Modelado de Tópicos

30 de marzo de 2015

Índice general

Capítulo 1

Modelado de tópicos

1.1. En qué consiste el modelado de tópicos

El modelado de tópicos es una herramienta que permite manejar un gran número de textos o documentos electrónicos para analizarlos, resumirlos, conocer su idea principal y archivarlos.

La motivación principal del modelado de tópicos es que en las últimas décadas los avances informáticos y tecnológicos han traido consigo que los textos y documentos sean cada vez más numerosos y aparezcan más frecuentemente en formato electrónico. Esto imposibilita que la fuerza humana pueda ser capaz de analizarlos todos y cada uno de ellos, principalmente por la enorme cantidad de tiempo que se requiere invertir para procesar esta gran cantidad de información. Para solucionar este problema, se recurre a automatizar este proceso.

En este sentido, muchos investigadores se han dedicado a desarrollar el *modelado de tópicos o modelado de tópicos probabilístico*, que consiste en una serie de algoritmos que analizan grandes colecciones de documentos con alguna temática en particular. En otras palabras, el modelado de tópicos es un método estadístico que permite analizar las palabras de los documentos, aglomerarlas en tópicos y ver cuál es la relación entre ellas, incluso permite determinar si estas cambian en el tiempo.

Es importante acotar que no se requiere información previa sobre los documentos, la inferencia de los tópicos surge del análisis de los documentos mismos. Para una mejor descripción de lo que es el modelado de tópicos se puede revisar el artículo [1].

1.2. Modelos probabilísticos

El propósito básico del modelado de tópicos es estudiar la condición de similitud que entre sí guarda un grupo grande de documentos. Este conjunto grande de documentos se denomina *corpus*.

Para simplificar, supongamos que la longitud de todos los documentos del corpus que estaremos estudiando, cuyos textos están formados por combinaciones de palabras de un mismo vocabulario, es constante e igual a seis. Es decir, todos los textos de este corpus tienen la misma longitud de seis palabras. Supongamos que el vocabulario a partir del cual están formados los textos de este corpus tiene solo tres palabras: sol, luna, planeta. Además, supongamos que el corpus tiene un número finito K de documentos. Para cada documento del corpus escogeremos seis palabras al azar y este procedimiento lo realizaremos K veces.

A lo largo de este documento seguiremos la siguiente nomenclatura, tomada de [3], donde cada símbolo indica un término:

- Palabra: es la unidad básica, definida como un ítem para un vocabulario y se desingará con el símbolo w_i, donde el subíndice i indica la i-ésima palabra del vocabulario.
- Vocabulario: es una colección de palabras. Se define como $\mathcal{W}=\{w_1, w_2, ..., w_M\}$.
- Tópico: es la distribución de palabras de un vocabulario fijo y se denotará como z_k, donde $z_k \in \mathcal{Z}=\{z_1, z_2, ..., z_K\}$, donde el subíndice k indica el k-ésimo tópico en la distribución de tópicos.
- Documento: es una secuencia de palabras, definido como d_j, donde el subíndice j indica el j-ésimo documento del corpus.
- Corpus: es una colección de documentos, definido como $\mathcal{D}=\{d_1, d_2, ..., d_N\}$.

Existen distintas formas para escoger al azar esas seis palabras de cada documento. A cada una de estas formas las llamaremos modelos y podemos distinguir entre los siguientes:

1. Supongamos que tenemos una caja con muchas pelotas etiquetadas con las palabras del vocabulario y que repetimos seis veces el experimento de extraer una pelota de la caja. En cada extracción estaríamos determinando una de las palabras de uno de los documentos del corpus. Si repetimos N veces (número de documentos) el procedimiento anterior, entonces estaríamos generando todo el corpus. Este modelo es llamado ***modelo de unigrama*** [2], donde la probabilidad de cada documento sería la distribución multinomial (ver apéndice A.1):

$$p(d) = \prod_{i=1}^{M} p(w_i), \tag{1.1}$$

   donde M=6, $p(d)$ es la distribución de probabilidad del documento d y w_i es cada una de las palabras que componen ese documento. El lado derecho de la ecuación quiere decir que se multiplican todas las probabilidades de la ocurrencia de esas seis palabras.

   Esto quiere decir que dentro del modelo de unigrama, las palabras de cada documento son extraidas de forma independiente.

