Guía Teórica del Proyecto Modelado de Tópicos

CENDITEL

22 de julio de 2016

Índice general

Capítulo 1

Introducción al 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 métodos 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) [5]:

- 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 finita de palabras. Se define como $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$.
- Tópico: es la distribución multinomial de probabilidad de palabras de un vocabulario y se denotará como z_k, donde $z_k \in \mathcal{Z} = \{z_1, z_2, \ldots, z_K\}$, donde el subíndice k indica el k-ésimo tópico en la colección 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 finita de documentos y está denotado como $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$.

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

1.2. Aprendizaje de los modelos

En el minado de datos de texto, el aprendizaje (o machine learning) es el proceso de automatizar y clasificar los conceptos o expresiones que tienen significados parecidos y agruparlos.

Para realizar este proceso varios algoritmos son utilizados. Estos algoritmos se clasifican en supervisados, semi supervisados y no supervisados, según sea el método utilizado para la clasificación de los documentos.

1.2.1. Aprendizaje no supervisado

La característica general de los algoritmos de aprendizaje no supervisado es que no requieren ninguna información previa sobre los documentos y que pueden ser aplicados a cualquier documento nuevo. Los principales métodos no supervisados son el *clustering* y el *modelado de tópicos*.

El clustering divide un conjunto de objetos en grupos o cúmulos (clusters) que presentan características similares. Para el minado de texto se utiliza el "bag of words" que contiene múltiples ítems idénticos.

El objetivo del clustering es ubicar los objetos similares en el mismo grupo y así, asignar objetos diferentes a diferentes grupos. Para buscar similitud entre las palabras se toman en cuenta las palabras vecinas inmediatas. De esta manera, cada palabra forma su propio cúmulo. En cada paso del algoritmo, dos cúmulos que sean similares se fusionan en un cúmulo nuevo.

Para descubrir si las palabras son similares se toman en cuenta los vecinos inmediatos izquierdo y derecho. Esto implementa la idea de que se puede caracterizar una palabra por la ocurrencia de las palabras vecinas. Entonces, la similitud de las palabras se mide como el grado de solapamiento en las distribuciones de los vecinos. Dos palabras son similares si los vecinos son similares.

En el modelado de tópicos se usa un modelo probabilístico para determinar la probabilidad de membresía de los documentos en grupos determinados. El modelado de tópicos se considera como un proceso de clustering con un modelo generativo probabilístico.

Cada documento puede ser expresado como combinación probabilística de diferentes tópicos, así, los tópicos se pueden considerar como una especie de cúmulo y la membresía del documento en ese tópico tiene naturaleza probabilística.

1.2.2. Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado requieren de un conjunto de datos entrenados, los cuales consisten en un conjunto de entrada y otro de respuesta que son usados para entrenar una base de datos y para ajustar un modelo que puede predecir los valores de la variable dependiente dada la estimación previa (arrojada por los datos entrenados).

Estos algoritmos funcionan con un conjunto entrenado de objetos, cada uno etiquetado con una o más clases que se codifican con un modelo de representación de datos (el documento es representado como un vector de conteo de palabras). Con esto se define una clase modelo y un procedimiento de entrenamiento.

La clase modelo pertenece a una familia de clasificadores y el procedimiento de entrenamiento selecciona un clasificador de esta familia.

El aprendizaje supervisado puede verse análogo a una función de ajuste, en donde se busca el mejor conjunto de parámetros que ajustan una función a los datos.

En la mitad de estos dos tipos de algoritmo (supervisados y no supervisados) se encuentran los algoritmos de aprendizaje **semi supervisado**.

1.2.3. Aprendizaje semi supervisado

Adicionalmente a los datos no etiquetados, estos algoritmos necesitan un conjunto de datos entrenados, pero no para todos los casos. Comúnmente, la información tomada de los datos en-

trenados se usan como objetivo asociados a algunos casos. Aquí, los datos son divididos en dos partes: una parte son los datos que no han sido etiquetados y los datos con etiquetas conocidas (entrenados).

Los algoritmos de aprendizaje semi supervisado pueden verse bien como un algoritmo de aprendizaje no supervisado guiado con restricciones o como un algoritmo de aprendizaje supervisado con información adicional en la distribución de los datos.

1.3. 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, es decir, un *corpus*.

Para simplificar, supongamos que la longitud N 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. Además, supongamos que el corpus tiene un número finito M de documentos. Para cada documento del corpus escogeremos seis palabras al azar y este procedimiento lo realizaremos M 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 M veces el procedimiento anterior, entonces estaríamos generando todo el corpus. Este modelo es llamado ***modelo de unigrama*** [3], donde la probabilidad de cada palabra dentro de cada documento se podría describir usando una distribución multinomial (ver apéndice **??**):

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

donde N=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.

Los modelos probabilísticos generativos se pueden ilustrar (bajo ciertas condiciones) usando gráficas en forma de "placas" [1]. En esta notación gráfica, se sombrean o no los nodos que corresponden a las variables para indicar si estas observadas u ocultas respectivamente.

Las flechas indican dependencias condicionales mientras que las placas de variables (las cajas en la figura) se refieren a la repetición de los pasos del algoritmo sobre la variable que se muestra en la esquina inferior derecha de la caja. Podemos representar gráficamente el modelo de unigrama de la siguiente forma:

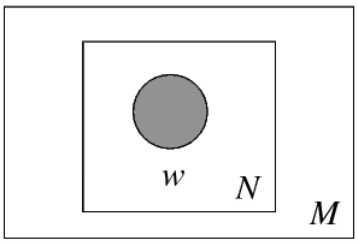

Figura 1.1: Modelo de Unigrama

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, distintos documentos no son necesariamente generados de la misma caja. Este modelo tiene la siguiente representación gráfica:

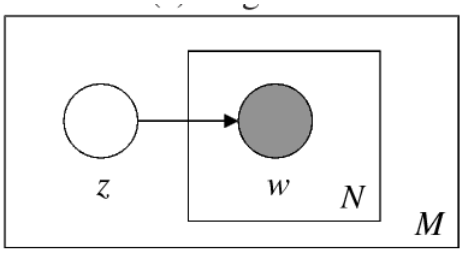

Figura 1.2: Modelo de Mixtura de Unigramas

Este modelo nos permite introducir la noción de tópico. Note que, en este modelo, la variable w depende de la variable z escogida en un nivel previo. En el ejemplo, el tópico es

[1] ver Buntine, 1994, para una introducción

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.

Así, en este modelo, denominado ***mixtura de unigramas*** (ver [3] sección 4.2), cada documento es generado a partir de un tópico, donde su probabilidad sería:

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

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 evento, las cuales también son conocidas junto con sus probabilidades individuales, también conocidas (ver apéndice **??**), la probabilidad total de que ocurra el suceso es:

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

pero

$$p(d|z) = p(w_1, w_2, w_3, \ldots, w_N|z) = \prod_{i=1}^{N} p(w_i|z), \tag{1.4}$$

siempre suponiendo que la condición de igualdad en la ecuación se debe a que los w_i son intercambiables gracias al Teorema De Finnetti, esto quiere decir que no importa el orden en que los w_i se encuentren.

Ahora, si se sustituye esta última ecuación en 1.3 obtenemos

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

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 N palabras independientemente, a partir de la distribución multinomial condicional $p(w|z)$ (ver apéndice **??** 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 [5] 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 [5]:

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

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). Simplificando la ecuación anterior (y aplicando el Teorema de Bayes) se obtiene la distribución de probabilidad conjunta[2]:

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

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). Esto se ve claramente el el modelo gráfico del PLSA

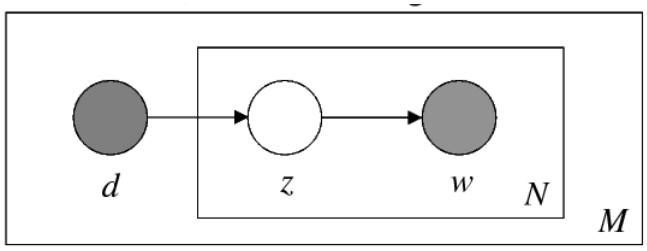

Figura 1.3: PLSA

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[3] hayan y el

[2]Ver apéndice **??**.
[3]Con información previa.

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[4]. 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 parámetros individuales que son explícitamente enlazados con documentos entrenados.

1.4. Interpretación Geométrica del Modelado Probabilístico de Tópicos

Para dar otro punto de vista al modelado de tópicos, podemos considerar los elementos geométricos del espacio sobre el que trabajamos las variables ocultas, mediante los cuales podemos representar los documentos del corpus. Primero definiremos algunos conceptos geométricos que nos ayudarán en este proceso.

Empecemos por el símplex, en geometía, un *Símplex* o n-símplex (o símplice) es el análogo en n dimensiones de un triángulo. Más formalmente, un símplex es la envoltura convexa de un conjunto de $(n+1)$ puntos independientes afines en un espacio euclídeo de dimensión n o mayor, es decir, el conjunto de puntos tal que ningún m-plano contiene más que $(m+1)$ de ellos. Pero para dado el objeto didáctico de esta guía es suficiente pensar en un símplex como la generalización de un triángulo.

Por ejemplo, un 0-símplex es un punto; un 1-símplex un segmento de una línea; un 2-símplex un triángulo; un 3-símplex es un tetraedro; y un 4-símplex es un pentácron (en cada caso, con su interior) como vemos en la figura 1.4.

Ahora bien, una propiedad particularmente útil en un símplex $\mathcal{S}$ es que todos sus puntos cumplen las siguientes características:

[4]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.

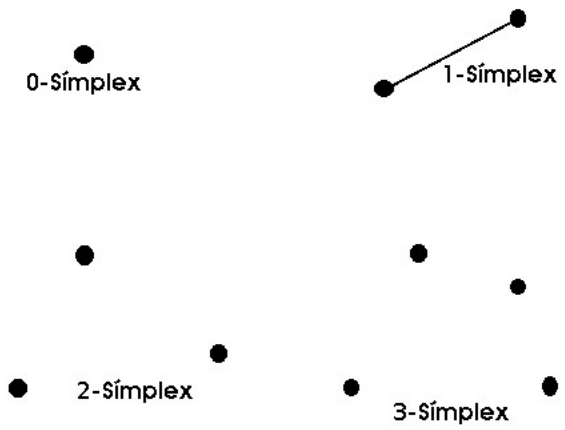

Figura 1.4: Interpretación Geométrica del Modelado Probabilístico de Tópicos

$$x \geq 0 \ , \forall x \in \mathcal{S},$$
$$\sum_{\mathcal{S}} x = 1$$

Con lo cual, en probabilidad, puede ser interpretado como el conjunto de todas funciones de probabilidad "admitibles" para una variable multinomial. Cada valor $x_i \in \mathcal{S}$ es entonces, la probabilidad de que la variable observada pertenezca a la clase i. Así, la distribución de Dirichlet es definida en un símplex de manera natural, siendo esta la conjugada a priori de la distribución multinomial.

Por otra parte, con un vocabulario que contenga V palabras distintas, se puede construir un espacio de dimensión V donde cada eje represente la probabilidad de observar un tipo de palabra particular. Luego, el símplex de dimensión $V-1$, al que llamaremos símplex de las palabras, representa todas las distribuciones de probabilidad sobre las palabras, es decir, todos los tópicos. Supongamos, por un momento, que se tiene un vocabulario de 3 palabras, podemos representar, usando un símplex de dos dimensiones (sobre un espacio de 3 dimensiones) todos los tópicos como vemos en la figura 1.5 en la que el símplex viene dado por la región sombreada.

Como una distribución de probabilidad sobre palabras, cada documento en el corpus puede ser representado como un punto en el símplex. Del mismo modo, cada tópico también puede ser representado como un punto en el símplex. Por último, cada documento que se genera por el modelo es una combinación convexa de los T tópicos que no sólo establecen todas las distribuciones de palabras generadas por el modelo como puntos en el símplex, sino también como puntos del símplex $(T-1)$-dimensional generado por los tópicos. Por ejemplo, en la figura 1.5, los dos tópicos generan un 1-símplex (representado por la línea punteada) y cada documento generado se encuentra dentro de él.

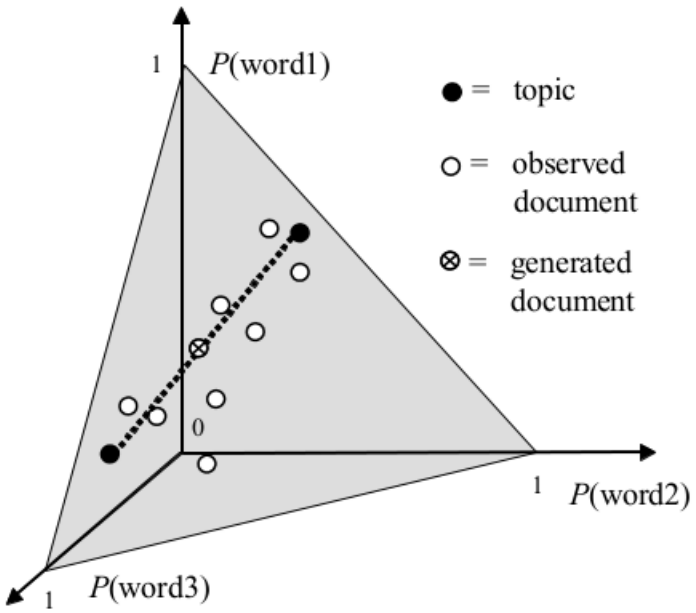

Figura 1.5: Interpretación Geométrica del Modelado Probabilístico de Tópicos

Cada modelo visto antes (unigrama, mixtura de unigramas y PLSA) opera en un espacio de distribuciones de palabras, y cada una de estas distribuciones puede verse como un punto en el símplex de las palabras. De esta manera, podemos describir dichos modelos en términos geométricos como sigue:

Unigrama El modelo de unigrama fija un punto en el símplex de las palabras y establece que todas las palabras en el corpus provienen de la distribución a la que el punto corresponde. Los modelos de variables ocultas consideran k puntos en el símplex de las palabras y forman un "sub-símplex" generado por ellos al que llamaremos el símplex de los tópicos. Note que cualquier punto en el símplex de los tópicos es también un punto en el tópico de las palabras. Diferentes modelos de variable oculta usan de diferenta manera el símplex de los tópicos para generar un documento.

