Lecture 12: Models for documents

Machine Learning 4F13, Spring 2015

Carl Edward Rasmussen and Zoubin Ghahramani

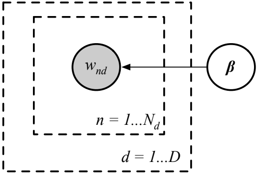
CUED

http://mlg.eng.cam.ac.uk/teaching/4f13/

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

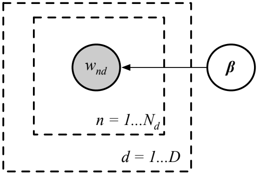


[1] It's a categorical distribution if we observe the sequence of words in the document, it's a multinomial if we only observe the counts.

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:

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

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

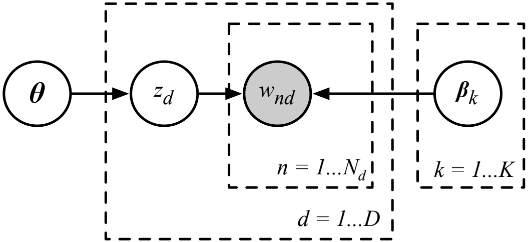
$$\boxed{\hat{\beta}_m = \frac{c_m}{N} = \frac{c_m}{\sum_{\ell=1}^{M} c_\ell}}$$

- $N = \sum_{d=1}^{D} N_d$: 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.

Limitations of the really simple document model

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

A mixture of categoricals model



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

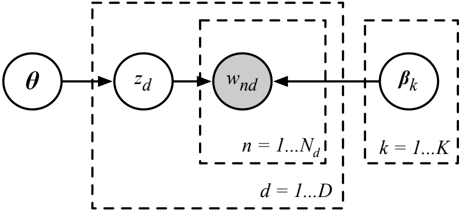
We want to allow for a mixture of K categoricals parametrised by $\beta_1, \ldots, \beta_K$. Each of those categorical distributions corresponds to a *document category*.

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

A mixture of categoricals model: the likelihood



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

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

The Expectation Maximization (EM) algorithm

Given a set of observed (visible) variables V, a set of unobserved (hidden / latent / missing) variables H, and model parameters θ, optimize the log likelihood:

$$\mathcal{L}(\theta) = \log p(V|\theta) = \log \int p(H, V|\theta) dH, \quad (1)$$

where we have written the marginal for the visibles in terms of an integral over the joint distribution for hidden and visible variables.

Using *Jensen's inequality* for any distribution of hidden states $q(H)$ we have:

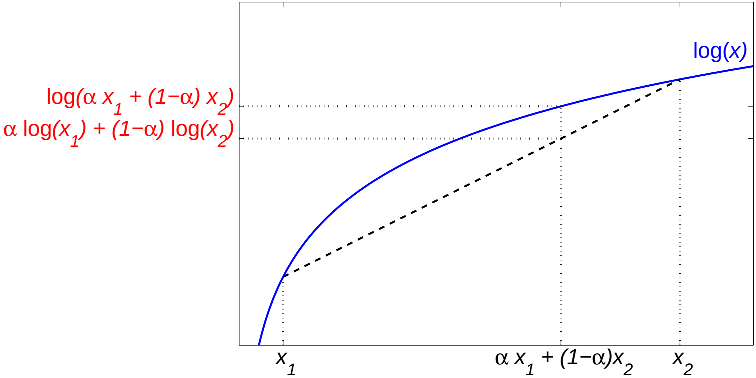
$$\mathcal{L}(\theta) = \log \int q(H) \frac{p(H, V|\theta)}{q(H)} dH \geqslant \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH = \mathcal{F}(q, \theta), \quad (2)$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt q and θ, and we can prove that this will never decrease $\mathcal{L}(\theta)$.

Jensen's Inequality

For any concave function, such as $\log(x)$



For $\alpha_i \geqslant 0$, $\sum_i \alpha_i = 1$ and any $\{x_i > 0\}$

$$\log\Big(\sum_i \alpha_i x_i\Big) \geqslant \sum_i \alpha_i \log(x_i)$$

Equality if and only if $\alpha_i = 1$ for some i (and therefore all others are 0).

