Bayesian (Nonparametric) Approaches to Language Modeling

Frank Wood

unemployed (and homeless)

January, 2013
- Whole Distributions As Random Variables
- Hierarchical Modeling
- Embracing Uncertainty
Challenge Problem

Predict What Comes Next

<table>
<thead>
<tr>
<th>01001001</th>
<th>01101110</th>
<th>00100000</th>
<th>01110100</th>
<th>01101000</th>
<th>01110010</th>
<th>01100101</th>
<th>01100101</th>
</tr>
</thead>
<tbody>
<tr>
<td>01110000</td>
<td>01101000</td>
<td>01100010</td>
<td>01100101</td>
<td>01111001</td>
<td>01110011</td>
<td>01100000</td>
<td>01111001</td>
</tr>
<tr>
<td>01101111</td>
<td>01110101</td>
<td>01110010</td>
<td>00100000</td>
<td>01101000</td>
<td>01100011</td>
<td>01110011</td>
<td>00100000</td>
</tr>
<tr>
<td>01100001</td>
<td>01110001</td>
<td>01101000</td>
<td>01100001</td>
<td>01101110</td>
<td>01100010</td>
<td>01101001</td>
<td>01110110</td>
</tr>
<tr>
<td>01100101</td>
<td>01100000</td>
<td>01110111</td>
<td>01110010</td>
<td>00100000</td>
<td>01100011</td>
<td>01100011</td>
<td>01110100</td>
</tr>
<tr>
<td>01101100</td>
<td>01101100</td>
<td>00100000</td>
<td>01100011</td>
<td>01100011</td>
<td>01100011</td>
<td>01101000</td>
<td></td>
</tr>
</tbody>
</table>

...
.6 < Entropy\(^1\) of English < 1.3 bits/byte [Shannon, 1951]

Sequence memoizer with life-long learning computational characteristics [Bartlett, Pfau, and Wood, 2010]

Wikipedia : 26.8GB → 4GB (≈ 1.2) bits/byte

- gzip : 7.9GB
- PAQ : 3.8GB

\(L\) is number of states in SM, data Wikipedia dump 2010 head, depth 16

\[\text{Entropy} = -\frac{1}{T} \sum_{t=1}^{T} \log_2 P(x_t | x_{1:t-1}) \text{ (lower is better)}\]
the fastest way to make money is . . .

*being player*. *bucketti wish i made it end in the reducers and assemblance smart practices to allow city is something to walk in most of the work of agriculture*. *i ’d be able to compete with an earlier goals : words the danger of conduct in southern california, has set of juice of the fights lack of audience that the eclasm beverly hills*. ” she companies or running back down that the book, ” grass. and the coynes, exaggerate between 1972. the pad, a debate on emissions on air political capital crashing that the new obviously program ” – irock price, ” coach began refugees, much and²

²constant-space, byte-SM, NY Times Corpus (50MB), log-loss 1.49 bits/byte, 2 million rest.’s
Let $\Sigma$ be a set of symbols, let $\mathbf{x} = x_1, x_2, ..., x_T$ be an observed sequence, $x_t, s, s' \in \Sigma$, and $u \in \mathbf{W} = \{w_1, w_2, \ldots w_K\}$ with $w_k \in \Sigma^+$. 

$$G_u(s) = \frac{N(us)}{\sum_{s' \in \Sigma} N(us')}$$

is the maximum likelihood estimator for the conditional distribution $P(X = s|u) \equiv G_u(s)$. 
Notation and iid Conditional Distribution Estimation

Let $\Sigma$ be a set of symbols, let $x = x_1, x_2, ..., x_T$ be an observed sequence, $x_t, s, s' \in \Sigma$, and $u \in W = \{w_1, w_2, \ldots, w_K\}$ with $w_k \in \Sigma^+$. Then

$$G_u(s) = \frac{N(us)}{\sum_{s' \in \Sigma} N(us')}$$

is the maximum likelihood estimator for the conditional distribution $P(X = s|u) \equiv G_u(s)$.

Estimating a finite set of conditional distributions corresponding to a finite set of “contexts” $u$ generally yields a poor model. Why?

- Long contexts occur infrequently
- Maximum likelihood estimates given few observations are often poor
- Finite set of contexts might be sparse
Use independent Dirichlet priors, i.e.

\[ G_u \sim \text{Dir}(\frac{\alpha}{|\Sigma|}) \]

\[ x_t | x_{t-k}, \ldots, x_{t-1} = u \sim G_u \]
Use independent Dirichlet priors, i.e.

\[ G_u \sim \text{Dir} \left( \frac{\alpha}{\Sigma} \right) \]

\[ x_t | x_{t-k}, \ldots, x_{t-1} = u \sim G_u \]

and embrace uncertainty about \( G_u \)'s:

\[ P(x_{T+1} = s | x = u) = \int P(x_{t+1} = s | G_u) P(G_u | x) dG = \mathbb{E}[G_u(s)], \]

yielding, for instance,

\[ \mathbb{E}[G_u(s)] = \frac{N(us) + \frac{\alpha}{\Sigma}}{\sum_{s' \in \Sigma} N(us')} + \alpha \]
Smooth $G_u$ using a similar distribution $G_{\sigma(u)}$, i.e.

