Modelado de Tópicos

****

3 de junio 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 contenido 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*, 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 que permite analizar las palabras de los documentos, aglomerarlas en tópicos y ver cuál es la relación entre palabras y tópicos, incluso permite determinar si estos cambian en el tiempo.

Dentro del conjunto de modelos para modelar tópicos, están aquellos que utilizan la teoría de probabilidad para modelar la incertidumbre en los datos y son llamados *modelos probabilísticos de tópicos*. Estos modelos describen un conjunto de distribuciones de probabilidades posibles para un conjunto de datos observados y el objetivo es utilizar los datos observados para determinar la distribución que mejor describa estos datos.

A lo largo de este documento seguiremos la siguiente nomenclatura, tomada de Hofmann (1999) [4]:

- Palabra: es la unidad básica definida como un ítem de 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 probabilidad de palabras de un vocabulario 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 y está denotado como d_j, donde el subíndice j indica el j-ésimo documento del corpus.
- Corpus: es una colección de documentos y está denotado como $\mathcal{D}=\{d_1, d_2, ..., d_N\}$.

Dependiendo del método utilizado para generar el modelo, estos se pueden dividir en modelos *supervisados* y *no supervisados*.

Los modelos supervisados incorporan etiquetas en el proceso de aprendizaje del algoritmo. Este utiliza documentos etiquetados o entrenados con la finalidad de clasificar el cuerpo entero de documentos. El método supervisado puede explotar la información de datos o documentos enrenados que pueden ser utilizados para calcular predicciones de datos no observados.

Por otro lado, los modelos no supervisados no requieren de ningún tipo de documentos entrenados, por lo tanto puede ser aplicado a cualquier texto del que no se tenga ninguna información previa. Es importante acotar que ya 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 se puede revisar el artículo [1].

1.2. Modelos probabilísticos

Como se había dicho, los modelos probabilísticos de tópicos utilizan la teoría de probabilidad para definir la distribución que mejor se ajusta a los datos observados y cuyo propósito básico 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: vaca, gandola, vieja. 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.

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 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:

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

   donde $p(d)$ es la distribución de probabilidad conjunta para todos los documentos y se fundamenta en la ley de probabilidad total. Veamos una pequeña demostración.

   La ley de probabilidad total dice que dado un suceso conocido con probabilidades condicionadas a un evendo, las cuales también son conocidas junto con sus probabilidades individuales, también conocidas (ver apéndice A.5), la probabilidad total de que ocurra el suceso es:

$$p(d) = \sum_{z} p(z)p(d|z), \tag{1.3}$$

   pero

$$p(d|z) = p(w_1, w_2, w_3, ..., w_M|z) = \prod_{i=1}^{M} p(w_i|z). \tag{1.4}$$

Si se sustituye esta última ecuación en (1.3) obtenemos

$$p(d) = \sum_{z} p(z) \prod_{i=1}^{M} p(w_i|z), \tag{1.5}$$

que es la misma ecuación (1.2).

Entonces, 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 [4] 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 [4]:

$$p(d, w) = p(d)p(w|d) \quad donde \quad p(w|d) = \sum_{z \in Z} p(w|z)p(z|d). \tag{1.6}$$

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

$$p(d, w) = \sum_{z \in \mathcal{Z}} p(z)p(d|z)p(w|z). \tag{1.7}$$

   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

[1] Ver apéndice A.3.
[2] Con información previa.

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.

[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

Interpretación geométrica del modelado de tópicos

Para visualizar el modelado de tópicos, podemos considerar los elementos geométricos del espacio sobre el que trabajamos las variables latentes. A partir de allí, es posible representar los documentos del corpus. Definiremos algunos conceptos geométricos que nos ayudarán en esta interpretación.

Emepecemos por el símplex. Un *símpex* de dimensión m, que denotaremos Δ_m, es una colección de puntos en $\mathbb{R}^m$ que satisfacen la siguiente condición

$$\Delta_m := \{\alpha \in \mathbb{R}^m : \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0 \text{ para } i = 1, \ldots, M\} \tag{2.1}$$

El símplex puede ser visto como la generalización a varias dimensiones de un triángulo en el plano.

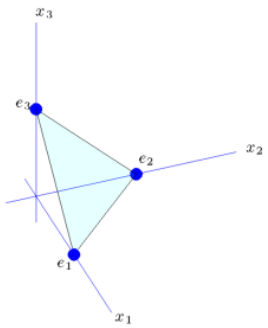

Figura 2.1: Símplex de dos dimensiones en $\mathbb{R}^3$.

En particular, una función de probabilidad sobre un espacio muestral finito $\Omega = \{\omega_1, \ldots, \omega_k\}$ se puede caracterizar por un vector de valores probables $\overline{\theta} = (\theta_1, \ldots, \theta_k)$ satisfaciendo, para cada $i = 1, \ldots, k$, $\theta_i \geq 1$ y $\sum \theta_i = 1$. Si definimos la función de probabilidad por la ecuación

$$P(A) = \sum_{i \in A} \theta_i$$

el símplex Δ_k parametriza el conjunto de todas las distribuciones sobre el espacio muestral Ω de tamaño k. Es decir, existe una correspondencia biunívoca entre los elementos del símplex y las diferentes distribuciones que existen sobre dicho espacio muestral

En $\mathbb{R}^m$ también podemos definir a un conjunto convexo. Un conjunto $S \subset \mathbb{R}^n$ se dice que es un *conjunto convexo* si $\alpha x + (1 - \alpha)x'$ pertenece a S siempre que x y x' están también en S, y α varía libremente en el intervalo real $[0, 1]$. Esto significa es otras palabras que dados dos puntos x y x' pertenecientes a S, el segmento de recta que los une también está enteramente contenida en dicho conjunto.

