

EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Asignación Latente Dirichlet.

Objetivo: En esta sección veremos un modelo que entra en la frontera de inferencia Bayesiana y Aprendizaje de Máquina: modelo de asignación latente. El cual puede ser entendido como un método de segmentación probabilística. Las aplicaciones de este modelo son variadas y en la clase lo discutiremos en términos de procesamiento de documentos escritos. *Disclaimer:* Parte del material fue tomado del curso en métodos Bayesianos para Bioestadística impartido por Jeff Miller en Harvard en la escuela de Salud Pública, materiales [aquí](#).

Lectura recomendada: El modelo de asignación latente Dirichlet lo puedes encontrar en [6] (pronto saldrá la nueva edición). Los artículos originales también son una buena referencia, [3].

1. INTRODUCCIÓN

- VI es una estrategia para poder hacer inferencia Bayesiana por medio de aproximaciones a la distribución posterior.
- La idea es:
 1. Escoger una familia de distribuciones \mathcal{Q} .
 2. Encontrar el elemento $q \in \mathcal{Q}$ mas cercano a la distribución posterior.
 3. Utilizar q^* para resolver los problemas de inferencia.
- Aplicación de inferencia variacional en problemas de aplicación probabilística y *machine learning*.
- Modelo para una colección de documentos.
- Cada documento es una colección de palabras donde cada palabra se extrae de un tema en particular.
- Los temas definen las palabras que se utilizarán para escribir el documento.
- Los documentos pueden tener mas de un tema.
- Las palabras son intercambiables.
- Con asignación latente Dirichlet (LDA), modelamos los tópicos de una colección de observaciones.
- Usualmente se utiliza para datos no-estructurados.
 - Imágenes, datos genómicos o redes sociales.
- Se puede utilizar para datos en *stream*.
- Catalogación automática de objetos.

2. EL MODELO DE LDA

Asumimos que cada documento puede hablar de distintos temas al mismo tiempo.

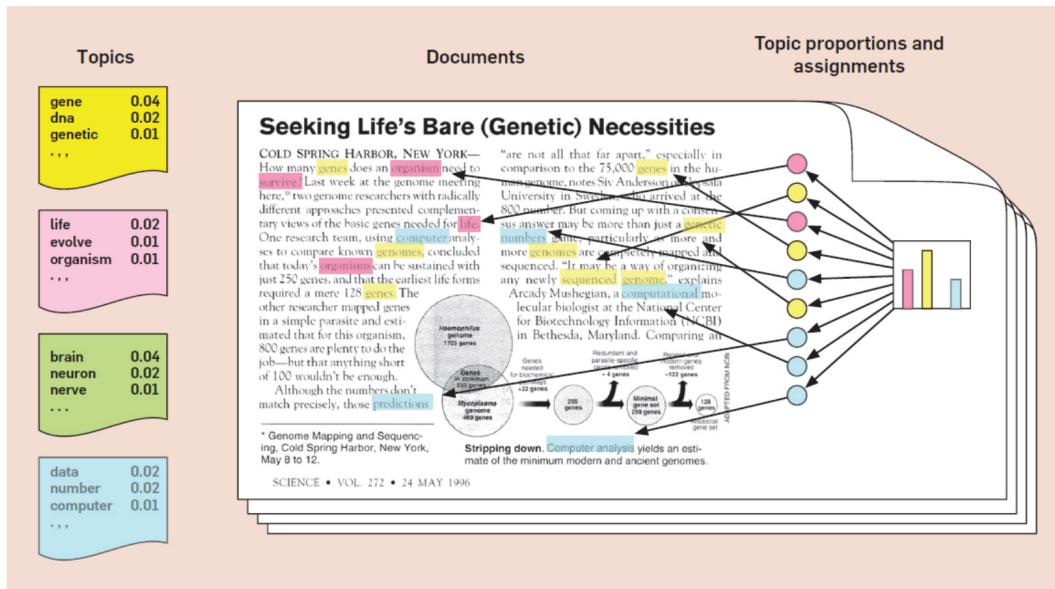


FIGURA 1. Imagen tomada de [1].