$$G_u \sim \text{Dir}(cG_{\sigma(u)})$$

or

$$G_u \sim \text{DP}(cG_{\sigma(u)})$$

with, for instance,

$$\sigma(x_1x_2x_3 \ldots x_t) = x_2x_3 \ldots x_t$$

(suffix operator)

---

3 [MacKay and Peto, 1995]
4 [Teh et al., 2006]
Is it possible to simultaneously estimate all distributions in such a model, even in the infinite-depth case?
Is it possible to simultaneously estimate all distributions in such a model, even in the infinite-depth case?

YES!

**Sequence memoizer (SM), Wood et al. [2011]**
The SM is a compact $O(T)$, unbounded-depth, hierarchical, Bayesian-nonparametric distribution over discrete sequence data.

**Graphical Pitman-Yor process (GPYP), Wood and Teh [2008]**
The GPYP is a hierarchical sharing prior with directed acyclic dependency.

Figure: Binary Sequence Memoizer Graphical Model
Notation

\[
G_{\varepsilon | U_{\Sigma}, d_0} \sim \text{PY}(d_0, 0, U_{\Sigma})
\]
\[
G_{u | G_{\sigma(u)}, d_{|u|}} \sim \text{PY}(d_{|u|}, 0, G_{\sigma(u)}) \quad \forall u \in \Sigma^+
\]
\[
x_t | x_1, \ldots, x_{t-1} = u \sim G_u
\]

Here \(\sigma(x_1 x_2 x_3 \ldots x_t) = x_2 x_3 \ldots x_t\) is the suffix operator, \(\Sigma = \{0, 1\}\), and \(U_{\Sigma} = [.5, .5]\).

Encodes natural assumptions:

- Recency assumption encoded by form of coupling related conditional distributions ("back-off")
- Power-law effects encoded by the choice of hierarchical modeling glue: the Pitman-Yor process (PY).
A Pitman-Yor process PY(c, d, H) is a distribution over distributions with three parameters

- A discount $0 \leq d < 1$ that controls power-law behavior
  - $d = 0$ is Dirichlet process (DP)
- A concentration $c > -d$ like that of the Dirichlet process (DP)
- A base distribution $H$ also like that of the DP

Key properties:

If $G \sim PY(c, d, H)$ then \textit{a priori}

- $\mathbb{E}[G(s)] = H(s)$
- $\text{Var}[G(s)] = (1 - d)H(s)(1 - H(s))$
Power-law scaling of word frequencies in English text. Relative word frequency is plotted against each words’ rank when ordered according to frequency. There are a few very common words and a large number of relatively rare words.

Figure from Wood et al. [2011].
Space. And time.
Coagulation and Fragmentation

PY properties used to achieve linear space sequence memoizer.

**Coagulation**

If

\[ G_2 | G_1 \sim PY(d_1, 0, G_1) \]

and

\[ G_3 | G_2 \sim PY(d_2, 0, G_2) \]

then

\[ G_3 | G_1 \sim PY(d_1d_2, 0, G_1) \]

with \( G_2 \) marginalized out.

**Fragmentation**

Suppose

\[ G_3 | G_1 \sim PY(d_1d_2, 0, G_1) \]

Operations exist to draw

\[ G_2 | G_1 \sim PY(d_1, 0, G_1) \]

and

\[ G_3 | G_2 \sim PY(d_2, 0, G_2) \]

---

\[ ^5 \text{[Pitman, 1999, Ho et al., 2006, Wood et al., 2009]} \]
$O(T)$ Sequence Memoizer\footnote{Wood et al, 2009} Graphical Model for $x = 110100$
SM versus $n^{th}$ order Markov models with hierarchical PYP priors as $n$ varies. In red is computational complexity.
Problem

Nonparametric $\rightarrow O(T)$ space $\implies$ doesn’t work for life-long learning
Towards Lifelong Learning : Incremental Estimation

Sequential Monte Carlo inference [Gasthaus, Wood, and Teh, 2010]

- Wikipedia 100MB → 20.8MB (1.66 bits/byte) 
  - ≈ PAQ, special-purpose, non-incremental, Hutter-prize leader

Still $O(T)$ storage complexity.

\[
- \frac{1}{T} \sum_{t=1}^{T} \log_2 P(x_t | x_{1:t-1})
\]

is average # of bits to transmit one byte under optimal entropy encoding (lower is better)
**Towards Lifelong Learning: Constant Space Approximation**

"Forgetting Counts" [Bartlett, Pfau, and Wood, 2010]