Mixtura de Unigramas El modelo de mixtura de unigramas establece que para cada documento, uno de los puntos del símplex de los tópicos es elegido aleatoriamente y todas las palabras del documento son extraídas de la distribución a la que corresponde dicho punto.

PLSA El modelo PLSA establece que cada palabra de cada documento entrenado proviene de un tópico elegido aleatoriamente. Los tópicos son, además, extraídos de una distribución específica de tópicos dentro de los documentos. Es decir, existe una distribución de tópicos para cada documento, el conjunto de los documentos entrenados define entonces una distribución empírica en el símplex de documentos.

La figura 1.6 muestra un símplex de tres tópicos dentro de un símplex de tres palabras. Las esquinas del símplex de palabras corresponden a las tres distribuciones donde cada palabra (respectivamente) tiene probabilidad 1. Los tres puntos dentro del símplex de tópicos corresponden a

tres distribuciones diferentes sobre palabras.

El modelo de mixtura de unigramas sitúa cada documento en una esquina del símplex de los tópicos, mientras que el modelo PLSA induce una distribución empírica (denotada por las x's) en el símplex de tópicos.

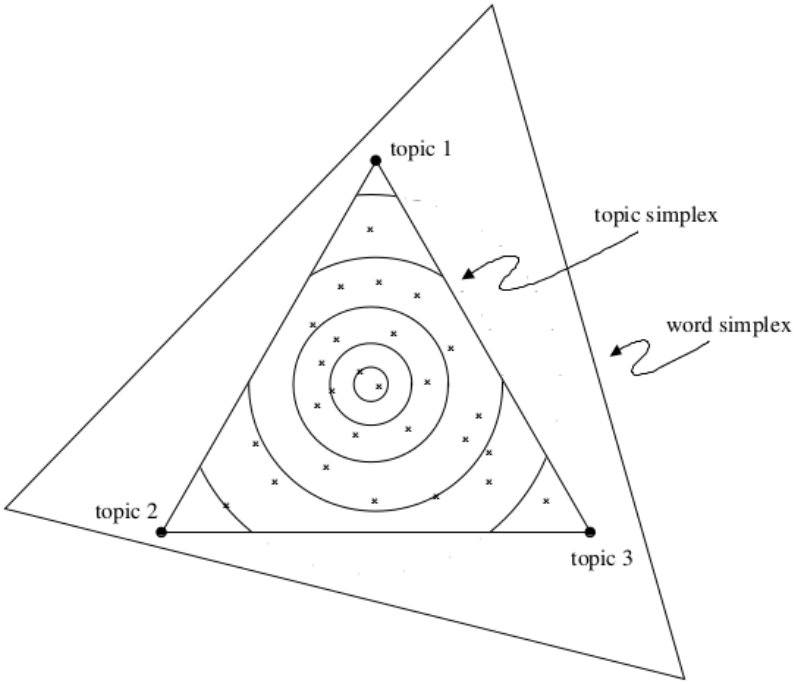

Figura 1.6: Representación de los distintos modelos dentro de un símplex

Capítulo 2

Latent Dirichlet Allocation

2.1. 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 [2] 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 [2], 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, el modelo asume, que el corpus se genera de la siguiente forma:

- 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.

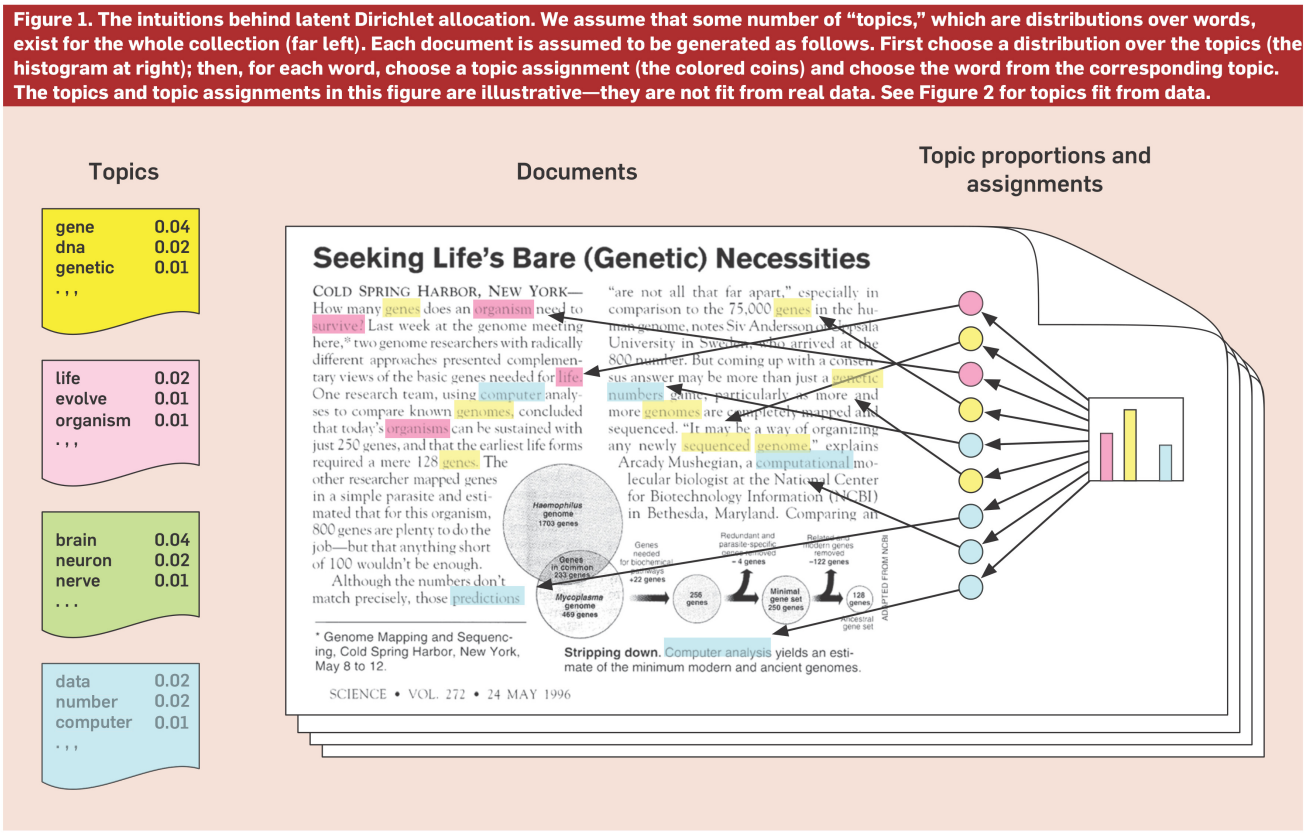

Figura 2.1: Figura 1 de Blei 2012 [2]

Con el objeto de garantizar 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, se tiene un único documento, donde se resaltan algunas palabras con diferentes colores, el color con que la palabra es resaltada representa el tópico asignado a esa palabra (círculos de colores). Contando cuántos círculos de cada color hay se va construyendo el histograma de la derecha de la figura donde, de nuevo, cada color representa un tópico diferente y la altura de cada rectángulo representa cuántas palabras fueron asisgnadas a ese color. Este histograma no es más que una representación de la distribución de Dirichlet (ver apéndice **??**).

La frecuencia con que las palabras se repiten (si se considera el mismo color) forma lo que es el tópico que está representado en la parte izquierda de la figura 2.1 y que no es más que una distribución sobre las palabras.

Entonces, el LDA asume que el documento se genera primero escogiendo una distribución sobre los tópicos (histograma de la derecha - distribución de Dirichlet), luego, cada palabra del documento se asigna a un tópico (un color) y luego se escoge una palabra de ese tópico de acuerdo a la distribución de palabras correspondiente, es decir, cada palabre tiene una probabilidad específica

de ocurrir. En este proceso, se supone que el número de tópicos es conocido.

El ejemplo de la figura 2.1 es reducido, ya que en la realidad el LDA se extiende a cientos o miles de documentos sobre los cuales se emplean los algoritmos.

Es importante resaltar que, como dice Blei 2012 [2], los algoritmos no tienen información del 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. LDA como modelo probabilístico

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: $p(w|\beta) \sim M(\beta)$, 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.

Como se puede ver en la figura 2.2 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 $\theta \sim Dirichlet(\alpha)$.

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 θ : $p(z) \sim M(\theta)$. 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: $p(w|z_i, \beta)$. Con esto, se genera el siguiente modelo gráfico:

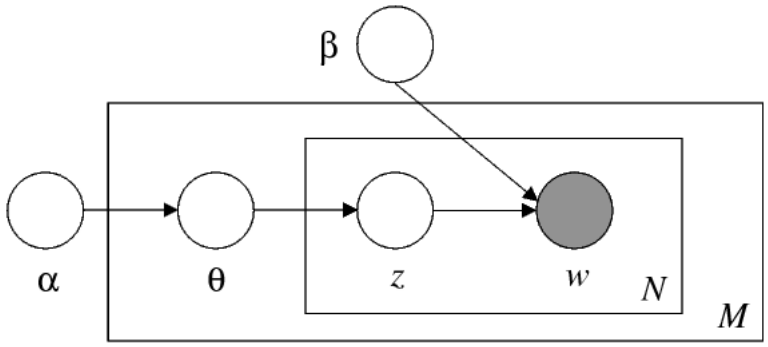

Figura 2.2: Modelo Gráfico del LDA

Como la figura deja claro, hay tres niveles para la representación LDA. Los parámetros α y β son parámetros del nivel del corpus. las variables θ_d son variables de nivel de los documentos, muestreadas una vez por documento. Finalmente, las variables z_{dn} y w_{dn} son variables de nivel de las palabras y se muestrean una vez por cada palabra dentro de cada documento. De esta manera, el LDA implica tres niveles de aleatoriedad, y en particular, se muestrea el nodo del tópico repetidamente dentro del documento. Por lo tanto, bajo este modelo, los documentos pueden estar asociados con múltiples tópicos.

Por otra parte, volviendo a la figura 1.6 ahora podemos apreciar que el LDA establece una distribución uniforme en el símplex de los tópicos indicada por las curvas de nivel.

2.3. Conclusión: Cómo Funciona el LDA

En los modelos probabilísticos generativos, los datos son tratados como el resultado de un proceso generativo que incluye variables ocultas. Este proceso generativo define una distribución de probabilidad conjunta tanto sobre las variables aleatorias observadas como sobre las variables ocultas. Luego, se lleva a cabo el análisis de datos mediante el uso de tal distribución conjunta para calcular la distribución condicional de las variables ocultas dadas las variables observadas.

Esta distribución condicional se conoce como la distribución posterior. Calcular esta es el mayor reto del LDA.

Las variables observadas son las palabras de la documentos, las variables ocultas corresponden a la estructura de tópicos que da lugar al corpus. El problema de cálculo de inferir la estructura oculta de los tópicos dentro de los documentos, es equivalente al problema de calcular la distribución posterior.

Se puede pensar en los procesos que se llevan a cabo dentro del LDA en dos grandes partes, la primera, el proceso generativo, descrito detalladamente en el capítulo 3, establece la forma en la que el LDA modela la generación de un corpus, mientras que el proceso de inferencia, descrito en el capítulo 4, finalmente en un caso particular del proceso de inferencia, se lleva a cabo un sub-algoritmo al que llamamos Estimación de parámetros que se atiende en el capítulo 4.

Capítulo 3

Proceso Generativo

3.1. Introducción

En principio, es importante aclarar que el proceso generativo del LDA no genera documentos reales. Supongamos que tenemos un corpus. Como se ha explicado antes, los documentos han sido generados por un proceso complejo subyacente, que no nos es conocido. El objetivo del LDA en este paso 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. (de acuerdo a una distribución multinomial(θ))
     - 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 3.1.1 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.

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

El proceso genérico de generar un corpus 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 3.2.1 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,

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

(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*. Por último note que cada tópico representa una distribución multinomial sobre las palabras del documento[2].

Observación 3.2.2 Por otra parte, note que puede pensar en cada tópico como una bolsa donde las palabras se repiten según su probabilidad, por ejemplo, la bolsa que corresponda al *tópico 1* deberá contener 2 veces la palabra *eléctrico,* tres veces la palabra *CENDITEL,* tres veces la palabra *tecnología* y una vez la palabra *servicio.* De esta manera, es claro que no importa el orden de las palabras al construir los documentos.

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 (en el sentido de que se llenarán una a una y no porque exista alguna relación de cada palabra con su posición dentro del documento).

Documento 1 (d_1):

1. Se elige una distribución de probabilidades de los tópicos para d_1[3].

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

   Observación 3.2.3 Podemos usar la idea intuitiva de ver la distribución de tópicos en el documento como una bolsa grande, que generará al documento, en la que se guardan los tópicos según su probabilidad, por ejemplo, la bolsa que generará al documento 1, contiene dos veces al *tópico 1,* y una vez a cada uno de los otros dos tópicos. Note que, ya que los tópicos también se representaban mediante bolsas, podemos ver estas bolsas generadoras como 'bolsas' que contienen 'bolsas' que contienen palabras.

2. Para cada casilla vacía, se elige un tópico[4].

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

[2]En el sentido en el que en cada topico existe la posibilidad de que ocurra más de un 'evento' (palabra).
[3]La suma de las probabilidades de todos los tópicos dentro de un documento debe ser igual a 1.
[4]La distribución establecida en el paso 1. debe verse reflejada en la asignación de los tópicos a las casillas.

Esto es, de la bolsa generadora, se saca aleatoriamente una bolsa de tópico.

3. Para cada casilla vacía, se elige una palabra (De acuerdo al tópico asignado a la casilla en el paso 2.)[5].

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

Esto es, de la bolsa de tópico obtenida en el paso anterior, se saca, aleatoriamente una palabra.

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

Observación 3.2.4 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;

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

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.

