Estadística Bayesiana

INF-396

Prof: Juan G. Pavez S.

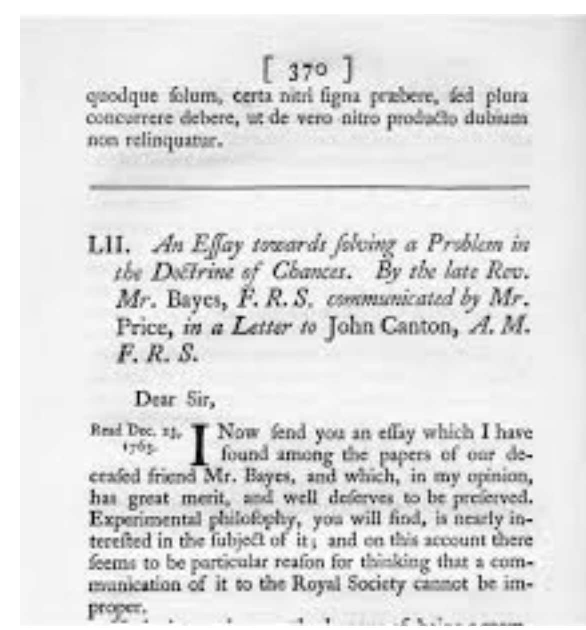
Teorema de Bayes

- Bayes propuso el teorema con su nombre en un articulo póstumo de 1736.

- De hecho, el era un reverendo y quería probar que los milagros existían.

- El teorema de Bayes es una de las formulas mas cruciales de la historia.

- En ese tiempo el problema se conocía como **probabilidad inversa**: Conocer la probabilidad de las causas dado los efectos.



[370]

quodque ſolum, certa nitri ſigna præbere, ſed plura concurrere debere, ut de vero nitro producto dubium non relinquatur.

LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

Credit: The Book of Why. J. Pearl & D. Mackenzie

Teorema de Bayes

Cliente	Te	Berlín	Cliente	Te	Berlín
1	Si	Si	7	Si	No
2	No	Si	8	Si	Si
3	No	No	9	Si	No
4	No	No	10	Si	No
5	Si	Si	11	No	No
6	Si	No	12	Si	Si

- ¿Que fracción de la clientela ordena té y berlín?
  - Dos tercios ordenaron té y una mitad de esos ordenaron berlín (1/2) x (2/3) = 1/3
  - Cinco doceavos ordenaron berlín y cuatro quintos de esos ordenaron té: (4/5) x (5/12) = 1/3
  - En probabilidades:
    - P(B) = Probabilidad de ordenar berlín   P(T) = Probabilidad de ordenar té
    - P(B|T) = Probabilidad de ordenar berlín dado que se ordeno té.
    - **La observación importante es que: P(B AND T) = P(B|T)P(T) = P(T|B)P(B)**

Créditos: The Book of Why. J. Pearl & D. Mackenzie

Teorema de Bayes

- La observación importante:
  - $P(B \text{ AND } T) = P(B|T)P(T) = P(T|B)P(B)$
- Y con algo de algebra:

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

- La formula es una forma de actualizar nuestra **creencia** en una **hipótesis**. Dado que ordené berlín, ¿es más probable que ordene té también?
- Mientras la evidencia sea más sorprendente, mas convencido estaremos de su causa.

Credit: The Book of Why. J. Pearl & D. Mackenzie

Teorema de Bayes

Verosimilitud: Cuan probable es la evidencia dada la hipótesis

A priori: Cuan probable es la hipótesis antes de la evidencia

$$P(S \mid T) = \frac{P(T \mid S)P(S)}{P(T)}$$

Posterior: Cuan probable es la hipótesis dada la evidencia observada.

Marginal: Cuan probable es la nueva evidencia, bajo todas las hipótesis.

$$P(T) = \sum_i P(T \mid S = i)P(S = i)$$

Teorema de Bayes

- Otro ejemplo

- Probabilidad de cancer de mama dado un test positivo?

- P(T=1|D=1) : Sensibilidad del test.