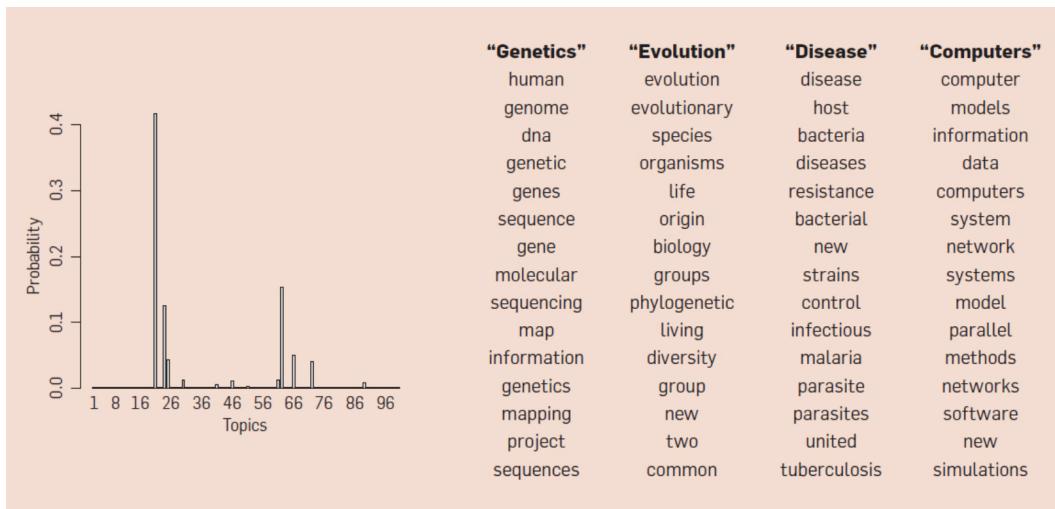


FIGURA 2. Imagen tomada de [1].

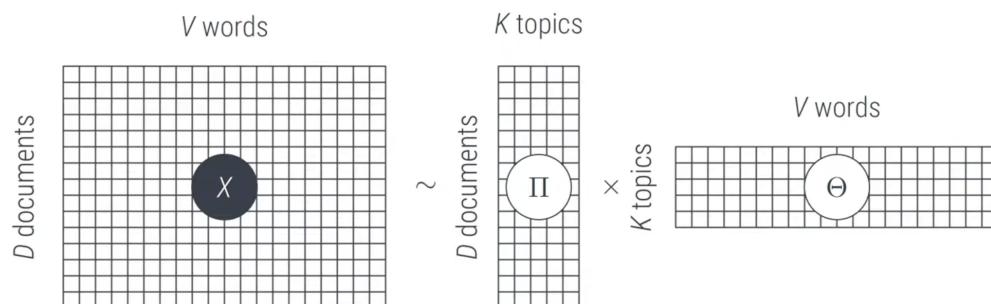


FIGURA 3. Modelo de tópicos sobre documentos en términos de contenido. Imagen tomada del curso de Probabilistic ML de Phillip Hennig.

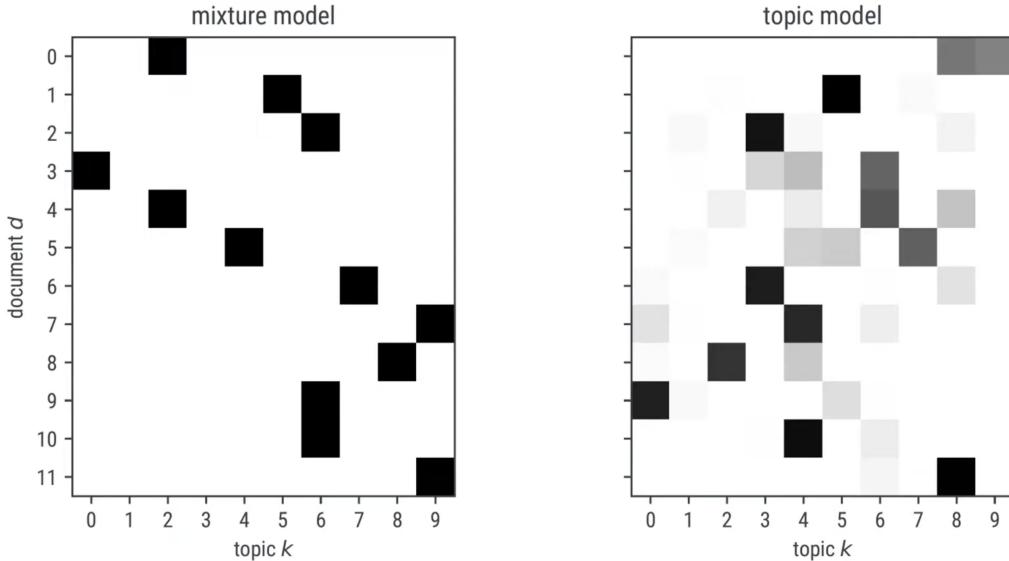


FIGURA 4. Diferencia entre modelo de mezcla y modelo de tópicos sobre documentos en términos de contenido. Imagen tomada del curso de Probabilistic ML de Phillip Hennig.