The E and M steps of EM

The lower bound on the log likelihood:

$$\mathcal{F}(q,\theta) \;=\; \int q(H)\log\frac{p(H,V|\theta)}{q(H)}\,dH \;=\; \int q(H)\log p(H,V|\theta)\,dH + \mathcal{H}(q), \qquad (3)$$

where $\mathcal{H}(q) = -\int q(H)\log q(H)\,dH$ is the entropy of q. We iteratively alternate:

E step: maximize $\mathcal{F}(q,\theta)$ wrt the distribution over hidden variables given the parameters:

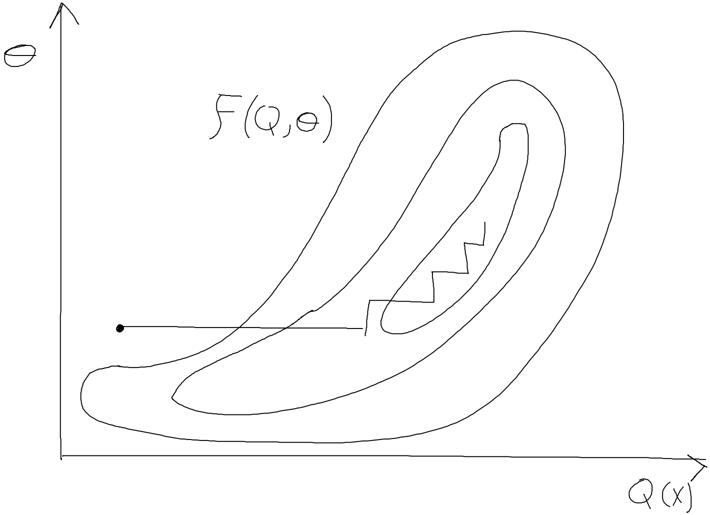
$$q^{(k)}(H) := \underset{q(H)}{\operatorname{argmax}}\; \mathcal{F}\big(q(H), \theta^{(k-1)}\big). \qquad (4)$$

M step: maximize $\mathcal{F}(q,\theta)$ wrt the parameters given the hidden distribution:

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}}\; \mathcal{F}\big(q^{(k)}(H), \theta\big) = \underset{\theta}{\operatorname{argmax}} \int q^{(k)}(H)\log p(H,V|\theta)\,dH, \qquad (5)$$

which is equivalent to optimizing the expected complete-data likelihood $p(H,V|\theta)$, since the entropy of $q(H)$ does not depend on θ.

EM as Coordinate Ascent in $\mathcal{F}$



The EM algorithm never decreases the log likelihood

The difference between the objective functions:

$$\begin{aligned}\mathcal{L}(\theta) - \mathcal{F}(q,\theta) &= \log p(V|\theta) - \int q(H) \log \frac{p(H,V|\theta)}{q(H)} dH \\ &= \log p(V|\theta) - \int q(H) \log \frac{p(H|V,\theta)p(V|\theta)}{q(H)} dH \\ &= -\int q(H) \log \frac{p(H|V,\theta)}{q(H)} dH = \mathcal{KL}\big(q(H), p(H|V,\theta)\big),\end{aligned}$$

is called the Kullback-Liebler divergence; it is non-negative and zero if and only if $q(H) = p(H|V,\theta)$ (thus this is the E step). Although we are optimising a lower bound, $\mathcal{F}$, the likelihood $\mathcal{L}$ is still increased in every iteration:

$$\mathcal{L}\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leqslant} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leqslant} \mathcal{L}\big(\theta^{(k)}\big),$$

where the first equality holds because of the E step, and the first inequality comes from the M step and the final inequality from Jensen. Usually EM converges to a local optimum of $\mathcal{L}$ (although there are exceptions).