- P(T) = Probabilidad de test positivo en la población.

- P(D) = Probabilidad de cancer en el paciente (dado factores predisponentes).

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)}$$

Likelihood Ratio

$$P(D \mid T) = \frac{P(T \mid D)}{P(T)} P(D)$$

Posterior **Prior**

Teorema de Bayes

P(D) = Hipótesis: Probabilidad de cancer
P(T) = Evidencia: Probabilidad de test positivo

Distribuciones Condicionales

- Intelligence(I): i0,i1
- Difficulty (D): d0,d1
- Grade (G): g1,g2,g3

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Créditos: Probabilistic Graphical Models (Coursera). D. Koller

Distribuciones Condicionales

- Intelligence(I): i0,i1
- Difficulty (D): d0,d1
- Grade(G): g1,g2,g3
- 12 parámetros en la distribución.
- 11 parámetros independientes.

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Distribuciones Condicionales

- **Condicionando:**
- Condicionando en g1

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Distribuciones Condicionales

- **Condicionando:**

- Condicionando en g1

I	D	G	Prob.
i0	d0	g1	0,126
i0	d1	g1	0,009
i1	d0	g1	0,252
i1	d1	g1	0,06

p(I, D, g1)

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Distribuciones Condicionales

- **Condicionando:**
- Condicionando en g1

I	D	G	Prob.
i0	d0	g1	0,126
i0	d1	g1	0,009
i1	d0	g1	0,252
i1	d1	g1	0,06

p(I, D| g1) = P(I,D,g1)/0.447

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Distribuciones Condicionales

- **Condicionando:**
- Condicionando en g1

I	D	G	Prob.
i0	d0	g1	0,126
i0	d1	g1	0,009
i1	d0	g1	0,252
i1	d1	g1	0,06

p(I, D| g1) = P(I,D,g1)/p(g1)

I	D	G	Prob.
i0	d0	g1	0,126
i0	d0	g2	0,168
i0	d0	g3	0,126
i0	d1	g1	0,009
i0	d1	g2	0,045
i0	d1	g3	0,126
i1	d0	g1	0,252
i1	d0	g2	0,0224
i1	d0	g3	0,005
i1	d1	g1	0,06
i1	d1	g2	0,036
i1	d1	g3	0,024

Distribuciones Condicionales

- **Marginalización:**

- Marginalizando I

I	D	Prob.
i0	d0	0,282
i0	d1	0,02
i1	d0	0,564
i1	d1	0,134

D	Prob.
d0	0,846
d1	0,154

Distribuciones Condicionales

- **Marginalización:**

$$P(x) = \sum_y P(x, y) = \sum_y P(x \mid y) p(y)$$

$$P(x) = \int_y dy P(x, y)$$

Distribuciones Condicionales

- **Normalización:**

$$p(W = s \mid T = c) = \frac{P(W = s, T = c)}{P(t = c)}$$

$$= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$

$$= \frac{0.2}{0.2 + 0.3} = 0.4$$

T	W	Prob.
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

W	Prob.
sun	0.4
rain	0.6

$$P(x) = \sum_y P(x, y) = \sum_y P(x \mid y) p(y)$$

Distribuciones Condicionales

- **Normalización:**

$$p(W = s \mid T = c) = \frac{P(W = s, T = c)}{P(t = c)}$$

$$= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$

$$= \frac{0.2}{0.2 + 0.3} = 0.4$$

T	W	Prob.
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

W	Prob.
sun	0.4
rain	0.6

$$P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(x, y)}{\sum_x P(x, y)}$$

Distribuciones Condicionales

- **Regla de la cadena:**

$$
\begin{aligned}
P(x_1, x_2, \ldots, x_n) &= P(x_1)P(x_2 \,|\, x_1)P(x_3 \,|\, x_2, x_1) \ldots P(x_n \,|\, x_1, x_2, \ldots x_{n-1}) \\
&= P(x_1)\prod_{i=2}^{n} P(x_i \,|\, x_1, \ldots x_{i-1})
\end{aligned}
$$

