Gibbs sampling for Mixture of Categoricals

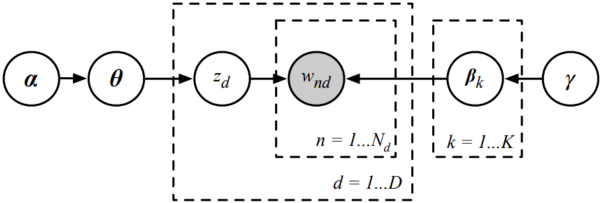
David Mareček

November 9, 2023

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

Bayesian Mixture of Categoricals Model



$$z_d \sim Cat(\vec{\theta})$$
$$\vec{\theta} \sim Dir(\vec{\alpha})$$
$$w_{nd} | z_d, \vec{\beta} \sim Cat(\vec{\beta}_{z_d})$$
$$\vec{\beta}_k \sim Dir(\vec{\gamma})$$

An alternative, Bayesian treatment infers these parameters starting from priors, e.g.:

- $\vec{\theta} \sim Dir(\vec{\alpha})$ is a symmetric Dirichlet over category probabilities,
- $\vec{\beta}_k \sim Dir(\vec{\gamma})$ are symmetric Dirichlets over vocabulary probabilities.

Collapsed sampling for Bayessian Mixture of Categoricals

We want to employ Gibbs Sampling to sample the model variables z_d, β, and θ.

Collapsed Gibbs Sampler: We will sample only the latent variables z_d. The model parameters β and θ are marginalized (integrated out).
In each step, we sample one latent variable z_d conditioned by all the other latent variables z_{-d}, all the documents w, and our hyperparameters γ and α.

$$p(z_d = k | \{w\}, \{z_{-d}\}, \gamma, \alpha)$$

We rewrite it using Bayes theorem.

$$= \frac{p(z_d = k | \{z_{-d}\}, \gamma, \alpha)\ p(\{w\} | z_d = k, \{z_{-d}\}, \gamma, \alpha)}{p(\{w\} | \{z_{-d}\}, \gamma, \alpha)}$$

The denominator is constant (does not depend on category k), the parts in the nominator also do not depend on both the hyperparameters.

$$\propto p(z_d = k | \{z_{-d}\}, \alpha)\ p(\{w\} | z_d = k, \{z_{-d}\}, \gamma)$$

Collapsed sampling for Bayessian Mixture of Categoricals [2]

We have:

$$p(z_d = k|\{w\}, \{z_{-d}\}, \gamma, \alpha) \;\propto\; p(z_d = k|\{z_{-d}\}, \alpha)\; p(\{w\}|z_d = k, \{z_{-d}\}, \gamma)$$

Probability of the document collection $p(\{w\})$ may be rewritten as $p(w_d|w_{-d})p(w_{-d})$. However $p(w_{-d})$ does not depend on z_d, so:

$$p(z_d = k|\{w\}, \{z_{-d}\}, \gamma, \alpha) \;\propto\; p(z_d = k|\{z_{-d}\}, \alpha)\; p(\{w_d\}|w_{-d}, z_d = k, \{z_{-d}\}, \gamma)$$

$$\propto p(z_d = k|\{z_{-d}\}, \alpha)\; \prod_{n=1}^{N_d} p(w_{nd}|\{w_{-d}\}, z_d = k, \{z_{-d}\}, \gamma)$$

For computing $p(z_d|z_{-d})$ and $p(w_d|w_{-d})$, we integrate over all possible parameters θ and γ respectively.

$$\propto \int p(z_d = k|\theta)p(\theta|z_{-d}, \alpha)d\theta \prod_{n=1}^{N_d} \int p(w_{nd}|\beta_k)p(\beta_k|\{w_{-d}\}, \{z_{-d}\}, \gamma)d\beta_k$$

Collapsed sampling for Bayessian Mixture of Categoricals [3]

We have:

$$p(z_d = k|\{w\}, \{z_{-d}\}, \gamma, \alpha) \propto \int p(z_d = k|\theta)p(\theta|z_{-d}, \alpha)d\theta \prod_{n=1}^{N_d} \int p(w_{nd}|\beta_k)p(\beta_k|\{w_{-d}\}, \{z_{-d}\}, \gamma$$

Both the integrals are expected values of Dirichlet distributions, therefore:

$$p(z_d = k|\{w\}, \{z_{-d}\}, \gamma, \alpha) \; \propto \; \frac{\alpha + c_d[k] - 1}{K\alpha + D - 1} \prod_{n=1}^{N_d} \frac{\gamma + c_w[w_{nd}][k]}{M\gamma + \sum_{m=1}^{M} c_w[m][k]}$$

- $c_d[k]$ ... How many documents are assigned to topic k.
- $c_w[m][k]$ ... How many times the word m is in a document assigned to topic k.

Algorithm for Bayessian Mixture of Categoricals

initialize z_d *randomly* $\forall d \in 1..D$;
compute initial counts $c_d[k]$, $c_w[m][k]$, $c[k]$ $\forall k \in 1..K$, $\forall m \in 1..M$;
for $i \leftarrow 1$ **to** I **do**
    for $d \leftarrow 1$ **to** D **do**
        $c_d[z_d]$−−;
        for $n \leftarrow 1$ **to** N_d **do**
            $c_w[w_{nd}][z_d]$−−; $c[z_d]$−−;
        end
        for $k \leftarrow 1$ **to** K **do**
            $p[k] = \frac{\alpha + c_d[k]}{K\alpha + D - 1} \prod_{n=1}^{N_d} \frac{\gamma + c_w[w_{nd}][k]}{M\gamma + c[k]}$;
        end
        sample k *from probability distribution* $p[k]$;
        $z_d \leftarrow k$; $c_d[k]{+}{+}$;
        for $n \leftarrow 1$ **to** N_d **do**
            $c_w[w_{nd}][z_d]{+}{+}$; $c[z_d]{+}{+}$;
        end
    end