Proceso Generativo de la Asignación Latente de Dirichlet (LDA)

10 de mayo de 2016

1. Introducción

En principio, es importante aclarar que el proceso generativo del LDA no genera documentos reales. Supongamos que tenemos un corpus. Los documentos han sido generados por un proceso complejo subyacente, que no nos es conocido. El objetivo del LDA es modelar el proceso generativo real por uno sintético, que se aproxime al real, y tratar de encontrar parámetros para éste, que se ajusten bien (o lo mejor posible) a los datos. Este proceso de síntesis se conoce como el proceso generativo del LDA.

Ahora bien, el proceso supone que un documento se genera como mezclas de palabras de tópicos con cierta probabilidad. En concreto, el LDA asume el siguiente proceso generativo para el corpus $\mathcal{D}$:

1. Se establece el vocabulario a usar.
2. Se determinan un número (fijo) K de tópicos, con su respectiva distribución de palabras (Distrubución multinomial).
3. Se establece el número M de documentos que tendrá el corpus.
4. Para cada uno de los M documentos:
   - 4.1. Se establece el número N de palabras que tendrá el documento (Por ejemplo de acuerdo a una distribución Poisson(ξ)).
   - 4.2. Se elige una distribución θ de tópicos para el documento (de acuerdo a una distribución Dirichlet(α) sobre el conjunto fijo de K tópicos).
   - 4.3. Para cada una de las N palabras:
     - 4.3.1. Se selecciona un tópico z_n.
     - 4.3.2. Se selecciona una palabra del tópico (de acuerdo a la distribución de palabras en el tópico establecida en el paso 2)

Observación 1.1 Claramente, este no es el proceso real por el cual se genera un documento, La idea de que los documentos son producidos por los discursos en lugar de los autores es ajena al sentido común, sin embargo, la aproximación obtenida es razonable. Note que si se usa este proceso para generar un documento, se obtendrá un texto ilegible.

Este proceso define una distribución de probabilidad conjunta sobre ambas, las variables ocultas y las variables observadas. El análisis de los datos se construye usando la distribución de probabilidad conjunta para calcular la distribución condicional de las variables ocultas dadas y las variables observadas. Esta distribución condicional es lo que en estadística bayesiana se llama distribución a posteriori.

2. Ejemplo del Proceso Generativo de un Modelado Probabilístico de Tópicos

El proceso general de generar un documento se puede describir de forma sencilla (sin atender a las distintas distribuciones de probabilidad involucradas) usando la siguiente idea:

Consideremos el vocabulario

$$V = \{\text{arte, música, eléctrico, CENDITEL, tecnología, servicio}\}$$

como un conjunto ordenado, y definamos tres tópicos t_1, t_2 y t_3. Los cuales tienen una probabilidad de ocurrencia dada para cada palabra del vocabulario como sigue:

$$\begin{aligned} t_1 &= \left\{x_{11} = 0, x_{12} = 0, x_{13} = \frac{2}{9}, x_{14} = \frac{1}{3}, x_{15} = \frac{1}{3}, x_{16} = \frac{1}{9}\right\} \\ t_2 &= \left\{x_{21} = \frac{2}{9}, x_{22} = \frac{1}{3}, x_{23} = 0, x_{24} = \frac{5}{18}, x_{25} = 0, x_{26} = \frac{1}{6}\right\} \\ t_3 &= \left\{x_{31} = 0, x_{32} = \frac{1}{6}, x_{33} = \frac{2}{9}, x_{34} = \frac{1}{6}, x_{35} = \frac{4}{9}, x_{36} = 0\right\} \end{aligned}$$

Donde $x_{ij} = p(pal_j)$ en t_i y pal_j es la j-ésima palabra del volabulario.

Observación 2.1 En este punto ya se tienen fijos *el vocabulario, la cantidad de tópicos* y *la distribución de palabras en cada tópico*[1]. Note que, las palabras pueden *pertenecer* a sólo un tópico, (por ejemplo, *arte*), a dos (*eléctrico*) o bien a los tres tópicos (*CENDITEL*). Así podremos decir, intuitivamente, que aquellos documentos en los que aparezca la palabra *arte*, serán más fáciles de clasificar que aquellos en los que aparezca la palabra *CENDITEL*.

A continuación, construiremos un corpus ($\mathcal{D}$) de 6 documentos de longitud constante igual a 4. Entonces podemos pensar en el documento en blanco como una hilera de 4 casillas (ordenadas) vacías.

