Lec 23
4/23/15

parameter learning – the case of HMMs

So far: given a Bayesian Network (including the associated ~~condition~~ conditional distributions), how can we do inference

parameters $\Theta$ =

ex: compute $P($~~some state s~~ state $S$ has value $j$, state $S'$ has value $l$ )

Today: learning the parameters from data

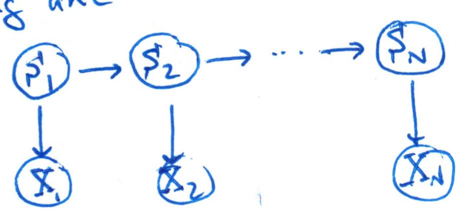
Focus: EM on HMMs. – ~~from last time~~ we've: HMMs are trees, ~~so inference is ea~~ and "half" of their nodes are leaves
⇒ inference is "easy" (compared to arbitrary BNs).

prep for A3

~~Setting & notation~~

– also: HMMs are nice models for "time series" (sequential data)

## Setting and notation



} unobserved. Each $S_t$ takes values from $1 \ldots K$
"state @ time $t$"

} observed

our usual notation for "hidden compressed dimensionality

example: (instead of modeling a student as in the Koller/Friedman BN) now going to model a lecturer

~~lecture~~ states ~~lecture state~~ values: ~~at time t~~ "announcement" $=1$, "important content" $=2$, ~~"aside"~~, "optional content" $=3$, "joke" $=4$.
(let's not say "unimportant")

~~obs~~
observed: relative loudness, modeled as a Gaussian with (hidden) mean & variance
– we do relative loudness so that $\Theta$ #s are OK.

~~so, parameter~~

parameters $\Theta$ =

Trans$(j \leftarrow i)$ = prob, given in state $i$, that next state is $j$ (joke ← important content). Notation = like cond prob

Out$(x \uparrow i)$ = prob, given in state $i$, that volume is $x \in \mathbb{R}$
notation: like cond prob.
controlled by $\mu_i$ & $\sigma_i^2$, mean & variance for state $i$
(very loud ↑ joke)

Start$(i)$ = prob the first state is $i$.

~~(for all of these, the sum~~
Task: you weren't listening to the words, just her volume. Can you infer her model?

Once you've got a model, you can infer things ~~like eyeing~~ b/c it would be convenient to do so for future lectures like:

~~given~~ I've been kind of zoned out in lecture until know, ~~what's the~~ but attuned to the volume: what's the prob that she's about to say sthg important ; I should tune in?

or, what's the prob that she just made a joke so I should humor her ; laugh?

<~~no end~~

---

$1^{st}$ attempt: find $\theta^*$ maximizing log-likelihood of observeds $x_1 \ldots x_N = \vec{x}$

$$\theta^* = \arg\max_{\substack{Trans(\cdot\cdot)\\ \mu_1..\mu_k\\ \sigma_1^2..\sigma_k^2\\ Start(1),..,Start(k)}} \left[\log\left(\sum_{N\text{-length paths } pa = i_1, i_2, \ldots, i_N} Start(i_1)\, Out(x_1|i_1) \prod_{t=2}^{N} Trans(i_t \leftarrow i_{t-1})\, Out(x_t|i_t)\right)\right]$$

$-$ lagrange constraints i.e., $\lambda_i \sum_j Trans(j \leftarrow i) = 1 \quad \forall i$

can the students work this out themselves? clicker-ize? ✓ (maybe in words) instead of the whole formula!

Most students had significant trouble: they could not get started. Karthik afterwards suggested: tell them: $\prod_n P(n \mid parents(n))$ <that seems like a good idea>.

<he also noted that maybe the notation shift was hard>

So, what would we do?

- Take the ~~Keep~~ partial derivative w.r.t., say, $Trans(4 \leftarrow 3)$, set to 0, and solve, ...

$$\frac{1}{(\text{giant sum w/ all of } \theta \text{ in it})} \; \frac{\partial}{\partial Trans(4 \leftarrow 3)} \sum_{pa} (\text{monomial in } Trans(4 \leftarrow 3))$$

$$\frac{\partial}{\partial Trans(4 \leftarrow 3)} (Const)\, Trans(4 \leftarrow 3)^m$$

$$(const)\, m\, Trans(4 \leftarrow 3)^{m-1}$$

so that's easy

$$= \frac{(const)\, m\, Trans^m}{Trans}$$

that's bad.

this is impossible: we want to solve for Trans independent of all the other variables

$-\lambda_i$

But thinking back over techniques that have helped us before:

EM: when guessing the hidden structure would make life easier.

