Análisis de sensibilidad para procesos stick-breaking con divergencia Kullback-Leibler

Seminario de Teoría de la Información,

Machine Learning y Estadística.

IIMAS - UNAM

12 de abril del 2023

Trabajo conjunto con:

Mario Diaz y Ramsés H. Mena

Contexto

Estadística bayesiana

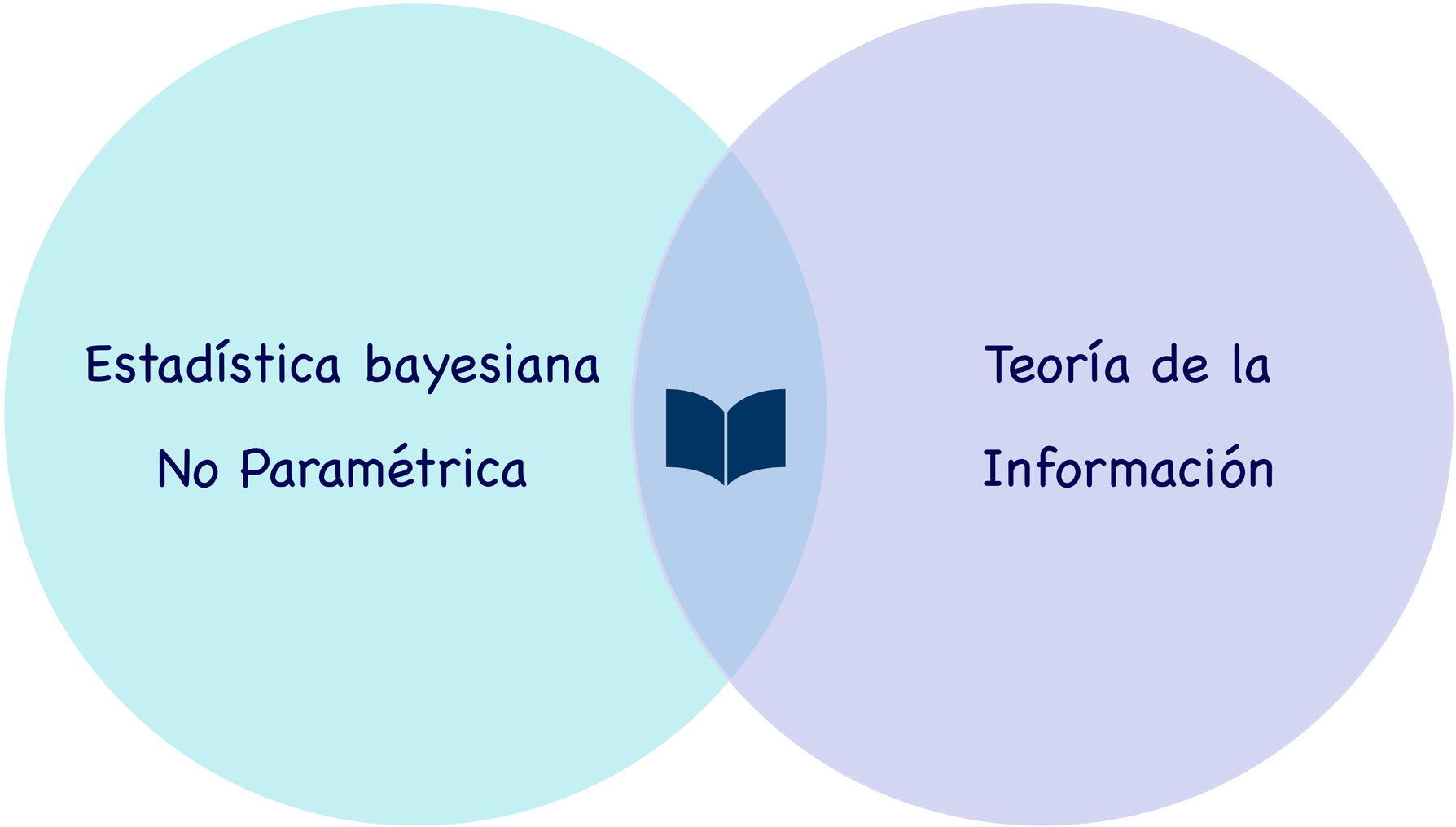
No Paramétrica

- Análisis de sensibilidad
- Procesos stick-breaking

Contexto

- Divergencia Kullback-Leibler

Contexto


Estadística bayesiana
No Paramétrica
Teoría de la
Información

Agenda

I. Preliminares

- Estadística bayesiana (no paramétrica)
- Divergencia Kullback-Leibler

II. Análisis de sensibilidad de procesos stick-breaking

III. Análisis de sensibilidad de procesos stick-breaking intercambiables

Estadística bayesiana

- Alternativa a la estadística clásica donde los parámetros son aleatorios con un distribución de probabilidad asociada y los datos son tratados como valores fijos.