2. Aumentemos un poco el modelo anterior, en el cual, la distribución de probabilidad de cada documento es exactamente la misma pues se usa la misma caja para generar todos los documentos. Supongamos ahora que no existe una única caja sino un número K de cajas (K representa el número de tópicos) donde cada una de ellas tiene una proporción distinta de pelotas etiquetadas con las palabras de nuestro vocabulario experimental.

   Para generar cada documento, escogemos al azar una de las varias cajas con pelotas, luego extraemos al azar las seis palabras del documento en cuestión. De este modo, cada documento es generado no necesariamente de la misma caja.

   Este modelo nos permite introducir la noción de tópico. En este ejemplo, el tópico es representado por la escogencia de cada una de la cajas, denotada por la distribución $p(z)$, que representa la probabilidad de que un documento sea generado a partir de un tópico determinado.

   Nótese que en este modelo, denominado ***mixtura de unigramas*** (ver [2] sección 4.2), cada documento es generado a partir de un tópico, donde su probabilidad sería:

$$p(d) = \sum_{z} p(z) \prod_{i=1}^{M} p(w_i|z). \tag{1.2}$$

   En otras palabras, en el modelo de mixtura de unigramas, cada documento es generado escogiendo primero un tópico z y luego generando las M palabras independientemente, a partir de la distribución multinomial condicional $p(w|z)$ (ver apéndice A.2 para una explicación sencilla de lo que es la probabilidad condicional).

3. Ahora pensemos en un modelo que permita generar un corpus en el que cada documento pueda estar compuesto por más de un tópico. Cada uno de los N documentos tiene determinada probabilidad de contener un tópico z_k de los K tópicos del corpus, donde cada z_k es una distribución multinomial sobre el vocabulario del corpus.

   Definamos dos dominios, uno para las palabras y otro para los documentos y preguntemos cuál es la probabilidad de que ocurran simultaneamente un elemento de cada dominio, condicionando dicha coocurrencia mediante una variable latente (u oculta) z con K posibles valores (ver [3] sección 3.1). En otras palabras, ¿Cuál es la probabilidad de que la palabra w_i ocurra en el documento d_j dado que dicha palabra proviene del tópico z_k.

   Formalizando esta propuesta tendríamos la siguiente ecuación cuyo desarrollo conduce al modelo ***Probabilistic Latent Semantic Analysis*** o PLSA [3]:

$$p(d, w) = p(d)p(w|d) \quad donde \quad p(w|d) = \sum_{z \in Z} p(w|z)p(z|d). \tag{1.3}$$

   Este modelo introduce un nuevo concepto de dependencia condicional asumida, donde el documento d y la palabra w son condicionalmente independientes de la variable latente (u oculta). Parametrizando la ecuación anterior se obtiene la distribución de probabilidad conjunta[1]:

[1]Distribución de probabilidad conjunta (joint probability distribution): dadas dos variables aleatorias x, y que son definidas en un espacio de probabilidades, la distribución de probabilidad conjunta es una distribución que da la probabilidad de que cada x, y caiga en un rango articular o conjunto discreto de valores específicos para

$$p(d, w) = \sum_{z \in \mathcal{Z}} p(z)p(d|z)p(w|z). \tag{1.4}$$

La ecuación anterior quiere decir, en palabras simples, que dado un documento d y una palabra w, los cuales son condicionalmente independientes, $p(d, w)$ es la probabilidad de la ocurrencia de esa palabra dentro de ese documento, dada una variable oculta z (tópico).

Este modelo trata de generalizar la suposición del modelo de mixtura de unigramas, donde cada documento es generado solamente por un tópico, asumiendo la posibilidad de que cada documento pueda contener varios tópicos.

Sin embargo, este modelo tiene dos grandes desventajas. Una de ellas es que d es una variable aleatoria multinomial con tantos valores posibles como documentos entrenados[2] hayan y el modelo aprende la mixtura de tópicos $p(z|d)$ solo para aquellos documentos que hayan sido entrenados, por tanto no hay una forma natural de asignar probabilidades a documentos que no hayan sido previamente examinados. Entonces, cada vez que se incorpora un nuevo documento al conjunto entrenado debe recalcularse todo el modelo.