Como caso particular podemos notar que un símplex es un conjunto convexo.

Otro concepto que necesitaremos en nuestro ejercicio de interpretación geométrica es el de cápsula convexa. La cápsula convexa de un conjunto no vacío S, denotado $\operatorname{co} S$ es la intersección de todos los conjuntos convexos que contienen a S. En otras palabras, es el conjunto convexo más pequeño que contiene a S, dado que la intersección de conjuntos convexos es también un conjunto convexo. Formalizando estas ideas

$$\begin{aligned} \operatorname{co} S &:= \cap\{C : C \text{ es convexo y contiene a } S\} \\ &= \{x \in \mathbb{R}^n : \text{ para } k \in \mathbb{N} \text{ existen } x_1, \ldots, x_k \in S \text{ y } \overrightarrow{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \Delta_k : \sum_{i=1}^{k} x_i = x\} \end{aligned}$$

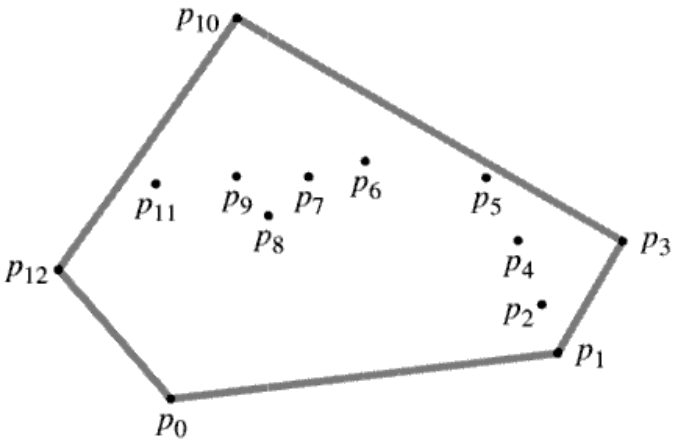

Figura 2.2: Cápsula convexa del conjunto de puntos $S = \{p_1, \ldots, p_{12}\}$.

Capítulo 3

Latent Dirichlet Allocation

3.1. Conceptualizando el LDA

El LDA pertenece al tipo de modelos estadísticos 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.

Para visualizar esto, tomemos como ejemplo a Blei 2012 [1] y su figura 1, que en este documento estará etiquetada como figura 3.1. 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.
- 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.

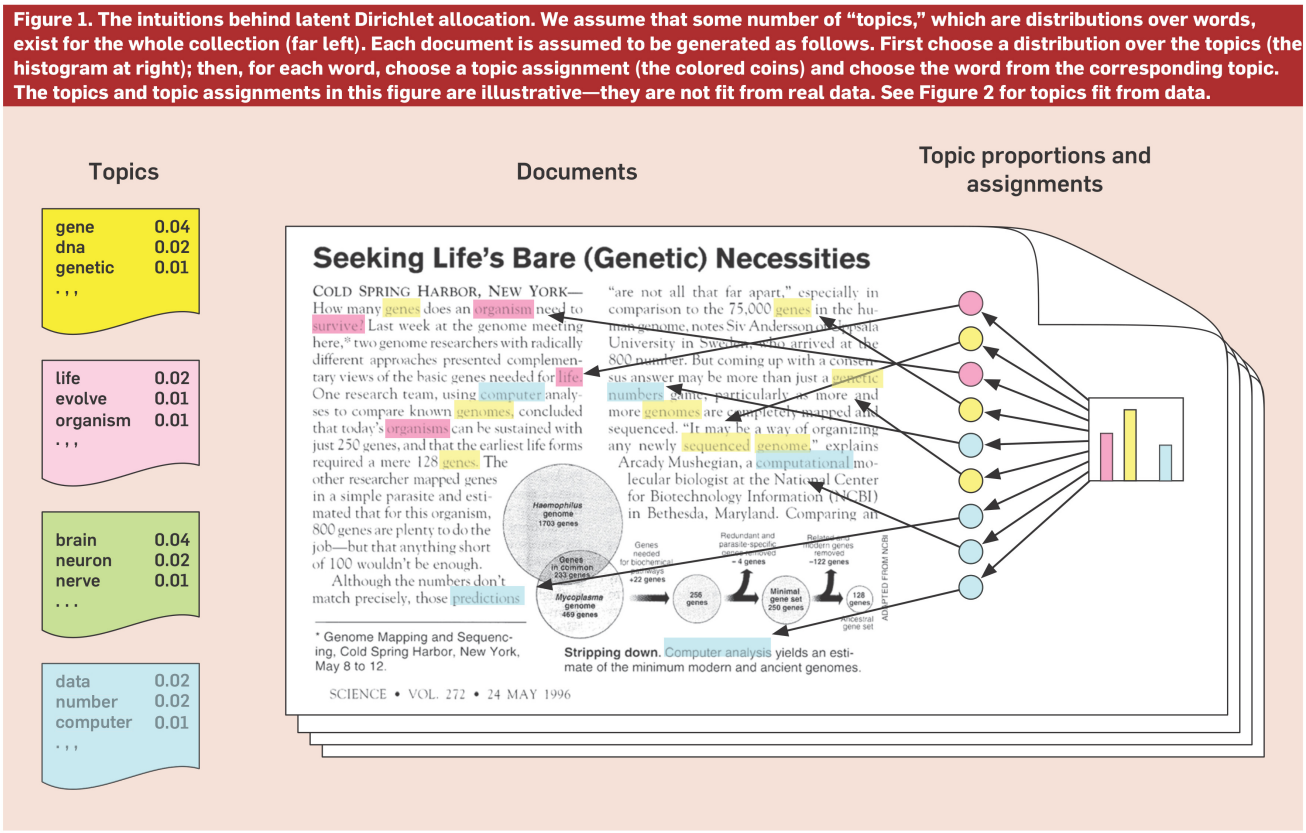

Figura 3.1: Figura 1 de Blei 2012 [1]

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.