Buscamos una representación con pocas entradas activas pues cada documento no podría hablar de todos los temas de conversación.

En particular usaremos distribuciones Dirichlet pues nos permiten tener una representación **rala** de los componentes que estarán activos en nuestras observaciones.

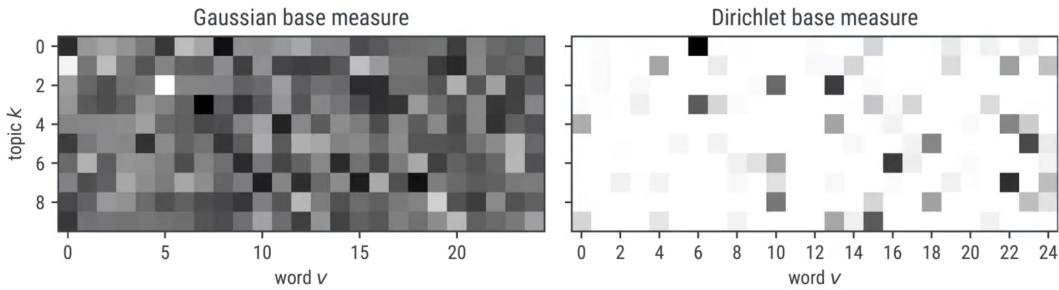


FIGURA 5. Diferencia entre modelos de tópicos usando una distribución **densa** y una distribución **rala**. Imagen tomada del curso de Probabilistic ML de Phillip Hennig.

- Supongamos que existen K temas, n documentos, L_i palabras en el documento i , y V palabras en el vocabulario.
- Cada documento tiene:
 - w_{ik} : la proporción del documento que proviene del tema k .
 - $z_{i\ell}$: el tema de la palabra ℓ .
 - $x_{i\ell}$: la palabra en la posición ℓ .
- De manera global definimos β_{kv} : la frecuencia con la que aparece la palabra v en el tema k .

2.1. Modelo generativo

Consideremos un modelo donde las palabras tienen una asignación de tema y, además, cada palabra es una realización aleatoria de acuerdo al tópico y a la colección de posibles palabras que se utilizan en dicho tema

$$Z_{i\ell}|w \sim \text{Categorical}(w_i), \quad (1)$$

$$x_{i\ell}|\beta, Z_{i\ell} = k \sim \text{Categorical}(\beta_k), \quad (2)$$

de manera independiente para cada $i \in \{1, \dots, n\}$ y $\ell \in \{1, \dots, L_i\}$.

Nota que

$$w_i = (w_{i1}, \dots, w_{iK})^\top, \quad \beta_k = (\beta_{k1}, \dots, \beta_{kV})^\top. \quad (3)$$

La distribución previa es

$$w_i \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K), \quad (4)$$

$$\beta_k \sim \text{Dirichlet}(\lambda_1, \dots, \lambda_V). \quad (5)$$

*2.1.1. Definición (**Distribución Dirichlet**):* Decimos que un vector aleatorio $w \in \mathbb{R}^K$ tiene una distribución Dirichlet(α) con $\alpha \in \mathbb{R}_+^K$ si su función de densidad es

$$\pi(w|\alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \cdot \prod_k w_k^{\alpha_k - 1}, \quad (6)$$

y $\sum_k w_k = 1$.

La **distribución Dirichlet** es una generalización para la **distribución Beta**. Es usual utilizar una distribución Dirichlet para el vector de probabilidades de un **modelo multinomial**. En este sentido la distribución inicial Dirichlet y el modelo Multinomial forman un **modelo conjugado**.

Aplicaciones clásicas de éstos modelos también se pueden encontrar en **modelos de mezclas** donde los pesos en la mezcla se consideran realizaciones aleatorias de un distribución Dirichlet.

Además, nota que el vector aleatorio es un vector de longitud fija. Si quisieramos modelar un vector donde el número de entradas es aleatoria entonces podemos considerar un **proceso Dirichlet**.

