Mixture of Categoricals
Expectation Maximization

David Mareček

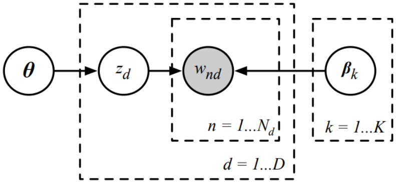
October 26, 2022

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)

A mixture of categoricals model



$$z_d \sim Cat(\vec{\theta})$$
$$w_{nd}|z_d \sim Cat(\vec{\beta}_{z_d})$$

We want to allow for a mixture of K categoricals parametrised by $\vec{\beta}_1, \ldots, \vec{\beta}_K$.
Each of those categorical distributions corresponds to a document category.

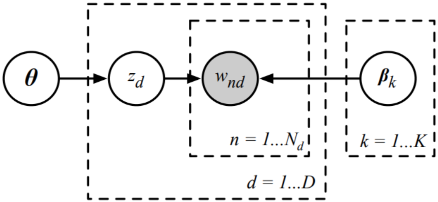
- $z_d \in 1, \ldots, K$ assigns document d to one of the K categories.
- $\theta_k = p(z_d = k)$ is the probability any document d is assigned to category k.
- so $\vec{\theta} = [\theta_1, \ldots, \theta_K]$ is the parameter of a categorical distribution over K categories.

We have introduced a new set of hidden variables z_d.

- How do we fit those variables? What do we do with them?
- Are these variables interesting? Or are we only interested in $\vec{\theta}$ and $\vec{\beta}$?

A mixture of categoricals model: the likelihood

$$
\begin{aligned}
p(w|\vec{\theta}, \vec{\beta}) &= \prod_{d=1}^{D} p(w_d|\vec{\theta}, \vec{\beta}) \\
&= \prod_{d=1}^{D} \sum_{k=1}^{K} p(w_d, z_d = k|\vec{\theta}, \vec{\beta}) \\
&= \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_d = k|\vec{\theta}) p(w_d|z_d = k, \vec{\beta}_k) \\
&= \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_d = k|\vec{\theta}) \prod_{n=1}^{N_d} p(w_{nd}|z_d = k, \vec{\beta}_k)
\end{aligned}
$$



$$z_d \sim Cat(\vec{\theta})$$

$$w_{nd}|z_d \sim Cat(\vec{\beta}_{z_d})$$

w: all the words in all the documents,
w_d: all the words in a document d,
w_{nd}: the n-th word in document d.

Expectation Maximization and Mixture of Categoricals

We want to maximize the likelihood of the data:

$$p(w|\vec{\theta}, \vec{\beta}) = \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_d = k|\vec{\theta}) \prod_{n=1}^{N_d} p(w_{nd}|z_d = k, \vec{\beta}_k)$$

However, the latent variables (document categories) are unknown.

Expectation-Maximization algorithm:

1. Initialize $\vec{\theta}$ and $\vec{\beta}$ randomly.
2. *E-step*: For each d and k, compute responsibilities r_{kd} as probabilities $q(z_d = k|\vec{\theta}, \vec{\beta})$
3. *M-step*: Maximize the likelihood of the model with weighted by the responsibilities r_{kd} from step 2 and update the parameters $\vec{\beta}$ and $\vec{\theta}$.
4. Repeat steps 2 and 3 until convergence.