FALTA

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.

3.2. Proceso generativo del LDA

(Tomado de Blei 2012, 2003 [1, 2])

Los datos, que incluyen las variables ocultas, surgen del proceso generativo.

El proceso generativo 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 las variables observadas. Esta distribución condicional es lo que en estadística bayesiana se llama distribución a posteriori.

El LDA asume el siguiente proceso generativo para cada documento d en el corpus $\mathcal{D}$:

- Se escoge M según la distribución de Poisson (ξ)[1].
- Se escoge θ según la distribución de Dirichlet(α)[2].
- Para cada una de las M palabras:
  1. Se escoge un tópico z_k según una distribución multinomial (θ).
  2. Se escoge una palabra w_m a partir de $p(w_m|z_k, \beta)$ la cual es una distribución de probabilidad condicionada al tópico z_k.

El LDA supone lo siguiente:

- El hiperparámetro β es una matriz que contiene la probabilidad de las palabras, donde cada componente de la matriz es $\beta_{ij} = p(w^j = 1|z^i = 1)$. Estas se suponen cantidades fijas a ser estimadas.
- El número de tópicos es conocido. Así, la dimensionalidad de la distribución de Dirichlet es conocida y fija.
- La distribución de Poisson define la distribución de longitudes de los documentos. Esta puede ser cambiada si es necesario.

La siguiente densidad de probabilidad, es la distribución de Dirichlet de k dimensiones con una variable aleatoria θ (ecuación (1) de Blei 2003 [2]):

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}, \tag{3.1}$$

El hiperparámetro α es un vector con componentes $\alpha_i > 0$ (falta: ¿con qué está relacionado?)

¿Por qué se usa la distribución de Dirichlet? Eso es porque es de la familia de las exponenciales, es de dimensionalidad finita y es conjugada de la distribución multinomial. Estas características hacen más manejable el procedimiento del método bayesiano que subyace en el LDA (inferencia y estimación de parámetros).

CONTINUAR

[1] M es el número de palabras.
[2] θ es la porción de tópicos.

Capítulo 4

El proceso Generativo del LDA

Supongamos que las palabras de un vocabulario determinado que pudieran aparecer en un texto, asumiento que el orden de las palabras en el texto no importa, se distribuye multinomialmente: $P(w|\beta)\ M(\beta)$, donde β es un vector con tantas componentes como palabras haya en el vocabulario, cuya i-ésima componente representa la probabilidad de que la i-ésima palabra del vocabulario ocurra w_i veces en el texto. Nótese que el parámetro β variará dependiendo del contexto temático del cual provenga el texto en cuestión, haciendo más probable la aparición de ciertas palabras y menos probable la aparición de otras. Por ejemplo, si el texto proviene del campo de las artes tal vez sea menos probable encontrar en él, por decir algo, la palabra guerra que la palabra color.

En el contexto del razonamiento Bayesiano cada punto x observado —en nuestro caso, la frecuencia w_i de cada palabra del vocabulario en cada documento— es una oportunidad para mejorar nuestro modelo, y para esto se ajustan sus parámetros con cada observación. Si observamos x entonces modificamos los parámetros en función de incorporar la nueva observación, esto es, en función de que el modelo "aprenda". Por ejemplo, si en los textos observados hasta el momento aparecen con frecuencia las palabras color, pincel y belleza tal vez convenga entonces modificar nuestro parámetro β para incorporar lo aprendido, por ejemplo aumentando la probabilidad de que aparezcan también otras palabras afines al discurso artístico. En general, el parámetro β podría variar dependiendo de los tópicos tratados en los textos encontrados hasta el momento.

En lo que sigue supondremos que cada documento trata un único tópico, para lo cual estudiaremos en clásico modelo de agrupamiento Dirichlet-Multinomial[1], el cual propone solamente dos niveles aleatoriedad. Este modelo nos servirá de introducción al modelo LDA y sus tres niveles de aleatoriedad.

Entonces, ¿cuánto y cómo deben modificarse los parámetros en función de los datos observados? Para precisarlo usualmente se supone que los parámetros —en nuestro caso, el parámetro β de la multinomial $P(w|\beta)$— son aleatorios y que su distribución de probabilidad depende de las nuevas observaciones. Por ejemplo, podemos suponer que β se distribuye aleatoriamente dependiendo de los tópicos tratados en los textos. El primer paso entonces consiste en precisar la distribución de probabilidad de ese parámetro aleatorio condicionada por la ocurrencia de cada observación w:

[1] En [2], en la página 997, se advierte sobre la diferencia entre el clásico agrupamiento Dirichlet-Multinomial y el LDA.

$P(\beta|w)$. De acuerdo con el teorema de Bayes, esa distribución sería:

$$p(\beta|w) = \frac{P(w|\beta P(\beta))}{\int P(w|\beta)P(\beta)d\beta} \tag{4.1}$$

En esta fórmula conocemos $P(w|\beta)$, que es nuestro modelo multinomial inicial. El segundo paso es precisar la distribución del parámetro β. Para esta distribución, por varias razones que valdrá la pena estudiar en otro momento, suele escogerse una conjugada a priori. De acuerdo con este criterio, para el caso de un modelo multinomial podríamos escoger la distribución de Dirichlet, tal que $\beta\ D(\alpha)$. Operando en la ecuación (4.1) tendríamos la siguiente expresión:

$$p(\beta|w, \alpha) = K \prod_{i=0}^{N} \beta_i^{w_i+\alpha_i-1}$$

Este desarrollo nos permite enriquecer nuestro modelo multinomial inicial por dos razones principales:

(a) incorpora la intuición de que la frecuencia de las apariciones de las palabras de un cierto vocabulario en un texto determinado están condicionadas por el tema tratado en cada documento, lo cual permite realizar un agrupamiento o categorización de los datos, v.g. los textos;

(b) mediante el hiperparámetro α, el cual podremos ajustar a priori, podemos incorporar al modelo cierta idea de nuestra "confianza" en el valor actual del parámetro β, pues para valores grandes de α la distribución de $P(\beta|w)$ variará relativamente poco.

