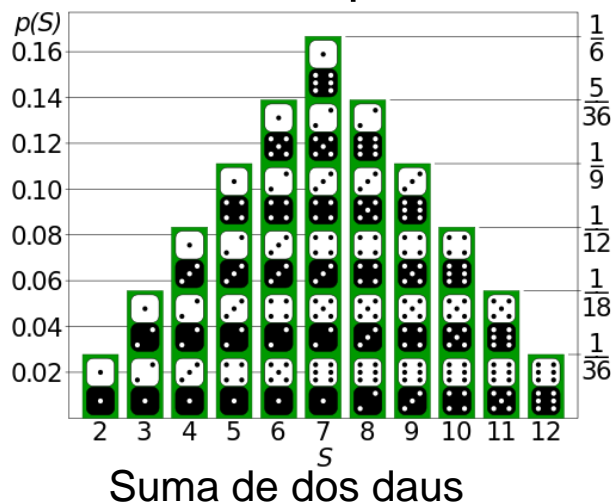
- Continues

$$f(x) \geq 0, \forall x$$

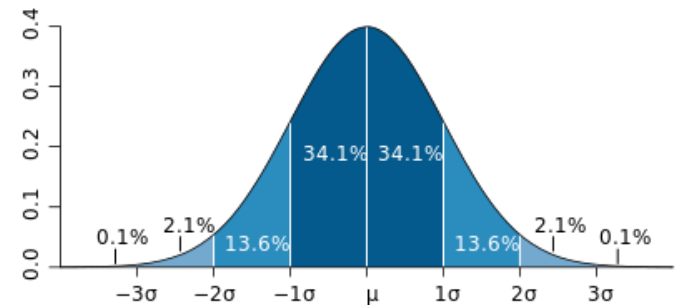
$$\int_{-\infty}^{+\infty} f(x) = 1$$

$$\Pr[a \leq X \leq b] = \int_a^b f(x) dx$$

Distribució de probabilitat



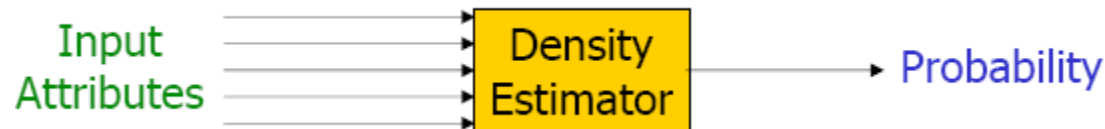
Densitat de probabilitat



Alçada població, ...

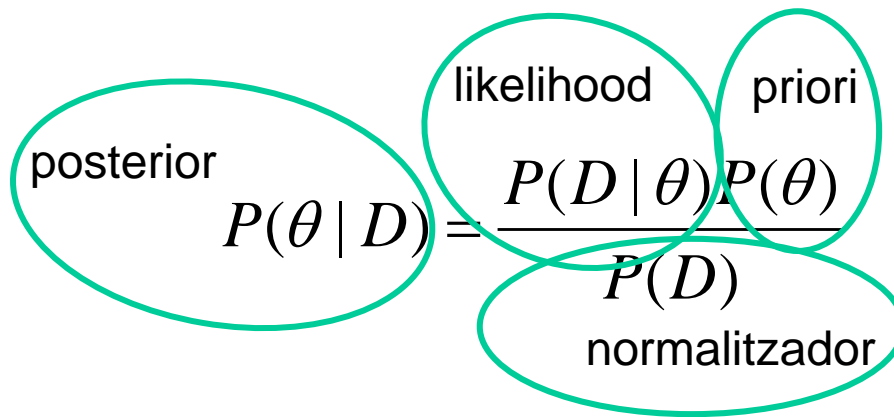
Estimació de la densitat de probabilitat

- Un estimador de Densitat apren un mapejat a partir d'un conjunt d'atributs cap a una probabilitat



- Sovint ho coneixem com a estimador de parametres si s'especifica la forma de la distribució
 - Binomial, Gaussian ...
- A tenir en compte:
 - Natura de les dades (iid, correlacionades, ...)
 - Funció objectiu (MLE, MAP, ...)
 - Algorisme (algebra simple, mètodes del gradient, EM, ...)
 - Esquema d'avaluació (likelihood sobre les dades de test, consistència, ...)