- Derivación:

-

Distribuciones Condicionales

- **Ejercicio:**

- Sabemos: $A \in \{1{,}2\}$ $T \in \{1{,}2{,}3\}$ $P \in \{1{,}2\}$

  $P(A = i, T = 1, P = 1)\ \forall i$

  $P(A = i)\ \forall i$

  $P(P = i)\ \forall i$

  $P(T = i \,|\, P = 1, A = 1)\ \forall i$

  $P(A, P) = P(A)P(P)$

- Queremos: $P(A = 1 \,|\, T = 1, P = 1)$

$$P(x) = \sum_y P(x, y) = \sum_y P(x \,|\, y)p(y)$$ p(x|y) = p(y|x)p(x)/p(y)

Distribuciones Condicionales

- **Independendendencia:**

$$
\begin{aligned}
P(X, Y) &= P(X)P(Y) \\
P(X|Y) &= P(X) \\
P(Y|X) &= P(Y)
\end{aligned}
$$

- **Independendencia condicional:**

$$
\begin{aligned}
P(X, Y|Z) &= P(X|Z)P(Y|Z) \\
P(X|Y, Z) &= P(X|Z) \\
P(Y|X, Z) &= P(Y|Z)
\end{aligned}
$$

Lanzando monedas

- **Enfoque Frecuentista:**

- Lanzar la moneda N veces y obtenemos m caras
- Queremos estimar θ: La probabilidad de cara.
- Conocemos la distribución que modela esto (**Bernoulli**)

$$p(x;\theta) = \theta^x(1-\theta)^{1-x}$$

- Como estimar θ de los datos: N i.i.d casos $X = \{x\}_{i=1}^N$

Lanzando monedas

- **Enfoque Frecuentista:**

- Lanzar la moneda N veces y obtenemos m caras
- Queremos estimar θ: La probabilidad de cara.
- Conocemos la distribución que modela esto (**Bernoulli**)

$$p(x;\theta) = \theta^x(1-\theta)^{1-x}$$

- Como estimar θ de los datos: N i.i.d casos $\quad X = \{x\}_{i=1}^{N}$

- Usamos MLE (Estimador de máxima verosimilitud)

$$L(\theta;X) = \prod_{i}^{N} p(x_i \,|\, \theta) = \prod_{i}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Lanzando monedas

- **Enfoque frecuentista:**

\- Solo en caso de que no recuerden:

$$\theta_{mle} = argmax_{\theta} \sum_{i}^{N} \ln p(x_i \,|\, \theta)$$

$$= argmax_{\theta} \sum_{i}^{N} x_i \ln \theta + (1 - x_i) \ln(1 - \theta)$$

\- De donde

$$\theta_{mle} = \frac{\sum_i x_i}{N} = \frac{N_1}{N}$$

Lanzando monedas

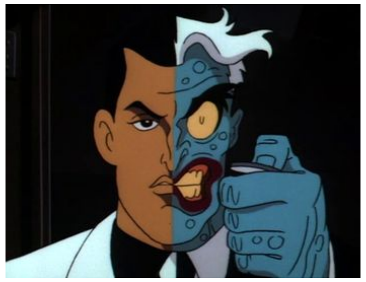
- **Enfoque frecuentista:**

Considerar que lanzamos 3 monedas y obtenemos 3 caras.

El MLE por la probabilidad de cara es 1.

¿Es decir que la moneda sólo tiene lados cara?:

Poco probable, a menos que seamos:



El problema es que estamos haciendo una estimación muy optimista, dado que no estamos considerando nuestro conocimiento previo acerca de las monedas.

Lanzando monedas

- **Enfoque bayesiano:**

- Lo que queremos hacer es codificar nuestro conocimiento a priori acerca de las monedas.
- Usando el teorema de Bayes podemos escribir:

$$P(\theta|X) \propto p(X|\theta)p(\theta)$$

- Notar que dejamos fuera el marginal.
- La verosimilitud:

$$P(X|\theta) = \prod_{i}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{N_1}(1-\theta)^{N_0}$$

Créditos This sections is based on Machine Learning. A probabilistic perspective. K. Murphy.

