Derivation of pLSA

CanoY

Derive the Co-occurrence Model of pLSA using EM algorithm.

Prove:

$$
\begin{aligned}
L &= \prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i, d_j)^{n(w_i,d_j)} \\
LL &= \sum_i \sum_j n(w_i, d_j) \log P(w_i, d_j) \\
&= \sum_i \sum_j n(w_i, d_j) \log \sum_{k=1}^{K} P(w_i, d_j, z_k)
\end{aligned}
$$

E step:

$$
\begin{aligned}
LL &\geq \sum_i \sum_j n(w_i, d_j) \sum_k P^{(t+1)}(z_k|w_i, d_j) \log P(w_i, d_j, z_k) \\
P^{(t+1)}(z_k|w_i, d_j) &= \frac{P^{(t)}(w_i|z_k)P^{(t)}(d_j|z_k)P^{(t)}(z_k)}{\sum_k P^{(t)}(w_i|z_k)P^{(t)}(d_j|z_k)P^{(t)}(z_k)}
\end{aligned}
$$

M step:

$$
\begin{aligned}
\min \quad & P := \sum_i \sum_j n(w_i, d_j) \sum_k P^{(t+1)}(z_k|w_i, d_j) \log[P(z_k)P(w_i|z_k)P(d_j|z_k)] \\
\text{s.t.} \quad & \sum_i P(w_i|z_k) = 1, \quad z = 1, 2, 3, \cdots, K \\
& \sum_j P(d_j|z_k) = 1, \quad z = 1, 2, 3, \cdots, K \\
& \sum_i P(z_k) = 1
\end{aligned}
$$

$$\Lambda = P + \sum_k u_k(\sum_i P(w_i|z_k) - 1) + \sum_k v_k(\sum_j P(d_j|z_k) - 1) + \lambda(\sum_k P(z_k) - 1)$$

$$\frac{\partial \Lambda}{\partial P(z_k)} = \frac{\sum_i \sum_j n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j)}{P(z_k)} + \lambda = 0$$

$$\Rightarrow P^{(t+1)}(z_k) = \frac{R_k}{N_w}$$

$$R_k = \sum_i \sum_j n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j), \quad N_w = \sum_i \sum_j n(w_i, d_j)$$

$$\frac{\partial \Lambda}{\partial P(w_i|z_k)} = \frac{\sum_j n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j)}{P(w_i|z_k)} + u_k = 0$$

$$\Rightarrow P^{(t+1)}(w_i|z_k) = \frac{R_{ik}}{R_k}$$

$$R_{ik} = \sum_j n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j)$$

$$\frac{\partial \Lambda}{\partial P(d_j|z_k)} = \frac{\sum_i n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j)}{P(d_j|z_k)} + v_k = 0$$

$$\Rightarrow P^{(t+1)}(w_i|z_k) = \frac{R_{jk}}{R_k}$$

$$R_{jk} = \sum_i n(w_i, d_j) P^{(t+1)}(z_k|w_i, d_j)$$