Estadística bayesiana

- Alternativa a la estadística clásica donde los parámetros son aleatorios con un distribución de probabilidad asociada y los datos son tratados como valores fijos.

- Considerada como una forma de hacer estadística más coherente y libre de contradicciones[1].

[1] Berger, J.O. & Wolpert, R. (1988). The Likelihood Principle. Hayward: Institute of Mathematical Statistics.

Estadística bayesiana

- Alternativa a la estadística clásica donde los parámetros son aleatorios con un distribución de probabilidad asociada y los datos son tratados como valores fijos.

- Considerada como una forma de hacer estadística más coherente y libre de contradicciones[1].

- Podemos dar respuesta a: ¿Qué valor de θ es más plausible dados los datos?

[1] Berger, J.O. & Wolpert, R. (1988). The Likelihood Principle. Hayward: Institute of Mathematical Statistics.

Aprendizaje bayesiano

Thomas Bayes

Aprendizaje bayesiano

Thomas Bayes

$$f(\theta \mid \mathbf{X}) \propto \quad f(\mathbf{X} \mid \theta) \quad f(\theta)$$

Aprendizaje bayesiano

Thomas Bayes

Distribución posterior

Distribución a priori

$$f(\theta \mid \mathbf{X}) \propto f(\mathbf{X} \mid \theta) \quad f(\theta)$$

Verosimilitud

Aprendizaje bayesiano

- Independencia física **no** implica que exista una independencia estocástica.

- El aprendizaje estadístico demanda dependencia estocástica entre las variables aleatorias que son réplicas del mismo fenómeno.

- Independencia física **solo** implica una simetría con la ley conjunta de las variables aleatorias, i.e., las etiquetas son no informativas.

Intercambiabilidad

Definición (Intercambiabilidad finita)

Un conjunto finito de v.a.'s $X_1, \ldots, X_n$ se dice intercambiable (finito) si

$$\left(X_1, \ldots, X_n\right) \stackrel{d}{=} \left(X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$$

para toda permutación σ de $\{1, \ldots, n\}$.

Intercambiabilidad

Definición (Intercambiabilidad finita)

Un conjunto finito de v.a.'s $X_1, \ldots, X_n$ se dice intercambiable (finito) si

$$\left(X_1, \ldots, X_n\right) \stackrel{d}{=} \left(X_{\pi(1)}, \ldots, X_{\pi(n)}\right)$$

para toda permutación σ de $\{1,\ldots,n\}$.

Definición (Intercambiabilidad)

Una colección de variables aleatorias $\{X_i\}_{i=1}^{\infty}$ se dice intercambiable si toda sub-colección finita es intercambiable.

Intercambiabilidad

Teorema de representación de De Finetti[2]

Sea $\mathbb{X}$ un espacio polaco dotado con su sigma álgebra de Borel $\mathscr{X}$ y denotemos por $\mathscr{P}_{\mathbb{X}}$ el espacio de medidas de probabilidad sobre $(\mathbb{X}, \mathscr{X})$. Entonces una colección de v.a.'s $\{X_i\}_{i=1}^{\infty}$ $\mathbb{X}$-valuadas es intercambiable si existe una y sólo una medida $\mathbb{Q}$ sobre $\mathscr{P}_{\mathbb{X}}$ tal que

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathscr{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i)\mathbb{Q}(dP)$$

[2]Enunciado para variables aleatorias por De Finetti (1931) y generalizado por Hewitt y Savage (1955)

Intercambiabilidad

Teorema de representación de De Finetti[2]

Sea $\mathbb{X}$ un espacio polaco dotado con su sigma álgebra de Borel $\mathscr{X}$ y denotemos por $\mathscr{P}_{\mathbb{X}}$ el espacio de medidas de probabilidad sobre $(\mathbb{X}, \mathscr{X})$. Entonces una colección de v.a.'s $\{X_i\}_{i=1}^{\infty}$ $\mathbb{X}$-valuadas es intercambiable si existe una y sólo una medida $\mathbb{Q}$ sobre $\mathscr{P}_{\mathbb{X}}$ tal que