[5]Es importante observar aquí que entre más largo sea el documento mejor se verán representadas en él, tanto la distribución de tópicos en el documento como la distribución de palabras en el tópico.

Documento 2			
dist. tópicos	$p(t_1\|d_2) = 0$	$p(t_2\|d_2) = \frac{1}{4}$	$p(t_3\|d_2) = \frac{3}{4}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 3	tópico 3	tópico 2	tópico 3
música	CENDITEL	arte	servicio

Documento 3			
dist. tópicos	$p(t_1\|d_3) = \frac{5}{9}$	$p(t_2\|d_3) = \frac{1}{9}$	$p(t_3\|d_3) = \frac{1}{3}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 3	tópico 3	tópico 1	tópico 1
tecnología	eléctrico	eléctrico	servicio

Documento 4			
dist. tópicos	$p(t_1\|d_4) = 0$	$p(t_2\|d_4) = \frac{8}{9}$	$p(t_3\|d_4) = \frac{1}{9}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 2	tópico 2	tópico 2	tópico 2
arte	música	arte	arte

Documento 5			
dist. tópicos	$p(t_1\|d_5) = \frac{1}{3}$	$p(t_2\|d_5) = \frac{1}{3}$	$p(t_3\|d_5) = \frac{1}{3}$
casilla 1	casilla 2	casilla 3	casilla 4
tópico 1	tópico 2	tópico 1	tópico 3
eléctrico	arte	servicio	tecnología

Documento 6			
dist. tópicos	$p(t_1\|d_6) = \frac{1}{2}$	$p(t_2\|d_6) = \frac{1}{4}$	$p(t_3\|d_6) = \frac{1}{4}$
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;

$$
\begin{aligned}
d_1 &= \text{« CENDITEL CENDITEL servicio tecnología »} \\
d_2 &= \text{« musica CENDITEL arte tecnología »} \\
d_3 &= \text{« tecnología eléctrico eléctrico servicio »} \\
d_4 &= \text{« arte música arte arte »} \\
d_5 &= \text{« eléctrico arte servicio tecnología »} \\
d_6 &= \text{« CENDITEL servicio tecnología CENDITEL »}
\end{aligned}
$$

Capítulo 4

Métodos de Inferencia

Tras su publicación en 2003 [3], la Asignación Latente de Dirichlet (LDA, por sus siglas en inglés), se ha usado para el modelado de tópicos, uno de los paradigmas más populares y de mayor éxito del aprendizaje supervisado y no supervisado.

El problema clave en el modelado de tópicos es la inferencia a posteriori. Esto se refiere a invertir el proceso generativo definido y el aprendizaje de las distribuciones posteriores de la variables ocultas en el modelo dados los datos observados. En el LDA, esto equivale a la solución de la siguiente ecuación:

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{4.1}$$

Donde cada $\phi^{(k)}$ para $k \in \{1, \ldots, K\}$ es una distribución discreta de probabilidades sobre un vocabulario fijo, que representa el k-ésimo tópico y K es el número de tópicos ocultos en el corpus. Cada θ_d es una distribución específica de los tópicos en el documento d. Cada z_i representa el tópico que generó la palabra w_i, y α y β son hiperparámetros de la distribución simétrica de Dirichlet de donde se extraen las distribuciones discretas.

Desafortunadamente, esta distribución tiene un costo de calculo muy alto, particularmente, el factor de normalización $p(w|\alpha, \beta)$, no se puede calcular con exactitud. Sin embargo, hay una serie de técnicas de inferencia aproximadas disponibles que podemos aplicar al problema como la inferencia variacional (tal como se utiliza en el artículo original) o el Muestreo de Gibbs.

4.1. Método de Muestreo: Muestreo de Gibbs

El algoritmo de muestreo de Gibbs es una solución a una pregunta básica e importante: ¿Cómo obtener los valores de muestra de una distribución de probabilidad?

Por ejemplo, suponga que su distribución tiene una única variable X que toma dos valores, a saber, $p(X = 0) = 0,5$ y $p(X = 1) = 0,5$. ¿Cómo se obtiene una muestra de los valores de X? Sencillo, se lanza una moneda. Si obtiene cara, entonces, X = 1, de lo contrario, X = 0. Note que la anterior representa una distribución binomial.

Ahora bien, digamos que se quiere modelar una distribución multinomial: $p(X = i) = \frac{1}{6}$, para $i \in \{1, \ldots, 6\}$, para esto basta con lanzar un dado.

Más aún, ¿Qué pasa si se tiene una distribución multinomial de más de una variable: $p(X_1, X_2, \ldots, X_n)$. Si las variables son independientes, se puede factorizar la distribución multivariable como un producto de distribuciones univariantes y muestrear de cada una de ellas por separado, obteniendo

$$p(X_1, X_2, \ldots, X_n) = P(X_1) \cdot P(X_2) \cdot \ldots \cdot P(X_n)$$

Sin embargo, para el caso más general, en el que se tiene una distribución de probabilidad en la que es difícil calcular directamente de la distribución conjunta, es decir, $p(X_1, X_2, \ldots, X_n)$, es evidente la inutilidad de nuestro enfoque anterior. La distribución puede incluso no tener una forma "buena" como una distribución binomial o multinomial, y puede no tener ninguna factorización en distribuciones "buenas".

El muestreo de Gibbs proporciona un método eficiente para aproximar esta distribución conjunta bajo una condición: es necesario ser capaz de calcular fácilmente la distribución condicional de cada X_i dados valores fijos para los restantes, es decir,

$$P(X_i | X_1, \ldots, X_{(i-1)}, X_{(i+1)}, \ldots, X_n)$$

Para complementar estas ideas, una explicación visual podría ayudar. Digamos que nuestro objetivo es tomar muestras de una distribución $p(X, Y)$

Lo cual, en un contexto Bayesiano, puede ser una distribución posterior con alguna dificultad analítica. Sin embargo, supongamos que las distribuciones condicionales univariadas ($p(X|Y)$ y $p(Y|X)$) son sencillas de manejar analíticamente y digamos que la distribución condicional $p(X|Y)$ se "parece" a una mezcla univariada como esta:

Entonces podemos realizar muestras de esta distribución univariada para obtener una muestra de X, y luego realizar muestras a partir de $p(Y|X)$ para obtener una muestra de Y e iterar.

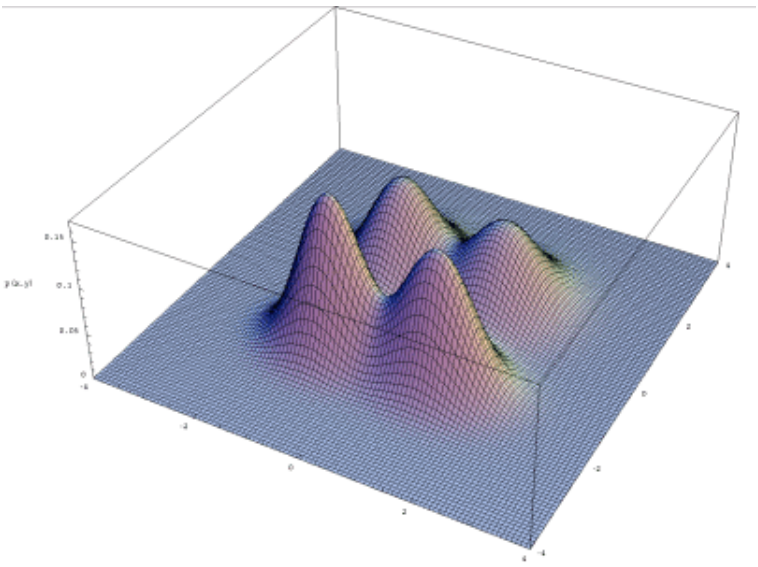

Figura 4.1: Distribución Conjunta Multivariada

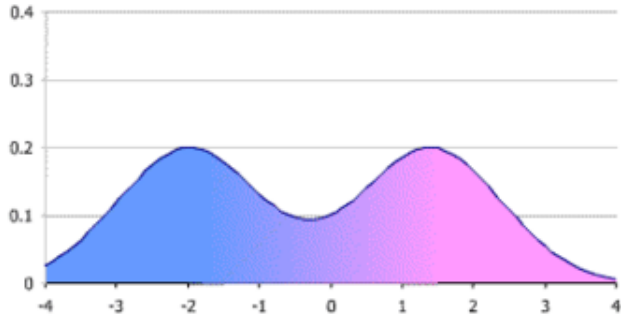

Figura 4.2: Distribución Univariada

Finalmente, por simetría, el mismo argumento se mantiene durante el muestreo a partir de $p(X|Y)$. Por lo tanto, iterando este procedimiento es razonable garantizar que, en el largo plazo nos aproximaremos a la distribución $p(X, Y)$ [1].

4.1.1. Algoritmo de Gibbs en el LDA

Como hemos venido viendo, la idea básica del muestreo de Gibbs es reestimar la probabilidad de una variable, asumiendo que la probabilidad asignada a todas las otras variables sea correcta. Específicamente, en el LDA el muestreo de Gibbs fija una palabra y reasigna un tópico a la misma suponiendo que los tópicos asignados a todas las otras palabras (en todo el corpus) son correctos.

[1] Hablamos de aproximaciones porque otras condiciones (periodicidad e irreductibilidad) deben ser verificadas. Para los detalles matemáticos, véase Tierney, Luke. Markov Chains for Exploring Posterior Distributions. Ann. Statist. 22 (1994), no. 4, 1701–1728.

A continuación presentaremos un ejemplo de una aplicación del Muestreo de Gibbs que ilustre el proceso, para posteriormente explicar los fundamentos matemáticos que respaldan el algoritmo (para un lector interesado en introducirse en el tema es suficiente leer con detenimiento el ejemplo).

Ejemplo

Aplicaremos un proceso sintetizado del muestreo de Gibbs al ejemplo que hemos estado trabajando antes, como entrada se necesitará un número de tópicos (usaremos 3) y un corpus, usaremos el obtenido en el ejemplo del proceso generativo:

$$\begin{aligned} d_1 &= \text{« CENDITEL CENDITEL servicio tecnología »} \\ d_2 &= \text{« musica CENDITEL arte tecnología »} \\ d_3 &= \text{« tecnología eléctrico eléctrico servicio »} \\ d_4 &= \text{« arte música arte arte »} \\ d_5 &= \text{« eléctrico arte servicio tecnología »} \\ d_6 &= \text{« CENDITEL servicio tecnología CENDITEL »} \end{aligned}$$

Luego, el vocabulario con el que se trabajará será:

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

Notación diremos pal_i para referirnos a la i-ésima palabra del vocabulario.

Observación 4.1.1 Ya que la longitud del corpus y del vocabulario son bastante pequeñas el resultado será una aproximación burda de las distribuciónes de probabilidades. Como antes, con este ejemplo sólo se busca ilustrar el proceso sin prestar mucha atención a las probabilidades.

En primer lugar, el muestreo de Gibbs asigna de manera aleatoria (pseudoaleatoria) un tópico a cada palabra de cada documento. Como se ha dicho, en este ejemplo se considarán tre tópicos, digamos t_a, t_b y t_c. Supongamos que en nuestro corpus se realiza la siguiente asignación:

Observación 4.1.2 Note que esta asignación establece una primera distribución de los tópicos dentro de los documentos y de las palabras dentro de los tópicos. La idea del algoritmo es refinar las distribuciones en cada iteración.

Según el algoritmo, se fija el primer documento d_1, que hasta ahora tiene la siguiente distribución:

y se fija la primera palabra, «CENDITEL» para reasignar su tópico. Se calculan las probabilidades:

d_1	CENDITEL	CENDITEL	servicio	tecnología
	tópico a	tópico c	tópico a	tópico c
d_2	música	CENDITEL	arte	tecnología
	tópico b	tópico a	tópico a	tópico b
d_3	tecnología	eléctrico	eléctrico	servicio
	tópico a	tópico c	tópico c	tópico c
d_4	arte	música	arte	arte
	tópico b	tópico a	tópico b	tópico c
d_5	eléctrico	arte	servicio	tecnología
	tópico c	tópico b	tópico a	tópico a
d_6	CENDITEL	servicio	tecnología	CENDITEL
	tópico a	tópico a	tópico c	tópico b

d_1	CENDITEL	CENDITEL	servicio	tecnología
	tópico a	tópico c	tópico a	tópico c

$$\begin{aligned} p(t_i|d_1) &= \{p(t_a|d_1), p(t_b|d_1), p(t_c|d_1)\} = \left\{\frac{1}{2}, 0, \frac{1}{2}\right\} \\ p(\text{CENDITEL}|t_i) &= \{p(pal_4|t_a), p(pal_4|t_b), p(pal_4|t_c)\} = \left\{\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right\} \end{aligned}$$

De donde se sigue que:

$$p(t_i|d_1) \cdot p(\text{CENDITEL}|t_i) = \left\{\frac{3}{10}, 0, \frac{1}{10}\right\}$$

Luego, el *tópico a* se mantiene en este caso.

Se fija ahora la segunda palabra, «CENDITEL» y se repite el proceso:

$$\begin{aligned} p(t_i|d_2) &= \left\{\frac{1}{2}, 0, \frac{1}{2}\right\} \\ p(\text{CENDITEL}|t_i) &= \left\{\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right\} \\ p(t_i|d_1) \cdot p(\text{CENDITEL}|t_i) &= \left\{\frac{3}{10}, 0, \frac{1}{10}\right\} \end{aligned}$$

Luego, para la segunda palabra, se asignará el *tópico a*.

d_1	CENDITEL	CENDITEL	servicio	tecnología
	tópico a	tópico a	tópico a	tópico a
d_2	música	CENDITEL	arte	tecnología
	tópico b	tópico a	tópico b	tópico a
d_3	tecnología	eléctrico	eléctrico	servicio
	tópico a	tópico c	tópico c	tópico c
d_4	arte	música	arte	arte
	tópico b	tópico b	tópico b	tópico b
d_5	eléctrico	arte	servicio	tecnología
	tópico a	tópico a	tópico c	tópico b
d_6	CENDITEL	servicio	tecnología	CENDITEL
	tópico a	tópico a	tópico a	tópico a

Completando todas las palabras de todos los documentos, obtenemos el primer refinamiento:

Que genera la siguiente distribución de probabilidad de las palabras dentro de los tópicos:

$$
\begin{aligned}
t_a &= \left\{ x_{11} = \frac{1}{13}, x_{12} = 0, x_{13} = \frac{1}{13}, x_{14} = \frac{5}{13}, x_{15} = \frac{4}{13}, x_{16} = \frac{2}{13} \right\} \\
t_b &= \left\{ x_{21} = \frac{4}{7}, x_{22} = \frac{2}{7}, x_{23} = 0, x_{24} = 0, x_{25} = \frac{1}{7}, x_{26} = 0 \right\} \\
t_c &= \left\{ x_{31} = 0, x_{32} = 0, x_{33} = \frac{1}{2}, x_{34} = 0, x_{35} = 0, x_{36} = \frac{1}{2} \right\}
\end{aligned}
$$

Donde $x_{ij} = p(pal_j)$ en t_i.