EM and Mixtures of Categoricals: Overview

We will use EM to learn a mixture of categoricals models, with observed data $V \to \mathbf{w}$, hidden variables $H \to \mathbf{z}$, and parameters $\theta \to (\boldsymbol{\theta}, \boldsymbol{\beta})$.

In this mixture model, the likelihood "$p(V|\theta)$" is:

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

The joint distribution "$p(H, V|\theta)$" is

$$p(\mathbf{w}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\beta}) \;=\; \prod_{d=1}^{D} p(z_d|\boldsymbol{\theta}) \prod_{n=1}^{N_d} p(w_{nd}|z_d, \boldsymbol{\beta})$$

The "$q(H)$" will be categorical over the K categories for each document:

$$q(\mathbf{z}) = \prod_{d} q(z_d)$$

E-step will optimize $q(\mathbf{z})$; M-step will optimise $\boldsymbol{\theta}, \boldsymbol{\beta}$.

EM and Mixtures of Categoricals: E-step

Remember that in the E-step we set the hidden variable distribution to the posterior, "$q(H) = P(H|V, \theta)$" since this minimises the KL-divergence and so maximises the lower bound.

$$\begin{aligned} q(\mathbf{z}) &= p(\mathbf{z}|\mathbf{w}, \theta, \boldsymbol{\beta}) \\ \prod_d q(z_d) &= \prod_d p(z_d|\mathbf{w}_d, \theta, \boldsymbol{\beta}) \propto \prod_d p(z_d|\theta)\, p(\mathbf{w}_d|z_d, \boldsymbol{\beta}) \end{aligned}$$

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

$$\begin{aligned} q(z_d = k) &\propto p(z_d = k|\theta) \prod_{n=1}^{N_d} p(w_{nd}|\beta_{k,w_n}) \\ &= \theta_k \text{ Mult}(c_{1d}, \ldots, c_{Md}|\boldsymbol{\beta}_k, N_d) \stackrel{\text{def}}{=} r_{kd} \end{aligned}$$

We call the r_{kd} the "responsibility" of category k for document d. It is a normalised product of a prior term θ_k and a multinomial likelihood term.

EM and Mixtures of Categoricals: M-step

The M-step maximises "$\int q(H) \log P(H, V|\theta) dH$" w.r.t. parameters. Here the log joint is:

$$\begin{aligned}\log p(\mathbf{w}, \mathbf{z}|\theta, \boldsymbol{\beta}) &= \log \prod_{d=1}^{D} p(z_d|\theta) \prod_{n=1}^{N_d} p(w_{nd}|z_d, \boldsymbol{\beta}) \\ &= \sum_d \log p(z_d|\theta) + \sum_{n,d} \log p(w_{nd}|z_d, \boldsymbol{\beta})\end{aligned}$$

Taking expectations w.r.t. each of the $q(z_d)$, using $r_{kd} \stackrel{\text{def}}{=} q(z_d = k)$, we get:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{w}, \mathbf{z}|\theta, \boldsymbol{\beta}) = \sum_{d,k} r_{kd} \log p(z_d = k|\theta) + \sum_{n,d,k} r_{kd} \log p(w_{nd}|z_d = k, \boldsymbol{\beta})$$

Plugging in $\theta_k = p(z_d = k|\theta)$ and the categorical likelihood, $\prod_{m=1}^{M} \beta_{km}^{c_{md}}$:

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

M-step: Maximize $F(R, \theta, \boldsymbol{\beta})$ w.r.t. $\theta, \boldsymbol{\beta}$.

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.

