Document models

Carl Edward Rasmussen

November 15th, 2017

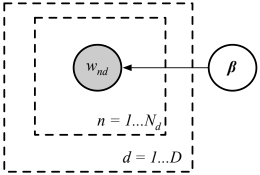
Key concepts

- a simple document model
- a mixture model for document
- fitting the mixture model with EM

A really simple document model

Consider a collection of D documents from a vocabulary of M words.

- N_d: number of words in document d.
- w_{nd}: n-th word in document d ($w_{nd} \in \{1 \dots M\}$).
- $w_{nd} \sim \text{Cat}(\underline{\beta})$: each word is drawn from a discrete categorical distribution with parameters $\underline{\beta}$
- $\underline{\beta} = [\beta_1, \dots, \beta_M]^\top$: parameters of a categorical / multinomial distribution[1] over the M vocabulary words.



small box is one document

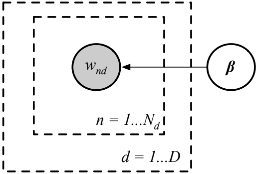
Summing beta_1 to beta_M gives unity.

[1]It's a categorical distribution if we observe the sequence of words in the document, it's a multinomial if we only observe the counts.

A really simple document model

Modelling D documents from a vocabulary of M unique words.

- N_d: number of words in document d.
- w_{nd}: n-th word in document d ($w_{nd} \in \{1 \ldots M\}$).
- $w_{nd} \sim \text{Cat}(\boldsymbol{\beta})$: each word is drawn from a discrete categorical distribution with parameters $\boldsymbol{\beta}$



We can fit $\boldsymbol{\beta}$ by maximising the likelihood:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \prod_{d=1}^{D} \prod_{n}^{N_d} \text{Cat}(w_{nd}|\boldsymbol{\beta})$$

$$= \text{argmax}_{\boldsymbol{\beta}} \text{Mult}(c_1, \ldots, c_M|\boldsymbol{\beta}, N)$$

$$\boxed{\hat{\beta}_m = \frac{c_m}{N} = \frac{c_m}{\sum_{\ell=1}^{M} c_\ell}}$$

- $N = \sum_{d=1}^{D} N_d$: total number of words in the collection.
- $c_m = \sum_{d=1}^{D} \sum_{n}^{N_d} \mathbb{I}(w_{nd} = m)$: total count of vocabulary word m.

Maximum Likelihood and Lagrange multipliers

In maximum likelihood learning, we want to maximize the (log) likelihood

(Notice the absence of multinomial coefficient, hence modelling sequence but not count.)

$$p(\mathbf{w}|\boldsymbol{\beta}) = \prod_{n=1}^{D} \prod_{n=1}^{N_d} \beta_{w_{nd}} = \prod_{m=1}^{M} \beta_m^{c_m}, \quad \text{or} \quad \log p(\mathbf{w}|\boldsymbol{\beta}) = \sum_{m=1}^{M} c_m \log \beta_m,$$

subject to the normalizing constraint that $\sum_{m=1}^{M} \beta_m = 1$.
An easy way to do this optimization is to add the Lagrange multiplier to the cost

$$\mathsf{F} = \sum_{m=1}^{M} c_m \log \beta_m + \lambda(1 - \sum_{m=1}^{M} \beta_m),$$

taking derivatives and setting to zero, we obtain

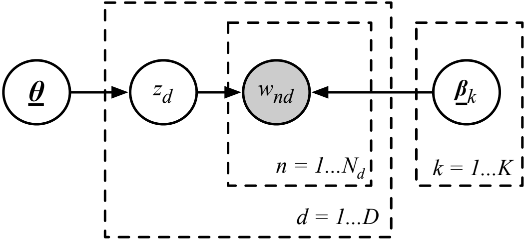
$$\frac{\partial \mathsf{F}}{\partial \beta_m} = \frac{c_m}{\beta_m} - \lambda = 0 \Rightarrow \beta_m = \frac{c_m}{\lambda} \text{ and } \frac{\partial \mathsf{F}}{\partial \lambda} = 0 \Rightarrow \sum_{m=1}^{M} \beta_m = 1,$$

which we combine to $\beta_m = c_m/n$, where n is the total number of words.

Limitations of the really simple document model

- Document d is the result of sampling N_d words from the categorical distribution with parameters β.
- β estimated by maximum likelihood reflects the aggregation of all documents.
- All documents are therefore modelled by the global word frequency distribution. bad!
- This generative model does not specialise.
- We would like a model where different documents might be about different *topics*.