Lanzando monedas

- **Enfoque bayesiano:**

- Un pequeño pie de página:

$$P(X|\theta) = \prod_i^N \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{N_1}(1-\theta)^{N_0}$$

- Dado $N_1 = \sum_i^N I(x_i = 1),\ N_0 = \sum_i^N I(x_i = 0)$

- N_1, N_0 son estadísticas suficientes - todo lo que necesitamos saber acerca de los datos para estimar los parámetros:

$$p(\theta|D) = p(\theta|S(D))$$

Lanzando monedas

- **Enfoque bayesiano:**

- Un pequeño pie de página:

$$P(X|\theta) = \prod_i^N \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{N_1}(1-\theta)^{N_0}$$

- Dado $N_1 = \sum_i^N I(x_i = 1),\ N_0 = \sum_i^N I(x_i = 0)$

- N_1, N_0 son estadísticas suficientes - todo lo que necesitamos saber acerca de los datos para estimar los parámetros:

$p(\theta|D) = p(\theta|S(D))$ Es N una estadística suficiente?

y N_1 y N?

Lanzando monedas

- **Enfoque bayesiano:**

$$P(X\,|\,\theta) = \theta^{N_1}(1-\theta)^{N_0}$$
$$\propto Bin(N_1\,|\,\theta, N_0 + N_1)$$

Porque?

- Debemos encontrar un a priori en el soporte [0,1] para el parámetro.

- Para esto propondremos un a priori con la misma forma que la verosimilitud, por razones que veremos a continuación

$$Beta(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Lanzando monedas

- **Beta Distribution:**

$$Beta(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{(\beta-1)}$$

$$B(\alpha,\beta) = \int_0^1 x^{(\alpha-1)}(1-x)^{(\beta-1)}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$E[X] = \frac{\alpha}{\alpha+\beta} \qquad Mode[x] = \frac{\alpha-1}{\alpha+\beta-1}$$

$$V[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

http://mathlets.org/mathlets/beta-distribution/

Lanzando monedas

- **Enfoque bayesiano:**

Likelihood Prior

$$P(X|\theta) = \theta^{N_1}(1-\theta)^{N_0} \qquad Beta(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior:

$$\begin{aligned} P(\theta|X) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N_1}(1-\theta)^{N_0} \\ &= \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1} \end{aligned}$$

- Si el a priori y el posteriori tienen la misma forma => **Conjugate Prior**

- Simplifica los cálculos y fácil de entender

https://seeing-theory.brown.edu/bayesian-inference/index.html#section3

Lanzando monedas

- **Enfoque Bayesiano**

- $$\begin{aligned} P(\theta|X) &\propto x^{\alpha-1}(1-x)^{\beta-1}x^{N_1}(1-x)^{N_0} \\ &= x^{N_1+\alpha-1}(1-x)^{N_0+\beta-1} \\ &\propto Beta(\theta|N_1+\alpha, N_0+\beta) \end{aligned}$$

- Es lo mismo actualizar en secuencia que en batch?

- Considerar dos datasets D', D'' con estadísticas

$$N_1', N_0', N_1'', N_0''$$

- Y las estadísticas de un dataset combinado

$$N_1 = N_1' + N_1'', N_0 = N_0' + N_0''$$

Lanzando monedas

- **Enfoque Bayesiano:**

- Sí!, pero se deben cumplir las condiciones de intercambiabilidad:

$$P(D'', D' | \theta) = P(D'' | \theta) P(D' | \theta)$$

- Entonces

$$\begin{aligned} P(\theta | D) &\propto Bin(N_1 | \theta, N_1 + N_0) Beta(\theta | \alpha + \beta) \\ &= Beta(\theta | N_1 + a, N_0 + b) \end{aligned}$$