Otra desventaja importante es que como utiliza una distribución añadida de documentos entrenados, el número de parámetros que deben ser estimados crecen linealmente con el número de documentos entrenados. Esto sugiere que el modelo es propenso a sobreajustarse[3]. Esto es un grave problema ya que los modelos que tienden a sobreajustarse tienen un comportamiento predictivo pobre.

Con el objetivo de eliminar estos problemas surge el modelo LDA o ***Latent Dirichlet Allocation***, ya que trata el peso de la mixtura de tópicos como una variable aleatoria oculta y no como un conjunto grande de de parámetros individuales que son explícitamente enlazados con documentos entrenados.

esas variables. Si se trata de dos variables se llama función bivariada, si son más de dos variables se llama función multivariada.

[2]Con información previa.

[3]El sobreajuste (overfitting) ocurre cuando el modelo tiende también a ajustar los errores, reconoce estos como información verdadera y no como errores. Por lo general, sucede en modelos complejos con muchos parámetros.

Capítulo 2

Latent Dirichlet Allocation

2.1. Conceptualizando el LDA

El LDA es un modelo estadístico de colecciones de documentos que trata de capturar la escencia de estos, encontrando palabras relacionadas con ciertos tópicos y definiendo en qué proporción están estas mezcladas. El LDA refleja la intuición de que los documentos contienen diferentes tópicos y cada documento contiene estos tópicos en diferentes proporciones.

La característica principal del LDA es que es capaz de distinguir, no solo los tópicos de los que está compuesto el documento, sino también de discernir las diferentes proporciones de tópicos de los que cada documento está compuesto, aunque todos los documentos compartan los mismos tópicos.

Para visualizar esto, tomemos como ejemplo a Blei 2012 [1] y su figura 1, que en este documento estará etiquetada como figura 2.1. En esta, se han seleccionado palabras que han sido asignadas a ciertos tópicos y resaltadas con los colores amarillo, rosado y azul, dependiendo del tópico asignado. Sigamos con el ejemplo de Blei 2012 [1], donde en la figura se han resaltado las siguentes palabras:

Palabras	Tópico	Color
computer, prediction	data analysis	azul
life, organism	evolutionary biology	rosado
sequenced, genes	genetics	amarillo

Es importante señalar que se descartan las palabras con poco contenido, por ejemplo, los artículos (la, los, un, unos, etc), las preposiciones (a, con, por, en, para, etc) y los conjuntivos (cuando, porque, aunque, etc).

Entonces, para cada documento, se generan las palabras en dos pasos:

- Se escoge de forma aleatoria una distribución de tópicos.