Ahora bien, para generar datos partiendo de este modelo, es decir, para generar cada texto del corpus que estamos modelando tendríamos que, primero, extraer un valor aleatorio de β de la distribución $P(\beta|\alpha)\ D(\alpha)$. Luego, debemos generar cada una de las palabras del texto mendiante la distribución multinomial de cada documento $P(w|\beta)\ M(\beta)$. Recordemos que w es una vector con tantas componentes como palabras haya en el vocabulario, cuyas i-ésima componente w_i representa el número de veces que la i-ésima palabra del vocabulario aparece en el texto en cuestión. Vale recordar en este punto que la posibilidad de usar la distribución multinomial, condicionada por el parámetro aleatorio β, para modelar la frecuencia con la que aparecen las palabras en los textos radica en la suposición de que las palabras de los textos son intercambiables y por tanto condicionalmente independientes sobre una variable aleatoria oculta, en este caso, el parámetro β.

La probabilidad (total o absoluta, es decir, sin condicionantes) de un texto sería entonces:

$$p(w) = \int P(\beta|\alpha)P(w|\beta)d\beta$$

Ahora bien, el modelo discutido hasta ahora tiene dos niveles de aleatoriedad: primero se determina el parámetro aleatorio β usando la distribución de Dirichlet y su parámetro α, y luego se determinan las palabras del texto que esté siendo generado usando la distribución multinomial

con parámetro β. Infiriendo el valor de α que mejor se ajuste a un determinado corpus de textos y a otros criterios complementarios[2], podremos encontrar los grupos temáticos de dicho corpus (recordando que este modelo de dos niveles admite un único tópico por documento), los cuales corresponderán a las regiones del símplex sobre el cual se define la distribución de β $D(\alpha)$, en los cuales ocurra una mayor densidad de probabilidad.

El modelo LDA propone un nivel adicional de aleatoriedad que permite introducir la intuición de que un texto puede estar asociado a más de un tópico. Para esto el modelo supone que cada texto está conformado por una mixtura aleatoria de tópicos, representada por una multinomial de parámetro θ, y que el parámetro de esta mixtura es aleatorio tal que θ $Dirichlet(\alpha)$. Entonces, en la generación de textos usando el modelo LDA, para cada texto primero se determina aleatoriamente el parámetro θ a partir de una distribución de Dirichlet de parámetro α. Este parámetro θ es un vector con tantas componentes como tópicos se deseen en el modelo, en donde la i-ésima componente es una medida de la probabilidad con la cual el i-ésimo tópico condicionará cada una de las palabras usadas en el texto que esté siendo generado. Es decir, cada palabra de cada texto es generada, primero, determinando un tópico a partir de una multinomial con parámetro $\theta : P(z)$ $M(\theta)$. Luego, la probabilidad de que la i-ésima palabra del vocabulario ocurra w_i veces, dado que el tópico z_i fue escogido previamente de acuerdo con $M(\theta)$, será a su vez una distribución multinomial: $P(w|z_i, \beta)$.

[2]Como la divergencia entre las regiones de alta densidad de probabilidades. Vale la pena precisar rigurosamente estos criterios.

Capítulo 5

Métodos de inferencia

5.1. Métodos de muestreo

5.2. Métodos variacionales

5.2.1. Inferencia Variacional Bayesiana

Capítulo 6

Medidas de evaluación del rendimiento

6.1. Método de máxima verosimilitud

Es un método estándar para ajustar un modelo y encontrar parámetros [6]. Para la medición de parámetros del modelo mediante la colección de documentos D se maximiza la verosimilitud (credibilidad) mediante la siguiente fórmula:

$$p(D) = C \prod_{d \in D} \prod_{w \in d} p(d,w)^{n_d w} = \prod_{d \in D} \prod_{w \in d} p(w|d)^{n_d w} C p(d)^{n_d w} \longrightarrow \text{máx} \tag{6.1}$$

Donde C es un factor normalizador que depende sólo del número n_{dw}. En miras a maximizar en (6.2) puede obviarse el término $Cp(d)^{n_d w}$ ya que es constante en toda la colección D y recordamos la fórma de $p(w|d)$ en (1.6), donde escribiremos $\theta_{td} = p(t|d)$ y $\varphi_{wt} = p(w|t)$. Luego, después de pasar logaritmos obtenemos

$$\mathcal{L}(D; \Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log_a \sum_{t \in T} \varphi_{wt} \theta_{tw} \longrightarrow \max_{\Phi, \Theta} \tag{6.2}$$

Donde a, la base del logaritmo, generalmente se toma como $a = 2$ si el cálculo es binario.

Es usada para encontrar el número de tópicos mostrando la capacidad de generalización de un modelo sobre datos no observados. No requiere una categorización previa. Valores pequeños de la perplejidad indican una mayor capacidad de generalización del modelo. Es el criterio estándar más usado para evaluar la calidad del modelo [6].

Formalmente, es una medida de discrepancia del modelo $p(w|d)$ con el término w observado en los documentos de la colección D. Una manera de definir la perplejidad es a través de la verosimilitud:

$$\mathcal{P}(D; \Phi, \Theta) = a^{-\frac{1}{n}\mathcal{L}(D;\Phi,\Theta)} = a^{-\frac{1}{n}\sum_{d \in D}\sum_{w \in d} n_{dw} \log_a p(w|d)} \tag{6.3}$$

Mientras esta medida es menor, mejor el modelo $p(w|d)$ predice la aparición de los términos w en el documento d.