Recordar



The diagram illustrates Bayes' theorem with its components highlighted by green circles:

- posterior**: $P(\theta | D)$
- likelihood**: $P(D | \theta)$
- priori**: $P(\theta)$
- normalitzador**: $P(D)$

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

Aprenentatge de paràmetres sobre dades iid

Objectiu: estimar els parametres de la distribució θ a partir d'un conjunt de dades de N casos d'aprenentatge **independents, identicament distribuïts (iid), completament observats**

$$D = \{x_1, \dots, x_N\}$$

Maximum likelihood estimation (MLE)

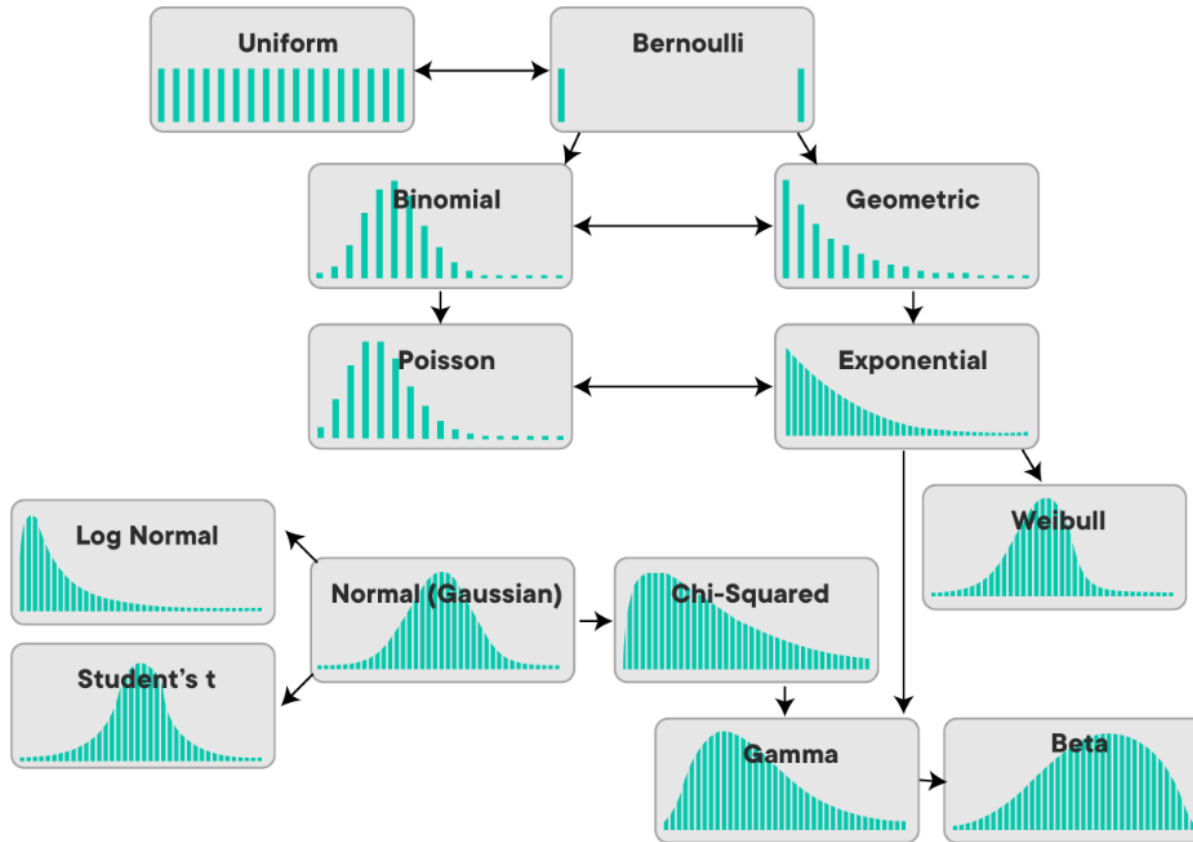
1. Un dels estimadors més comuns
2. Amb asumcions de iid i observació completa, $L(\theta)$ és la likelihood de les dades:

$$\begin{aligned}
 L(\theta) &= P(x_1, x_2, \dots, x_N \mid \theta) \\
 &= P(x_1 \mid \theta) P(x_2 \mid \theta), \dots, P(x_N \mid \theta) \\
 &= \prod_{i=1}^N P(x_i \mid \theta)
 \end{aligned}$$

3. Triar el conjunt de paràmetres que més plausiblement han generat les dades que tenim:

$$\theta^* = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \log L(\theta)$$