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathscr{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i) \mathbb{Q}(dP)$$

- Las v.a.'s son condicionalmente iid dado P

[2] Enunciado para variables aleatorias por De Finetti (1931) y generalizado por Hewitt y Savage (1955)

Ejemplo: caso paramétrico

- En términos paramétricos

$$f\left(X_1, \ldots, X_n\right) = \int_{\Theta} f(\theta) \prod_{i=1}^{n} f(X_j \mid \theta) d\theta$$

Ejemplo: caso paramétrico

- En términos paramétricos

$$f\left(X_1, \ldots, X_n\right) = \int_{\Theta} f(\theta) \prod_{i=1}^{n} f(X_j \mid \theta) d\theta$$

- O de forma jerárquica

$$X_i \mid \theta \overset{iid}{\sim} f(\cdot \mid \theta)$$
$$\theta \sim f(\theta)$$

Ejemplo: caso paramétrico

- En términos paramétricos

$$f\left(X_1, \ldots, X_n\right) = \int_{\Theta} f(\theta) \prod_{i=1}^{n} f(X_j \mid \theta) d\theta$$

- O de forma jerárquica

$$X_i \mid \theta \overset{iid}{\sim} f(\cdot \mid \theta)$$
$$\theta \sim f(\theta)$$

- ¿ Cómo especificar $f(\theta)$?

Distribución a priori

- Modelos conjugados, i.e., elegir $f(\theta)$ tal que $f(\theta|\mathbf{X})$ pertenezca a la misma familia.

Distribución a priori

- Modelos conjugados, i.e., elegir $f(\theta)$ tal que $f(\theta|\mathbf{X})$ pertenezca a la misma familia.

- Elegir una priori no informativa para "dejar a los datos hablar por ellos mismos", e.g.

Distribución a priori

- Modelos conjugados, i.e., elegir $f(\theta)$ tal que $f(\theta \mid \mathbf{X})$ pertenezca a la misma familia.

- Elegir una priori no informativa para "dejar a los datos hablar por ellos mismos", e.g.

  1. Distribución uniforme (impropia) en un intervalo acotado (no acotado).

Distribución a priori

- Modelos conjugados, i.e., elegir $f(\theta)$ tal que $f(\theta \mid \mathbf{X})$ pertenezca a la misma familia.

- Elegir una priori no informativa para "dejar a los datos hablar por ellos mismos", e.g.

  1. Distribución uniforme (impropia) en un intervalo acotado (no acotado).

  2. Distribución de Jeffrey, i.e., $f(\theta) \propto |I(\theta)|^{\frac{1}{2}}$

Distribución a priori

- Modelos conjugados, i.e., elegir $f(\theta)$ tal que $f(\theta \mid \mathbf{X})$ pertenezca a la misma familia.

- Elegir una priori no informativa para "dejar a los datos hablar por ellos mismos", e.g.

1. Distribución uniforme (impropia) en un intervalo acotado (no acotado).
2. Distribución de Jeffrey, i.e., $f(\theta) \propto |I(\theta)|^{\frac{1}{2}}$
3. Distribución de referencia[3], i.e., buscar $f(\theta)$ maximice

$$\int f(t) \int f(\mathbf{X} \mid \theta) \log \left(\frac{f(\mathbf{X} \mid \theta)}{f(\theta)} \right) d\theta dt$$

[3]Berger, J.O. & Bernardo, J. & Sun. D (2009). The Formal Definition of Reference Priors. The Annals of Statistics.

Estadística bayesiana no paramétrica

- Recordando el Teorema de representación

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathscr{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i)\mathcal{Q}(dP)$$

donde

- P es una medida de probabilidad aleatoria (MPA)
- $\mathcal{Q}$ es la distribución de De Finetti (priori de P)

Estadística bayesiana no paramétrica

- Recordando el Teorema de representación

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{\mathscr{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i)\mathcal{Q}(dP)$$

donde

- P es una medida de probabilidad aleatoria (MPA)
- $\mathcal{Q}$ es la distribución de De Finetti (priori de P)

- O de forma jerárquica

$$X_i \,|\, P \overset{iid}{\sim} P$$
$$P \sim \mathcal{Q}$$

Estadística bayesiana no paramétrica

Objetivo: Construir distribuciones $\mathcal{Q}$ para medidas de probabilidad aleatorias P

Estadística bayesiana no paramétrica

Objetivo: Construir distribuciones $\mathbb{Q}$ para medidas de probabilidad aleatorias P

Método 1: Especificar $\mathbb{Q}$ e.g.