~~takeout~~

log-~~likelihood~~ $P_\theta(\vec{x}: x_1, \dots, x_N,\ \text{pa}: \cancel{i_1, \dots, i_N} \mid \theta)$

~~add in~~ hallucinate the hidden structure

b/c that would sure make things easier

$$\sum_{\text{pa}} P(\text{pa}: \cancel{i_1, \dots, i_N} \mid \vec{x}: x_1, \dots, x_N, \theta^{old})$$

since I don't know the hidden structure I'll take an expectation over all the possibilities

But how am I spsd to know what this distribution is?

$\theta^{old}$ - ~~gues~~ use a previous guess!

Is taking the partial derivative of this going to be easier?

this portion is constant in Trans(2←1). (it has terms like $\text{Trans}^{old}(2\leftarrow 1)$, but so what ☺)

So! $$\frac{\partial}{\partial\, \text{Trans}(4\leftarrow 3)} \left[ \sum_{\text{pa}} P(\text{pa} \mid \vec{x}, \theta^{old}) \log P(\text{pa}, \vec{x} \mid \theta) - \lambda_1 \left( \sum_j \text{Trans}(j\leftarrow 3) - 1 \right) - \text{(stuff irrelevant to Trans}(4\leftarrow 3)) \right]$$

this we saw

$$= \left[ \sum_{\text{pa}} P(\text{pa} \mid \vec{x}, \theta^{old}) \cdot \frac{\frac{\partial}{\partial\, \text{Trans}(4\leftarrow 3)} P(\text{pa}, \vec{x} \mid \theta)}{P(\text{pa}, \vec{x} \mid \theta)} \right] - \lambda_1$$

= the great thing about monomials, as seen above:

$$= \left[ \sum_{\text{pa}} P(\text{pa} \mid \vec{x}, \theta^{old}) \cdot \frac{1}{P(\text{pa}, \vec{x} \mid \theta)} \cdot \frac{P(\text{pa}, \vec{x} \mid \theta) \cdot \#(4\leftarrow 3 \text{ in pa})}{\text{Trans}(4\leftarrow 3)} \right] - \lambda_1 .$$

Set to 0, solve:

$$\text{Trans}(4\leftarrow 3) = \frac{1}{\lambda_1} \sum_{\text{pa}} P(\text{pa} \mid \vec{x}, \theta^{old}) \cdot \#(4\leftarrow 3 \text{ in pa}) .$$

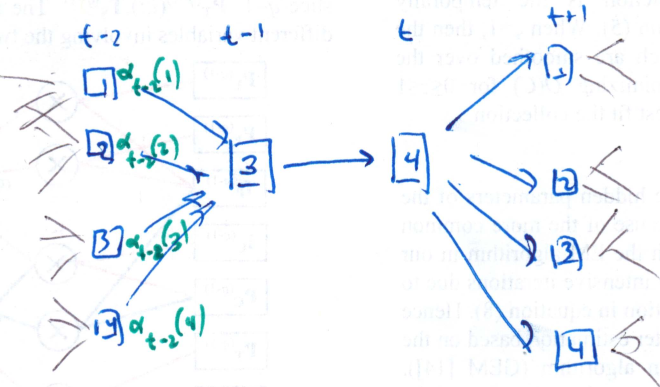
normalizing constant, which is expected # of times you'd see j←i which makes sense!

this is, though not impossible, still bad

= expectation over all paths of ~~$S_{t-1}=3, S_t=4$~~ ~~wrt $\theta^{old}$~~. Now, b/c these are HMMs:

which we can ~~reconcep~~ reconceptualize as this:

$$\frac{1}{P(\vec{x},\theta^{old})} \sum_{t=2}^{N} \underbrace{P(x_1,\dots,x_{t-1}, S_{t-1}=3 \mid \theta^{old})}_{\text{called this "Sridharan's } \alpha\text{"}} \; Trans^{old}(4\leftarrow 3)\, Out^{old}(x_t \mid 4) \cdot \underbrace{P(x_{t+1},\dots,x_N \mid S_t=4)}_{\text{called this "Sridharan's } \beta\text{"}}$$




Now this certainly looks better, b/c we syntactically got rid of the sum over an exponential # of things.