Tipus de funcions de densitat de probabilitat



Tipus de funcions de densitat de probabilitat

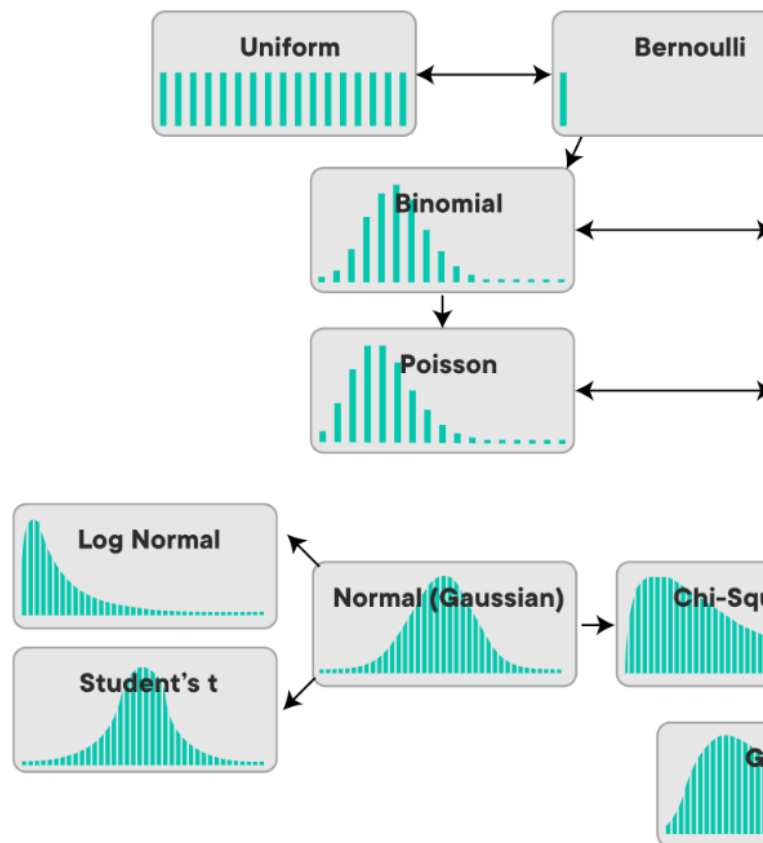


Table of Discrete and Continuous distributions

Distribution	Type	Mass/density function $f(x)$	Mean μ	Variance σ^2
UNIFORM(n)	D	$1/n$, for $x = 1, 2, \dots, n$	$(n+1)/2$	$(n^2-1)/12$
UNIFORM(a, b)	C	$\frac{1}{b-a}$, for $x \in [a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
BERNOULLI(p)	D	$f(0) = 1-p, f(1) = p$	p	$p(1-p)$
BINOMIAL(n, p)	D	$\binom{n}{x} p^x (1-p)^{n-x}$, for $x = 0, 1, \dots, n$	np	npq
GEOMETRIC(p)	D	$q^{x-1}p$, for $x = 1, 2, \dots$	$1/p$	$(1-p)/p^2$
NEGATIVEBINOMIAL(p, r)	D	$\binom{x-1}{r-1} p^r q^{x-r}$, for $x = r, r+1, \dots$	r/p	$r(1-p)/p^2$
HYPERGEOMETRIC(N, M, n)	D	$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$, for $x = 0, 1, \dots, n$	np	$np(1-p)$
POISSON(λt)	D	$\frac{1}{x!} (\lambda t)^x e^{-\lambda t}$, for $x = 0, 1, \dots$	λt	λt
EXPONENTIAL(λ)	C	$\lambda e^{-\lambda x}$, for $x \in [0, \infty)$	$1/\lambda$	$1/\lambda^2$
GAMMA(λ, r)	C	$\frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x}$	r/λ	r/λ^2
GAMMA(α, β)		$= \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$, for $x \in [0, \infty)$	$= \alpha\beta$	$= \alpha\beta^2$
BETA(α, β)	C	$\frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$, for $0 \leq x \leq 1$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
NORMAL(μ, σ^2)	C	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, for $x \in \mathbf{R}$	μ	σ^2
CHISQUARED(ν)	C	$\frac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2} \Gamma(\nu/2)}$, for $x \geq 0$	ν	2ν
T(ν)	C	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu} \Gamma(\frac{\nu}{2})} (1+x^2/\nu)^{-(\nu+1)/2}$, for $x \in \mathbf{R}$	0	$\nu/(\nu-2)$
F(ν_1, ν_2)	C	$\frac{1}{B(\frac{\nu_1}{2}, \frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1} (1+\frac{\nu_1}{\nu_2}x)^{-(\nu_1+\nu_2)/2}$, for $x > 0$	$\frac{\nu_2}{\nu_2-2}$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$