<table>
<thead>
<tr>
<th>States</th>
<th>Naive Log Loss (bits)</th>
<th>Finite Depth Log Loss (bits)</th>
<th>Random Deletion Log Loss (bits)</th>
<th>Greedy Deletion Log Loss (bits)</th>
<th>Sequence Memoizer Log Loss (bits)</th>
<th>gzip Log Loss (bits)</th>
<th>bzip2 Log Loss (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,164</td>
<td>1.8</td>
<td>2</td>
<td>2.2</td>
<td>2.4</td>
<td>2.6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>41,860</td>
<td>2</td>
<td>2.2</td>
<td>2.4</td>
<td>2.6</td>
<td>2.8</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td>89,479</td>
<td>2.2</td>
<td>2.4</td>
<td>2.6</td>
<td>2.8</td>
<td>3.0</td>
<td>2.6</td>
<td>2.8</td>
</tr>
<tr>
<td>155,623</td>
<td>2.4</td>
<td>2.6</td>
<td>2.8</td>
<td>3.0</td>
<td>3.2</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>300,490</td>
<td>2.6</td>
<td>2.8</td>
<td>3.0</td>
<td>3.2</td>
<td>3.4</td>
<td>3.0</td>
<td>3.2</td>
</tr>
<tr>
<td>467,502</td>
<td>2.8</td>
<td>3.0</td>
<td>3.2</td>
<td>3.4</td>
<td>3.6</td>
<td>3.2</td>
<td>3.4</td>
</tr>
<tr>
<td>632,448</td>
<td>3.0</td>
<td>3.2</td>
<td>3.4</td>
<td>3.6</td>
<td>3.8</td>
<td>3.4</td>
<td>3.6</td>
</tr>
<tr>
<td>776,547</td>
<td>3.2</td>
<td>3.4</td>
<td>3.6</td>
<td>3.8</td>
<td>4.0</td>
<td>3.6</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Constant-state-count SM approximation

Data: Calgary corpus, States = number of states used in 3-10 grams

"Improvements to the SM" [Gasthaus and Teh, 2011]

Constant-size representation of predictive distributions
SM with life-long learning computational characteristics.

Wikipedia : 26.8GB $\rightarrow$ 4GB ($\approx 1.2$) bits/byte

$L$ is number of states in SM, data Wikipedia dump 2010 head, depth 16,
SM = powerful Bayesian nonparametric sequence model

- Intuition: \( \lim_{n \to \infty} \) smoothing \( n \)-gram
- Byte predictive performance \emph{in human range}
- Bayesian-nonparametric \( \implies \) incompatible with lifelong learning
- Non-trivial approximations preserve performance
- Resulting model compatible with lifelong learning

Widely useful

- Language-modeling
- Lossless compression
- Plug-in replacement for Markov model component of any complex graphical model
- Direct plug-in replacement for \( n \)-gram language model
  - Software: \url{http://www.sequencememoizer.com/}
Can We Do Better?

Problem:
- No long-range coherence
- Domain specificity

Solution:
- Develop models that more efficiently encode long-range dependencies
- Develop models that use domain-specific context
  - Graphical Pitman-Yor process (GPYP)
Simultaneous estimation = automatic domain adaptation
\[ G_D^\{\} \sim \text{PY}(d_D^0, \alpha_D^0, \lambda_D^0 \mathcal{U} + (1 - \lambda_D^0) G_H^\{\} ) \]
\[ G_D^{\{w_{t-1}\}} \sim \text{PY}(d_D^1, \alpha_D^1, \lambda_D^1 G_D^\{\} + (1 - \lambda_D^1) G_H^{\{w_{t-1}\}} ) \]
\[ \vdots \]
\[ G_D^{\{w_{t-j}:w_{t-1}\}} \sim \text{PY}(d_D^j, \alpha_D^j, \lambda_D^j G_D^{\{w_{t-j+1}:w_{t-1}\}} + (1 - \lambda_D^j) G_H^{\{w_{t-j}:w_{t-1}\}} ) \]
\[ X | w_{t-n+1} : w_{t-1} \sim G_D^{\{w_{t-n+1}:w_{t-1}\}}. \]

\[ [\text{Wood and Teh, 2009}] \]
Domain Adaptation Effect (Copacetic)

---

Training corpora
- **Small**
  - State of the Union (SOU)
    - ~370k tokens
    - ~13k types
- **Big**
  - Brown
    - 1967
    - ~1m tokens
    - ~50k types

Test corpus
- Johnson’s SOU speeches
  - 1963-1969
  - ~37k tokens

---

Mixture : [Kneser and Steinbiss, 1993]
Union : [Bellegarda, 2004]
Domain Adaptation Effect (Not)

Training corpora
- Small
  - Augmented Multi-Party Interaction (AMI)
    - 2007
    - ∼ 800k tokens
    - ∼ 8k types
- Big
  - Brown
    - 1967
    - ∼ 1m tokens
    - ∼ 50k types

Test corpus
- AMI excerpt
  - 2007
  - ∼ 60k tokens
Realistic scenario
- Increasing Brown Corpus size imposes computational cost
  - $  
- Increasing SOU corpus size imposes data acquisition cost
  - $$$
Whole Distributions As Random Variables
Hierarchical Modeling
Embracing Uncertainty
The End

Thank you.


