Pedagogical goal: more practice w/ developing generative stories

Outline:

1. review mixture of multinomials notation (model, notation)
   longitudinal
2. new setting: "streamed" data (analogy w/ documents), when same person can have different types?
   choice of multinomials:
   from fixed set → distribution over multinomials: the Dirichlet

- the LDA story

(For handout)

Last time: data X has raws like:

\[ x \begin{bmatrix} 3 & 0 & 6 \end{bmatrix} \]

Generative story: mixture of multinomials

For a given customer, \( x_t \) (corresponding to the \( t \)-th row):

\[ x_t = [x_{t[1]}, x_{t[2]}, \ldots, x_{t[d]}] \]

picks among given customer types w/ prob

\[ \pi[1], \ldots, \pi[d] \]

\# of customer types

A customer type is represented by parameters of a multinomial:

\[ \phi_{j}^{[1]} \quad \text{(prob of picking a pepsi)} \]
\[ \phi_{j}^{[2]} \quad \text{(prob of picking a coke)} \]
\[ \vdots \]

\[ \phi_{j}[d] \]

Ex: an anything-but-coke

might have

\[ \phi[1] = .6, \phi[2] = .05, \]
\[ \phi[3] = .35. \]

We let \( e_t \) the type that customer \( t \) picked.

\( x_t \) then picks their \( m \) purchases according to the \( \phi_{j}^{[e]} \) they picked:

The prob of \( x_t \):

\[ \prod_{i=1}^{d} \pi_i \prod_{j=1}^{d} \phi_{j}^{[e_i]} \]

Task: given X, recover the best:

\[ \prod (k \text{ unknowns}, k-1 \text{ degrees of freedom}) \]

\[ \phi_{j}^{[k]} \quad \text{each has d unknowns, d-1 d.o.f.} \]

\[ e_t \quad \text{(a unknown meaning)} \]

ex: "20% of my population are sports fans, type (0.05, 0.05, .9)"
our mixture of multinomials story is a pretty good one. seems reasonable.

we can use EM to choose "good" values of our unknown parameters. so, not that computationally complicated.

new scenario: x_i represents purchase t's purchases over an extended time period... (so as to get independence)

- tastes/needs can change.

⇒ one p_j isn't a good fit.

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

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

- instead of one p_j for coke-haters, we have:

  - p_j: when picking softdrinks, anything but coke
  - occasionally

  - plus a diff p_j for when picking cereal, favorite is Froot Loops, etc.

⇒ re-select a (possibly different) p_j for each purchase t makes.

⇒ use a metric on sequences in which purchases were made.
so: data matrix is now:
\[ X: \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \]

- \( n \) - product type bought as \( q \)th purchase (e \{1, \ldots, d\})

\[ x_{t[q]} \]

\[ m \] - each customer makes exactly \( m \) purchases (for simplicity)

- task: recover values for:
  - hidden variables: the \( K \) different preference types \( \phi \) (d-dimensional)
  - the prob \( \phi \) of picking preference types (K-dimensional)
  - the choices \( \phi \) of the customer \( t \) of which pref. type customer \( t \) was using to pick their \( q \)th purchase. (\( \phi \) is m-dimensional)

\( X \) entries run from 1 to \( m \).

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 these purchases.

given the hidden variables

the likelihood for vector \( x_t \) now looks something like this:

\[
\prod_{q=1}^{m} \left[ \prod_{\phi \in \{1, \ldots, K\}} \phi_{\phi} \left[ x_t[q] \right] \right] \quad \text{<ignoring normalization>}
\]

altho' if we wanted to marginalize over all the possible \( c_t[q] \) (assignments),

we'd get:

\[
\prod_{q=1}^{m} \sum_{\phi \in \{1, \ldots, K\}} \phi_{\phi} \left[ x_t[q] \right]
\]

- hidden structure makes likelihood simpler.
how should we model this choice? 

We'd like to say this person is mostly a pepsi-bayer, but sometimes is a bananas-or-anges buyer.

That is, their prob of 

being a pepsi-buyer 

is higher than their prob of 

being a banana-or-anges buyer.

But this person, say, never buys softdrinks at all, and is only a 

banana-or-anges buyer.

Another design choice: each customer has their own probabilities of choosing among the 

$\mathcal{F}$ for each purchase:

- one prefers much more often than when
- another prefers slightly more

$\mathcal{F}$ with prob 50%, $\mathcal{W}$ with prob 50% at each purchase

Instead of a fixed $\mathcal{F}(\rho)$ for each customer $t$, we have a set $\mathcal{F}_t$ of preference types $\rho$ for each customer $t$.

Think of $\pi_t$ as user's profile over preference types $\rho$.

Task so far: (just say, don't need on the board)

Given $X$, recover:

- the various preference types $\rho$.
- the various user profiles $\pi_t$.
- the preferences responsible for each purchase $\mathcal{F}(\rho)$.

We've fixed the # of preference types each user makes.

A (graphical-model-like) summary:

\[ \text{shading} = \text{observed} \]
So, if we've significantly expanded the search space of hidden variables, 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 trade-off 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 \( \pi_k \)'s to search over, or know that certain \( \pi \)'s are more likely, this would make things easier for us.

What prior can we put over multinomials \( \pi_k \)?

<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're given you some intuitions on the handout.
About the Dirichlet distribution

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

Ex: \( d=3 \)

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

Value of \( \varphi[3] \):

\[
\varphi[1] = 0.4, \quad \varphi[2] = 0.4, \quad \text{so} \quad \varphi[3] = 0.2 \quad \text{by arithmetic}
\]

\( \varphi[1] = \varphi[2] = 0, \quad \text{so} \quad \varphi[3] = 1 \quad \text{by arithmetic} \)

(show 2nd slide 1st):

heat map. Bottom-right corner = multinomials that put most of their probability mass on the first product. The "redness" that is 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?)

<than do 2nd slide: go over it>