Y la siguiente distribución de tópicos dentro de los documentos:

documento	porción de t_a	porción de t_b	porción de t_c
d_1	1	0	0
d_2	$\frac{1}{2}$	$\frac{1}{2}$	0
d_3	$\frac{1}{4}$	0	$\frac{3}{4}$
d_4	0	1	0
d_5	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
d_6	1	0	0

En una segunda corrida se obtuvo

d_1	CENDITEL	CENDITEL	servicio	tecnología
	tópico a	tópico a	tópico a	tópico a
d_2	música	CENDITEL	arte	tecnología
	tópico b	tópico a	tópico b	tópico a
d_3	tecnología	eléctrico	eléctrico	servicio
	tópico a	tópico c	tópico c	tópico c
d_4	arte	música	arte	arte
	tópico b	tópico b	tópico b	tópico b
d_5	eléctrico	arte	servicio	tecnología
	tópico c	tópico b	tópico c	tópico b
d_6	CENDITEL	servicio	tecnología	CENDITEL
	tópico a	tópico a	tópico a	tópico a

De donde se sigue que:

$$
\begin{aligned}
t_a &= \left\{ x_{11} = 0, x_{12} = 0, x_{13} = 0, x_{14} = \frac{5}{11}, x_{15} = \frac{4}{11}, x_{16} = \frac{2}{11} \right\} \\
t_b &= \left\{ x_{21} = \frac{5}{8}, x_{22} = \frac{1}{4}, x_{23} = 0, x_{24} = 0, x_{25} = \frac{1}{8}, x_{26} = 0 \right\} \\
t_c &= \left\{ x_{31} = 0, x_{32} = 0, x_{33} = \frac{3}{5}, x_{34} = 0, x_{35} = 0, x_{36} = \frac{2}{5} \right\}
\end{aligned}
$$

Donde $x_{ij} = p(pal_j)$ en t_i.

Y se obtiene la siguiente distribución de tópicos dentro de los documentos:

documento	porción de t_a	porción de t_b	porción de t_c
d_1	1	0	0
d_2	$\frac{1}{2}$	$\frac{1}{2}$	0
d_3	$\frac{1}{4}$	0	$\frac{3}{4}$
d_4	0	1	0
d_5	0	$\frac{1}{2}$	$\frac{1}{2}$
d_6	1	0	0

Note que esta vez se obtuvieron bastantes menos reasignaciones, en otras palabras el sistema se acerca a una estructura estable, en las próximas iteraciones no se realizó ninguna reasignación, por lo tanto el proceso concluye aquí.

Para efectos de comparación se presentará una tabla con la distribución de tópicos por documento con la que se construyó el corpus:

documento	porción de t_1	porción de t_2	porción de t_3
d_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
d_2	0	$\frac{1}{4}$	$\frac{3}{4}$
d_3	$\frac{5}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
d_4	0	$\frac{8}{9}$	$\frac{1}{9}$
d_5	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
d_6	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Observación 4.1.3 Como se predijo, la aproximación obtenida en este proceso es bastante pobre, en este caso es debido al tamaño del corpus y de la longitud de los documentos.

Fundamentos Teóricos

El Muestreo de Gibbs es un miembro de una familia de algoritmos conocida como Algoritmos de Tipo Cadena de Markov Monte Carlo (MCMC, por sus siglas en inglés). Los algoritmos MCMC tienen como objetivo construir una cadena de Markov que tiene la distribución posterior de destino como su distribución estacionaria. En otras palabras, después de una serie de iteraciones paso a paso a través de la cadena, el muestreo de la distribución debe converger para estar cerca de muestreo de la posterior deseada.

Para LDA, estamos interesados en las porciones de documentos relacionados con los tópicos ocultos θ_d, las distribuciones tópico-palabra $\phi^{(z)}$ y las asignaciones de tópico para cada palabra z_i. Ahora bien, mientras que las distribuciones condicionales (y, por tanto, un algoritmo de muestreo de Gibbs para el LDA) se pueden derivar para cada una de estas variables ocultas, note que tanto θ_d, como $\phi^{(z)}$, pueden calcularse utilizando sólo las asignaciones de tópicos z_i (es decir, $\mathbf{z}$ es una estadística suficiente para estas dos distribuciones[2]).

Sin embargo, un algoritmo más sencillo se puede utilizar si integramos los parámetros multinomiales y simplemente muestreamos z_i. Este es llamado un Muestreo de Gibbs colapsado. El muestreo de Gibbs colapsado para LDA necesita para calcular la probabilidad de que un tópico sea asignado a una palabra en específico w_i, teniendo en cuenta todas las otras asignaciones tema a todas las demás palabras. Un poco más formalmente, estamos interesados en el cálculo de la siguiente posterior:

$$p(z_i|\mathbf{z}_{-i}, \alpha, \beta, \mathbf{w}) \tag{4.2}$$

Donde $\mathbf{z}_{-i}$ se refiere a todas las asignaciones de tópicos *excepto* z_i.

[2]
$$\theta_{d,z} = \frac{n(d,z) + \alpha}{\sum_{|Z|} n(d,z) + \alpha} \text{ , } \phi_{z,w} = \frac{n(z,w) + \beta}{\sum_{|W|} n(z,w) + \beta}$$

Para empezar, las reglas de la probabilidad condicional indican que:

$$p(z_i|\mathbf{z}_{-i},\alpha,\beta,\mathbf{w}) = \frac{p(z_i,\mathbf{z}_{-i},\mathbf{w}|\alpha,\beta)}{p(\mathbf{z}_{-i},\mathbf{w}|\alpha,\beta)} \propto p(z_i,\mathbf{z}_{-i},\mathbf{w}|\alpha,\beta) = p(\mathbf{z},\mathbf{w}|\alpha,\beta) \tag{4.3}$$

Luego tenemos que:

$$p(\mathbf{w},\mathbf{z}|\alpha,\beta) = \int\int p(\mathbf{z},\mathbf{w},\theta,\phi|\alpha,\beta)d\theta d\phi \tag{4.4}$$

Siguiendo el modelo definido en el proceso generativo [3], se obtiene:

$$p(\mathbf{w},\mathbf{z}|\alpha,\beta) = \int\int p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z)d\theta d\phi \tag{4.5}$$

Luego, agrupando los términos que tienen variables dependientes se sigue que:

$$p(\mathbf{w},\mathbf{z}|\alpha,\beta) = \int p(\theta|\alpha)p(\mathbf{z}|\theta)d\theta \int p(\phi|\beta)p(\mathbf{w}|\phi_z)d\phi \tag{4.6}$$

Note que ambos términos son multinomiales Dirichlet a priori. Como la distribución de Dirichlet es conjugada de la distribución multinomial, el trabajo se simplifica al multiplicar ambas distribuciones de Dirichlet con un parámetro ajustado. empecemos con el primer término:

$$\begin{aligned}
\int p(\theta|\alpha)p(\mathbf{z}|\theta)d\theta &= \int \prod_i \theta_{d,z_i} \frac{1}{B(\alpha)} \prod_k \theta_{d,k}^{\alpha_k} d\theta_d \\
&= \frac{1}{B(\alpha)} \int \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k} d\theta_d \\
&= \frac{B(n_{d,\cdot}+\alpha)}{B(\alpha)}
\end{aligned} \tag{4.7}$$

donde $n_{d,k}$ es el número de veces que una palabra del documento d fue asignada al tópico k, el $\cdot$ indica una sumatoria sobre ese índice, y $B(\alpha)$ es la función distribución beta, dada por $B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$. Análogamente, para el segundo término se sigue que:

[3] $$p(\mathbf{z},\mathbf{w},\theta,\phi|\alpha,\beta) = p(\phi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\phi_z)$$

$$\begin{aligned}\int p(\phi|\beta)p(\mathbf{w}|\phi_z)d\phi &= \int \prod_d \prod_i \phi_{z_{d,i},w_{d,i}} \prod_k \frac{1}{B(\beta)} \prod_w \phi_{k,w}^{\beta_w} d\phi_k \\ &= \prod_k \frac{1}{B(\beta)} \int \prod_w \phi_{k,w}^{n_{k,w}+\beta_w} d\phi_k \\ &= \prod_k \frac{B(n_{k,\cdot}+\beta)}{B(\beta)} \end{aligned} \tag{4.8}$$

Combinando las ecuaciones 4.7 y 4.8, la distribución conjunta extendida es entonces:

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \prod_d \frac{B(n_{d,\cdot}+\alpha)}{B(\alpha)} \prod_k \frac{B(n_{k,\cdot}+\beta)}{B(\beta)} \tag{4.9}$$

la ecuación para el muestreo de Gibbs para el LDA, puede ser ahora derivada, usando la regla de la cadena. (Note que el superíndice $(-i)$ significa que el i-ésimo término, fue dejado fuera del cálculo:

$$\begin{aligned} p(z_i|\mathbf{z}^{(-i)}, \mathbf{w}) &= \frac{p(\mathbf{w}, \mathbf{z})}{p(\mathbf{w}, \mathbf{z}^{(-i)})} = \frac{p(\mathbf{z})}{p(\mathbf{z}^{(-i)})} \cdot \frac{p(\mathbf{w}|\mathbf{z})}{p(\mathbf{w}^{(-i)}|\mathbf{z}^{(-i)})p(w_i)} \\ &\propto \prod_d \frac{B(n_{d,\cdot}+\alpha)}{B(n_{d,\cdot}^{(-i)}+\alpha)} \prod_k \frac{B(n_{k,\cdot}+\beta)}{B(n_{k,\cdot}^{(-i)}+\beta)} \\ &\propto (n_{d,k}^{(-i)}+\alpha_k) \frac{n_{k,w}^{(-i)}+\beta_w}{\sum'_w n_{k,w'}^{(-i)}+\beta_{w'}} \end{aligned} \tag{4.10}$$

4.2. Método Variacional: Inferencia Variacional

De manera sencilla, el proceso de inferencia variacional funciona de la misma forma que una ponencia. Alguien de la audiencia pregunta al presentador una respuesta muy difícil, que este no puede responder. El presentador convenientemente replantea la cuestión de una manera más fácil y da una respuesta exacta a esta pregunta reformulada en lugar de responder a la pregunta original.

Como vimos en la sección anterior, Los métodos basados en Cadenas de Markov Monte Carlo (como el muestreo de Gibbs) son una buena opción para obtener la distribución a posteriori exacta, sin embargo, la convergencia puede ser prohibitivamente lenta si se tienen muchos parámetros. Aquí es donde la inferencia variacional tiene su motivación. La inferencia variacional pretende aproximar la distribución posterior, $P(Z|X)$, mediante una distribución Q, que se puede calcular con más rapidez.

A continuación, se presentará un ejemplo ilustrativo. Para los lectores interesados en una descripción más técnica, la sección 4.2.3 presenta los fundamentos de la inferencia variacional aplicada al LDA, mientras que la sección 4.3 compara esta con el muestreo de Gibbs.

4.2.1. Ejemplo

De nuevo, durante el ejemplo aplicaremos un proceso sintetizado del muestreo de la Inferencia Variacional al ejemplo que hemos estado trabajando antes, como entrada se necesitará un número de tópicos (usaremos 3, denotados por t_A, t_B y t_C) y un corpus, usaremos el obtenido en el ejemplo del proceso generativo:

$$\begin{aligned} d_1 &= \text{« CENDITEL CENDITEL servicio tecnología »} \\ d_2 &= \text{« musica CENDITEL arte tecnología »} \\ d_3 &= \text{« tecnología eléctrico eléctrico servicio »} \\ d_4 &= \text{« arte música arte arte »} \\ d_5 &= \text{« eléctrico arte servicio tecnología »} \\ d_6 &= \text{« CENDITEL servicio tecnología CENDITEL »} \end{aligned}$$

El vocabulario con el que se trabajará:

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

Notación diremos pal_i para referirnos a la i-ésima palabra del vocabulario.

Necesitaremos también una distribución a priori de las palabras dentro de los tópicos, esta vez la presentaremos en forma de matriz como sigue:

$$\beta = \begin{bmatrix} \text{arte} & \text{música} & \text{eléctrico} & \text{CENDITEL} & \text{tecnología} & \text{servicio} \\ 0{,}1 & 0{,}5 & 0{,}1 & 0{,}1 & 0{,}1 & 0{,}1 \\ 0{,}1 & 0{,}1 & 0{,}1 & 0{,}5 & 0{,}1 & 0{,}1 \\ 0{,}15 & 0{,}15 & 0{,}4 & 0{,}1 & 0{,}1 & 0{,}1 \end{bmatrix}$$

Donde la entrada en la columna i, fila j de la matriz β representa la probabilidad de la palabra pal_i en el tópico j.

Y una distribución a priori de los tópicos dentro de los documentos, que esta vez presentaremos en forma de vector:

$$\gamma = (p(t_A|d), p(t_B|d), p(t_C|d))$$

Que se irá actualizando en cada corrida. En particular, para simplificar los cálculos de este ejemplo, se tomará $\gamma = (2, 2, 2)$ para todos los documentos. Finalmente necesitamos un parámetro de concentración α.