A mixture of categoricals model



$$
\begin{aligned}
z_d &\sim \mathrm{Cat}(\underline{\theta}) \\
w_{nd} | z_d &\sim \mathrm{Cat}(\underline{\beta}_{z_d})
\end{aligned}
$$

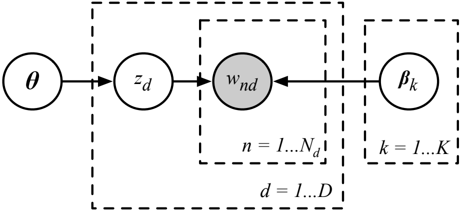
We want to allow for a mixture of K categoricals parametrised by $\underline{\beta}_1, \ldots, \underline{\beta}_K$. Each of those categorical distributions corresponds to a *document category*.

- $z_d \in \{1, \ldots, K\}$ 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 $\underline{\theta} = [\theta_1, \ldots, \theta_K]$ 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 $\underline{\theta}$ and $\underline{\beta}$?

A mixture of categoricals model: the likelihood



$$z_d \sim \mathrm{Cat}(\theta)$$
$$w_{nd}|z_d \sim \mathrm{Cat}(\beta_{z_d})$$

prob that the observed words appear this way, given the model parameters.

$$p(\mathbf{w}|\theta,\boldsymbol{\beta}) = \prod_{d=1}^{D} p(\mathbf{w}_d|\theta,\boldsymbol{\beta})$$

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} p(\mathbf{w}_d, z_d = k|\theta,\boldsymbol{\beta})$$ sum rule of prob.

beta_k theta

$$= \prod_{d=1}^{D}\sum_{k=1}^{K} p(z_d = k|\theta)\,p(\mathbf{w}_d|z_d = k,\boldsymbol{\beta}_k)$$ beta_k and theta are independent.

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

Assume all words in document d are independent.

EM and Mixtures of Categoricals

In the mixture model, the likelihood is:

sum over latent variable z.

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

E-step: for each d, set q to the posterior (where $c_{md} = \sum_{n=1}^{N_d} \mathbb{I}(w_{nd} = m)$):

$$q(z_d = k) \propto \underbrace{p(z_d = k|\theta)}_{\text{prior}} \underbrace{\prod_{n=1}^{N_d} p(w_{nd}|\beta_{k,w_n})}_{\text{likelihood}} = \theta_k \operatorname{Mult}(c_{1d}, \ldots, c_{Md}|\beta_k, N_d) \stackrel{\text{def}}{=} r_{kd}$$

M-step: Maximize

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

Why has d become a summation?

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

$$= \sum_{k,d} r_{kd} \Big(\sum_{m=1}^{M} c_{md} \log \beta_{km} + \log \theta_k \Big) \stackrel{\text{def}}{=} F(R, \theta, \beta)$$

EM: M step for mixture model

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

Need Lagrange multipliers to constrain the maximization of F and ensure proper distributions.

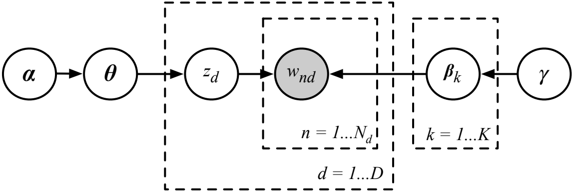
$$\hat{\theta}_k \leftarrow \operatorname{argmax}_{\theta_k} \; F(R, \theta, \beta) + \lambda \Big(1 - \sum_{k'=1}^{K} \theta_{k'}\Big)$$

$$= \frac{\sum_{d=1}^{D} r_{kd}}{\sum_{k'=1}^{K} \sum_{d=1}^{D} r_{k'd}} = \frac{\sum_{d=1}^{D} r_{kd}}{D}$$

$$\hat{\beta}_{km} \leftarrow \operatorname{argmax}_{\beta_{km}} \; F(R, \theta, \beta) + \sum_{k'=1}^{K} \lambda_{k'} \Big(1 - \sum_{m'=1}^{M} \beta_{k'm'}\Big)$$

$$= \frac{\sum_{d=1}^{D} r_{kd} c_{md}}{\sum_{m'=1}^{M} \sum_{d=1}^{D} r_{kd} c_{m'd}}$$

A Bayesian mixture of categoricals model



$$\begin{aligned} \theta &\sim \mathrm{Dir}(\alpha) \\ \beta_k &\sim \mathrm{Dir}(\gamma) \\ z_d|\theta &\sim \mathrm{Cat}(\theta) \\ w_{nd}|z_d, \beta &\sim \mathrm{Cat}(\beta_{z_d}) \end{aligned}$$

With the EM algorithm we have essentially estimated θ and β by maximum likelihood. An alternative, Bayesian treatment infers these parameters starting from priors, e.g.:

- $\theta \sim \mathrm{Dir}(\alpha)$ is a symmetric Dirichlet over category probabilities.
- $\beta_k \sim \mathrm{Dir}(\gamma)$ are symmetric Dirichlets over vocabulary probabilities.

What is different?

- We no longer want to compute a point estimate of θ or β.
- We are now interested in computing the *posterior* distributions.