Es evidente que la perplejidad, calculada para la colección particular D para la cual fue construido el modelo Φ-Θ, puede estar sujeta a los efectos del recálculo y dar valores distintos para distintas ejecuciones de los algoritmos de cálculo de la misma.

6.2. Entropía

La Entropia mide la incertidumbre de una fuente de información. C. Shannon (1916-2001), la consideró inicialmente como la cantidad de información promedio que contienen los símbolos usados para generar un mensaje: Los símbolos con menor probabilidad son los que aportan mayor información [8].

Shannon estudió la generación de mensajes bajo un esquema probabilistico

$$A = \begin{pmatrix} a_1, & \dots, & a_i, & \dots \\ p(a_1), & \dots, & p(a_i), & \dots \end{pmatrix}, \quad \sum_{i=1}^{n} p_i = 1 \quad 0 \leq p_i \leq 1$$

En relación con la cantidad de información contenida en un mensaje T de longitud l, compuesto por un esquema de eventos independientes, se expresa bajo las siguientes axiomas:

(a) Un mensaje vacío no contiene información.

(b) La cantidad de información contenida en un mensaje es proporcional a su longitud.

Así, la cantidad de información del mensaje $T = a_1 \dots a_l$ es igual a $I(T) = lH$ donde

$$H = -\sum_{i=1}^{n} p(a_i) \log(p(a_i))$$

es la **entropía**.

La entropía puede ser usada para medir la calidad de los tópicos, revela el desorden del sistema. A menor entropía usualmente el sistema es mejor en el sentido de la relevancia de los tópicos descubiertos.

Habiendo definido la entropia, podemos decir que la perplejidad puede expresarse como 2 elevado a la entropía.

6.3. Divergencia de Kullback-Leibler

Es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad [7]. Mide la cantidad esperada de información extra en muestras de la distribución p cuando se usa la distribución q.

De manera formal definimos así: dadas dos distribuciones p y q sobre una variable aleatoria discreta la divergencia KL está dada por

$$D_{KL}(p\|q) = \sum_i p(i) \log \frac{p(i)}{q(i)}$$

De manera alternativa, podemos escribir

$$\begin{aligned} D_{KL}(p\|q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \\ &= -H(p) + H(p,q) \end{aligned}$$

donde $H(p)$ es la entropía de la distribución p y $H(p,q)$ es la entropía cruzada de p y q.

La divergencia KL se usa para medir la calidad de los tópicos en términos de una pseudo-distancia entre ellos. Mide la relación entre dos tópicos: mayor distancia entre los tópicos es un indicador de mayor calidad del sistema.

Una baja entropía y alta divergencia entre los tópicos es un indicador de dispersión del sistema, esto nos dice la capacidad de generalización del sistema.

Propiedades de la Divergencia KL:

1. La divergencia KL es no negativa.
2. Si $\Omega_p = \Omega_q$, la divergencia KL es igual a 0 si y sólo si las distribuciones son iguales $p_i \equiv q_i$.
3. Si P es una distribución empírica y $Q(\alpha)$ es una familia parametrizada de distribuciones, entonces minimizar la divergencia KL es equivalente a maximizar la verosimilitud:

$$D_{KL}(P\|Q(\alpha)) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i(\alpha)} \longrightarrow \min_{\alpha} \Longleftrightarrow \sum_{i=1}^{n} p_i \log q_i(\alpha) \longrightarrow \max_{\alpha}$$

Capítulo 7

Ejemplo de aplicación del LDA

Tomemos el vocabulario $W = \{\text{arte}, \text{música}, \text{eléctrico}, \text{cenditel}, \text{tecnología}, \text{servicio}\}$.

Definamos tres tópicos t_1, t_2, t_3, los cuales tienen una probabilidad de ocurrencia dada para cada palabra del vocabulario y se distribuyen multinomial:

$$\begin{aligned} t_1 &= \left\{x_1 = \frac{1}{5}; x_2 = \frac{1}{5}; x_3 = \frac{1}{5}; x_4 = \frac{1}{10}; x_5 = \frac{1}{10}; x_6 = \frac{1}{5}\right\} \\ t_2 &= \left\{x_1 = \frac{1}{20}; x_2 = \frac{1}{20}; x_3 = \frac{1}{5}; x_4 = \frac{1}{4}; x_5 = \frac{1}{4}; x_6 = \frac{1}{5}\right\} \\ t_3 &= \left\{x_1 = \frac{7}{20}; x_2 = \frac{1}{4}; x_3 = \frac{1}{10}; x_4 = \frac{1}{20}; x_5 = \frac{1}{20}; x_6 = \frac{1}{5}\right\} \end{aligned}$$

donde $x_i = p(w_i)$

La distribución multinomial tiene función de densidad dada por

$$f(x_1, \ldots, x_k) = \frac{n!}{\prod x_i!} \prod p_i^{x_i}$$

Apéndice A

Conceptos básicos de estadística

A.1. Distribución multinomial

En probabilidad, una distribución multinomial se refiere a cuando un número finito 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.

Esto es lo mismo que decir que una distribución binomial con parámetros n y p es la distribución de probabilidad discreta del número de sucesos en una secuencia de n experimentos independientes (verdadero/falso), cada uno de los sucesos con probabilidad de ocurrencia p. Esto se escribe matemáticamente como:

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{A.1}$$

para k=0,1,2,3...,n. Donde:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{A.2}$$

Esta ecuación (A.2), es conocida como coeficiente 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.