- Y

$$P(\theta | D', D'') \propto Beta(\theta | N_1 + a, N_0 + b)$$

Lanzando monedas

- **Demostración:**
  Asumiendo las condiciones de intercambiabilidad

$$P(D'', D' \mid \theta) = P(D'' \mid \theta) P(D' \mid \theta)$$

Lanzando monedas

- **Enfoque Bayesiano:**

- Sí!, pero se deben cumplir las condiciones de intercambiabilidad:

$$P(D'', D' | \theta) = P(D'' | \theta)P(D' | \theta)$$

- Entonces

$$P(\theta | D) \propto Bin(N_1 | \theta, N_1 + N_0)Beta(\theta | \alpha + \beta)$$
$$= Beta(\theta | N_1 + a, N_0 + b)$$

- Y

$$P(\theta | D', D'') \propto Beta(\theta | N_1 + a, N_0 + b)$$

Permite aprendizaje online

Lanzando monedas

- **Enfoque Bayesiano**

- Ahora podemos estimar los parámetros, en estadística bayesiana usamos el MAP (máximo a posteriori) que es el modo de la distribución posteriori.

$$\hat{\theta}_{MAP} = argmax_{\theta} P(\theta | X)$$

- De la distribución posteriori Beta:

$$\hat{\theta}_{MAP} = \frac{\alpha + N_1 - 1}{\alpha + \beta + N_0 + N_1 - 2}$$
$$= \frac{\alpha + N_1 - 1}{\alpha + \beta + N - 2}$$

Lanzando monedas

- **Enfoque Bayesiano**

- Ahora podemos estimar los parámetros, en estadística bayesiana usamos el MAP (máximo a posteriori) que es el modo de la distribución posteriori.

$$\hat{\theta}_{MAP} = argmax_{\theta} P(\theta | X)$$

- De la distribución posteriori Beta:

$$\hat{\theta}_{MAP} = \frac{\alpha + N_1 - 1}{\alpha + \beta + N_0 + N_1 - 2}$$
$$= \frac{\alpha + N_1 - 1}{\alpha + \beta + N - 2}$$

Recordar que:

$$\theta_{mle} = \frac{\sum_i x_i}{N} = \frac{N_1}{N}$$

Que pasa sí: A priori uniforme?, Mucho datos?

Lanzando monedas

- **Enfoque bayesiano:**

- Para entender mejor el tradeoff entre a piori y verosimilitud, considerar lo siguiente:
- Sea $\alpha_0 = a + b$ que llamamos tamaño muestral equivalente del a priori.
- La media del a priori es $m_1 = a/\alpha_0$
- La media del a posteriori es: $E[\theta | D] = \dfrac{a + N_1}{a + b + N}$

Lanzando monedas

- **Enfoque bayesiano:**

- Para entender mejor el tradeoff entre a piori y verosimilitud, considerar lo siguiente:
- Sea $\alpha_0 = a + b$ que llamamos tamaño muestral equivalente del a priori.
- La media del a priori es $m_1 = a/\alpha_0$
- La media del a posteriori es: $E[\theta | D] = \frac{a + N_1}{a + b + N}$

$$
\begin{aligned}
E[\theta | D] &= \frac{a + N_1}{a + b + N} \\
&= \frac{\alpha_0 m_1 + N_1}{N + \alpha_0} \\
&= \frac{\alpha_0}{N + \alpha_0} m_1 + \frac{N}{N + \alpha_0} \frac{N_1}{N} \\
&= \lambda m_1 + (1 - \lambda) \hat{\theta}_{MLE}
\end{aligned}
$$

La media del a posteriori es una combinación convexa de la media del a priori y el MLE

A priori débil => menor λ

$$\lambda = \alpha_0 / (N + \alpha_0)$$

Lanzando monedas

- **Distribución predictiva a posteriori:**

- Considerar que queremos estimar la probabilidad de datos futuros.
- La probabilidad de cara en un sólo intento en el futuro es:

$$p(x = 1 \mid D) =$$

Lanzando monedas

- **Distribución predictiva a posteriori:**

- Considerar que queremos estimar la probabilidad de datos futuros.
- La probabilidad de cara en un sólo intento en el futuro es:

$$
\begin{aligned}
p(x=1|D) &= \int_0^1 p(x=1|\theta)p(\theta|D)d\theta \\
&= \int_0^1 \theta Beta(\theta|N_1+a, N_0+b) \\
&= E[\theta|D] = \frac{N_1+a}{N+a+b}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_0^1 p(\theta|D)\theta \\
&= E[\theta|D]
\end{aligned}
$$