[1] La suma de las probabilidades de todas las palabras dentro de un tópico debe ser igual a 1.

Documento 1 (d_1):

1. Se elige una distribución de probabilidades de los tópicos para d_1[2].

$$\begin{aligned} p(t_1|d_1) &= \frac{1}{2} \\ p(t_2|d_1) &= \frac{1}{4} \\ p(t_3|d_1) &= \frac{1}{4} \end{aligned}$$

2. Para cada casilla vacía, se elige un tópico[3].

casilla 1	casilla 2	casilla 3	casilla 4
tópico 1	tópico 3	tópico 2	tópico 1

3. Para cada casilla vacía, se elige una palabra (De acuerdo al tópico asignado a la casilla en el paso 2.).

casilla 1	casilla 2	casilla 3	casilla 4
tópico 1	tópico 3	tópico 2	tópico 1
CENDITEL	CENDITEL	servicio	tecnología

Así, se ha generado el siguiente documento: « CENDITEL CENDITEL servicio tecnología » al que llamaremos d_1.

Observación 2.2 Note aquí tres cosas:

- El documento generado es ininteligible.
- Dentro de un documento pueden haber palabras repetidas.
- La suma de las probabilidades de todas las palabras de documento de acuerdo al tópico asignado a la casilla, es igual a 1, Formalmente:

$$\sum_{i=1}^{4} p(w_i|z_i) = 1$$

Donde, w_i es la i-ésima palabra del documento y z_i es el tópico asignado a la i-ésima casilla. Esto es;

$$\begin{aligned} \sum_{i=1}^{4} p(w_i|z_i) &= p(\text{CENDITEL}|\text{tópico 1}) + p(\text{CENDITEL}|\text{tópico 3}) \\ &+ p(\text{servicio}|\text{tópico 2}) + p(\text{tecnología}|\text{tópico 1}) \\ &= \frac{1}{3} + \frac{1}{6} + \frac{1}{6} + \frac{1}{3} = 1. \end{aligned}$$

[2]La suma de las probabilidades de todos los tópicos dentro de un documento debe ser igual a 1.

[3]La distribución establecida en el paso 1. debe verse reflejada en la asignación de los tópicos a las casillas.

En lo sucesivo repetiremos el proceso para la generación de los 5 documentos restantes del corpus.

Naturalmente, queremos que la observación anterior sea válida para todos los documentos.

Documento 2			
dist. tópicos	$p(t_1\|d_2) = 0$	$p(t_1\|d_2) = \frac{1}{4}$	$p(t_1\|d_2) = \frac{3}{4}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 3	tópico 3	tópico 2	tópico 3
música	CENDITEL	arte	servicio

Documento 3			
dist. tópicos	$p(t_1\|d_3) = \frac{5}{9}$	$p(t_1\|d_3) = \frac{1}{9}$	$p(t_1\|d_3) = \frac{1}{3}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 3	tópico 3	tópico 1	tópico 1
tecnología	eléctrico	eléctrico	servicio

Documento 4			
dist. tópicos	$p(t_1\|d_4) = 0$	$p(t_1\|d_4) = \frac{8}{9}$	$p(t_1\|d_4) = \frac{1}{9}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 2	tópico 2	tópico 2	tópico 2
arte	música	arte	arte

Documento 5			
dist. tópicos	$p(t_1\|d_5) = \frac{1}{3}$	$p(t_1\|d_5) = \frac{1}{3}$	$p(t_1\|d_5) = \frac{1}{3}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 1	tópico 2	tópico 1	tópico 3
eléctrico	arte	servicio	tecnología

Documento 6			
dist. tópicos	$p(t_1\|d_6) = \frac{1}{2}$	$p(t_1\|d_6) = \frac{1}{4}$	$p(t_1\|d_6) = \frac{1}{4}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 3	tópico 2	tópico 1	tópico 1
CENDITEL	servicio	tecnología	CENDITEL

Con lo que finalmente se ha generado el corpus $\mathcal{C} = \{d_1, d_2, d_3, d_4, d_5, d_6\}$ donde;

Documentos del Corpus	
d_1	« CENDITEL CENDITEL servicio tecnología »
d_2	« música CENDITEL arte tecnología »
d_3	« tecnología eléctrico eléctrico servicio »
d_4	« arte música arte arte »
d_5	« eléctrico arte servicio tecnología »
d_6	« CENDITEL servicio tecnología CENDITEL »