Documentación LDA

Andrés Villaseñor

9 de noviembre de 2015

A través de modelación de tópicos es posible analizar grandes cantidades de datos sin categorizar, uno de los métodos más populares modelar tópicos es LDA (Latent Dirichlet Allocation) por sus cifras en ingles. Es imporante mencionar la propiedad de conjugación dadas las siguientes distribuciones: Si tenemos:

$$\phi \sim Dir(\alpha)$$
$$W \sim Mult(\phi)$$

y

$$n_k = |\{w_i : w_i = k\}|$$

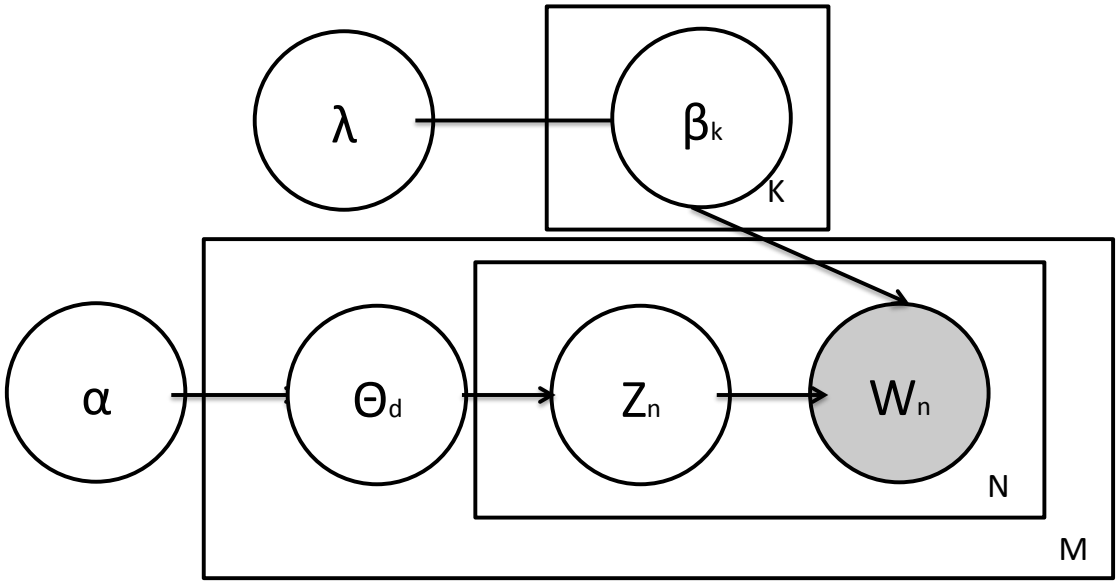
entonces

$$\begin{aligned} P(\phi|\alpha, W) &\propto P(W|\phi)P(\phi|\alpha) \\ &\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \\ &\propto \prod_k \phi^{\alpha_k + n_k - 1} \end{aligned}$$

Es decir son conjugadas, por lo que la distribución posterior tiene la misma forma que la distribución a priori y en nuestro caso sólo estamos agregando las observaciones. Ahora es más sencillo comprender la función de LDA, el cual es un modelo generativo que asume:

1. Cáda tópico $k\epsilon\{1, ..., K\}$ es una mezcla de palabras que tiene una distribución multinomial β_k la cual proviene de una ditribución de Dirichlet con parámetro λ.
2. Cáda documento $d\epsilon\{1, ..., M\}$ es una mezcla de tópicos con distribución multinomial σ_d la cual proviene de una distribución de Dirichlet con parámetro α.
3. Para cada palabra $n\epsilon\{1, ..., N\}$, se selecciona un tópico escondido (variable latente) Z_n, los cuales provienen de una distribución multinomial parametrizada por θ.

El modelo está representado por la siguiente Figura:



Por lo que es posible determinar de las variables no observadas (latentes) una palabra para cáda documento. Esto se logra a través de una asignación de tópicos a cada palabra dentro de los documentos y vamos a ir cambiando los tópicos de cada palabra resampleando utilizando *Gibbs* y se hace inferencia utilizando *Rao-Blackwellized* para conocer la probabilidad de la asignación de un tçopico a una palabra dado el cambio de asignación de tópicos de las demás palabras y dados los parámetros de las distribuciones de Dirichlet. Este proceso marginaliza la proporcion de tópicos en los documentos y por lo tanto asigna una probabilidad de tópicos a cada documento.