El modelo completo queda escrito como en Fig. 6 donde queda claro que la estructura condicional del modelo es bastante compleja pero que es relativamente sencillo resolver utilizando muestreo de Gibbs.

2.2. Observaciones del modelo

- El orden no afecta la composición del modelo.
- No es un buen modelo de lenguaje, pero ayuda a generar conocimiento de los documentos.
- El modelo es invariante al orden en el que estudiamos los documentos.

2.3. Modelo variacional

- La distribución objetivo es la posterior $\pi(z, w, \beta|x)$.

$$\begin{aligned}
p(C, \Pi, \Theta, W) &= p(\Pi | \alpha) \cdot p(C | \Pi) \cdot p(\Theta | \beta) \cdot p(W | C, \Theta) \\
&= \left(\prod_{d=1}^D p(\pi_d | \alpha_d) \right) \cdot \left(\prod_{d=1}^D \prod_{i=1}^{l_d} p(c_{di} | \pi_d) \right) \cdot \left(\prod_{d=1}^D \prod_{i=1}^{l_d} p(w_{di} | c_{di}, \Theta) \right) \cdot \left(\prod_{k=1}^K p(\theta_k | \beta_k) \right) \\
&= \left(\prod_{d=1}^D \mathcal{D}(\pi_d; \alpha_d) \right) \cdot \left(\prod_{d=1}^D \prod_{i=1}^{l_d} \left(\prod_{k=1}^K \pi_{dk}^{c_{dk}} \right) \right) \cdot \left(\prod_{d=1}^D \prod_{i=1}^{l_d} \left(\prod_{k=1}^K \theta_{kw_{di}}^{c_{dk}} \right) \right) \cdot \left(\prod_{k=1}^K \mathcal{D}(\theta_k; \beta_k) \right) \\
&= \left(\prod_{d=1}^D \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^K \pi_{dk}^{\alpha_{dk}-1+n_{dk}} \right) \cdot \left(\prod_{k=1}^K \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^V \theta_{kv}^{\beta_{kv}-1+n_{kv}} \right)
\end{aligned}$$

FIGURA 6. *Modelo completo en asignación de temas. Imagen tomada del curso de Probabilistic ML de Phillip Hennig.*

- Se consideran modelos

$$q(z, w, \beta) = q(z) q(w) q(\beta). \quad (7)$$

- El modelo variacional obtiene

$$q(w) = \prod_{i=1}^n \text{Dirichlet}(w_i | r_{i1}, \dots, r_{iK}), \quad (8)$$

$$q(\beta) = \prod_{k=1}^K \text{Dirichlet}(\beta_k | s_{k1}, \dots, s_{kV}), \quad (9)$$

$$q(z) = \prod_{i=1}^n \prod_{\ell=1}^{L_i} \text{Categorical}(z_{i\ell} | t_{i\ell}), \quad (10)$$

en donde cada término explota la estructura conjugada del modelo.

2.4. Observaciones del método variacional

- Nota que aunque hemos asumido una factorización del estilo $q(z, w, \beta) = q(z) q(w) q(\beta)$ el modelo en si obtiene

$$q(z, w, \beta) = \left(\prod_{i,\ell} q(z_{i\ell}) \right) \left(\prod_i q(w_i) \right) \left(\prod_k q(\beta_k) \right). \quad (11)$$

- La funciones de densidad óptimas (en KL) son distribuciones **Dirichlet**.

2.5. Aplicación: Associated Press

- Ejemplo original en [3].
- Contiene $n = 16,333$ artículos.
- Contiene $V = 23,075$ palabras.
- Se necesitan eliminar palabras sin contenido informativo (*stop-words*).
- Se define un número de tópicos $K = 100$.
- El artículo original solo usa VI en z, w .

3. EXTENSIONES DEL MODELO

- LDA y un modelo de estados ocultos: captura de dependencias en palabras cercanas.

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

FIGURA 7. Resultados de [3].

- Modelo no-paramétrico basado en un proceso Dirichlet.
- Modelo dinámico: cómo cambian los tópicos a lo largo del tiempo.
- Modelo jerárquico de tópicos (temas): de lo mas general a lo mas particular.
- Extensiones con meta-datos: autor, títulos de documentos, afiliaciones, etc.

4. MAS EXTENSIONES

- LDA con temas correlacionados, Blei and Lafferty [2].
- LDA en línea, Hoffman et al. [4].
- LDA en paralelo, Zhai et al. [7].
- LDA multilenguajes, Hu et al. [5].
- Inferencia automática (Infer.NET).