- Vía distribuciones infinito dimensionales con distribuciones finito dimensionales específica.
- Especificar la lay de la sucesión de v.a.'s mediante distribuciones predictiva.

Estadística bayesiana no paramétrica

Objetivo: Construir distribuciones $\mathbb{Q}$ para medidas de probabilidad aleatorias P

Método 1: Especificar $\mathbb{Q}$ e.g.

- Vía distribuciones infinito dimensionales con distribuciones finito dimensionales específica.
- Especificar la lay de la sucesión de v.a.'s mediante distribuciones predictiva.

Método 2: Construir directamente a P

- Transformación de procesos estocásticos.
- Modelos de muestreo de especies.

Distribución Dirichlet

- Distribución continua y generalización de la distribución beta

Distribución Dirichlet

- Distribución continua y generalización de la distribución beta

- Soporte el simplex n-1 dimensional, i.e.,

$$\left\{ \mathbf{x} \in [0,1]^n : \sum_i x_i = 1 \right\}$$

Distribución Dirichlet

- Distribución continua y generalización de la distribución beta

- Soporte el simplex n-1 dimensional, i.e.,

$$\left\{ \mathbf{x} \in [0,1]^n : \sum_i x_i = 1 \right\}$$

- Con densidad dada por:

$$f(\mathbf{x}) = \frac{\Gamma\left(\sum_{i=1}^{n} \alpha_i\right)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}$$

Distribución Dirichlet

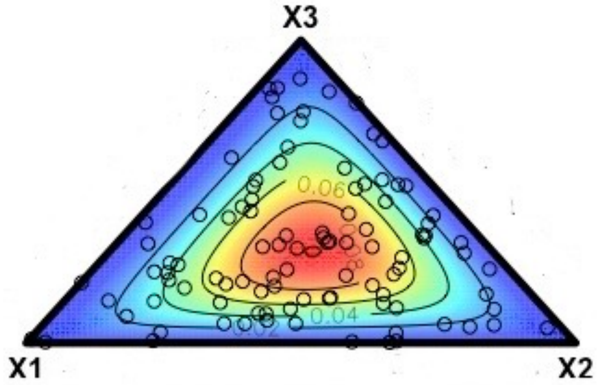
Construcción

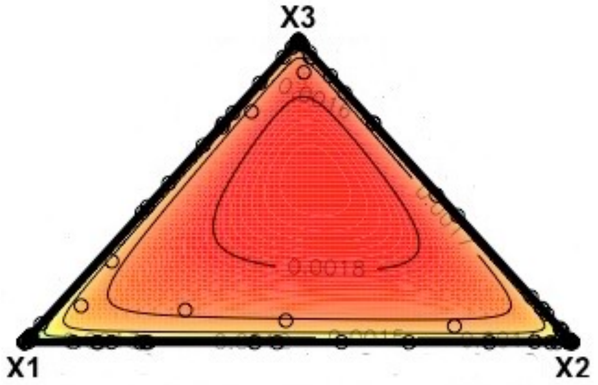
Sean $X_1, \ldots, X_n \overset{ind}{\sim} Ga(\alpha_1, \theta)$ entonces $V = \sum_i X_i \sim Ga(\alpha_0, \theta)$

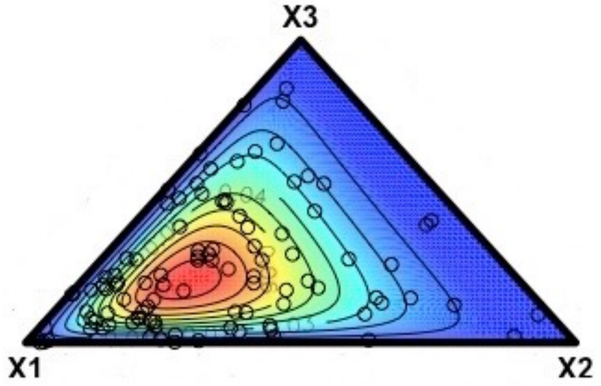
$$(Y_1, \ldots, Y_n) = \left(\frac{X_1}{V}, \ldots, \frac{X_n}{V} \right) \sim Dir(\alpha_1, \ldots, \alpha_n)$$