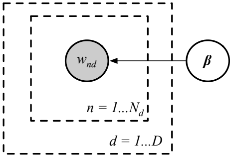
$$\hat{\theta}_k \leftarrow \text{argmax}_{\theta_k} \;\; F(R, \theta, \beta) + \lambda\Big(1 - \sum_{k'=1}^{K} \theta_{k'}\Big)$$

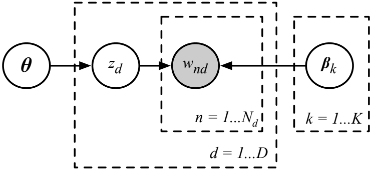
$$= \frac{\sum_{d=1}^{D} r_{kd}}{\sum_{k'=1}^{K} \sum_{d=1}^{D} r_{k'd}} = \frac{\sum_{d=1}^{D} r_{kd}}{D}$$

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

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

M-step for mixture compared to simple categorical





Recall the estimation equation for a simple single categorical model:

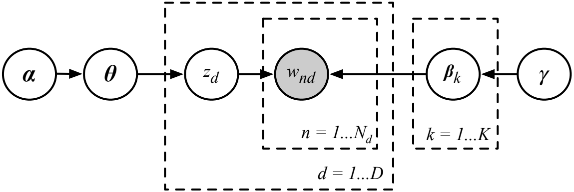
$$\hat{\beta}_m \leftarrow \frac{\sum_{d=1}^{D} c_{md}}{\sum_{m'=1}^{M} \sum_{d'=1}^{D} c_{m'd'}} = \frac{c_m}{\sum_{m'} c_{m'}} = \frac{c_m}{N}$$

Compare to the M-step for a *mixture* of categoricals:

$$\hat{\beta}_{km} \leftarrow \frac{\sum_{d=1}^{D} r_{kd}\, c_{md}}{\sum_{m'=1}^{M} \sum_{d'=1}^{D} r_{kd'}\, c_{m'd'}}$$

We see is that it's the same idea, but weighting the word counts by the responsibilities for each category.

A Bayesian mixture of categoricals model



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

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.:

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

Variational Bayesian Learning

Let the hidden latent variables be H, observed data V and the parameters θ.

We are going to generalise EM to do approximate Bayesian learning, by lower bounding the log marginal likelihood (Bayesian model evidence) using Jensen's inequality:

$$\begin{aligned}
\log P(V) &= \log \int dH\, d\theta\, P(V, H, \theta) \\
&= \log \int dH\, d\theta\, Q(H, \theta) \frac{P(V, H, \theta)}{Q(H, \theta)} \\
&\geqslant \int dH\, d\theta\, Q(H, \theta) \log \frac{P(V, H, \theta)}{Q(H, \theta)}.
\end{aligned}$$

Use a simpler, factorised approximation to $Q(H, \theta)$:

$$\begin{aligned}
\log P(V) &\geqslant \int dH\, d\theta\, Q_H(H) Q_\theta(\theta) \log \frac{P(V, H, \theta)}{Q_H(H) Q_\theta(\theta)} \\
&= \mathcal{F}(Q_H(H), Q_\theta(\theta), V).
\end{aligned}$$

Maximize this lower bound.

Variational Bayesian Learning . . .

Maximizing this lower bound, $\mathcal{F}$, leads to **EM-like** updates:

$$\begin{aligned} Q^*_H(H) &\propto \exp \langle \log P(H,V|\theta) \rangle_{Q_\theta(\theta)} && \textit{E-like step} \\ Q^*_\theta(\theta) &\propto P(\theta) \exp \langle \log P(H,V|\theta) \rangle_{Q_H(H)} && \textit{M-like step} \end{aligned}$$

Maximizing $\mathcal{F}$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $Q(\theta)Q(H)$ and the *true posterior*, $P(\theta, H|V)$.

$$\begin{aligned} \log P(V) - \mathcal{F}(Q_H(H), Q_\theta(\theta), V) &= \\ \log P(V) - \int dH\, d\theta\; Q_H(H) Q_\theta(\theta) \log \frac{P(V, H, \theta)}{Q_H(H) Q_\theta(\theta)} &= \\ \int dH\, d\theta\; Q_H(H) Q_\theta(\theta) \log \frac{Q_H(H) Q_\theta(\theta)}{P(H, \theta|V)} &= \mathrm{KL}(Q||P) \end{aligned}$$

Note that variational Bayesian learning is an alternative to MCMC.