Announcements: timings on competition; assignments

Lec 18, 4/7/15
- LDA (Latent Dirichlet Allocation) - An important generative story

Pedagogical goal: more practice w/ developing generative stories

Outline:

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

② new setting: "~~segmented~~ longitudinal" data [analogy w/ documents], where 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 ($x_t[1]$), 0 ($x_t[2]$), 6 ($x_t[3]$) . . . (pepsi, coke, sprite) — $\underbrace{\qquad}_{d \text{ possible products.}}$

Generative story: mixture of multinomials

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

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

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

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

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

$\vdots$

$\varphi_j[d]$

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

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

We let c_t = the type that customer t picked.

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

The prob (log) of x_t $\left(\pi[c_t] \left[\frac{m!}{x_t[1]!\, x_t[2]! \cdots x_t[d]!}\right] \prod_{\ell=1}^{d} \varphi_{c_t}[\ell]^{x_t[\ell]}\right)$

make this a clicker q: swap φ and π; swap φ and x; swap K and π; don't swap anything

Task: given X, recover the best: $\boxed{\pi}$ (k unknowns, $k-1$ degrees of freedom) — write as $\pi[1] \ldots \pi[k]$

$\boxed{\varphi_j}$, $1 \leq j \leq k$ (each has d unknowns, $d-1$ d.o.f.) — write as $\varphi_j[1], \ldots, \varphi_j[d]$

c_t (n unknowns)

ex: "20% of my population are sprite fans, meaning type (0.05, 0.05, .9)"

$\vdots$

~~that's not a~~

- our mixture-of-multinomials story is a pretty good one: (1) seems reasonable
  (2) we can use EM to ~~recover guess the~~ choose "good" values of our unknown parameters. (so, not that computationally "complicated")

  hidden vars: $\boxed{\pi}$, the $\boxed{\varphi_j}$s, and the c_t's: which type customer t picks ~~hidden vars: π, φ, c_t~~

- new scenario: x_t represents ~~purchases over~~ customer t's purchases over an extended time period ... (so as to get independence)
  - ~~tastes/or~~ needs (or tastes) can change.
    - one shopping trip, they need softdrinks
    - but on a different trip, they want cereal.
    - and on a third trip, they're choosing fruit.
  - => one φ_j isn't just a good fit, convenient ~~[illegible]~~.

  we probably don't want to have one multinomial for:

  people who hate-coke AND love Froot Loops AND love either bananas or oranges.

  - instead of one φ_j for = coke-haters who love Froot Loops and either bananas or oranges:

    have a φ for: when picking softdrinks, anything but coke (~~but occasionally~~)

    plus ~~plus~~ a diff φ for: when picking cereal, favorite is Froot Loops, etc.

=> re-select a (possibly different) φ for each purchase customer t makes.
- ~~have a~~ index on sequence in which purchases were made.

...nomials idea?
w.l.l. d fn, same # of ~~para~~

in the likelihood.

~~[illegible]: what the q-th purchase is is hidden structure, but it's [illegible] to parameters~~

so: data matrix is now:

$$X: \quad n\left[\; - \; x_t^T \; - \;\right]$$

$x_t[q]$ entry: product type bought as q^{th} purchase ($\in \{1, \dots, d\}$)

m ↳ each customer makes exactly m purchases (for simplicity)

task: recover values for:

hidden variables: the K different preference types φ_j (d-dimensional)

(– almost the same story) the prob π of picking preference types (K-dimensional)

the choices $c_t[q]$ ~~that customer t was~~ of which pref. type customer t was using to pick their q^{th} purchase. (c_t is m-dimensional) entries run from 1 to K.

definitely want to learn: what kind of preferences are there, and how likely is each of these preferences.

might want to know what 'frame of mind' (need) the user was in when making those purchases.

clicker q this:

the likelihood for vector x_t now looks something like this: (given the hidden variables)

$$\prod_{q=1}^{m} \boxed{\pi[c_t[q]]}\; \varphi_{c_t[q]}\left[x_t[q]\right]$$

<ignoring normalization>

(altho' if we wanted to marginalize over all the possible $c_t[q]$ (assignments), we'd get:

$$\prod_{q=1}^{m} \sum_j \boxed{\pi[j]}\; \varphi_j\left[x_t[q]\right]$$

)

– hidden structure makes likelihood simpler.

~~• We have new hidden variables $c_t[q]$ = what j s.t. the qth purchase by customer t was made according to customer preference type φ_j?~~

• how should we model this choice?

~~We'd like to say, this person is mostly buys pepsi, but sometimes buys bananas or oranges.~~

We'd like to say, this person (Prof Snidlow) is mostly a pepsi-buyer, but sometimes is a bananas-or-oranges buyer.

That is, their prob of being a pepsi-buyer is higher than their prob of being a bananas-or-oranges buyer.

But this other person, say, prof Lee, never buys softdrinks @ all, and is only a bananas-or-oranges buyer.

