CS4786/5786: Machine Learning for Data Science, Spring 2015
4/7/2015: Lecture 18 handout: Developing an important generative story: Latent Dirichlet allocation

---

Announcements *Tentative* upcoming workload schedule:

- Competition 1: Dataset already released. Instructions out today-ish? Tentative due date Wed April 22, 11:59pm.
- A3: Released relative soon-ish, probably due somewhere in the weeks of the 20th or the 27th.
- Competition 2: Released maybe around the weeks of the 20th or 27th, due Mon May 11th, 4:30pm.

Pedagogical goal: more practice w/ developing generative stories

LDA (Latent Dirichlet Allocation) — An important generative story

Outline:

(1) review mixture of multinomials ~~notation~~ (model, notation)

(2) new setting: "longitudinal" ~~segmented~~ data [analogy w/ documents], when same person can have different types?

(2a) choice of multinomials: from fixed set → distribution over multinomials: the Dirichlet

– the LDA story

(For handout:)

Last time: data X has n rows like x_t = 3 ($x_t[1]$, pepsi), 0 ($x_t[2]$, coke), 6 ($x_t[3]$, sprite) . . . — d possible products.

Generative story: mixture of multinomials

For a given customer, x_t (corresponding to the (transpose of the) tth row), $= (x_t[1], x_t[2], \ldots, x_t[d])^T$ — $x_t[1]$: count, like "3"

mother nature ~~they~~ picks among given customer types w/ prob $\pi[1], \ldots \pi[k]$ — k → # of customer types

A customer type j is represented by ~~the~~ d ~~parameters~~ values of ~~a multinomial~~ ϕ_j: a multinomial,

$\phi_j[1]$ (prob of picking a pepsi)

$\phi_j[2]$ (prob of picking a coke)

$\vdots$

$\phi_j[d]$

ex: an anything-but-coke type might have $\phi[1]=.6$, $\phi[2]=.05$, $\phi[3]=.35$.

$\sum_{\ell=1}^{d} \phi_j[\ell] = 1$, all $\phi_j[\ell] \geq 0$.

We let c_t = the type that customer t picked.

x_t then picks their m purchases according to the ϕ_j they picked.

I. Clicker question: notation review What is the (possibly unnormalized) probability of a single $\mathbf{x}_t$ according to our model, assuming we knew the hidden π, ϕ_js, and c_ts?

$$
\begin{aligned}
&(A) \qquad \pi[c_t]\frac{m!}{\mathbf{x}_t[1]!\mathbf{x}_t[2]!\cdots\mathbf{x}_t[d]!}\prod_{\ell=1}^{d}\phi_{c_t}[\ell]^{\mathbf{x}_t[\ell]} \\
&(B) \qquad \pi[c_t]\frac{m!}{\phi[1]!\phi[2]!\cdots\phi[d]!}\prod_{\ell=1}^{d}\mathbf{x}_{c_t}[\ell]^{\phi[\ell]} \\
&(C) \qquad \mathbf{x}_t[c_t]\frac{m!}{\phi[1]!\phi[2]!\cdots\phi[d]!}\prod_{\ell=1}^{d}\pi_{c_t}[\ell]^{\phi[\ell]} \\
&(D) \qquad \phi[c_t]\frac{m!}{\mathbf{x}_t[1]!\mathbf{x}_t[2]!\cdots\mathbf{x}_t[d]!}\prod_{\ell=1}^{d}\pi_{c_t}[\ell]^{\mathbf{x}_t[\ell]}
\end{aligned}
$$

II. Clicker question: new likelihood What is the likelihood of a single $\mathbf{x}_t$ under our second model (after mixture of multinomials, before full-blown each customer has different preference profile)?

$$
\begin{aligned}
&(A) \qquad \prod_{q=1}^{m}\phi[c_t[q]]\pi_{c_t[q]}[\mathbf{x}_t[q]] \\
&(B) \qquad \prod_{q=1}^{m}\pi[c_t[q]]\phi_{c_t[q]}[\mathbf{x}_t[q]]
\end{aligned}
$$

III. From Percy Liang and Dan Klein's 2007 tutorial, Structured Bayesian Nonparametric Models with Variational Inference (on next page)

Dirichlet distribution

A Dirichlet is specified by concentration parameters:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K),\ \alpha_z \geq 0$$

$$\text{Mean: } \left(\frac{\alpha_1}{\sum_z \alpha_z}, \ldots, \frac{\alpha_n}{\sum_z \alpha_z}\right)$$

Variance: larger $\boldsymbol{\alpha}$s $\rightarrow$ smaller variance

A Dirichlet draw $\boldsymbol{\phi}$ is written $\boldsymbol{\phi} \sim \text{Dirichlet}(\boldsymbol{\alpha})$,
which means $p(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) \propto {\phi_1}^{\alpha_1 - 1} \cdots {\phi_K}^{\alpha_K - 1}$

The Dirichlet distribution assigns probability mass to multinomials that are likely to yield pseudocounts $(\alpha_1 - 1, \ldots, \alpha_K - 1)$.

$$\text{Mode: } \left(\frac{\alpha_1 - 1}{\sum_z (\alpha_z - 1)}, \ldots, \frac{\alpha_n - 1}{\sum_z (\alpha_z - 1)}\right)$$

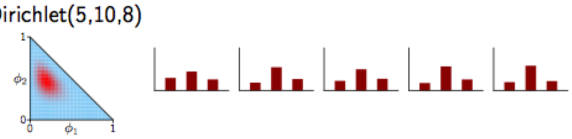
The full expression for the density of a Dirichlet is $p(\boldsymbol{\phi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{z=1}^K \alpha_z)}{\prod_{z=1}^K \Gamma(\alpha_z)} \prod_{z=1}^K {\phi_z}^{\alpha_z}$. Note that unlike the Gaussian, the mean and mode of the Dirichlet are distinct. This leads to a small discrepancy between concentration parameters and pseudocounts: concentration parameters $\boldsymbol{\alpha}$ correspond to pseudocounts $\boldsymbol{\alpha} - 1$.

Draws from Dirichlet distributions

Dirichlet(.5,.5,.5)

Dirichlet(1,1,1)

Dirichlet(5,10,8)



A Dirichlet$(1, 1, 1)$ is a uniform distribution over multinomial parameters. As the concentration parameters increase, the uncertainty over parameters decreases. Going in the other direction, concentration parameters near zero encourage sparsity, placing probability mass in the corners of the simplex. This sparsity property is the key to the Dirichlet process.