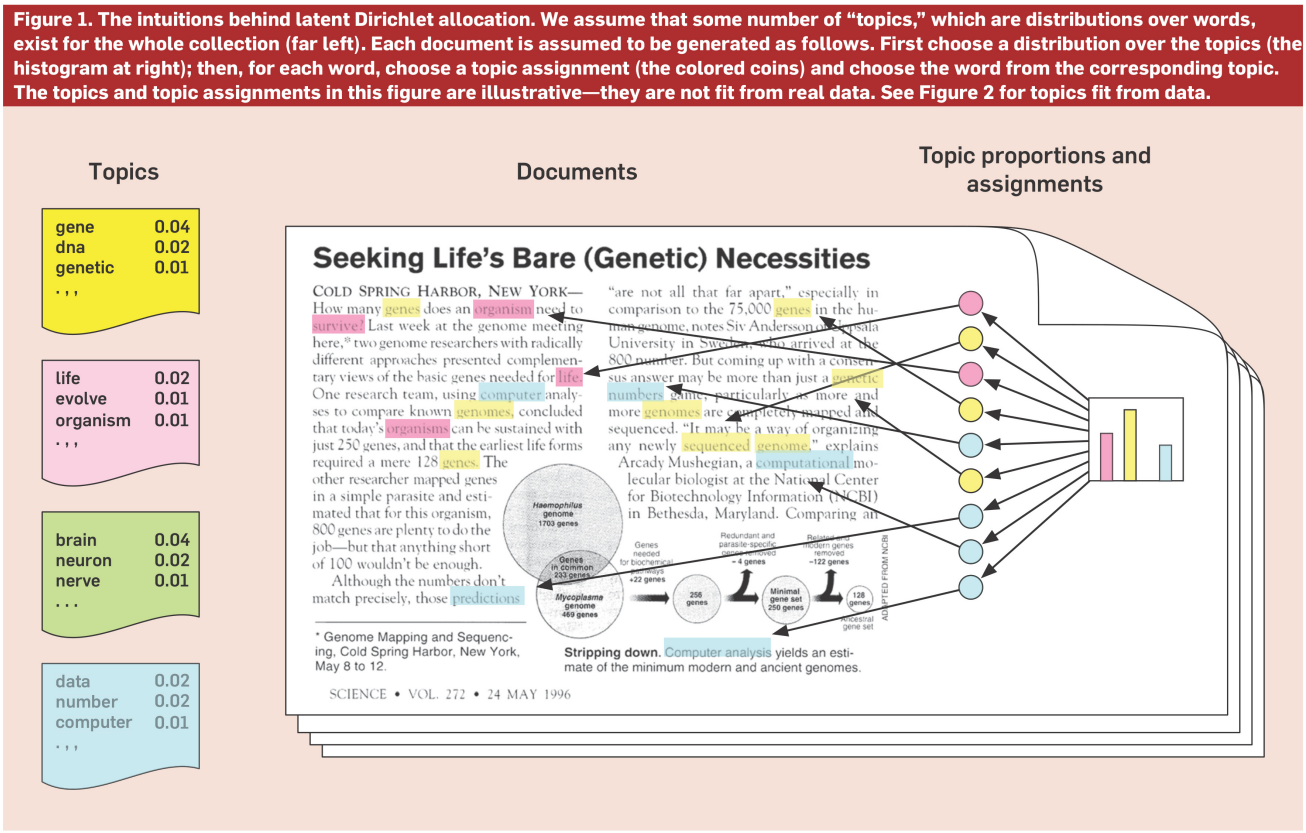

Figura 2.1: Figura 1 de Blei 2012 [1]

- Para cada palabra del documento (a) se elige de forma aleatoria un tópico de la distribución de tópicos, luego, (b) también de forma aleatoria, se escoge una palabra de la distribución de volcabulario del tópico correspondiente.

Esto garantiza que cada documento esté compuesto por tópicos en diferentes proporciones, ya que cada palabra en cada documento es extraida de un tópico, donde este último es escogido de la distribución de tópicos del documento.

En el ejemplo de la figura 2.1, los tópicos están representados en los recuadros de la izquierda y los colores representan la asignación de tópicos donde, contando cada palabra que haya sido asignada a cada tópico, se obtiene la distribución representada en el recuadro a la derecha. Esta distribución es lo que se llama distribución de Dirichlet.

Este ejemplo simplificado se puede expandir en el caso de tener cientos o miles de documentos, sobre el cual se emplean los algoritmos de modelado de tópicos para inferir la distribucón de tópicos, haciendo ńfasis en que cada documento no contendrá todos los tópicos sino unos pocos donde se podrán estudiar las palabra más probable en los tópicos más probables para cada documento.

Es importante resaltar que, como dice Blei 2012 [1], los algoritmos no tienen información sobre el tema sobre el cual los documentos están escritos y tampoco los documentos están etiquetados con los tópicos o palabras claves. La distribución de tópicos surge de analizar cuál es la estructura oculta más probable para generar la colección de documentos observada.

2.2. Proceso generativo del LDA

Capítulo 3

Métodos de inferencia

3.1. Métodos de muestreo

3.2. Métodos variacionales

3.2.1. Inferencia Variacional Bayesiana

Apéndice A

Conceptos básicos de estadística

A.1. Distribución multinomial

En probabilidad, una distribución multinomial se refiere a cuando un número finitos de procesos tienen la misma probablilidad de ocurrir. Esto es una generalización de la distribución binomial en donde existen solo dos probabilidades.

Por ejemplo, si se tira una moneda al aire exite la misma probabilidad de que caiga del lado de la cara o del sello. Si esa moneda se lanza muchas veces y se va anotando cuántas veces cae cara y cuántas cae sello se obtiene una distribución binomial.

Ahora, supongamos que en vez de una moneda tenemos una caja llena de muchas pelotas de colores (rojo, amarillo, azul, verde, blanco y negro) y que cada vez que sacamos una pelota, la sacamos de un color diferente. Si luego de sacar un número finito de pelotas contamos cúantas pelotas hay de cada color, obtenemos una distribución multinomial.

