11/25/14
lec 26

~~Schedule checking appts~~

~~(comparing)~~ (~~comparing~~) language models

- have dft comparisons:

[A] - what can you tell from them?
[B] how can you compare them?

[A] language models

which we spent some time on last time:
remember we're broadly construing these as:
(prob. distributions over strings <begin> V^* <end> for fixed non-empty vocab V).

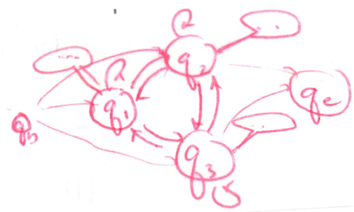
typically only implicitly noted.

Put here as a reminder to keep the generation model in mind.

e: ~~HMM-based lang. models~~,
hidden-markov-model based language models - "ngram" models as a special case,
probabilistic context-free-grammar based language models.

should: if "$P(xx)$" $= P(x)P(x|x) = P(x)P(x)$ then $P(x)+P(xx) =$ 2 for $P(x)=1$ for this x;
so don't confuse "base distribution" w/ distribution over all strings.

"structured" language model like a hidden Markov Model:



can be too big to look@.
[could, of course, just start w/ fewer states; but how satisfying is that, since you built the one w/ a larger state set.

(1) Fix length ℓ, compute top-k highest prob paths of length ℓ. [product of its transit'n & emission probs]
"what does this HMM 'like to do'?"

- dynamic programming.

[of course, you might want to vary ℓ]

(2) Fix sequence of interest s, compute top-k paths that assign highest prob to s.

reminder: can have more than one path generate the same sequence.

<why this instead of (1), sometimes?>

- ~~might not care ab~~ your LM might spend a lot of prob on sequences you don't care about, e.g., nonsense
- you might be particular interested in possible analyses of a test seq.
<ex: Polymath discussion>

[of course, you might want to vary s]

(3) look @ ~~distribution~~ transition or emission ~~prob~~ distribution for a particular state.

~~nb: this generalizes to n-gram mod~~

→ $p(x)$ over ~~some~~ some space $\mathcal{X}$.

~~For simplicity~~, assume $\mathcal{X} = \{x_i\}$ finite, and write θ_i for the prob of the i^{th} elt.

Notat'n: (b/c we will want these p's to be variables, soon).

constraints: $1 \geq \theta_i \geq 0$, $\sum \theta_i = 1$.

let $\vec{\theta}$ be the vector of θ_i's.

- note: n-gram models fit this: $p(y|x)$: prob given word x that next word is y.

y over some finite vocab

- each x induces its own distributi'n

ex: $\mathcal{X} = \{\overset{1}{\text{the}}, \overset{2}{\text{dog}}\}$

$p(\text{dog} \mid \text{the}) > p(\text{the} \mid \text{the})$.

$\theta_2^{\text{the}} > \theta_1^{\text{the}}$.

- what can we say about the θ_i's, besides just eyeballing them.

- measures of "diversity" or "spread"
  - are lots of things equally likely, or are there only a few things that are highly likely?

~~an important one:~~ the entropy, Gini-Simpson index (others possible, like L_1 or L_2 norms)

dumb in prob. simplex.

entropy $H(\vec{\theta})$ $\sum_i \theta_i \log\left(\frac{1}{\theta_i}\right)$.

small when θ_i big

"surprise" in seeing event i.

"back-integrat" as $p(x_i)$

= "expected surprise".

Convent'n: $0 \log 0 \triangleq 0$.

"by limiting arguments".

ex: $\theta_i^{\text{sharp}} = \begin{cases} 1, & i = 1 \\ 0 & \text{o.w.} \end{cases}$ $H(\vec{\theta}^{\text{sharp}}) = 0 + \sum_{i \neq 1} 0 = 0$

you are never surprised, b/c you will always see x_1

~~ex: $\theta_i^{\text{flat}} = \frac{1}{|\mathcal{X}|}$ $H(\vec{\theta}^{\text{flat}}) = \frac{1}{|\mathcal{X}|}\sum_i \log(|\mathcal{X}|) = \log(|\mathcal{X}|)$~~

so we know what has 0 entropy.
But could this be negative?
And what ~~could have the h~~ kind of distribution has the highest entropy?

Find $\underset{\vec{\theta}}{\arg\max}\; H(\vec{\theta}) \underbrace{- \lambda\left(\sum_i \theta_i - 1\right)}_{\text{constraint opt.}}$

(doing this exercise even if you already know the answer b/c useful later).

$$\frac{\partial}{\partial \theta_j}\left(\sum_i \theta_i \log\left(\frac{1}{\theta_i}\right) - \lambda\left(\sum_i \theta_i - 1\right)\right)$$

when $i \neq j$, the whole thing is a constant wrt θ_j.

$$= \frac{\partial}{\partial \theta_j}\left[\underbrace{\theta_j \log\left(\frac{1}{\theta_j}\right)}_{= -\theta_j \log \theta_j} - \lambda\theta_j\right]$$

$$= -\cancel{\theta_j}\cdot\frac{1}{\cancel{\theta_j}} - \log\theta_j - \lambda \quad \text{set to } 0$$

$$-1 - \log\theta_j - \lambda = 0$$

$$\log\theta_j = -(1-\lambda) \quad \text{constant}$$

. So all the θ_j's are the same.

