Lec. 17, 3/27/1

- More generative stories (and hence more EM) need for not?
- Easing into graphical models.

Announcements:

- Yovenn Paris' Friday 3/28 hours moved to today 8-9pm, location T3A.
- All spring break office hours are cancelled, but:
  - Mevlana Gemici will hold his Sunday hours if you request by email with 24 hours notice
  - Jack Hessel will hold his Monday hours if you request by email with 24 hours notice

Coursework scheduling updates:

- Givin time constraints, we are almost surely only going to assign one more assignment. As for the "competition"...
- The first "competition" dataset is still being tested and documented. We hope to release the dataset before Spring Break, so students can take a look if they want...

  ...but not release instructions, so that there's no pressure to do anything over Spring Break.
- The due date for the second competition is, as determined by the registrar, May 11th, 4:30 p.m.

A2 pedagogical note (see also Prof Sridharan's email):

What we want you to learn from this assignment is:

NOT that bizarre, degenerate, unrealistic datasets reveal instability for clustering algorithms,

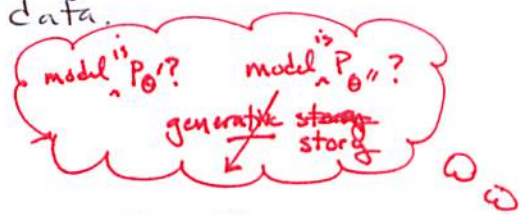
BUT ~~that what kinds of~~

Skip

So far: we've moved from ~~just~~ considering our data as "just given" (projection, clustering)

to thinking (imagining) that the data is generated by some (unknown member of) a family of probabilistic process

and regarding our task as ~~picking~~ selecting a model based on the data.



in your head

The model P_{θ^*} you pick is your "dimension-reduced" understanding of the data.

"These 10K datapoints are really just examples of 3 species of trees, where the oak trees ~~look~~ are usually on the east coast, the orange trees tend to concentrate in Florida".

* how to pick θ' vs. θ''?

maximum likelihood principle
maximum a posteriori principle
} either requires max'ing

~~which could be hard~~.
often hard to compute directly

EM often comes to the (11) rescue"

~~quotes b/c it often is only going to get you~~
quotes b/c it may only get you to a stationary point.

~~* very very general and powerful~~
* very general technique:
We're teaching it to you b/c it's important.
when your model's hidden structure ~~makes it easy~~ makes ~~taking derivatives~~ solving "derivative = 0" easy if only it weren't hidden

~~EM~~

~~let's demonstrate that EM work~~

let's further exemplify EM by applying it to another set of models.

Gaussian mixture models: apply to "numerical, spatially concentrated" data.

But there are lots of kinds of data that they don't apply to.

~~What about category~~

Contrasting example: purchase data. for market segmentation.

Mixture of multinomials

~~customer~~ ~~#bags of chips~~ ~~#boxes bars~~

product1 product2 ...

customer: 3 0 6

~~bottles of~~ pepsi | bottles of coke | ~~bottles of~~ spnk.

(don't normalize now - will confuse w/ the φ_i's)

(~~typically~~ ~~might be~~ ~~let's assume normalized~~)

~~these kinds of~~

this kind of 'count' data just doesn't look like Gaussian-generated.

You could go ahead & try to fit a GMM to it, but why not pick a more appropriate generative story?

↓

can't buy 6.2 bottles

notation for multinomials: {choice over D options}:

φ_j : the set of D parameters for the j^{th} multinomial.

↓

ex: ~~pepsi fans~~. anything-but-pepsi

$\varphi_j[\ell]$: prob of picking ~~the~~ product ℓ.

ex: for ~~pepsi fans~~: pepsi enemies:

$\varphi_j[1] = 0.05$ ⟶ desert island scenario

$\varphi_j[2] = .6$

$\varphi_j[3] = .35$ | ⟶ ~~regular like~~ pick competing product.

✓ ~~generative story~~:

[there are K types of customers.
assume each customer makes exactly m purchases.] → probably not an essential assumption

(I don't want to deal w/ ~~proper~~ normalization or the whole multinomial coeff stuff.

pick the customer type j according to π, distribution over types.

- pick the ~~###~~ m purchases by the distribution given by φ_j.

Task: given X, pick a "best" π, $\varphi_1, \ldots, \varphi_K$

↳ each is $\varphi_i[1], \ldots \varphi_i[d]$, ...

hope you might discover things like 'oh, 20% of my customer base is people who buy anything but pepsi'.

