Análisis probabilístico de la semántica latente (PLSA)

Salvador López Mendoza

Junio de 2018

Ideas

Un documento es una muestra de varios tópicos.

Dado un comjunto de tópicos $\{\theta_1, \ldots, \theta_k\}$, se considera a un documento como una cierta generación de texto relacionado con cada uno de los tópicos.

$$d = p(d|\theta_1) + \ldots + p(d|\theta_k)$$

El documento completo también contiene palabras que no pertenecen a los tópicos (el, un, una). Son el *LM de fondo*.

Ayudan a descubrir los tópicos que discriminan a los documentos.

A estas palabras se les trata como un tópico especial θ_B.

Entonces

$$d = p(d|\theta_1) + \ldots + p(d|\theta_k) + p(d|\theta_B)$$

PEUVI

Generación de texto con varios tópicos

¿Cuál es la probabilidad de que una cierta palabra w sea parte de un documento?

¿$p(w) =$?

Sea $p(\theta_b) = \lambda_B$

La palabra debe pertenecer a alguno de los tópicos.

Entonces,

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_b)p(\theta_1)p(w|\theta_1) + \ldots + (1 - \lambda_b)p(\theta_1)p(w|\theta_k)$$

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_b)\sum_{j=1}^{k} p(\theta_j)p(w|\theta_j)$$

Como $p(\theta_i) = \pi_{d,i}$ se tiene que

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_b)\sum_{j=1}^{k} \pi_{d,j}p(w|\theta_j)$$

PEUVI

PLSA

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_b) \sum_{j=1}^{k} \pi_{d,j} p(w|\theta_j)$$

- λ_b representa el porcentaje de palabras de fondo. Es un valor conocido.
- $p(w|\theta_B)$ es el LM de fondo. También es conocido.
- $\pi_{d,j}$ es la cobertura del tópico θ_j en el documento d.
- $p(w|\theta_j)$ es la probabilidad de que la palabra w se encuentre en el tópico θ_j.

Hay muchos parámetros ($\{\pi_{d,j}\}$, $\{\theta_j\}$, para $j = 1, \ldots, k$.

Se necesita determinar el conjunto de parámetros que hagan la mejor clasificación.

PEUVI