En términos generales, el parámetro de concentración es un parámetro numérico que, mientras más altos son sus entradas, más uniformemente distribuida es la distribución resultante. En cambio, mientras más pequeñas, la distribución es más esparcida, con la mayoría de los valores de probabilidad cercanos a cero, Esta idea se aclarará durante el ejemplo.

De nuevo con el objeto de simplificar las cuentas del ejemplo, tomaremos $\alpha = (0{,}1, 0{,}1, 0{,}1)$.

Fijados ya los datos, el objetivo será actualizar los valores del vector γ para cada documento, empecemos con el documento $d_1 =$ « CENDITEL CENDITEL servicio tecnología »:

Se debe calcular la probabilidad de cada palabra dentro del documento de pertenecer a cada tópico, usando la ecuación:

$$\phi_{\text{palabra } i,\text{tópico } j} \propto \beta_{ij} \times exp\left(\Psi(\gamma_j) - \Psi(\sum_k \gamma_k)\right)$$

Donde Ψ representa la función Digamma, que es la derivada logarítmica de la función Gamma. Gracias a que asumimos que el vector γ sería uniforme igual a 2 para todos los documentos, simplificaremos la ecuación $exp\left(\Psi(\gamma_j) - \Psi(\sum_k \gamma_k)\right)$ por

$$\begin{aligned} exp\left(\Psi(2) - \Psi(2+2+2)\right) &= exp\left(\Psi(2) - \Psi(6)\right) \\ &= 0{,}275 \end{aligned}$$

por ejemplo, las probabilidades de la palabra «CENDITEL» se calculan como sigue:

$$\begin{aligned} \phi_{\text{CENDITEL},A} &\propto 0{,}1 \times 0{,}275 = 0{,}027 \\ \phi_{\text{CENDITEL},B} &\propto 0{,}5 \times 0{,}275 = 0{,}137 \\ \phi_{\text{CENDITEL},C} &\propto 0{,}1 \times 0{,}275 = 0{,}027 \end{aligned}$$

de manera que hemos obtenido el vector $(0{,}027, 0{,}137, 0{,}027)$ asociado a la palabra CENDITEL que se normaliza como $(0{,}142, 0{,}716, 0{,}142)$.

Este proceso se repite para las otras palabras del documento obteniendo los siguientes vectores normalizados asociados:

$$\begin{aligned} \text{Vector normalizado asociado a la palabra servicio} &= (0{,}333, 0{,}333, 0{,}333) \\ \text{Vector normalizado asociado a la palabra tecnología} &= (0{,}333, 0{,}333, 0{,}333) \end{aligned}$$

Ahora, estamos en capacidad de actualizar los valores de γ_{d_1}, esto lo haremos mediante la siguiente tabla:

$\frac{\text{vector asociado}}{\text{palabra}}$	1era entrada	2da entrada	3ra entrada
CENDITEL	0.142	0.716	0.142
CENDITEL	0.142	0.716	0.142
servicio	0.333	0.333	0.333
tecnología	0.333	0.333	0.333
α	0.1	0.1	0.1
suma	1.05	2.198	1.05
normalización	0.245	0.510	0.245
expresión $\frac{x}{y}$ aprox.	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

es decir, los nuevos valores del vector que representa la proporción de tópicos dentro del documento d_1 son aproximadamente $\gamma_{d_1} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Análogamente obtuvimos los vectores γ correspondientes al resto de los documentos, que presentamos en la siguiente tabla:

documento	porción de t_A	porción de t_B	porción de t_C
d_1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
d_2	$\frac{7}{20}$	$\frac{3}{8}$	$\frac{7}{25}$
d_3	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
d_4	$\frac{2}{5}$	$\frac{6}{25}$	$\frac{7}{20}$
d_5	$\frac{7}{25}$	$\frac{7}{25}$	$\frac{2}{5}$
d_6	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Para efectos de comparación se presentará una tabla con la distribución de tópicos por documento con la que se construyó el corpus:

documento	porción de t_1	porción de t_2	porción de t_3
d_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
d_2	0	$\frac{1}{4}$	$\frac{3}{4}$
d_3	$\frac{5}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
d_4	0	$\frac{8}{9}$	$\frac{1}{9}$
d_5	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
d_6	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Observación 4.2.1 Note que en esta ocación, la aproximación a las distribuciones reales de tópicos dentro de los documentos es mejor, observe por ejemplo, el parecido del tópico 1 (original) con el tópico aproximado B.

4.2.2. Fundamentos de la inferencia Variacional

$$p(Z|X) \approx Q(Z|V) = \prod_i Q(Z_i|V_i)^{4}$$

Q se llama a la distribución varacional aproximada o simplemente distribución varacional. El término variacional se deriva del cálculo de variaciones, que se ocupa de los problemas de optimización que escogen la mejor función (en este caso, una distribución Q).

Usualmente, en la distribución posterior, las variables ocultas no son independientes dados los datos, pero si se limita la familia de distribuciones variacionales a una distribución que se factoriza sobre cada variable en Z[5], se simplifica el problema. Por otra parte, cada V_i se elige de manera que $Q(Z|V)$ sea lo más cercano a $P(Z|X)$ como sea posible cuando se mide mediante la divergencia Kullback Leibler (KL). Por lo tanto, el problema de interés es ahora la obtención de un $\tilde{V}$ tal que

$$\tilde{V} = \arg\min_V KL(Q(Z|V)||p(Z|X))$$

Cuando esto se escribe en términos de la fórmula de divergencia KL, se obtiene una suma de términos que contienen a V, que se pueden reducir al mínimo. Así que ahora el procedimiento de estimación se convierte en un problema de optimización.

Una vez que obtenido $\tilde{V}$, podemos utilizar $Q(Z|\tilde{V})$ como la aproximación de la distribución posterior.

Observe que el método no es trivial, ya que no se trata de simplemente desechar el modelo complejo $p(Z, X)$ por el uso de una más simple $Q(Z, X)$ en su lugar. Observe que nunca se definine algo como $q(Z, X)$, sólo $Q(Z)$ para una entrada dada X. El modelo complejo p todavía se utiliza para definir lo que estamos tratando de aproximar por $Q(Z)$, es decir, $p(Z|X)$, que puede ser diferente para cada entrada X.

4.2.3. Inferencia Variacional en el LDA

Para utilizar el método variacional de la inferencia en LDA, en principio, es necesario como antes definir una distribución Q tal que se aproxime a la distribución a posteriori original.
Un método sencillo para obtener una familia de distribuciones variacionales Q es de considerar modificaciones sencillas de la distribución original P que permitan considerar las variables de manera independiente mientras que cada variable en la distribución variacional, tenga una variable

[4]Para simplificar notación digamos que V es el conjunto de parámetros de la distribución variacional.
[5]Esto se llama una aproximación de campo medio.

correspondiente en la distribución original.

Note que cada palabra observada w tendrá una distribución variacional sobre los tópicos, esto permitirá que diferentes palabras esten relacionadas con diferentes tópicos. or otra parte, la distribución de los tópico dentro de los documentos tiene una distribución variacional generada por una distribución de Dirichlet diferente para cada documento, lo que permite diferentes documentos estén asignados a diferentes tópicos en diferentes proporciones.

Ahora bien, Dados los hiperparámetros α y β y siguiendo el esquema de inferencia variacional de antes, el objetivo de la inferencia variacional en el LDA es la creación de un problema de optimización que determine los valores de los parámetros variacionales γ y ϕ los cuales se encuentran minimizando la divergencia KL entre la distribución variacional $q(\theta, z|\gamma, \phi)$ y la posteriori verdadera $p(\theta, z|w, \alpha, \beta)$, mediante un proceso de optimización análogo al presentado en la sección 1, en el cual se específica la familia de distribuciones de probabilidades. El problema de optimización se reduce a obtener $(\tilde{\gamma}, \tilde{\phi})$ tales que:

$$(\tilde{\gamma}, \tilde{\phi}) = \arg\min D(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)). \tag{4.11}$$

Es decir, $(\tilde{\gamma}, \tilde{\phi})$ son los argumentos bajo los cuales se minimiza la divergencia KL entre la distribución variacional y la posteriori.

De este procedimiento, Blei 2003, obtiene el siguiente par de ecuaciones para los parámetros variacionales:

$$\phi \propto \beta_{iw_n} exp\left\{E_q[\log(\theta_i)|\gamma]\right\} \tag{4.12}$$

$$\gamma_i = \alpha_i + \sum_{n=i}^{N} \phi_{ni} \tag{4.13}$$

Donde E_q es la esperanza calculada en el paso-E del algoritmo EM (Esperanza - Maximización), es decir, la probabilidad máxima de los parámetros estimados, que está definida por:

$$E_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)$$

siendo Ψ la primera derivada de la función $\log\Gamma$ calculada según la aproximación de Taylor.

Las ecuaciones 4.12 y 4.13 de las distribuciones de los parámetros variacionales ϕ y γ son, de hecho, distribuciones condicionales que dependen de w. Esto se puede ver en la ecuación 4.11 ya que los parámetros variacionales optimizados $(\tilde{\gamma}, \tilde{\phi})$ se construyen para un w fijo. De este modo, la

distribución variacional $q(\theta, z|\gamma, \phi)$ puede escribirse como $q(\theta, z|\tilde{\gamma}(w), \tilde{\phi}(w))$ donde se ve explícitamente la dependencia con w y se observa más claramente la relación con la posteriori $p(\theta, z|w, \alpha, \beta)$.

Esto quiere decir que los parámetros variacionales optimizados $(\tilde{\gamma}(w), \tilde{\phi}(w))$ son particulares para un documento. Por ejemplo, el parámetro variacional de Dirichlet $\tilde{\gamma}(w)$ representa un documento en un simplex de tópicos.

4.3. Comparación del Muestreo de Gibbs Y la Inferencia Variacional

"Inferencia Variacional es lo que implementas mientras esperas que tu muestreo de Gibbs converja"
David Blei

La velocidad es de hecho la razón principal para utilizar métodos variacionales. La inferencia variacional puede obtener mejores resultados que el muestreo de Gibbs para la misma cantidad de tiempo, además es más sencillo diagnosticar la convergencia durante la inferencia variacional.

Sin embargo, la inferencia variacional está irremediablemente sesgada, mientras que el margen de error de los métodos basados en MCMC, como el muestreo de Gibbs, se aproxima a 0 cuando se ejecuta la cadena de Markov durante más y más tiempo. (De hecho, puede hacer que el error sea exactamente 0 por medio de la muestra perfecta.) Así que si se tiene un número ilimitado de recursos computacionales, entonces la inferencia variacional perderá en términos de aproximación.

Por otra parte, la muestra de varianza del muestreo de Gibbs por lo general se aproxima a 0 a medida que calcula cada vez más muestras, Mientras que un cálculo variacional ya tiene varianza de la muestra de exactamente 0 porque es determinista. Este determinismo puede ser útil.

En particular, esto significa que se puede utilizar de propagación hacia atrás para determinar el gradiente de la estimación variacional con respecto a los parámetros del modelo (o las observaciones). Esto le permite sintonizar los parámetros del modelo para obtener las estimaciones correctas en los datos de entrenamiento. Sin embargo, eso no es un argumento convincente para la estimación variacional, ya que no es difícil de adaptar la idea de utilizar MCMC.

Capítulo 5

Estimación de Parámetros: Algoritmo de Esperanza Maximización

La estimación de parámetros se refiere al proceso de usar una muestra de datos para determinar los parámetros de la distribución seleccionada. El problema de la estimación de parámetros en la estadística bayesiana es encontrar los parámetros α y β que maximicen la distribución marginal de los datos. Una vez encontrados estos parámetros se utilizan para calcular, en el proceso de inferencia, los parámetros variacionales.

En el caso particular del LDA propuesto por Blei, se plantea la estimación de parámetros de la siguiente manera: dado un corpus $\mathcal{D} = \{w_1, w_2, \ldots, w_M\}$, el problema de la estimación de parámetros se reduce a encontrar los parámetros del modelo α y β de forma tal que la concordancia entre las observaciones y el modelo sea máxima.

El muestreo de Gibbs es una manera de aproximar la distribución posterior a través de un grupo de variables aleatorias. En caso de LDA, la estimación de la distribución posterior es equivalente a la búsqueda de los parámetros de la distribución, en otras palabras, Cuando es usado el muestreo de Gibbs para el proceso de inferencia para el LDA, no es necesario utilizar previamente algún método de estimación de parámetros.

En el siguiente ejemplo podemos fijar las intuiciones del funcionamiento del algoritmo EM, que se refinaran a lo largo del documento.

5.1. Ejemplo Ilustrativo

Suponga que un observador nunca ha visto ninguna fruta antes en su vida. Y que 50 frutas más o menos esféricas son presentadas en su mesa distribuidas uniformemente, junto con un único dato: hay 5 tipos de fruta.

Además, el observador sabe que estas frutas vienen de diferentes tipos de árboles y que los árboles no producen cosas al azar, sino que tienden a producir cosas similares. ¿Cómo puede proceder el observador sobre la organización de la fruta? Bueno, en realidad tiene dos problemas que resolver:

1. ¿Cómo asignar cada uno de los frutos individuales a un tipo de árbol en particular? Llamemos a esto el problema de asignación de valores de Z.
2. ¿Cuáles son las características del fruto del árbol de cada tipo? Llamemos a esto el problema de la estimación de los los parámetros desconocidos o θ.

Pero estos dos problemas están relacionados entre sí: es claro que se puede usar uno para ayudar a resolver el otro. En lo sucesivo se describirá la solución planteada por el Algoritmo de Esperanza-Maximización:

1. Escoger aleatoriamente una asignación de tipos a las frutas. Es decir, hacer una conjetura de los valores de Z, a la que llamaremos $Z(0)$. Inicialmente, la asignación no tiene por qué aproximarse de ninguna forma a la realidad, por ejemplo, un ramillete de uvas y una sandía pueden pertencer al mismo tipo.
2. Ahora, teniendo una asignación de tipos, es posible intentar responder a la segunda pregunta: ¿cuáles son las características de cada tipo de fruta suponiendo que proceden de un mismo árbol?

