language models

which we spent some time on last time:

remember we broadly constraining them as:

\( P_{\text{dist}}(\text{strings}) \) \( <\text{begin}>V^*<\text{end}> \) for

fixed non-empty vocab \( V \):

Hidden Markov Model (HMM) based language models - "n-gram" models as a special case.

Typically, only implicitly noted, R/Pi has a

Trek generation model in mind.

Show:

\[ P(z) = P(x)P(v|x) \]

\[ = P(x)P(v|x) \]

the \( P(x) + P(x) = 2 \) for

\( P(x) \leq 1 \) for

this.

so don't confuse

"best distribution

of distribution

over strings".

"Structured" language model

like a Hidden Markov Model:

can be too big to look at.

(could of course just start with

fewer states, but how

satisfying is that, since you built the ac to a larger state set.

1) Fix length \( l \), compute top-\( k \) highest prob path of length \( l \). [product of its

transition i omission p.e.s]

- dynamic programming

(of course, you might need to vary \( l \) ]

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

rememer: can have more than one path,

generate the same sequence.

< why this instead of 1, sometimes? >

- might not concern your LH, might spend a lot of prob on

sequences you don't care about, i.e., nonsense

- you might be particular interested in possible analogies of a test seq.

< ex Polymath discussion >

[of course, you might need to vary \( s \) ]
(3) look at **distribution** or emission **distribution** for a particular state.

Note this generalizes to n-gram model

\[ p(x) \text{ over some space } \mathbb{X} = \{ x_i \} \]

For simplicity, assume \( \mathbb{X} \) finite, and write \( \Theta_i \) for the prob of the \( i \)-th elt.

**Notation:**

(We will want these \( p_i \)'s to be variables, soon.)

\[ \sum \Theta_i = 1 \]

**Constraints:** \( \Theta_i > 0 \)

- Note: n-gram models fit this: \( p(x_1 \ldots x_n) \) prob given word \( x \) that went before it

\[ p(x_{12}) > p(x_1) \]

or over some finite vocab, induces its own distribution.

- What can we say about the \( \Theta_i \)'s, besides just eyeballing them?

- Measure 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)

\[ H(\Theta) = \sum_{i} \Theta_i \log \left( \frac{1}{\Theta_i} \right) \]

small when \( \Theta_i \) big. "Surprise" in seeing event \( i \). "Back-integrate" as \( p(x_i) \)

= expected surprise.

\[ \Theta_i = \begin{cases} 1, & i = 1 \\ 0, & \text{o.w.} \end{cases} \]

\[ H(\Theta) = 0 + \sum_{i=1}^{\infty} \Theta_i = 0 \]

You are never surprised, b/c you will always see \( x \).
so we know what has 0 entropy.
But could this be negative?

And what kind of distribution has the highest entropy?

Find \( \theta^* \) argmax \( H(\theta) = \lambda \sum_i \theta_i - 1 \) (constRAINTS opt).

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

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

\[
= -\theta_j \log \theta_j - \lambda \theta_j \text{ set to 0}
\]

\[
-1 - \log \theta_j - \lambda = 0
\]

\[
\log \theta_j = -1 - \lambda \text{ constant. So all the } \theta_j \text{'s are the same.}
\]

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

Same

Gini-Simpson index: \( G_S(\theta) = \sum_i \theta_i^2 \): prob that 2 samples will not be the same.

\[
\lambda G_S(\theta) = 1 - \sum_i \theta_i^2 \quad \text{set to } \theta_i \text{'s are the same.}
\]

\[
0 \quad G_S(\theta^{\text{sharp}}) = 1 - \sum_i \theta_i = 0
\]

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

\[
= \frac{\partial}{\partial \theta_j} \left( \theta_j^2 - \lambda \theta_j \right) = 2 \theta_j - \lambda
\]

set to 0, again, all \( \theta_j \)'s the same.
[B] comparing two LMs.

Could use $L_{2,1},$ etc.

Today, KL divergence: $D(\hat{\Theta} \parallel \hat{\Phi}) = \sum \Theta_i \log \frac{\Theta_i}{\hat{\Phi}_i}$ → some terms can be negative.

If $\hat{\Phi} = \hat{\Theta},$ $D(\hat{\Theta} \parallel \hat{\Theta}) = -\sum \Theta_i \log \Theta_i - H(\Theta).$

If $\hat{\Phi} = \hat{\Theta},$ $D(\hat{\Theta} \parallel \hat{\Theta}) = -\sum \Theta_i \log \Theta_i - H(\Theta) = H(\Theta) - H(\Theta) = 0.$

\[ \frac{\partial}{\partial \Theta_j} (\sum \Theta_i \log \Theta_i - H(\Theta)) - \lambda (\sum \Theta_i, -1) \]

\[ = -\Theta_j \frac{1}{\hat{\Phi}_j} - \lambda, \quad \text{set to 0}, \quad \Theta_j = \frac{\lambda \hat{\Phi}_j}{\lambda} \rightarrow \Theta_j = \frac{1}{\hat{\Theta}_j}. \]

Since $\Theta$'s already sum to 1,

so $\hat{\Theta} = \hat{\Phi}$ is at least a stationary point.

And note that if $\Theta = 0,$ $\hat{\Phi} = 0,$ $D(\hat{\Theta} \parallel \hat{\Phi}) = 0.$

Is that weird? Two 'finite' things having 'infinite' distance?

but it makes sense. 2 distributions when one says something is possible that another says impossible, they are irreconcilable.

... cool ... although in data, you think those things are necessarily impossible.

Solution: smooth your distributions. (If had time, would have done Kullback-Leibler)

(b) use Jensen-Shannon divergence:

(c) JS D($\hat{\Theta} \parallel \hat{\Phi}$) = $\frac{1}{2} \left( D(\hat{\Theta} \parallel \hat{\Phi}) + D(\hat{\Phi} \parallel \hat{\Theta}) \right)$

-or the skew divergence:

\[ D_s(\hat{\Theta}, \hat{\Phi}) = D(\hat{\Theta} \parallel \hat{\Phi}) + (1 - \alpha) H(\hat{\Theta}) \]

Cross entropy: $-\sum \Theta_i \log \hat{\Phi}_i$ (expected in ut $\Theta$ of surprise according to $\hat{\Phi}_i$).

For empirical unigram model:

\[ -\sum \frac{#(x_i)}{\text{dataset}} \log \hat{\Phi}_i = \frac{1}{\text{dataset}} \sum_\text{sample} \log \hat{\Phi}_i \]

what looks like the prob assigned to the sample.

E.g. in no country for old members, empirical = empirical lang. sample; $\hat{\Phi} = \text{user's input}.$

In general:

for large sample, do $\frac{1}{\text{dataset}} \hat{\Phi}_i$ (sample)