Formalmente, si se define x_i como una la variable aleatoria que indica el número de veces que se ha dado el resultado i sobre un número n de sucesos. El vector $\boldsymbol{x} = (x_1, ..., x_k)$ sigue una distribución multinomial con parámetros n y p, donde $\boldsymbol{p} = (p_1, ..., p_k)$.

La forma de la distribución de probabilidades multinomial será:

$$f(x_1, ..., x_k; n, p_1, ..., p_k) = \begin{cases} \dfrac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k} & \text{si } \sum_{i=1}^{k} x_i = n \\ \\ 0 & \text{en otros casos} \end{cases}, \tag{A.3}$$

donde $x_1, ..., x_k$ son enteros no negativos.

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:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{A.4}$$

El símbolo $\cap$ quiere decir intersecćion. 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:

[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

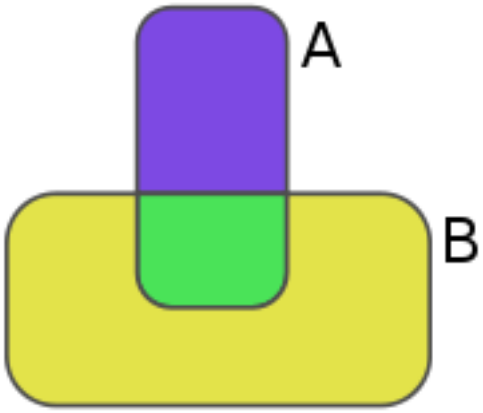

Figura A.1: Probabilidad condicional $P(A \mid B)$. Se puede pensar como en el espacio en el que B es verdadero (área amarilla) también se cumple que A es verdadero (área morada). Entonces $P(A \mid B)$ se representa en esta figura como el área verde.

$P(A) =$ fuma (0.5)

$P(B) =$ hipertensa (0.1)

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

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

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.

A.3. Distribución de probabilidad conjunta

La distribución de probabilidad conjunta (joint probability distribution, en inglés) se define dadas dos variables aleatorias x, y que son definidas en un espacio de probabilidades, la distribución que da la probabilidad de que cada x, y caiga en un rango particular o conjunto discreto de valores específicos para esas variables. Si se trata de dos variables se llama función bivariada, si se trata de más de dos variables se llama función multivariada.

Matemáticamente hablando, si las variables aleatorias x, y son discretas, la función de probabilidad conjunta viene dada por:

$$P(X = x \quad \text{y} \quad Y = y) = P(Y = y|X = x)P(X = x) = P(X = x|Y = y)P(Y = y), \tag{A.6}$$

donde:

$$\sum_i \sum_j P(X = x_i \quad \text{y} \quad Y = y_j) = 1. \tag{A.7}$$

Si las variables x, y son continuas, la *función de densidad conjunta* se escribe como:

$$f_{X,Y}(x, y) = f_{Y|X}(y, x) f_X(x) = f_{X|Y}(x, y) f_Y(y), \tag{A.8}$$

donde $f_{Y|X}(y, x)$ y $f_{X|Y}(x, y)$ son las distribuciones de probabilidad condicional y $f_X(x)$ y $f_Y(y)$ son las distribuciones marginales de X y Y respectivamente.

Ya que hablamos de distribuciones de probabilidad:

$$\int_x \int_y f_{X,Y}(x, y) dy dx = 1. \tag{A.9}$$

Veamos la figura A.2[3], aquí se ilustra graficamente la distribución conjunta de probabilidad de las variables x, y, la cual está representada en el óvalo verde, junto con las distribuciones marginales de X (gausiana azul) y Y (gausiana roja).

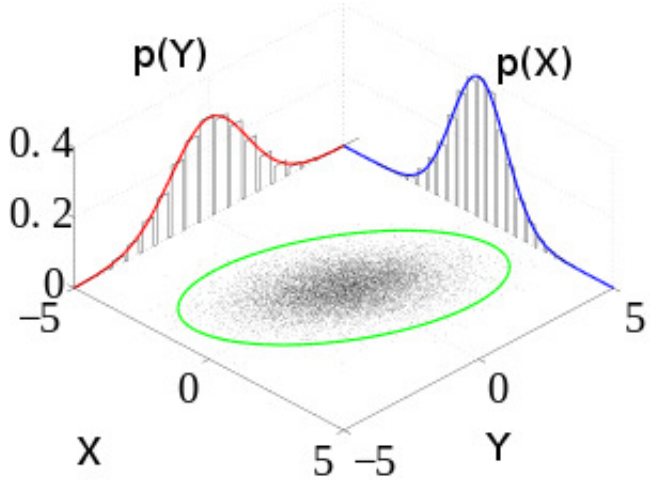

Figura A.2: Probabilidad conjunta.

A.4. 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 Dir($\boldsymbol{\alpha}$) y se define como diremos a continuación.

[3]La figura y el texto es tomado de http://en.wikipedia.org/wiki/Joint_probability_distribution

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:

$$f(x_1, ..., x_{K-1}; \alpha_1, ..., \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K}, x_i^{\alpha_i - 1} \tag{A.10}$$

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

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \tag{A.11}$$

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 probabilidad (A.10) establece que la probabilidad de ocurrencia de K eventos es x_i dado que cada evento se observó $\alpha_i - 1$ veces.

A.5. Ley de probabilidad total

El teorema de la probabilidad total permite calcular la probabilidad de un suceso a partir de probabilidades condicionadas. Dicho en otras palabras, dado un suceso A, con probabilidades condicionales conocidads dado cualquier evento B_n, $P(A|B_n)$, cada uno con probabilidades propias conocidas, $P(B_n)$ ¿Cuál es la probabilidad total de que A ocurra?

Esto se obtiene resolviendo $P(A)$, donde:

$$P(A) = \sum_{n} P(A|B_n)P(B_n). \tag{A.12}$$

La sumatoria puede ser interpretada como el promedio pesado y $P(A)$ es llamada, aveces, probabilidad promedio.