Lanzando monedas

- **La paradoja del cisne negro:**

- Considerar el estimador MLE:

$$p(x = 1 \,|\, D) = Ber(x \,|\, \hat{\theta}_{MLE})$$

- Como vimos, si sólo observamos tres caras, nuestra probabilidad de sello es 0. Esto se conoce como el problema **zero count**.
- Un enfoque bayesiano con a priori a = b = 1 nos dá:

$$p(x = 1 \,|\, D) = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

Que se conoce como add-one smoothing

El lanzamiento de dados

- **Lanzando un dado:**

- Previamente consideramos la probabilidad de que una moneda salga cara.

- Ahora consideraremos la probabilidad de que un dado de K caras salga en la cara k.

El lanzamiento de dados

- **Likelihood:**

- Considerar N lanzamientos de dados $D = \{x_1, \ldots, x_N\}$ donde $x_i \in \{1,...,K\}$

- Podemos modelar esta probabilidad como:

pdf: $$p(x_i = \theta_i \,|\, \theta) = \theta_i \qquad \sum_i \theta_i = 1$$

likelihood: $$p(D \,|\, \theta) = \prod_{k=1}^{K} \theta_k^{N_k} \qquad N_k = \sum_i^{N} I(x_i = k)$$

El lanzamiento de dados

- **Prior:**

- Necesitamos un a priori que satisfaga:

$$S_\theta = \{\theta : 0 \leq \theta_k \leq 1, \sum_{i=1}^{K} \theta_k = 1\}$$ (K-dimensional Simplex)

- Dirichlet Distribution:

$$Dir(\theta \,|\, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

El lanzamiento de dados

- **Dirichlet Distribution:**

$$Dir(x\,|\,\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

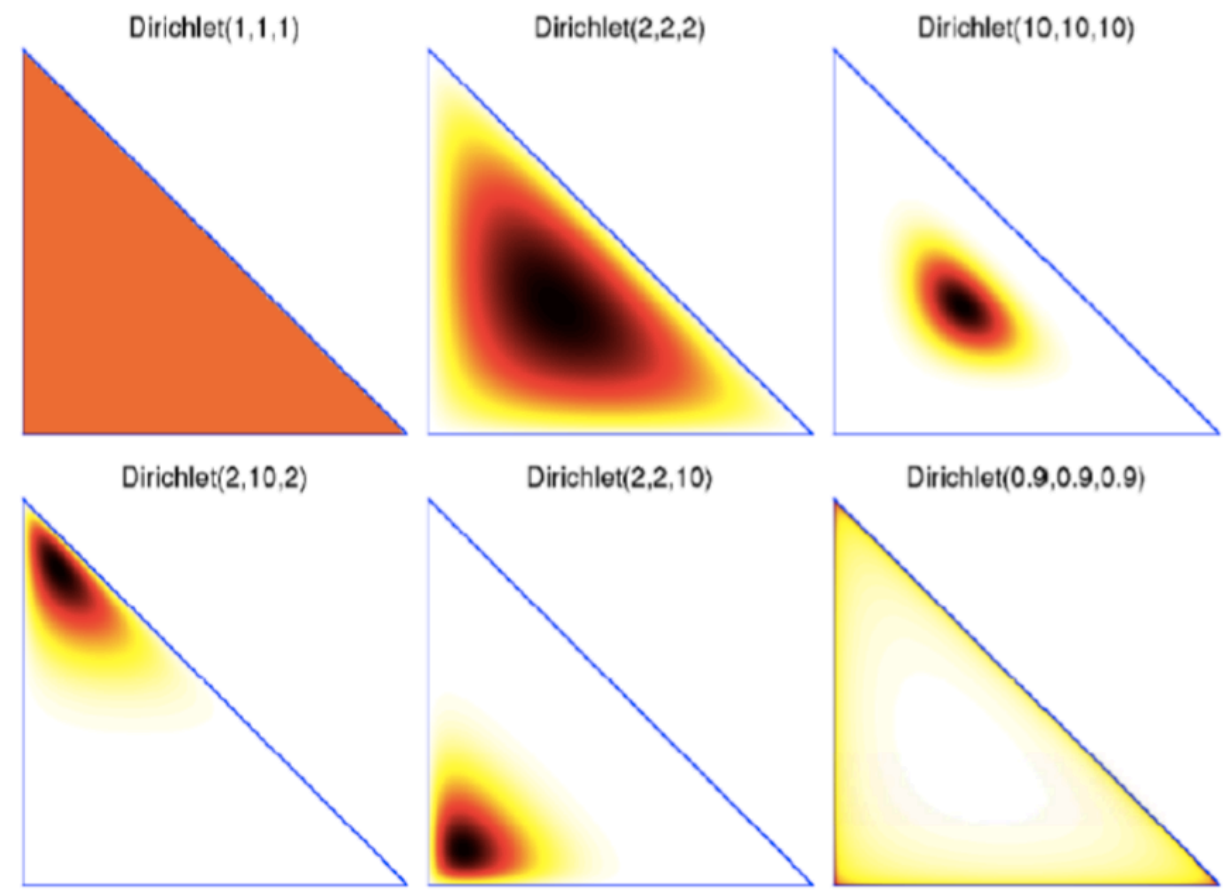
- Donde $B(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$ y $\alpha_0 = \sum_{k=1}^{K} \alpha_k$