But not so fast, you say – it's just hidden in the notation, right?

Ah, but here's the saving grace...

we've showed how to compute something like this in previous lectures!

Or, you can look @ the trellis graph.

notice that if we call $P(x_1,\dots,x_{t-1}, S_{t-1}=3 \mid \theta^{old}) \doteq \alpha_{t-1}(3)$

then $\alpha_{t-1}(3) = \sum_{i=1}^{k}$ ~~$P(x_1,\dots,x_{t-2}, S_{t-2}=i)\, Trans^{old}(3\leftarrow i)$~~

$$Trans^{old}(3\leftarrow i) \cdot Out^{old}(x_{t-1} \mid 3) \cdot \underbrace{P(x_1,\dots,x_{t-2}, S_{t-2}=i)}_{\alpha_{t-2}(i)}$$

alternately present as more wishful thinking

if we knew $P(x_1,\dots,x_{t-2}, S_{t-2}=i)$ for each $i$...

dynamic-program your way ~~to compute~~ left-to-right to compute an $\alpha$ per node.

polytime computation (poly # of nodes, linear work/node).

– and that's why we cared about those $\alpha$'s.

What about this? You also saw how to do this marginal in the last lecture Prof Sridharan did.

However, does the interior look familiar...?
that's like our inference stuff!

So, if only we could get rid of the need to sum over an exponential # of things...

Flashback: we've seen how, in HMMs, to magically get an ~~exponential~~ sum over exponential # of things to disappear.

– Prof Sridharan's lecture:

$P(s_t^{\vec{x}} = i \mid \theta, \theta^{\text{known}})$ = , mathematically sum over all paths pa for $\vec{x}$ of prob that pa has $i$ as $t$th state.

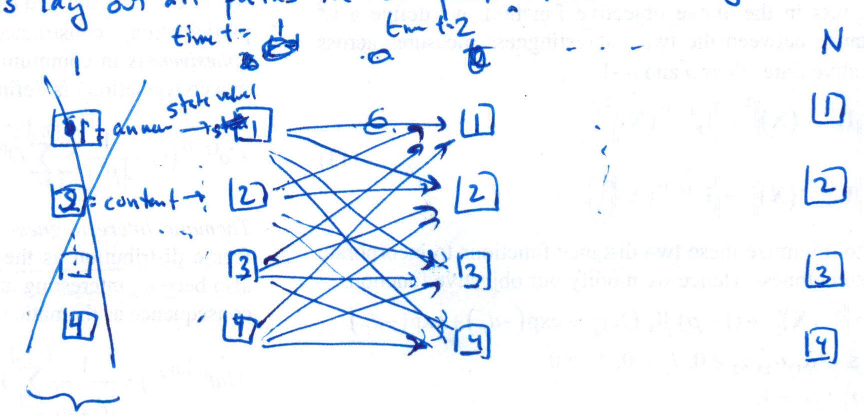
looks like a problem.

But we can reorganize as a dynamic program; given cond indep, stepping through time indices (those $\alpha$ & $\beta$ quantities)

So, here:

$$\sum_{pa} P(pa \mid \vec{x}, \theta^{old})\, \#(4 \leftarrow 3 \text{ in } pa)$$

let's lay out all paths in compact poly-size trellis diagram




again, reorganize by time steps:

$$\sum_{t=2}^{N} \sum_{pa} P(pa \mid \vec{x}, \theta^{old}) \cdot \mathbb{1}(\text{state at time } t-1 \text{ in } pa \text{ is } 3,\ \text{state at time } t \text{ in } pa \text{ is } 4).$$

remember we like marginals.

$$= \frac{1}{P(\vec{x}, \theta^{old})} \sum_{t=2}^{N} \sum_{pa} P(pa, \vec{x} \mid \theta^{old})\, \mathbb{1}(s_{t-1}^{pa} = 3,\ s_t^{pa} = 4)$$

} so, we are summing up the probs of all paths that have 3 → 4 @ time t-1 → t

You can similarly solve for $\mu_3, \sigma_3$.

in the log-likelihood w/ hidden structure, you get terms ~~that are~~ linear in $\mu_3$.