5. SET DE HERRAMIENTAS

En el curso hemos aprendido:

$$\int h(\theta) \pi(\theta) d\theta, \quad \pi(x, \theta) = \pi(x|\theta)\pi(\theta), \quad \pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}. \quad (12)$$

En términos de modelado:

- Modelos bayesianos.
- Modelos predictivos probabilísticos.

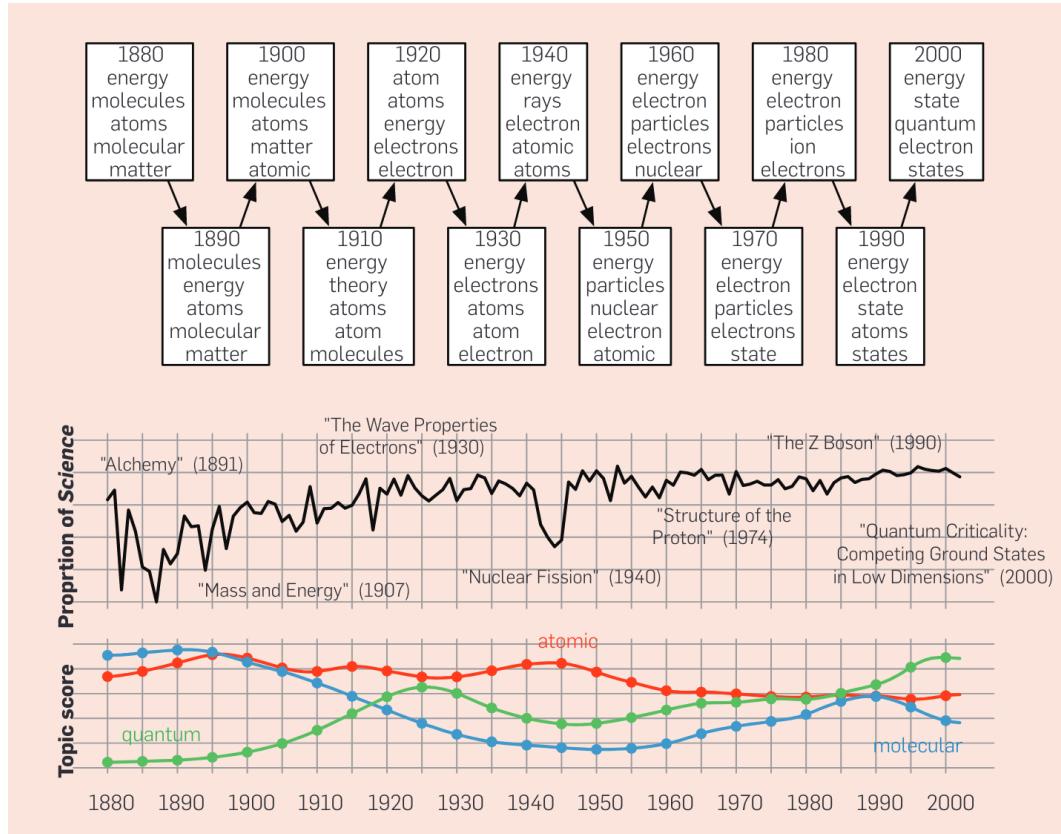


FIGURA 8. Imagen tomada de [1].

- Comparación de modelos.
- Crítica de modelos.

En términos computacionales:

- Monte Carlo.
- Monte Carlo vía Cadenas de Markov.
- Inferencia variacional.

REFERENCIAS

- [1] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, apr 2012. ISSN 0001-0782, 1557-7317. . 2, 7
- [2] D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1): 17–35, jun 2007. ISSN 1932-6157, 1941-7330. . 6
- [3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. ISSN ISSN 1533-7928. 1, 5, 6
- [4] M. Hoffman, F. Bach, and D. Blei. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. 6
- [5] Y. Hu, K. Zhai, V. Eidelman, and J. Boyd-Graber. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176, Baltimore, Maryland, jun 2014. Association for Computational Linguistics. . 6
- [6] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, 2012. ISBN 978-0-262-01802-9. 1
- [7] K. Zhai, J. Boyd-Graber, N. Asadi, and M. L. Alkhousa. Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce. In *Proceedings of the 21st International Conference on World Wide Web*, pages 879–888, Lyon France, apr 2012. ACM. ISBN 978-1-4503-1229-5. . 6