- α_0 controla la fuerza del peak y α_k controla la posición del peak k.

El lanzamiento de dados

- **Dirichlet Distribution:**

$$Dir(x|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$



\- α_0 controla la fuerza del peak y α_k controla la posición del peak k.

El lanzamiento de dados

- **Dirichlet Distribution:**

$$Dir(x \mid \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

- $E[x_k] = \dfrac{\alpha_k}{\alpha_0}$ $\quad mode[x_k] = \dfrac{\alpha_k - 1}{\alpha_0 - K}$ $\quad var[x_k] = \dfrac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$

- A priori común es $\alpha_k = \alpha / k$

El lanzamiento de dados

- **Posterior:**

- Dirichlet Distribution:

$$
\begin{aligned}
p(\theta | D) &\propto p(D | \theta) p(\theta) \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} \\
&= \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} \\
&= Dir(\theta | \alpha_1 + N_1, \ldots, \alpha_k + N_k)
\end{aligned}
$$

El lanzamiento de dados

Theta = [theta_1, theta_2, ….]
theta_-1 = [theta_2, ….]

- **Predictive Posterior:**

$$P(x=j|D) = \int_\theta P(x=j|\theta)p(\theta|D)d\theta$$

$$= \int_{\theta_j}\int_{\theta_{-j}} P(x=j|\theta)p(\theta|D)d\theta_j d\theta_{-j}$$

$$= \int_{\theta_j} p(x=j|\theta_j)\left[\int_{\theta_{-j}} p(\theta_j,\theta_{-j}|D)d\theta_{-j}\right]d\theta_j$$ **marginalización**

$$= \int_{\theta_j} \theta_j p(\theta_j|D)d\theta_j = E[\theta_j|D]$$

$$= \frac{\alpha_j + N_j}{\sum_k(\alpha_k + N_k)}$$

$$= \frac{\alpha_j + N_j}{\alpha_0 + N}$$

Modelamiento de Lenguaje

- **Un ejemplo más practico:**

- El modelamiento de lenguaje es la tarea de predecir la probabilidad de la próxima palabra dada una secuencia de palabras.

- Tiene importantes usos, como traducción automatizada, traducción automatizada, corrección de errores, reconocimiento de voz, etcétera.