Expectation Maximization and Mixture of Categoricals

E-step: For each document, compute posterior distribution over categories:

$$r_{kd} = q(z_d = k) \propto p(z_d = k|\vec{\theta}) \prod_{n=1}^{N_d} p(w_{nd}|z_d = k, \vec{\beta}_k) = \theta_k \prod_{m=1}^{M} \beta_{km}^{c_{md}}$$

M-step: Maximize the log-likelihood weighted by the responsibilities r_{kd}:

$$\sum_{d=1}^{D} \sum_{k=1}^{K} r_{kd} \log p(w, z_d) = \sum_{k,d} r_{kd} \log[p(z_d = k|\vec{\theta}) \prod_{n=1}^{N_d} p(w_{nd}|z_d = k, \vec{\beta}_k)]$$

$$= \sum_{k,d} r_{kd} (\log \theta_k + \log \prod_{m=1}^{M} \beta_{km}^{c_{md}})$$

$$= \sum_{k,d} r_{kd} (\log \theta_k + \sum_{m=1}^{M} c_{md} \log \beta_{km})$$

M-step (continued): We need Lagrange multipliers to constrain the maximization of the function ensure proper distributions.

$$L_1 = \sum_{k=1}^{K} \sum_{d=1}^{D} r_{kd} (\log \theta_k + \sum_{m=1}^{M} c_{md} \log \beta_{km}) + \lambda (1 - \sum_{k'=1}^{K} \theta_{k'})$$

$$\frac{\partial L_1}{\partial \theta_k} = \sum_{d=1}^{D} r_{kd} \frac{1}{\theta_k} - \lambda = 0 \quad \Rightarrow \quad \theta_k = \frac{\sum_{d=1}^{D} r_{kd}}{\lambda} = \frac{\sum_{d=1}^{D} r_{kd}}{\sum_{k'=1}^{K} \sum_{d=1}^{D} r_{k'd}} = \frac{\sum_{d=1}^{D} r_{kd}}{D}$$

$$L_2 = \sum_{k=1}^{K} \sum_{d=1}^{D} r_{kd} (\log \theta_k + \sum_{m=1}^{M} c_{md} \log \beta_{km}) + \sum_{k'=1}^{K} \lambda_{k'} (1 - \sum_{m'=1}^{M} \beta_{k'm'})$$

$$\frac{\partial L_2}{\partial \beta_{km}} = \sum_{d=1}^{D} r_{kd} \frac{c_{md}}{\beta_{km}} - \lambda_k = 0 \quad \Rightarrow \quad \beta_{km} = \frac{\sum_{d=1}^{D} r_{kd} c_{md}}{\lambda_k} = \frac{\sum_{d=1}^{D} r_{kd} c_{md}}{\sum_{m'=1}^{M} \sum_{d=1}^{D} r_{kd} c_{m'd}}$$

Expectation Maximization and Mixture of Categoricals

EM Algorithm:

1. Initialize $\vec{\theta}$ and $\vec{\beta}$ randomly.
2. *E-step*: For each d and k, compute responsibilities r_{kd} using current parameters $\vec{\theta}$ and $\vec{\beta}$.

$$r_{kd} = \frac{\theta_k \prod_{m=1}^{M} \beta_{km}^{c_{md}}}{\sum_{k'=1}^{K} \theta_k' \prod_{m=1}^{M} \beta_{k'm}^{c_{md}}}$$

3. *M-step*: Maximize the likelihood of the model with weighted by the responsibilities r_{kd} from step 2 and update the parameters $\vec{\theta}$ and $\vec{\beta}$.

$$\beta_{km} = \frac{\sum_{d=1}^{D} r_{kd} c_{md}}{\sum_{m'=1}^{M} \sum_{d=1}^{D} r_{kd} c_{m'd}}, \qquad \theta_k = \frac{\sum_{d=1}^{D} r_{kd}}{D}$$

4. Repeat steps 2 and 3 until convergence.

Exercises

1. Let's have $K = 2$, $M = \{a, b, c\}$ and observe the following set of documents

$$D_1 = \{a, b, b\}, \quad D_2 = \{a, c, c\}, \quad D_3 = \{a, b\}, \quad D_4 = \{c\}.$$

   Could you estimate the resulting $\vec{\theta}$ and $\vec{\beta}$?

2. What would happen if we initialize the parameters $\vec{\theta}$ and $\vec{\beta}$ uniformly?