   Pues bien, las frutas de el tipo 1 tienen este tamaño medio, y este color, y así sucesivamente. Este es el paso expectativa. Es decir, se estima $\theta(0)$ de esta manera.
3. Luego, teniendo $\theta(0)$, se puede encontrar una mejor asignación de frutas para cada tipo, ya que se sabe que los frutos de un mismo árbol son similares. Así que eventualmente: las uvas son más propensas a terminar en un grupo (caracterizado por su pequeño tamaño y ser suave), y las sandías en otro (las que se caracterizan por el tamaño grande y ser duro). Así, se genera $Z(1)$.
4. Se regresa al paso 2. Pero en lugar de $Z(0)$, se usa de $Z(1)$.

En algún momento, el sistema comenzará a estabilizarse, por ejemplo, $Z(11)$ será el mismo que $Z(12)$. Entonces se habrán obtenido buenas aproximaciones de Z y θ.

5.2. Algoritmo de Esperanza-Maximización

El Algoritmo de Esperanza-Maximización (EM) es un procedimiento determinista de Estimación de Máxima Verosimilitud (EMV). Por lo tanto, busca los parámetros "óptimos" de una

distribución hipotética para adaptarse a un conjunto de datos observados.

Dado cualquier conjunto de valores para los parámetros, la probabilidad de que estos produzcan los datos iniciales, se puede ver como el producto de la función de densidad de probabilidad evaluada en cada uno de los puntos de datos, por lo tanto, podemos escribir una fórmula para la densidad en función de los valores de los parámetros. De cálculo, sabemos que una función sólo puede tener un máximo local en un punto en donde la derivada se anule.

Este algoritmo proporciona una forma numérica para calcular los valores de los parámetros. Como esta estimación es computacionalmente costosa, en la práctica, es usual maximizar la probabilidad maximizando el logaritmo de la probabilidad, que es equivalente.

Procedimiento del Algoritmo EM

1. Inicializar los parámetros del modelo.
2. **Paso E** Estimar las probabilidades de la variable oculta mediante los valores de los parámetros actuales.
3. **Paso M** Reestimar los valores de los parámetros dadas las probabilidades actuales.
4. Repetir los pasos $2-3$ hasta obtener un sistema convergente.

Esencialmente, lo que hace la etapa M, es maximizar el valor esperado de $\log p(X, X|\theta)$ donde las expectativas se obtienen con respecto a la distribución posterior de las variables ocultas calculada en la etapa E.

Veamos una interpretación geométrica de los fundamentos del Algoritmo:

Suponga que se desea encontrar máximo local de $\log P(X|\theta)$, sin embargo, no se puede hacer esto directamente por gradiente de ascenso (o se puede, pero sería muy lento). Entonces el algoritmo EM hace lo siguiente:

1. fijar un valor θ^{old} en el dominio de $\ln P(X|\theta)$ y un margen de error ε.
2. hallar una curva $L(q, \theta)$ que se mantenga inferior a la función $\log P(X|\theta)$ en un entorno de θ^{old} y que $L(q, \theta^{\text{old}}) = \log p(x|\theta^{\text{old}})$. (Paso E).

   - Esto es, aproximar (por debajo) la curva original con una curva sencilla de calcular que toque a la curva original en el punto (valor) fijado en el paso anterior.
3. hallar un θ^{new} tal que $L(q, \theta)$ tenga el mayor valor posible. (Paso M).

   - Esto es, buscar el punto (valor) más alto en la curva trazada en el paso anterior, proyectarlo al dominio de la curva original y establecerlo como nuevo valor para la siguiente iteración.

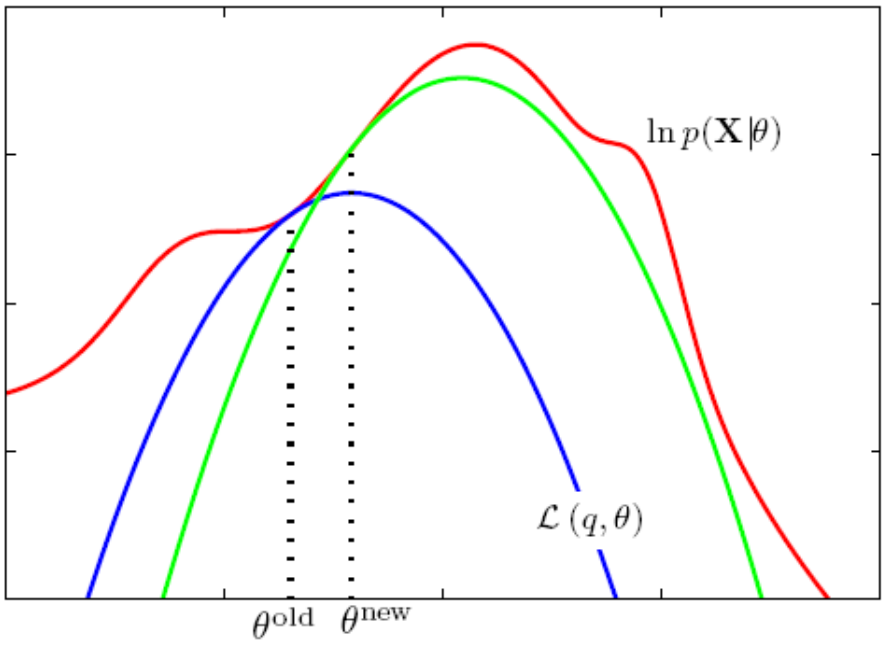

Figura 5.1: Interpretación gráfica del algoritmo EM

4. iterar 1 y 2, hasta que $\ln p(X|\theta) - L(q,\theta) < \varepsilon$ en un entorno de θ^{new}.

5.3. Estimación de Parámetros en el LDA

En particular, dado un corpus de documentos $\mathcal{D} = \{w_1, \ldots, w_M\}$, el objetivo es encontrar parámetros α y β tales que la concordancia entre las observaciones y el modelo sea máxima.

Así, podemos encontrar estimaciones aproximadas de Bayes para el modelo LDA a través de un procedimiento de *variacional* EM, que maximiza un límite inferior con respecto a los parámetros variacionales γ y ϕ y, a continuación, para valores fijos de los parámetros variacionales, maximiza el límite inferior con respecto a los parámetros α y β.

El algoritmo EM variacional para el LDA es el siguiente:

1. Para cada documento, encontrar los valores de la optimización de los parámetros variacionales $\left\{\tilde{\gamma}, \tilde{\phi} : d \in \mathcal{D}\right\}$.(Paso E).
2. Maximizar el resultado de límite inferior del logarítmo de la probabilidad con respecto al modelo de parámetros α y β. Esto corresponde a la búsqueda de estimaciones de máxima verosimilitud con estadísticas suficientes esperadas para cada documento bajo la aproximación de la posteriori que se calcula en el paso anterior. (Paso M).
3. Repetir los pasos 1 y 2 hasta que el sistema converja.

5.4. Alternativas al EM

El algoritmo EM converge típicamente a un valor óptimo local, no necesariamente al máximo global y no hay límite en la tasa de convergencia en general. Es posible que se pueda obtener una aproximación arbitrariamente pobre en dimensiones altas. Por lo tanto, hay una necesidad de técnicas alternativas, especialmente en entornos de alta dimensión.

Existen alternativas al EM con mejores garantías en términos de consistencia, que son conocidos como enfoques basados en el momento o técnicas espectrales. Los enfoques basados en el momento para el aprendizaje de los parámetros de un modelo probabilístico son de creciente interés recientemente, ya que gozan de garantías tales como la convergencia global en determinadas condiciones.

Apéndice A

Breve Introducción a la Estadística

A.1. Conceptos Básicos de Probabilidad

Para entender los procesos llevados a cabo en el LDA son necesarios algunos conocimientos matemáticos básicos, esta sección introduce al lector en ellos, primero, consideremos las siguientes definiciones de la teoría de la probabilidad:

Probabilidad Es el conjunto de posibilidades de que un evento ocurra o no en un momento y tiempo determinado. Dichos eventos pueden ser medibles a través de una escala de 0 a 1, donde el evento que no pueda ocurrir tiene una probabilidad de 0 (evento imposible) y un evento que ocurra con certeza es de 1 (evento cierto).

Experimento Es toda acción sobre la cual vamos a realizar una medición u observación, es decir cualquier proceso que genera un resultado definido. si los resultados de la acción no se determinan con certeza (por ejemplo, lanzar una moneda al aire) llamamos al experimento, *Experimento Aleatorio*.

Espacio Muestral Es el conjunto de todos los resultados posibles que se pueden obtener al realizar un experimento aleatorio. Por ejemplo, si el experimento consiste en lanzar un dado, el espacio muestral correspondiente es $S = \{1, 2, 3, 4, 5, 6\}$.

Punto Muestral o Muestra Es un elemento del espacio muestral de cualquier experimento dado.

Evento o Suceso Es todo subconjunto de un espacio muestral.

La probabilidad de que ocurra un evento, siendo ésta una medida de la posibilidad de que un suceso ocurra favorablemente, se determina principalmente de dos formas: empíricamente (de manera experimental) o teóricamente (de forma matemática).

1. Probabilidad empírica.- Si E es un evento que puede ocurrir cuando se realiza un experimento, entonces la probabilidad empírica del evento E, que a veces se le denomina definición de frecuencia relativa de la probabilidad, está dada por la siguiente fórmula:

$$p(E) = \frac{\text{Numero de veces que se realizo el experimento } E}{\text{Numero de veces que se realizo el experimento}}$$

2. Probabilidad teórica.- Si todos los resultados en un espacio muestral S finito son igualmente probables, y E es un evento en ese espacio muestral, entonces la probabilidad teórica del evento E está dada por la siguiente fórmula, que a veces se le denomina la definición clásica de la probabilidad, expuesta por Pierre Laplace en su famosa Teoría analítica de la probabilidad publicada en 1812:

$$p(E) = \frac{\text{Numero de resultados favorables}}{\text{Numero de resultados posibles}} = \frac{|E|}{|S|}$$

Durante todo el documento, al decir probabilidad de un evento E nos referiremos a este valor, al que denotaremos como $p(E)$.

Es importante aclarar la siguiente notación básica antes de continuar:

Sean A y B dos eventos;

Unión La unión de los dos eventos, denotada por $A \bigcup B$, es el espacio en el que cualquiera de los dos, o ambos podrían ocurrir.

Intersección La intersección de los dos eventos, denotada por $A \bigcap B$, es el espacio en el que ambos ocurren simultaneamente.

Complemento El complemento de un evento A, denotado por $\overline{A}$, es el espacio en el que el evento A no ocurre en absoluto.

note que $p(A) = 1 - p(\overline{A})$

Eventos Mutuamente Excluyentes A y B son mutuamente excluyentes si, y sólo si, no es posible que ambos se den a la vez. En lenguaje matemático esto se describe como $A \subset \overline{B}$.

Eventos Independientes A y B son dos eventos independientes, si el conocimiento de la incidencia de uno de ellos no tiene efecto en la probabilidad de ocurrencia del otro. En lenguaje matemático, se dice que A y B son independientes si, y sólo si, $p(A \bigcup B) = p(A) * p(B)$.

A.1.1. Reglas de la probabilidad

Sean A y B son dos eventos;

Regla de la Adición de Probabilidades Se aplica la siguiente regla para calcular la probabilidad de que el evento $A \bigcup B$ ocurra:

$$p(A \bigcup B) = p(A) + p(B) - p(A \bigcap B)$$

Note que si A y B son eventos mutuamente excluyentes $p(A \bigcap B) = 0$, por lo tanto $p(A \bigcup B) = p(A) + p(B)$.

Regla de la Multiplicación de Probabilidades Se aplica la siguiente regla para calcular la probabilidad de que el evento $A \bigcap B$ ocurra:

$$p(A \bigcap B) = p(A) * p(B|A)$$

A.1.2. Distribuciónes de Probabilidad

Definición A.1.1 *Una **Variable Aleatoria** es una variable cuyo valor es obtenido a partir de un fenómeno numérico aleatorio. Usualmente se denota por X, Y o Z. Y puede ser:*

- *Discreta: Una variable aleatoria que tiene una cantidad finita o infinita numerable de valores posibles.*
- *Continua: Una variable aleatoria que tiene un conjunto infinito de posibles valores.*

Una **Distribución de Probabilidad** de una variable aleatoria X representa todos los valores posibles de X y las probabilidades de que cada valor posible ocurra. Denotaremos la distribución de probabilidad de una variable aleatoria X como $P(X)$.

A.2. 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 sólo dos probabilidades.

Por ejemplo, si se tira una moneda al aire existe 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 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.3. 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:

1. Probabilidades independientes: $p(A)$, $p(B)$ es la probabilidad de que A y B ocurran de forma independiente una de la otra.
2. 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:

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

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).

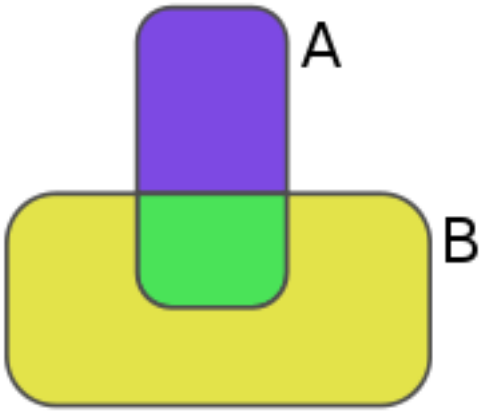

Figura A.1: Probabilidad condicional $p(A \mid B)$. Se puede pensar como en el espacio en el que B es verdadero (área amarilla) también se cumple que A es verdadero (área morada). Entonces $p(A \mid B)$ se representa en esta figura como el área verde.

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

Ilustremos la idea de probabilidad condicional con el siguiente ejemplo:

Si el 20 % de la población es hipertensa y el 10 % de la población fuma y es hipertensa, ¿Cuál es la probabilidad de que dada una persona hipertensa, esta sea fumadora?

Primero, fijemos el nombre de los eventos del problema, diremos que:

A Es el grupo de las personas fumadoras.

B Es el grupo de las personas hipertensas.