Considerar el próximo ejemplo:

Mary had a little lamb , little lamb , little lamb ,
Mary had a little lamb , its fleece as white as snow ***[?]***

Modelamiento de Lenguaje

- **Un ejemplo más práctico:**

Mary had a little lamb , little lamb , little lamb ,
*Mary had a little lamb , its fleece as white as snow **[?]***

- Para resolver el problema vamos a hacer el siguiente supuesto (probablemente erróneo): $X_i = \{1,...,K\}$
- Asumiremos que cada palabra $Cat(\theta)$ es muestreada independientemente de una distribución:
- Esto se conoce como la representación de bolsa de palabras:
- K es el número de palabras en el vocabulario:

mary: 1, lamb: 2, little: 3, big: 4, fleece: 5, white: 6, black: 7
snow: 8, rain: 9, unk: 10

- Notar que removimos la puntación y las palabras comunes (stop words). También el símbolo ***unk*** representa palabras que no están en el vocabulario.

Modelamiento de Lenguaje

- **Un ejemplo más práctico:**

Mary had a little lamb , little lamb , little lamb ,
Mary had a little lamb , its fleece as white as snow ***[?]***

\- Las estadísticas del texto que queremos modelar son:

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

- Ocuparemos un a priori de dirichlet, por lo que podemos representar la probabilidad de cada palabra como $\theta_1 = mary, \ldots, \theta_K = unk$

Pdf: $p(x_i = \theta_i \mid \theta) = \theta_i$ $\quad\quad$ $\sum_i \theta_i = 1$

Likelihood: $p(D \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$

Posterior: $p(\theta \mid D) = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} = Dir(\theta \mid \alpha_1 + N_1, \ldots, \alpha_K + N_K)$

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

- Ocuparemos un a priori de dirichlet, por lo que podemos representar la probabilidad de cada palabra como $\theta_1 = mary$, ..., $\theta_1 = mary$

Asumamos un a priori con $\alpha_i = 1$

Para encontrar el máximo a posteriori debemos calcular

$$max_\theta p(\theta | D) = max_\theta \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} = mode[Dirichlet(\alpha_1 + N_1, \ldots, \alpha_K + N_K)]$$

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Asumamos un a priori con $\alpha_i = 1$

Para encontrar el máximo a posteriori debemos calcular

$$max_\theta p(\theta | D) = max_\theta \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} = mode[Dirichlet(\alpha_1 + N_1, \ldots, \alpha_K + N_K)]$$

Podemos hacerlo calculando $\dfrac{\partial p(\theta | D)}{\partial \theta}$

Pero ya lo conocemos para la Dirichlet

$$mode_{\theta_k} = \frac{\alpha_k + N_k - 1}{\alpha_0 + N - 1}$$

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Asumamos un a priori con $\alpha_i = 1$

Para Mary

$$mode_{\theta_1} = \frac{1 + 2 - 1}{10 + 17 - 1} = 2/26$$

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Asumamos un a priori con $\alpha_i = 1$

Ahora si queremos predecir la próxima palabra debemos calcular la probabilidad $P(x = j \mid D)$ para cada palabra, es decir $P(x = mary \mid D), p(x = lamb \mid D), \ldots$ y elegir el máximo.

Este es el posterior productivo

$$P(x = j \mid D) = \int_\theta P(x = j \mid \theta) p(\theta \mid D) = E[\theta_j \mid D]$$

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Asumamos un a priori con $\alpha_i = 1$

Para el posterior dirichlet, calculamos que era

$$P(x = j \,|\, D) = E[\theta_j \,|\, D] = \frac{\alpha_j + N_j}{\sum_{j'} a_{j'} + 17} = \frac{1 + N_j}{10 + 17}$$

Entonces por ejemplo, la probabilidad de que la próxima palabra sea Mary es (1+2)(10+17)= 3/27

Modelamiento de Lenguaje

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

Asumamos un a priori con $\alpha_i = 1$

En general es

$$P(x \mid D) = (3/27, 5/27, 5/27, 1/27, 2/27, 2/27, 1/27, 2/27, 1/27, 5/27)$$

Es decir que la próxima palabra más probable es o lamb, little o unk