ML approach: find params maxing (log) likelihood of data:

Recall: for a given φ_i and a given $x_t = x_t[1], x_t[2], \ldots, x_t[d]$

$$P(x_t) = \frac{m!}{x_t[1]!\, x_t[2]! \cdots x_t[d]!} \cdot \varphi_i[1]^{x_t[1]} \cdot \varphi_i[2]^{x_t[2]} \cdots \varphi_i[d]^{x_t[d]}$$

($x_t[d]$: count; $\varphi_i[d]$: prob)

~~AHEE~~ ~~P(X)~~

$$\text{loglik}(X) = \log\left(\prod_t P(x_t \mid \theta)\right)$$

(θ → our π_i's, φ_i's)

$$= \sum_t \log P(x_t \mid \theta)$$

$$= \langle \text{generative story} \rangle$$

(use □ box to not confuse w/ product)

$$\sum_t \log\left(\sum_j \boxed{\pi_j} \cdot \frac{m!}{x_t[1]! \cdots x_t[d]!} \cdot \prod_{\ell=1}^{d} \varphi_j[\ell]^{x_t[\ell]}\right)$$

(could be any customer type, so sum over them all.)

... ~~ready~~

... ready to take some derivatives & and set to 0?

clicker question: how many derivatives must we ~~derive those to~~ compute?

(6) 29% (A) K (one for each φ_i) + K (for each $\boxed{\pi_i}$)

(10) 48% (B) K × ℓ (one for each $\varphi_i[\ell]$) + K (one for each $\boxed{\pi_i}$)

(2) 10% ~~(C) almost A, but off concave~~

(C) my answer is almost but not quite (A)

(3) 14% (D) my answer is almost but not quite (B).

↳ it's this one, but I only need $K \times (\ell - 1) + (K - 1)$

since $\sum_{\ell=1}^{d} \varphi_j[\ell] = 1$.

(I pick something w/ prob 1).

... now, if we try to take the derivative $\frac{\partial}{\partial \varphi_{j'}[\ell']}$

↳ enforce w/ Lagrange multiplier λ_i or just directly normalize by sum (this entire Pset's theme of Spark).

clicker q:

$$\sum_t \frac{\boxed{\pi_j} \prod_{\ell \neq \ell'} \varphi_j[\ell]^{x_t[\ell]} \cdot x_t[\ell'] \varphi_j[\ell']^{x_t[\ell']-1}}{\sum_j \boxed{\pi_j} \prod_\ell \varphi_j[\ell]^{x_t[\ell]}}$$

(all constant wrt j', ℓ')

OMG $\varphi_{j,\ell}[\ell^0]$ are everywhere, ~~exist~~

facts: $\frac{d}{dy} \log(zy^n) = \frac{(n-1)zy^{n-1}}{zy^n} \cdot y$. if $\ell \neq \ell'$ and $j \neq j'$, then $\varphi_j[\ell]$ looks like a constant to $\frac{\partial}{\partial \varphi_{j'}[\ell']}$

clicker question: ~~do we need EM?~~ should we use EM, instead of directly solving for the max. likelihood soln?

(a) 43% (A) Yes, because we can ~~get the equation~~ take the derivatives but can't solve for $\phi_j[l']$ independently of the other params (i.e., just in terms of the data)

(3) 14% (B) Yes, because we can't take derivatives of we need

(6) 29% (C) No, because we can take the derivatives and solve for $\phi_j[l']$

14% (D) No, because there's no hidden structure for EM to take advantage of.

---

so, what's the problem?

- sum over ~~i~~ i in denom: { keeping all the ϕ_i's in the equation wouldn't be necessary if we knew ~~that~~ that x_t "actually is of type j" (no longer a mixture prob) (unknown)

the c_t = which type x_t is. (e.g. "anti-pepsi").

EM idea: ~~hallucinate initial parameters~~

~~if~~ we don't know the c_t's, ... so just average (take expectation) over all their possible values

~~but if we can't know their distribution,~~

we can take the expectation over possible c_t values:

$\Rightarrow$ ~~instead of~~ find $\frac{\partial \ell}{\partial \phi_{j'}[l']}$ maxing $\sum_t \sum_{k=1}^{E} p(c_t = k)$ "γ" $\log P(x_t, c_t = k \mid \phi_j)$

for $k = j'$, $\log \frac{m!}{\prod_{l=1}^{E} x_t[l]!} \left(\prod_{l=1}^{E} \phi_{j'}[l]^{x_t[l]} \right) \cdot \phi_{j'}^{x_t[l']}$

↓ constant wrt $\phi_j[l']$

this was our $Q_t(c_t)$ from last ~~time~~ lecture.

↳ Q = "like a P".

~~get it~~ $= P(c_t = k \mid$ some parameters, $x_t)$

↑ use an initial guess!

where do we get $P(c_t = k \mid$ old params, $x_t)$? Our previous guess of $\pi[k]$!)

EM: just keep imagining what would be useful, and guess when you run out of other possibilities!

$$\frac{\partial}{\partial \varphi_j[\ell']}\left[\log Q_t(c_t) + \log\left(\text{big constant} \cdot \varphi_j[\ell']^{x_t[\ell']}\right)\right]$$

$$= \frac{Q_t(c_t)}{\varphi_j[\ell']}\,\frac{x_t[\ell']}{\varphi_j[\ell']} \qquad \text{set to } 0 \Rightarrow \varphi_j[\ell'] \cdot \propto x_t[\ell'] \cdot Q_t(c_t).$$

↑ how many times do we see ℓ' being used.