Además tenemos los siguientes datos:

$p(B) = 0{,}2$ Probabilidad de encontrar una persona hipertensa.

$p(A \cap B) = 0{,}1$ Probabilidad de encontrar una persona que fume y sea hipertensa.

La probabilidad de que una persona dada que se sabe que es hipertensa, fume es la siguiente:

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

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

Esto quiere decir que la probabilidad de que se escoga, entre las personas hipertensas, una persona al azar y esta sea fumadora es del 50 % (esto representará la zona verde en la figura A.1.)

A.4. 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:

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

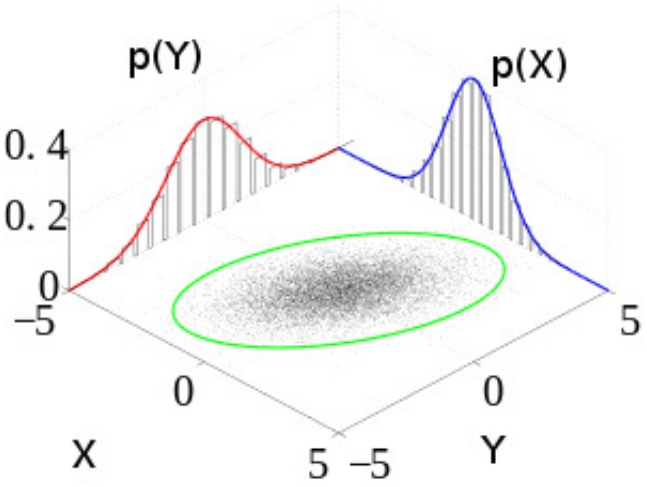

Figura A.2: Probabilidad conjunta.

Veamos la figura A.2[2], 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).

A.5. 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:

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

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.

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

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.6. 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:

$$p(A) = \sum_n p(A|B_n)p(B_n). \tag{A.12}$$

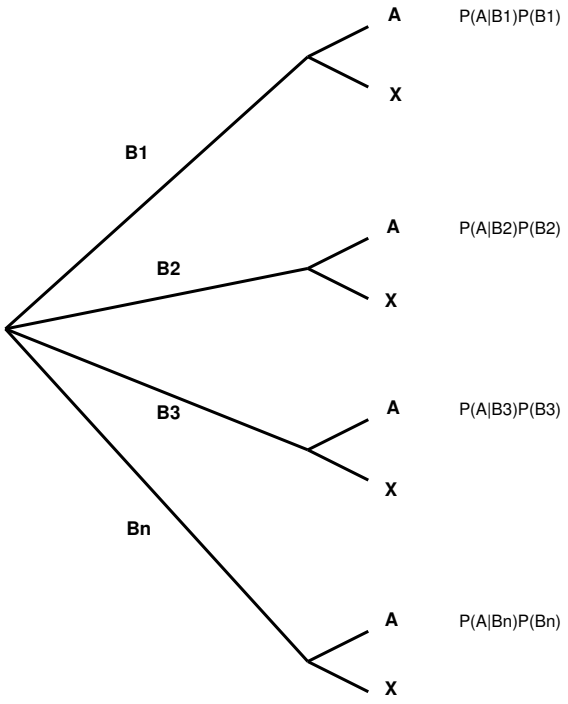

Figura A.3: Ley de probabilidad total representado en un diagrama de árbol.

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:

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

A.7. Teorema de representación de De Finetti

Una secuencia de variables aleatorias $(x_1, x_2, ..., x_n)$ es infinitamente intercambiable si y sólo 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) [7].

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 intercambia-

bilidad 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 [3].

A.8. Desigualdad de Jensen

Cuando no existe una relación de proporcionalidad entre dos variables, el promedio de la que se comporta como efecto resultará subestimado o sobreestimado si lo obtenemos a partir del promedio de la variable que funciona como causa.

De manera informal, se puede definir esta desigualdad así: cuando la relación que liga una variable dependiente "y" (o variable efecto) con una variable independiente "x" (o variable causa) no es lineal, se cumple siempre que el valor esperado de "y" correspondiente al promedio de "x" es diferente (mayor o menor, según la forma de la función: cóncava o convexa) del promedio de los valores observados de "y".

La desigualdad de Jensen generaliza el planteamiento de que en una función convexa, la línea secante permanece sobre el gráfico de la función, la cual es la desigualdad de Jensen para dos puntos: la línea secante consiste en la media pesada de la función convexa.

Como se muestra en la figura A.4, la línea secante viene dada por:

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

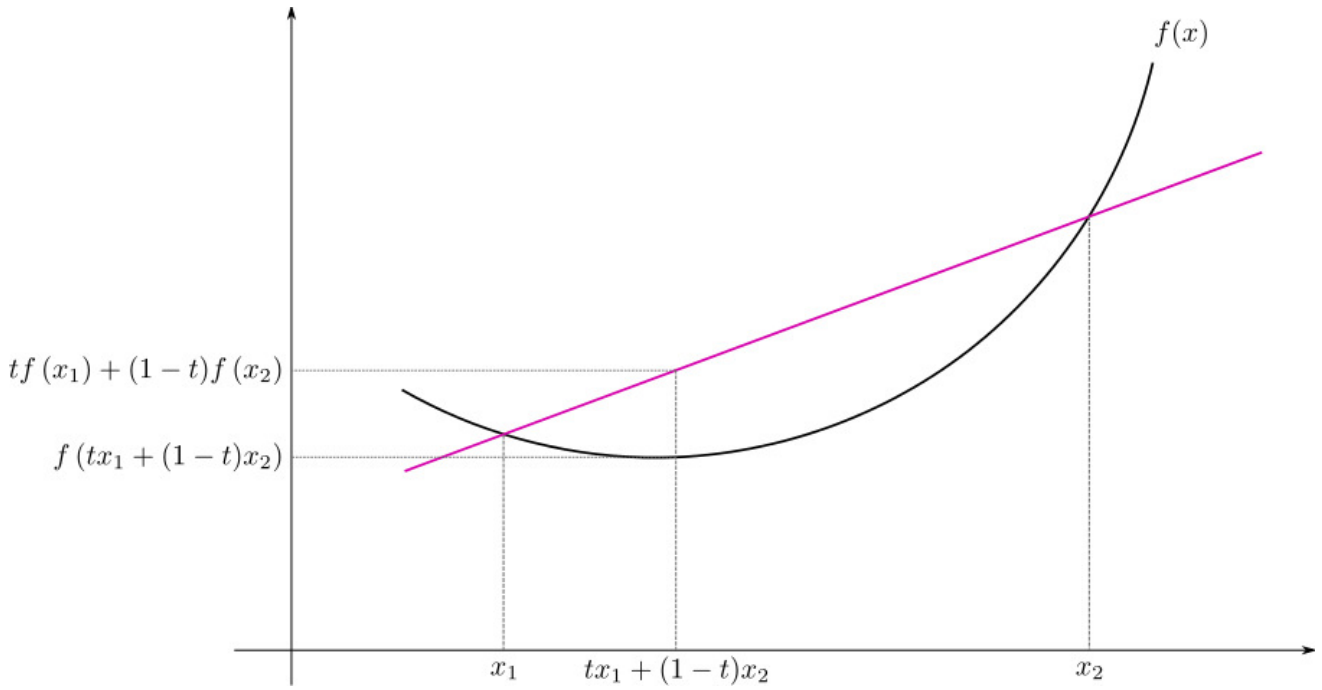

Figura A.4: Desigualdad de Jensen como representación de una línea secante (figura tomada de https://en.wikipedia.org/wiki/Jensen%27s_inequality).

$$tf(x_1) + (1-t)f(x_2), \tag{A.15}$$

Mientras que el gráfico de la función es la función convexa de la media pesada :

$$f\left(tx_1 + (1-t)x_2\right). \tag{A.16}$$

En la teoría de probabilidad, la desigualdad de Jensen es generalmente definida de la siguiente forma: si X es una variable aleatoria y φ es una función convexa, entonces:

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[\varphi(X)\right]. \tag{A.17}$$

Se puede ver https://en.wikipedia.org/wiki/Jensen%27s_inequality para más información sobre la desigualdad de Jensen.

Apéndice B

Estadística Bayesiana

B.1. Introducción: ¿Qué es la Estadística Bayesiana?

La Estadística Bayesiana es un enfoque particular de la aplicación de herramientas probabilísticas a problemas estadísticos, esta proporciona herramientas matemáticas para actualizar las estimaciones iniciales acerca de los acontecimientos aleatorios dados nuevos datos o pruebas sobre esos eventos.

En particular la inferencia bayesiana interpreta probabilidad como una medida de la verosimilitud o la confianza que un individuo puede poseer acerca de la ocurrencia de un evento en específico.

La idea básica es que se puede tener una estimación previa acerca de un evento, pero esta puede cambiar cuando nueva evidencia es obtenida. La estadística bayesiana nos da un medio matemático sólido de incorporar estimaciones previas, y pruebas, para producir nuevas estimaciones posteriores. Es decir, la estadística bayesiana nos proporciona herramientas matemáticas para actualizar racionalmente creencias subjetivas a la luz de nuevos datos o pruebas.

Esto está en contraste con otra forma de inferencia estadística, conocida como la estadística clásica o frecuentista, que supone que las probabilidades son la frecuencia de determinados eventos aleatorios que ocurren en un largo plazo de los ensayos repetidos.

Por ejemplo, a medida que lanzamos un dado no trucado (es decir, no ponderado) de seis caras en repetidas ocasiones, veríamos que cada número en el dado tiende a obtenerse $\frac{1}{6}$ de las veces.

La estadística frecuentista asume que las probabilidades son la frecuencia en un largo plazo de los acontecimientos al azar en los ensayos repetidos.

Al llevar a cabo la inferencia estadística, es decir, inferir información estadística de los sistemas probabilísticos, los dos enfoques - frecuentistas y bayesianos - tienen filosofías muy distintas.

La estadística frecuentista trata de eliminar la incertidumbre al proporcionar estimaciones generales, mientras que la estadística bayesiana intenta preservar y perfeccionar la incertidumbre mediante el ajuste de las estimaciones individuales cada nuevas evidencias.

En conclusión, el proceso bayesiano de análisis de datos puede ser idealizado dividiéndolo en los tres pasos siguientes:

1. La creación de un modelo de una distribución de probabilidad completo, para todas las magnitudes observables y no observables en un problema. El modelo debe ser coherente con el conocimiento sobre el problema científico subyacente y el proceso de recolección de datos.
2. El acondicionado en datos observados, mediante el cálculo y la interpretación de la distribución posteriori, es decir, la distribución de probabilidad condicional adecuada de las cantidades no observadas de interés, dados los datos observados.
3. Evaluar el ajuste del modelo y las implicaciones de la distribución posterior resultante: ¿qué tan bien el modelo se ajusta a los datos, son las conclusiones sustantivas razonables, y qué tan sensibles son los resultados de los supuestos del modelo en el paso 1? En respuesta, se puede alterar o ampliar el modelo y repetir los tres pasos.

B.2. Comparación entre Estadística Frecuentista y Bayesiana

Con el fin de hacer clara la distinción entre las dos filosofías diferentes estadísticas, vamos a considerar los siguientes dos ejemplos de sistemas probabilísticos:

1. Lanzamientos de una moneda: ¿Cuál es la probabilidad de que una moneda trucada salga cara?
2. Elección de un candidato en particular en unas eleciones presidenciales: ¿Cuál es la probabilidad de que un candidato que no ha participado en ninguna elección antes gane?

En la siguiente tabla se describen las aproximaciones frequentista y bayesiana a los problemas anteriores.

	Aprox. Frecuentista	Aprox. Bayesiana
Lanzamiento de una Moneda Trucada	La probabilidad de obtener una cara cuando se lanza una moneda trucada es la frecuencia relativa a largo plazo de ver una cara cuando repetidos lanzamientos de la moneda se llevan a cabo. Es decir, al llevar a cabo más lanzamientos de la moneda el número de caras que se obtengan en proporción a la cantidad total de lanzamientos tiende a la probabilidad "verdadera"de que la moneda salga cara.	Antes de cualquier lanzamiento de la moneda, un individuo puede creer que la moneda NO está cargada. Después de unos cuantos lanzamientos en los que se obtenga continuamente cara, la creencia previa acerca de la imparcialidad de la moneda es modificada para tener en cuenta el hecho de que se han obtenido, digamos, tres caras seguidas y por lo tanto la moneda podría estar trucada. Después de 500 lanzamientos, en los que se obtengan 400 caras, el individuo cree que la moneda está trucada. Es decir, la creencia ***posterior*** fue muy modificada de la creencia ***a priori*** sobre el estado la moneda.
Elecciones Presidenciales	Ya que el candidato sólo se presenta para esta elección en particular no pueden realizar "pruebas repetidas". En un entorno frequentista se construyen ensayos "virtuales" del proceso electoral. La probabilidad de que el candidato sea elegido como ganador se define como la frecuencia relativa de que el candidato gane en los ensayos "virtuales" en proporción a la cantidad total de ensayos.	Un individuo tiene una creencia previa de las posibilidades del candidato de ganar una elección y su confianza se puede cuantificar como una probabilidad. Sin embargo, otra persona también podría tener una creencia previa de que difiere por separado sobre las posibilidades de la misma candidatos. Con la llegada de nuevos datos, ambas creencias son (racionalmente) actualizada por el procedimiento bayesiano.

Así, en la interpretación bayesiana una probabilidad es un resumen de la opinión de un individuo. Un punto clave es que los diferentes individuos (inteligentes) pueden tener opiniones diferentes (y por tanto diferentes estimaciones a priori), ya que tienen diferentes acceso a los datos y las formas de interpretarlo. Sin embargo, en tanto que estos dos individuos obtengan ambos nuevos datos,

sus (potencialmente diferentes) estimaciones a ***priori*** darán lugar a estimaciones ***posteriores*** que comenzarán a converger una hacia la otra, en el marco del procedimiento de actualización racional de la inferencia bayesiana.