A.2. Distribución de probabilidad condicional

La distribución de probabilidad condicional se define como la probabilidad de que ocurra un evento A suponiendo que otro evento B es verdadero.

En términos generales, la probabilidad se escribe según la siguiente nomenclatura:

i) Probabilidades independientes: $P(A)$, $P(B)$ es la probabilidad de que A y B ocurran de forma independiente una de la otra.

ii) Probabilidades condicionales: $P(A \mid B)$ es la probabilidad de que A ocurra si B es verdadera y $P(B \mid A)$ es la probabilidad de que B ocurra si A es verdadera.

Formalmente, la probabilidad condicional se define como:

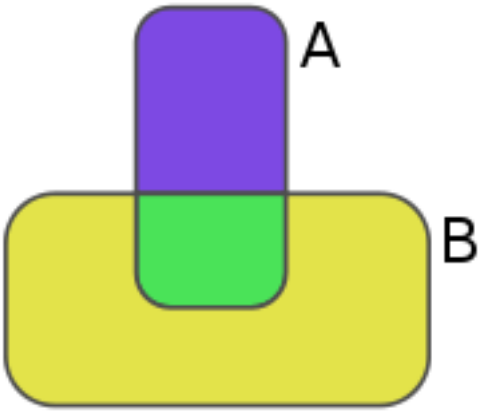

Figura A.1: Probabilidad condicional $P(A \mid B)$. Se puede pensar como en el espacio en el que B es verdadero (área amarilla) también se cumple que A es verdadero (área modada). Entonces $P(A \mid B)$ se representa en esta figura como el área verde.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{A.1}$$

El símbolo $\cap$ quiere decir interseccíon. La ecuación anterior quiere decir que la probabilidad de que A ocurra sabiendo que B es verdadero (lado izquierdo de la ecuación) es igual al espacio donde A y B se intersectan (ver figura A.1).

En la Figura A.1 [1] se puede ver una representación gráfica de lo que se define como probabilidad condicionada.

Un ejemplo sencillo de esto [2] sería el siguiente. Si el 50 % de la población fuma y el 10 % además de que fuma también es hipertensa:

$P(A) =$ fuma (0.5)

$P(B) =$ hipertensa (0.1)

La probabilidad de encontrarse con una persona que fuma y es hipertensa es:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}1}{0{,}5} = 0{,}2. \tag{A.2}$$

Esto quiere decir que la probabilidad de que se escoga una persona al azar y esta sea fumadora e hipertensa es del 20 % (esto representará la zona verde en la figura A.1.

[1] La figura A.1 es tomada de http://es.wikipedia.org/wiki/Probabilidad_condicionada

[2] Este ejemplo es tomado de http://www.hrc.es/bioest/Probabilidad_15.html

A.3. Distribución de Dirichlet

La distribución de Dirichlet es una familia de distribuciones de probabilidad multivariadas contínuas, parametrizadas por un vector $\boldsymbol{\alpha}$ de números reales positivos. Usualmente se denota por $\text{Dir}(\boldsymbol{\alpha})$ y se define como diremos a continuación.

Siendo la distribución de Dirichlet de orden $K \geq 2$ y parámetros $\alpha_1, ..., \alpha_K > 0$, la función de densidad de probabilidad viene siendo:

$$f(x_1, ..., x_{K-1}; \alpha_1, ..., \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K}, x_i^{\alpha_i - 1} \tag{A.3}$$

donde $B(\boldsymbol{\alpha})$ es la función Beta definida como:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \tag{A.4}$$

la cual está definida en función de la función Gamma Γ. Por tanto, la distribución de Dirichlet se puede ver como una versión multivariada de la distribución Beta.

Es importante acotar que la distribución de Dirichlet es usada comúnmente en estadística Bayesiana como distribuciíon previa a priori o prior (como es en el caso de la LDA).

La función de densidad de probabilidadón (A.3) regresa la convicción de que la probabilidad de ocurrencia de K eventos es x_i dado que cada evento se observó $\alpha_i - 1$ veces.

Bibliografía

[1] Blei, D. *Probabilistic topic models*. Communications of the ACM. 55, 4 2012.

[2] Blei, D., Y.Ng. A. & Jordan. M. *Latent Dirichlet Allocation*. Journal of Machine Learning Research. 3, 993 2003.

[3] Hofmann., T. *Probabilistic latent semantic analysis*. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Pág. 289-296, 1999.

.

.