En la figura A.3 se representa la Ley de probabilidad total en un diagrama de árbol. Si se quiere saber la probabilidad total de obtener un suceso A, se debe recorrer todas las ramas que llevan a A y sumarlas:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) + ... + P(A|B_n)P(B_n). \tag{A.13}$$

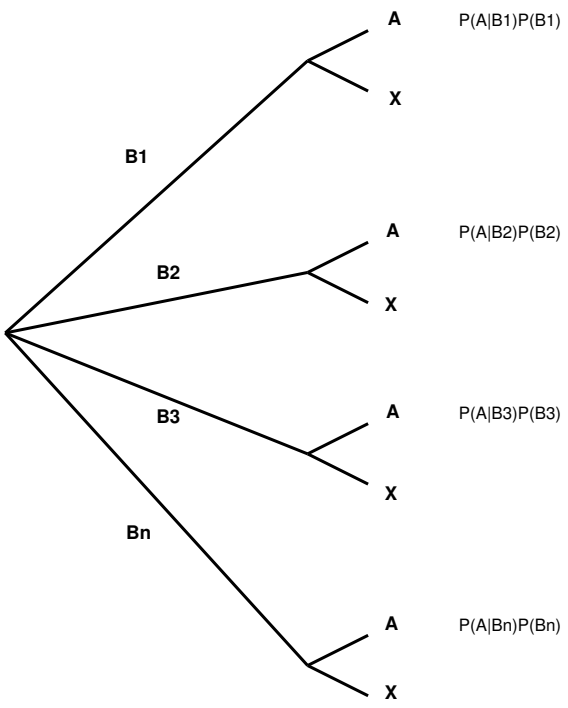

Figura A.3: Ley de probabilidad total representado en un diagrama de árbol.

A.6. Teorema de representación de De Finetti

Una secuencia de variables aleatorias $(x_1, x_2, ..., x_n)$ es infinitamente intercambiable si y solo si, para todo n se cumple que:

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta), \tag{A.14}$$

para alguna medida P en el parámetro θ.

Si la distribución de θ es una densidad (variable continua), entonces, $P(\theta) = p(\theta)d\theta$.

El producto $\prod_{i=1}^{n} p(x_i|\theta)$ es invariante. Esto quiere decir que no importa en qué orden estén los términos.

Entonces, cualquier distribución de secuencias que pueda ser escrita como $\int \prod_{i=1}^{n} p(x_i|\theta)P(d\theta)$ debe ser infinitamente intercambiable para todo n.

Para ver un resumen y la aplicabilidad del teorema de De Finetti en algunos casos dentro del modelado de tópicos ver Jordan (2010) [5].

Una suposición en muchos análisis estadísticos es que las variables aleatorias a estudiar son *independientes e idénticamente distribuidas* (iid).Una colección aleatoria de variables son iid si cada variable aleatoria tiene la misma distribución de probabilidad que la otra y todas son mutuamente

independientes.

La suposición de que las variables sean iid tiende a simplicar la matemática de fondo de muchos métodos estadísticos.

La noción general que comparte las principales propiedades de las variables iid son las variables aleatorias intercambiables definidas por el teorema de represenación de De Finetti. La intercambiabilidad significa que cualquier valor de una secuencia es tan probable como cualquier permutación de esos valores. Un ejemplo es la distribución de probabilidad conjunta, que es invariante ante un grupo simétrico.

Es importante acotar que todas las variables iid son intercambiables, pero no viceversa.

Entonces, si se tienen datos intercambiables:

- Debe existir un parámetro θ.
- Debe existir una probabilidad $p(x|\theta)$ (también llamada likelihood function).
- Debe existir una distribución P de θ.

Estas cantidades deben existir para que los datos $(x_1, x_2, ..., x_n)$ sean condicionalmente independientes.

La demostración del teorema de De Finetti es larga y rigurosa, pero si se está interesado en darle un vistazo se puede visitar página web que está en el pie de página [4].

[4] http://www.dpye.iimas.unam.mx/eduardo/MJB/node7.html

Apéndice B

Notas para mí (Fabiola): Bases de la estadística Bayesiana

B.1. introducción a la estadística bayesiana

La teoría de probabilidad busca cuantificar la ocurrencia de hechos aleatorios. Visto de otra manera, es el estudio del comportamiento de un sistema, a sabiendas de que el sistema está compuesto por partículas o componentes que se comportan de manera aleatoria.

Si se estudia la variación de un suceso a lo largo de muchas repeticiones de ese experimento, la estadística que se emplea es la *estadística frecuentista*. El punto de vista frecuentista estudia la probabilidad de un suceso a partir de muchas repeticiones.

Ahora, si se quiere estudiar la probabilidad de una variable no observada, la estadística frecuentista no es apropiada y se recurre a la *inferencia*. El método de asignar una distribución de probabilidad a variables no observadas es llamado *probabilidad Bayesiana*, donde la distribución de las variables no observadas dados los datos es llamada *likelihood function* y la distribución de las variables no observadas dados los datos y la *distribución a priori*, es la *distribución a posteriori*.

Este último método es llamado también probabilidad inversa, ya que estudia los parámetros (o condiciones) más probables que dieron como resultado ese suceso, en vez de estudiar muchas realizaciones repetidas del suceso para predecir la ocurrencia del mismo.

En ambos enfoques, frecuentista y bayesiano, se utilizan modelos con parámetros desconocidos y la recolección de datos es la base para estimación de dichos parámetros desconocidos.

Sin embargo, en el enfoque frecuentista los parámetros se conciben como valores fijos, pero desconocidos, mientras que en el enfoque bayesiano, los parámetros son variables aleatorias cuya distribución de probabilidad es estudiada a través del teorema de Bayes.