Exemple 1: distribució de Bernoulli

- Dades:
 - Observem N **iid** tirades de moneda: $D=\{1, 0, 1, \dots, 0\}$

- Representació:

r.v binaria: $x_n = \{0,1\}$

- Distribució de Bernoulli: $P(x) = \begin{cases} 1-\theta & \text{per } x=0 \\ \theta & \text{per } x=1 \end{cases} \Rightarrow P(x) = \theta^x (1-\theta)^{1-x}$

- Com expresem la likelihood d'una única observació x_i ?

$$P(x_i) = \theta^{x_i} (1-\theta)^{1-x_i}$$

- La likelihood de les dades $D=\{x_1, \dots, x_N\}$:

$$P(x_1, x_2, \dots, x_N \mid \theta) = \prod_{i=1}^N P(x_i \mid \theta) = \prod_{i=1}^N \left(\theta^{x_i} (1-\theta)^{1-x_i} \right) = \theta^{\sum_{i=1}^N x_i} (1-\theta)^{\sum_{i=1}^N 1-x_i} = \theta^{\text{\#head}} (1-\theta)^{\text{\#tails}}$$

MLE

- Funció objectiu:

$$\ell(\theta; D) = \log P(D | \theta) = \log \theta^{n_h} (1 - \theta)^{n_t} = n_h \log \theta + (N - n_h) \log(1 - \theta)$$

- Hem de maximitzar-ho respecte a θ
- Prenem la derivada respecte θ i busquem el 0

$$\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{N - n_h}{1 - \theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE} = \frac{n_h}{N} \quad \text{o} \quad \hat{\theta}_{MLE} = \frac{1}{N} \sum_i x_i$$

nota: $\frac{\partial \log \theta}{\partial \theta} = \frac{1}{\theta}$

Overfitting

- Recordar que per la distribució de Bernoulli, tenim

$$\hat{\theta}_{ML}^{head} = \frac{n^{head}}{n^{head} + n^{tail}}$$

- Que passa si tirem la moneda massa poques vegades de manera que tenim 0 cares?

Tindriem $\hat{\theta}_{ML}^{head} = 0$, i prediriem que la probabilitat que la següent tirada fos cara seria zero!!!

- solució:
 - On n' es coneix com additive smoothing o Laplace smoothing

$$\hat{\theta}_{ML}^{head} = \frac{n^{head} + n'}{n^{head} + n^{tail} + n'}$$

Exemple 2: distribució normal univariada

- Dades:
 - Observem N **iid** dades reals (continues):

$$D = \{-0.1, 10, 1, -5.2, \dots, 3\}$$

- Model:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Log likelihood:

$$\ell(\theta; D) = \log P(D | \theta) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_{n=1}^N \frac{(x_n - \mu)^2}{\sigma^2}$$

- MLE: prendre la derivada i buscar el zero per a maximitzar: **TENIM 2 PARÀMETRES**

$$\frac{\partial \ell}{\partial \mu} = (1/\sigma^2) \sum_n (x_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_n (x_n - \mu)^2$$



$$\mu_{MLE} = \frac{1}{N} \sum_n (x_n)$$

$$\sigma_{MLE}^2 = \frac{1}{N} \sum_n (x_n - \mu_{ML})^2$$

Com afegim informació a priori?

- Objectiu: estimar els parametres de la distribució θ a partir d'un conjunt de dades de N casos d'aprenentatge **independents**, **identicament distribuïts (iid)**, **completament observats**

$$D = \{x_1, \dots, x_N\}$$

- Maximum a posteriori (MAP)
 1. Un altre dels estimadors més comuns
 2. Amb assumcions de iid i observació completa, i aplicant la regla de Bayes:

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

3. De forma equivalent

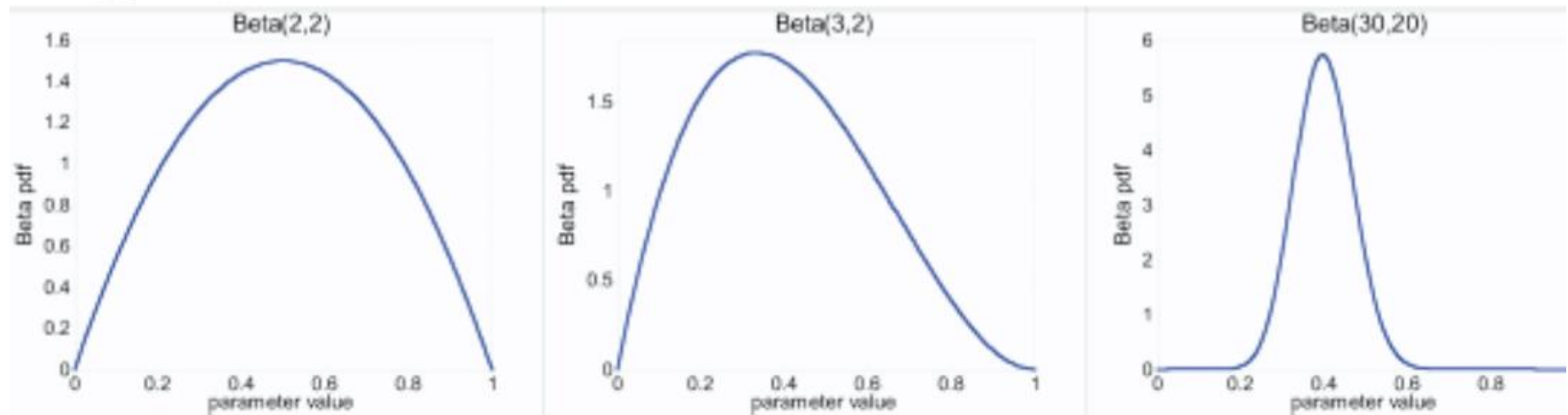
$$P(\theta | D) \propto P(D | \theta)P(\theta)$$

4. Triem el conjunt de parametres θ que maximitzin $P(\theta | D)$

$$\theta_{MAP} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} P(D | \theta)P(\theta)$$

Exemple 1: distribució de Bernoulli

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$



Exemple 1: distribució de Bernoulli

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$

Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$

$$P(\theta \mid x_1, \dots, x_N) = \frac{p(x_1, \dots, x_N \mid \theta)p(\theta)}{p(x_1, \dots, x_N)} \propto \theta^{\alpha_h}(1-\theta)^{\alpha_t} \times \theta^{\beta_h-1}(1-\theta)^{\beta_t-1} = \theta^{\alpha_h+\beta_h-1}(1-\theta)^{\alpha_t+\beta_t-1}$$

$$\theta_{MAP} = \frac{\alpha_h + \beta_h - 1}{\alpha_h + \beta_h - 1 + \alpha_t + \beta_t - 1}$$

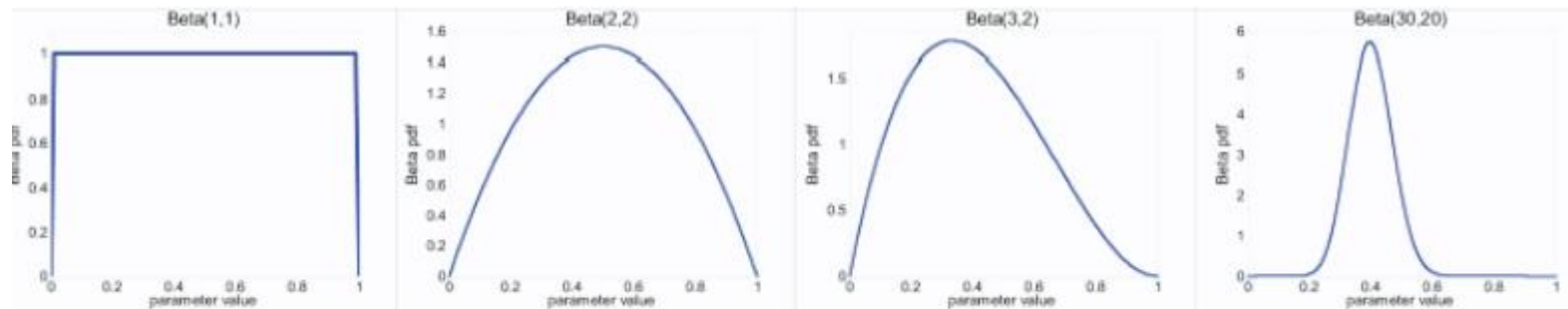
Exemple 1: distribució de Bernoulli

Prior: $Beta(\beta_H, \beta_T)$

Data: α_H heads and α_T tails

Posterior distribution:

$$P(\theta \mid \mathcal{D}) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



- Quan N creix el prior per importància
- Per poques dades és important

$$\theta_{MAP} = \frac{\alpha_h + \beta_h - 1}{\alpha_h + \beta_h - 1 + \alpha_t + \beta_t - 1}$$