~~Further~~

Another design choice: each customer has their own probabilities of choosing among the φ_j for each purchase.

ex: One shops for softdrinks much more often than for fruit ~~each purchase~~.

another shops for softdrinks with prob 50%, cereal w/ prob 50% ~~at each purchase~~

$\Rightarrow$ instead of a fixed $\boxed{\pi}[j]$, we have a $\boxed{\pi_t}$ for each customer t's probability,

$\boxed{\pi_t[j]}$, $j \in \{1, \ldots, K\}$ = prob of picking preference-type φ_j.
customer t

Think of $\boxed{\pi_t}$ as user's profile over pref types φ_j.

But we've now increased our # of hidden variables significantly

(just say, don't need on the board).

task so far:
Given X, recover: the various preference types φ_j (k d-dimensional vectors)
the various user profiles π_t (n k-dim vectors)
the preferences responsible for each purchase $c_t[q]$ (n m-dim vectors) — we've fixed the # of purchases each user makes.
each in $\{1, \ldots, K\}$

$n \begin{bmatrix} \\ \\ \end{bmatrix}$ with m columns

skip

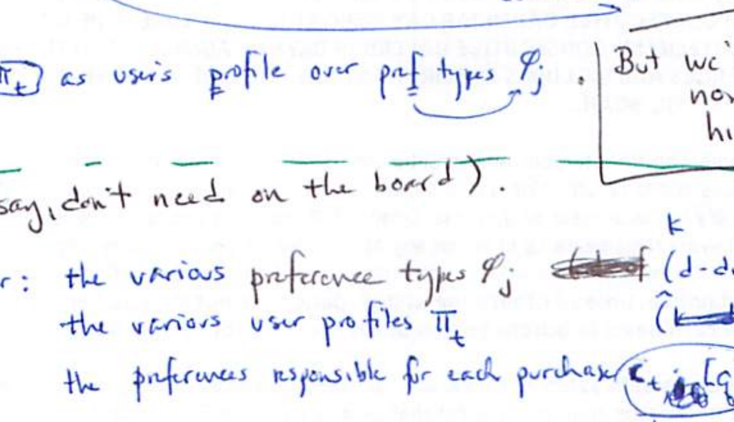
SKIP

~~This is a lot of unknowns.~~

should we skip this? It kind of interrupts the flow

A (graphical-model-like) summary:

shading = observed.



in corner = # of "versions"..

So, if we've significantly expanded the search space of hidden variables.
('for instance, we now have roughly $n \times k$ π values to fill in),
we might expect that even as powerful a tool as EM might get in trouble
(maybe too many stationary points or local maxima),

that is: greater model expressivity, @ cost of making search for hidden values harder.

- this kind of tradeoff is an important factor in developing your own generative stories.

What have we done in other such situations, when things seem complex or even impossible?

- in EM, we guessed what the right distribution would be
- when clustering proved impossible, we introduced constraints.

→ assume some prior info is given

- so, if we could a priori restrict the space of π_t's to search over, or know that certain π_t's are more likely, this would make things easier for us.

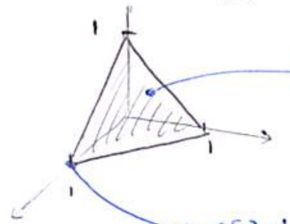
what prior can we put over multinomials π_t?
<at least one student flipped the handout over to look>
The probabilists & statisticians know a conjugate prior <you don't have to know what that is> for multinomials:

the Dirichlet prior - we've given you some ~~intuitio~~ geometric intuitions on the handout.

About the Dirichlet distribution

Each φ can be thought of as a point on the probability simplex in $\mathbb{R}^d$. (i.e, vectors φ s.t. $\sum_{\ell=1}^{d} \varphi[\ell] = 1$, $\varphi[\ell] \geq 0$.)

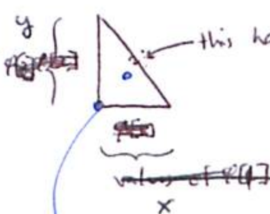
Ex: $d=3$



$\varphi[1] = .2$
$\varphi[2] = .4$
$\varphi[3] = .4$

$\varphi[1] = 1$
$\varphi[2] = 0$
$\varphi[3] = 0$

Given that $\varphi[1] = 1 - (\varphi[2] + \varphi[3])$, we can "project" this simplex into 2-d by looking at it from the z-axis: (put your eye right there ⊙), like you're sighting down a pool cue!



this has $\varphi[1] = .4$, $\varphi[2] = .4$, so $\varphi[3] = .2$ by arithmetic

$\varphi[1] = \varphi[2] = 0$, so $\varphi[3] = 1$ by arithmetic

---

<show 2nd slide 1st>:

$\alpha = (.5, .5, .5)$

heat map. bottom-right corner = multinomials that put most of their probability mass on the first product.
The "redness" though = this Dirichlet prior puts high prob on those kinds of multinomials.

- the Dirichlet (1,1,1) is a uniform prior on multinomials.

- the Dirichlet (5, 10, 8) isn't "symmetric" like the other two, and you see this in the fact that "the red splotch" isn't "symmetric".

(q: why are (.5, .5, .5) and (1, 1, 1) different?

<then do 1st slide; go over it>