En la estadística bayesiana se ha de tener una distribución subjetiva de los parámetros antes de ver los datos (priori) que se modificará en función de los datos que se hayan observado y así tener una distribución a posteriori, que resume todo el conocimiento que se tiene sobre los parámetros

dados los datos y sus creencias a priori.

The main goal of a typical Bayesian statistical analysis is to obtain the pos- terior distribution of model parameters. The posterior distribution can best be understood as a weighted average between knowledge about the parameters before data is observed (which is represented by the prior distribution) and the information about the parameters contained in the observed data (which is represented by the likelihood function.

Once the posterior distribution has been obtained, one can compute point and interval estimates of parameters, prediction inference for future data, and probabilistic evaluation of hypotheses

B.2. Teorema de Bayes

Definamos x como un suceso cualquiera y el parámetro θ como la confidencia (belief) de que x ocurra.

La confidencia o el parámetro θ de que x ocurra debe ser actualizada con cada nueva observación. Cada vez que se reestima θ se debe reestimar qué tanto se confía en ese valor, esto es, se calcula una nueva distribución de probabilidad sobre los valores posbiles de θ. Esta nueva distribución será $P(\theta|x)$ y se calcula según el teorema de Bayes:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}. \tag{B.1}$$

La expresión generalizada del teorema de Bayes es:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}, \tag{B.2}$$

que al aplicarle la ley de probabilidad total al denominador resulta en la ecuación B.1.

El objetivo es encontrar el mejor conjunto de parámetros que se puedan usar para hacer predicciones. Para esto se escoge el parámetro θ que maximiza $P(\theta|x)$. Generalmente se usa el algoritmo E-M (expectation - maximization) que garantiza la convergencia local. Este método se usa para distribuciones sencillas, para distribuciones complicadas, este método no es factible, por esto es importante mantener $P(x|\theta)$ y $P(\theta)$ lo más simple posible. Además, modelos complicados pueden ocupar mucho tiempo y espacio computacional.

La pregunta está en cómo escoger $P(\theta)$ (priori) y la mejor forma de escogerla es que sea conjugada de $P(x|\theta)$. $P(\theta)$ es conjugada de $P(x|\theta)$ si multiplicando estas dos distribuciones y normalizando, se obtiene una distribución de la misma familia.

B.3. Axiomas de la teoría de probabilidad

(Clases de Thomas Loredo - XXVI IAC-WS)

Al evaluar las hipótesis calculando sus probabilidades condicionales $P(H_i|...)$ sobre información conocida o presumida (incluyendo los datos obervados), se usan las siguientes reglas de la teoría de probabilidad:

Regla de la suma (OR)

$$P(H_1 \vee H_2|I) = P(H_1|I) + P(H_2|I) - P(H_1, H_2|I). \tag{B.3}$$

Regla del producto (AND)

$$P(H_1, D_{obs}|I) = P(H_1|I)P(D_{obs}|H_1, I) = P(D_{obs}|I)P(H_1|D_{obs}, I). \tag{B.4}$$

(NOT)

$$P(\overline{H_1}|I) = 1 - P(H_1|I). \tag{B.5}$$

B.4. Modelo general de mixtura

Se tienen n variables aleatorias iid $(X_1, X_2, ..., X_n)$ con observaciones $(x_1, x_2, ..., x_n)$, las cuales siguen el modelo de mixtura con K componentes.

Cada uno de los k-ésimo componentes es una distribución que sigue una familia de distribuciones con parámetros θ_k y tiene la forma $F(x|\theta_k)$, donde π_k es el peso del k-ésimo componente y denota la probabilidad de que una observación sea generada a partir del componente y cumple con la condición $\pi \geq 0 \;\; y \;\; \sum_k \pi_k = 1$. Entonces la probabilidad de la observación x_i es escrita como:

$$p(x_i) = \sum_{k=1}^{K} \pi_k f(x_i|\theta_k), \tag{B.6}$$

donde $f(x_i|\theta_k)$ es la función de masa (caso discreto) o de densidad (caso continuo) para $F(x|\theta_k)$.

La función de probabilidad conjunta para todas las observaciones es:

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(x_i|\theta_k). \tag{B.7}$$

(Nótese el parecido de esta ecuación con la ecuación (1.2).)

Si $Z_i \in \{1, 2, 3, ..., K\}$ es la etiqueta oculta para X_i, la función de probabilidad puede ser vista como la sumatoria sobre toda la distribución conjunta de todos los X_i y Z_i

$$p(x_i) = \sum_{Z_i} p(x_i, z_i) = \sum_{Z_i} p(x_i|Z_i = z_i)p(z_i), \tag{B.8}$$

donde $X_i|Z_i \sim F(x_i|\theta_{z_i})$ y $Z_i \sim M_k(1; \pi_1, ..., \pi_k)$ es la distribución multinomial de k dimensiones y 1 observación.

Z_i se refiere a la variable auxiliar que identifica la etiqueta con la observación x_i.

Esto es tomado del libro de minado de datos de texto [3].

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] Charu, A y ChengXiang, Z. (Editores)*Mining Text Data* Editorial Springer 2012. ISBN 978-1-4614-3222-7.

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

[5] Jordan, M., *Lecture1: History and De Finetti's Theorem*. Bayesian modeling and inference. 2010

[6] Manning C. D. and Schütze H. *Foundations of Statistical Natural Language Processing*. The MIT Press. 1999.

[7] Daud A., Li J. Zhou L., Muhammad F. *Knowledge discovery through directed probabilistic topic models: a survey* Higher Education Press and Springer-Verlag. 2009

[8] Dujin A. A. *Teoriya Informatzii*. Gelios ARV. Moskva, 2007.