– max "surprise" (least concentration) when you have all possibilities equally likely.

~~Same~~
~~What about Gini-S~~

Gini-Simpson index $GS(\vec{\theta}) = 1 - \sum_i \theta_i^2$: prob that 2 samples will not be the same.

$$= 1 - \sum_i \underbrace{\theta_i}_{p(x_i)} \theta_i$$

→ big when θ_i big
→ expectation of θ_i,
(cf. entropy).

- $GS(\vec{\theta}^{\,sharp}) = 1 - 1 - \sum_{i \neq 1} 0 = 0$ ✓

max? $\frac{\partial}{\partial \theta_j}\left(1 - \sum_i \theta_i^2 - \lambda\left(\sum_i \theta_i - 1\right)\right)$

$$= \frac{\partial}{\partial \theta_j}\left(\theta_j^2 - \lambda\theta_j\right) = 2\theta_j - \lambda.$$

set to 0, again, all θ_j's the same.

[B] comparing two LMs.

Could use L_2, L_1, etc.

Today, KL divergence: $D(\vec{\theta} \| \vec{\varphi}) = \boxed{\sum \theta_i \log \frac{\theta_i}{\varphi_i}}$ ← some terms can be negative.

$$= -\sum \theta_i \log \varphi_i - H(\vec{\theta})$$

if $\vec{\varphi} = \vec{\theta}$, $D(\vec{\theta} \| \vec{\theta}) = -\sum \theta_i \log \theta_i - H(\theta_i) = H(\vec{\theta}) - H(\vec{\theta}) = 0$.

$$\frac{\partial}{\partial \varphi_j}\left(-\sum_i \theta_i \log \varphi_i - \cancel{H(\vec{\theta})} - \lambda\left(\sum_i \varphi_i - 1\right)\right)$$

$= -\theta_j \frac{1}{\varphi_j} - \lambda$. set to 0, $-\frac{\theta_j}{\varphi_j} = \lambda$ or $-\frac{\theta_j}{\lambda} = \varphi_j$.

Since θ_i's already sum to 1, $\lambda = -1$.

so, $\vec{\theta} = \vec{\varphi}$ is at least a ~~saddle point~~ critical point.

2^{nd} partials: $\frac{\partial}{\partial \varphi_i \varphi_k}(\ ) = 0$, $\frac{\partial}{\partial \varphi_i^2} \geq 0$, @ least one is > 0. So, Hessian is ~~negative~~ positive definite, and we have a minimum.

[there are ~~ergo~~ better arguments using Jensen's inequality or other log properties, but I wanted to stick w/ using the same techniques throughout the lecture].

And note that if $\exists j$ s.t. $\theta_j > 0, \varphi_j = 0$, $D(\vec{\theta} \| \vec{\varphi}) = \infty$ ...

~~And, $\theta_i \log \varphi_i \to 0$ for φ_i a prob (btwn θ_i, 1)~~

is that weird? Two "finite" things having "infinite" distance?

but it makes sense: 2 distributions where one says something is possible that another says impossible, they are irreconcilable!

- ~~...~~ ... altho' zeroes in ~~estimate~~ data $\not\Rightarrow$ you think those things are necessarily impossible ...

- solutions: (a) smooth your distributions <if had time, would have done Kneser-Ney>

(b) use Jensen-Shannon divergence: $\pi_i \neq 0$ if θ_i or $\varphi_i \neq 0$

$\vec{\pi} = \frac{1}{2}(\vec{\theta} + \vec{\varphi})$ ~~– never 0 if~~

(θ, φ)

$$JSD(\vec{\theta}, \vec{\varphi}) = \frac{1}{2}\left(D(\vec{\theta} \| \vec{\pi}) + D(\vec{\varphi} \| \vec{\pi})\right)$$

- or the skew divergence:

$$D_\alpha(\vec{\theta}, \vec{\varphi}) = D(\vec{\theta} \| \alpha\vec{\varphi} + (1-\alpha)\vec{\theta}).$$

---

cross-entropy: $-\sum \theta_i \log \varphi_i$ (expectat'n wrt $\vec{\theta}$ of surprise according to $\vec{\varphi}_i$).

for empirical unigram model:

$$-\sum_i \frac{\#(x_i)}{\text{datasize}} \log \varphi_i = \frac{-1}{\text{datasize}} \log \prod \varphi_i^{\#(x_i)}$$

, which looks like the "prob" assigned to the sample.

E.g. in no country for old members, empirical = empirical lang. sample, $\vec{\varphi}$ = user's l.u.

In general:

for large sample, do $\frac{1}{\text{sample size}} P_{\vec{\theta}}(\text{sample})$.