Con el fin de llevar a cabo la inferencia bayesiana, es necesario utilizar un famoso teorema de probabilidad conocido como el Teorema de Bayes. En la siguiente sección, se deriva dicho teorema usando la definición de probabilidad condicional. Sin embargo, no es esencial seguir la derivación con el fin de utilizar métodos bayesianos, por lo que, un lector que desee introducirse en el tema puede pasar directamente a la sección B.4.

B.3. Teorema de Bayes

Comenzamos considerando la definición de probabilidad condicional, lo que nos da una regla para determinar la probabilidad de un suceso A, dada la ocurrencia de otro evento B. Un ejemplo de pregunta en este sentido podría ser "¿Cuál es la probabilidad de que llueva hoy dado que hay nubes en el cielo?"

La definición matemática de la probabilidad condicional es el siguiente:

$$p(A|B) = \frac{p(A \bigcap B)}{p(B)} \tag{B.1}$$

Esto simplemente indica que la probabilidad de A dado que ocurra B, es igual a la probabilidad de que ambos ocurran, entre la probabilidad de que B ocurra.

O en el idioma del ejemplo anterior: La probabilidad de llueva hoy dado que hemos visto nubes, es igual a la probabilidad de que llueva y hayan nubes al mismo tiempo, entre la probabilidad de que hayan nubes.

De la ecuación B.1, se sigue que:

$$p(B) * p(A|B) = p(A \bigcap B) \tag{B.2}$$

Ahora bien, note que uno podría hacerse la pregunta exactamente opuesta: "¿Cual es la probabilidad de ver nubes dado que está lloviendo?" La cual corresponde al valor $p(B|A)$, que usando la fórmula análoga de B.1 es igual a:

$$p(B|A) = \frac{p(B \bigcap A)}{p(A)} \tag{B.3}$$

De donde se sigue,

$$p(A) * p(B|A) = p(B \bigcap A) \tag{B.4}$$

Luego, ya que $p(B \bigcap A) = p(A \bigcap B)$, igualando B.2 y B.4 se obtiene que:

$$p(B) * P(B|A) = p(A) * p(A|B) \tag{B.5}$$

De donde se sigue el famoso Teorema de Bayes:

$$P(B|A) = \frac{p(A) * p(A|B)}{p(B)} \tag{B.6}$$

Sin embargo, para un uso posterior de la regla de Bayes, será útil modificar el denominador, $p(B)$ en términos de $p(B|A)$. De hecho, usando la regla de la probabilidad total, podemos escribir:

$$p(B) = \sum p(A) * p(B|A) \tag{B.7}$$

con lo cual, mejoramos la ecuación B.6 como sigue:

$$P(B|A) = \frac{p(A) * p(A|B)}{\sum p(A) * p(B|A)} \tag{B.8}$$

Cabe destacar que la ecuación anterior no es propiamente el Teorema de Bayes, pero de aquí en adelante (y en un abuso de notación) nos referiremos a ella como tal. Ahora que hemos derivado la regla de Bayes somos capaces de aplicarlo a la inferencia estadística.

B.4. Inferencia Bayesiana

Como se dijo al principio de este apartado, la idea básica de la inferencia bayesiana es actualizar continuamente las estimaciones previas acerca de los eventos cada vez que se presenten nuevas pruebas. Esta es una manera muy natural de pensar acerca de los eventos probabilísticos.

Consideremos, por ejemplo, la estimación previa de que la Luna va a colisionar con la Tierra. Por cada noche que pasa, la aplicación de la inferencia bayesiana tenderá a corregir nuestra estimación previa a la estimación posterior de que es cada vez menos probable que la Luna echoque con la Tierra, ya que esta permanece en órbita.

Con el fin de demostrar un ejemplo numérico concreto de inferencia bayesiana es necesario introducir alguna nueva notación.

En primer lugar, debemos tener en cuenta el concepto de parámetros y modelos. Un parámetro podría ser la ponderación de una moneda injusta, que podríamos etiquetar como θ. Por lo tanto $\theta = P(c)$ describiría la distribución de probabilidad de que al lanzar la moneda se obtenga cara. El modelo es el medio real que codifica este lanzamiento matemáticamente. En este caso, el lanzamiento de moneda puede ser modelado como un ensayo de Bernoulli.

Definición B.4.1 *Un **Ensayo de Bernoulli** es un experimento aleatorio con sólo dos salidas, usualmente etiquetadas como "éxito" o "fracaso", en las que la probabilidad de éxito es exactamente igual cada una de las veces que se lleve a cabo el experimento. La probabilidad de éxito es denotada por θ, el cual es un valor entre 0 y 1.*

En el transcurso de la realización de algunos experimentos cara o cruz (repetido ensayos de Bernoulli) se generarán un conjunto de datos, $\mathcal{D}$, sobre los resultados de los lanzamientos (cara o cruz). Una pregunta natural sería "¿Cuál es la probabilidad de ver 3 caras en 8 lanzamientos (8 ensayos de Bernoulli), dada una moneda con $\theta = 0,5$)?".

Un modelo permite determinar la probabilidad de obtener $\mathcal{D}$, dado un valor del parámetro θ mediante el valor $P(\mathcal{D}|\theta)$.

Sin embargo, existe una pregunta alternativa cuya respuesta es un poco más compleja de obtener: "¿Cuál es la probabilidad de que la moneda esté o no trucada, dado que se ha visto una secuencia particular de cara y cruz?".

En este caso, el interés se centra en la distribución de probabilidad que refleja nuestra estimación acerca de los diferentes valores posibles de θ, Dado que hemos observado un conjunto de valores $\mathcal{D}$. La respuesta se describe con el valor $P(\theta|\mathcal{D})$. Note que este es el opuesto del viejo conocido $P(\mathcal{D}|\theta)$.

Pues bien, como el lector atento anticipará, el vínculo entre estas dos distribuciones viene dado por el Teorema de Bayes, como veremos en el apartado siguiente.

B.4.1. Teorema de Bayes en la Inferencia Bayesiana

$$P(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)P(\theta)/P(\mathcal{D}) \tag{B.9}$$

Dónde:

$P(\theta)$ **es la distribución a priori** Este valor representa nuestra estimación previa de θ sin tener en cuenta la evidencia $\mathcal{D}$. Es decir, es nuestra creencia de si la moneda está o no trucada.

$P(\theta|\mathcal{D})$ **es la distribución posterior** Este es el valor (refinado) la estimación de θ, una vez la evidencia se ha tenido en cuenta. Es la nueva creencia sobre la imparcialidad de la moneda dado que se han realizado 8 experimentos y se han obtenido 4 caras.

P $(\mathcal{D}|\theta)$ Esta es la probabilidad de ver los datos $\mathcal{D}$ como el resultado generado por un modelo con el parámetro θ. Si supiéramos, por ejemplo, que la moneda no está trucada, este valor indicaría la probabilidad de ver un número de caras en un determinado número de lanzamientos.

P $(\mathcal{D})$ **es la evidencia** Esta es la probabilidad de los datos, determinada mediante una suma (o una integral) sobre todos los valores posibles de θ, cada uno, multiplicado por por la intensidad con que creemos en esos valores particulares. Por ejemplo, si tuviéramos múltiples puntos de vista de si la moneda está o no trucada, o en qué medida está trucada (sin estar seguros de ninguno en particular), entonces $P(\mathcal{D})$ indicaría la probabilidad de ver una cierta secuencia de lanzamientos para todas los valores (que creemos) posibles sobre la imparcialidad de la moneda.

El objetivo de la inferencia bayesiana es que nos proporcione un procedimiento racional y matemáticamente racional para la incorporación de nuestras estimaciones previas junto a la evidencia, con el fin de producir una estimación posterior actualizada y eficiente.

Lo que hace a esta una técnica tan valiosa es que las estimaciones posteriores, pueden ser utilizadas nuevamente como estimaciones previas en virtud de la generación de nuevos datos. De ahí que la inferencia bayesiana nos permite ajustar continuamente las estimaciones bajo nuevos datos aplicando repetidamente la regla de Bayes.

Con el objetivo de ayudar a aclarar y fijar las ideas teóricas que se plantearon en las últimas dos secciones, se presenta a continuación un ejemplo concreto de la inferencia bayesiana mediante la más clásica herramienta de los estadísticos: el lanzamiento de una moneda.

B.5. Ejemplo: Lanzamiento de una Moneda

En este ejemplo se considerarán múltiples lanzamientos de una moneda cuya imparcialidad se desconoce. se usará la inferencia bayesiana para actualizar nuestras estimaciones acerca de si la moneda está o no trucada a medida que más datos (es decir, más lanzamientos de la moneda) se realizan. En principio, al no haber realizado ningún ensayo, no deberíamos tener estimaciones previas acerca de el peso de la moneda, es decir, podemos decir que cualquier nivel de imparcialidad es igualmente probable.

Empezaremos por realizar N ensayos de Bernoulli repetidos, con $\theta = 0,5$ que modelaran los primeros N lanzamientos de moneda. se utilizará una distribución uniforme como medio de carac-

terizar nuestra creencia previa de que no estamos seguros acerca de la imparcialidad. Esto indica que tenemos en cuenta cada valor de θ para ser igualmente probable.

Vamos a utilizar un procedimiento de actualización bayesiana para actualizar la estimación previa a la posterior a medida que se observen nuevos lanzamientos. No vamos a entrar en detalles sobre los procedimientos matemáticos explícitos que se llevan a cabo durante el proceso, sin embargo se explicará el proceso general.

En la figura B.1 podemos ver 6 momentos particulares durante la serie de ensayos de Bernoulli (los lanzamientos de la moneda).

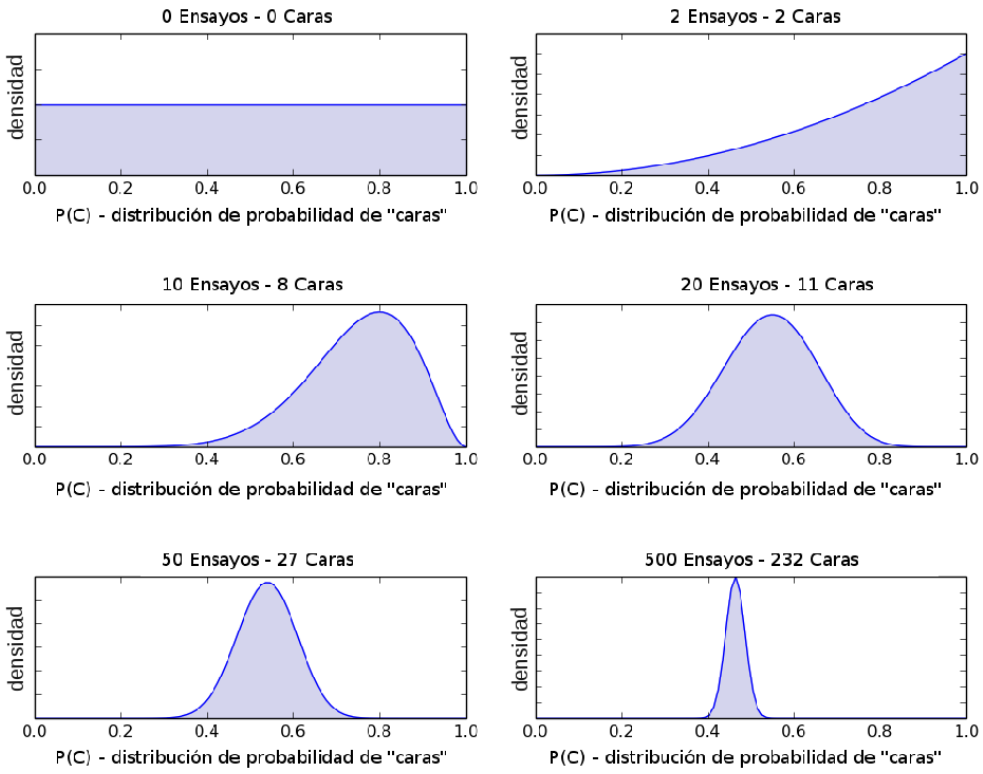

Figura B.1: Inferencia bayesiana sobre la imparcialidad de una moneda

En la primera cuadícula, no se ha llevado a cabo aún ningún ensayo, y por lo tanto la función de distribución de probabilidad (es decir, la distribución a priori) es la distribución uniforme, que afirma que se tiene la misma creencia en todos los valores posibles de θ.

La segunda cuadrícula, muestra 2 ensayos llevados a cabo en los cuales se obtuvo cara ambas veces. Nuestro procedimiento bayesiano usando las distribuciones beta conjugadas ahora nos permite actualizar a una distribución posterior. Observe cómo el peso de la distribución está des-

plazado hacia la parte derecha de la tabla. Esto indica que la estimación previa (de que moneda no estaba trucada), junto con 2 nuevos datos, nos lleva a pensar que es más probable que la moneda esté trucada (hacia caras).

Los siguientes dos paneles muestran los ensayos 10 y 20, respectivamente. Nótese que a pesar de que hemos visto 2 sellos en los primeros 10 ensayos estamos todavía (razonablemente) sesgados a pensar que es muy probable que la moneda esté trucadada hacia caras. Después de 20 ensayos, se han obtenido una mayor cantidad de sellos, por lo que la distribución de probabilidad se ha desplazado ahora más cerca de $\theta = 0,5$. con lo cual, estamos empezando a creer que es posible que la moneda sea justa.

Después de 50 y 500 ensayos (respectivamente las cuadrículas 5 y 6), creemos que es muy probable que la moneda sea imparcial, por lo que la densidad de la distribución de probabilidad se aproxima, cada vez más al valor $\theta = 0,5$.

Esto se indica por la reducción del ancho de la función de distribución de probabilidad, que ahora se agrupa apretadamente alrededor de $\theta = 0,46$ en el panel final. Si tuviéramos que llevar a cabo otros 500 ensayos (ya que la moneda es en realidad justo) veríamos esta densidad de probabilidad aún más centrada y agrupada al rededor de $\theta = 0,5$.

Bibliografía

[1] Abramowitz and Stegun, editors. Handbook of Mathematical Functions. Dover, New York,1970.

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

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

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

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

[6] Jordan, M. et al., *An Introduction to Variational Methods for Graphical Models*. Machine Learning, 37, 183–233, 1999.

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

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

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

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