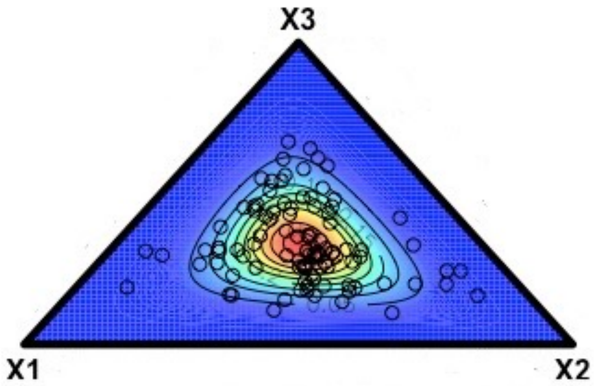
donde $\alpha_0 = \sum_i \alpha_i$

Distribución Dirichlet









Distribución Dirichlet

Considerar $\mathbb{X} = \{1, \ldots, k\}$ y un modelo multinomial - Dirichlet, i.e.

Distribución Dirichlet

Considerar $\mathbb{X} = \{1, \ldots, k\}$ y un modelo multinomial - Dirichlet, i.e.

- $X | (p_1, p_2, \ldots, p_n)$ tienen densidad proporcional a

$$\prod_{i=1}^{n} p_n^{\delta_i(X_i)}$$

Distribución Dirichlet

Considerar $\mathbb{X} = \{1, \ldots, k\}$ y un modelo multinomial - Dirichlet, i.e.

- $X | (p_1, p_2, \ldots, p_n)$ tienen densidad proporcional a

$$\prod_{i=1}^{n} p_n^{\delta_i(X_i)}$$

- La verosimilitud es proporcional a

$$\prod_{i=1}^{n} p_n^{\sum_{i=1}^{N} \delta_i(X_i)}$$

Distribución Dirichlet

Considerar $\mathbb{X} = \{1, \ldots, k\}$ y un modelo multinomial - Dirichlet, i.e.

- $X | (p_1, p_2, \ldots, p_n)$ tienen densidad proporcional a

$$\prod_{i=1}^{n} p_n^{\delta_i(X_i)}$$

- La verosimilitud es proporcional a

$$\prod_{i=1}^{n} p_n^{\sum_{i=1}^{N} \delta_i(X_i)}$$

- La posterior es

$$\text{Dir}\left(\alpha_1 + \sum_{i=1}^{N} \delta_1(X_i), \ldots, \alpha_n + \sum_{i=1}^{N} \delta_n(X_i)\right)$$

Proceso Dirichlet

- Vía distribuciones infinito dimensionales con distribuciones finito dimensionales específicas, i.e.,

Definición

Sea $\alpha > 0$ una medida finita sobre un espacio Polaco. Se dice que una MPA P tiene una distribución Dirichlet si para toda partición medible $\left(B_1, \ldots, B_n\right)$ de $\mathbb{X}$,

$$\left(P(B_1), \ldots, P(B_n)\right) \sim \text{Dir}\left(\alpha(B_1), \ldots, \alpha(B_n)\right)$$

Proceso Dirichlet

- Vía distribuciones predictivas

$$\mathbb{P}(X_1 \in dx_1, \ldots, X_n \in dx_n) = \prod_{i=1}^{n} \frac{\alpha(dx_i) + \sum_{j=1}^{i-1} \delta_{x_j}(dx_i)}{\alpha(\mathbb{X}) + i - 1}$$

- Una consecuencia directa es que

$$\mathbb{P}(X_i = X_j) = \frac{1}{\theta + 1}$$

Proceso Dirichlet

- Vía modelo de muestre de especies ("stick-breaking")

$$P(B) = \sum_{i=1}^{\infty} \omega_i \delta_{z_i}(B)$$

donde

$$w_1 = v_1$$
$$w_n = v_n \prod_{i<n} (1 - v_i)$$

y $v_i \sim Be(1,\theta)$

Procesos stick-breaking

- Muchas y muy variadas formas de definirlas a través de particulares elecciones de la colección v_i, e.g.:

  - Procesos independientes (proceso Dirichlet)
  - Procesos dependientes (proceso geométrico) donde $v_i = v \sim Be(a, b)$.
  - Procesos intercambiables, donde

$$v_i | \nu \sim \nu$$
$$\nu \sim Dir(\beta, \nu_0)$$

¡Gracias!