Pitman-Yor Process in statistical language models

J Li

September 28, 2011

Pitman-Yor Process

- General stick-breaking prior: $\mathcal{P}(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.)$:

Pitman-Yor Process

- General stick-breaking prior: $\mathcal{P}(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.)$:
  - Generating values: $Z_k \sim \mathcal{H}$
  - Assigning weights:

$$p_k = \prod_{i=1}^{k-1} (1 - V_i) V_k \qquad V_k \sim Beta(a_k, b_k)$$

Pitman-Yor Process

- General stick-breaking prior: $\mathcal{P}(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.)$:
  - Generating values: $Z_k \sim \mathcal{H}$
  - Assigning weights:

$$p_k = \prod_{i=1}^{k-1}(1 - V_i)V_k \qquad V_k \sim Beta(a_k, b_k)$$

- Pitman-Yor as a special case: $\mathcal{PY}(a, b, \mathcal{H}), a \in [0, 1), b > -a$
  - $V_k \sim Beta(1 - a, a + bk)$
  - Weights $\{p_k\}_{k=1}^{N}$ induces a power-law distribution:

$$P(n_w) \propto n_w{}^{-(1+a)}$$

Why power-law?

- Think of the proportions broken off the remaining stick, $V_k \sim Beta(1 - a, b + ka)$:

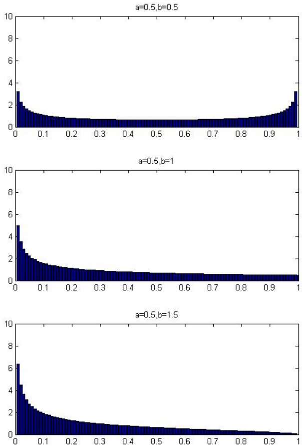
$$E[V_k] = \frac{1 - a}{1 + b + (k - 1)a}$$

- $a = 0$ reduces to DP

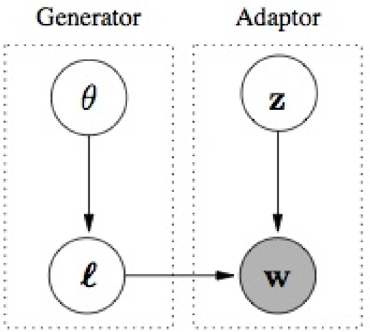
- Think of the proportions broken off the remaining stick, $V_k \sim Beta(1-a, b+ka)$:

$$E[V_k] = \frac{1-a}{1+b+(k-1)a}$$

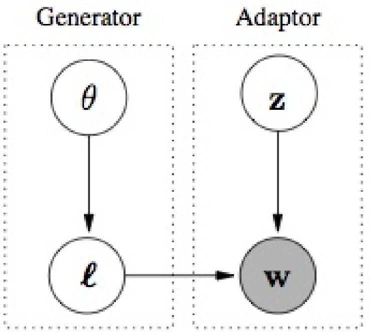
- $a = 0$ reduces to DP
- Suppose we set $a = 0.5, b = 0$, the pdfs of successive V_k (on the right):



Two stage language model



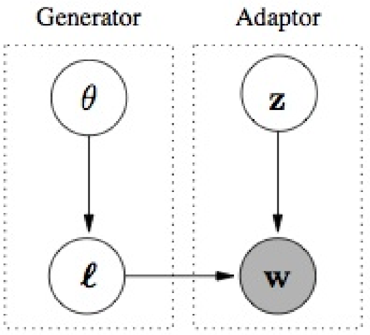
Two stage language model



Using the CRP analogy:

- The "generator" labels tables with *word types*: $l_k \sim \pi(l|\theta)$
- The "adaptor" assigns customers to tables: $z_k \sim \mathcal{PY}(a, b)$
- The outcome is a steam of *word tokens*: $w_1 = l_{z_1}, w_2 = l_{z_2}, \cdots$

Two stage language model



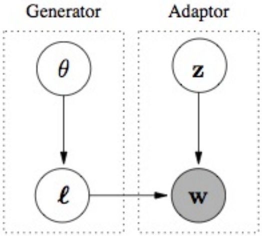
Using the CRP analogy:

- The "generator" labels tables with *word types*: $l_k \sim \pi(l|\theta)$
- The "adaptor" assigns customers to tables: $z_k \sim \mathcal{PY}(a, b)$
- The outcome is a steam of *word tokens*: $w_1 = l_{z_1}, w_2 = l_{z_2}, \cdots$

note: this is not necessarily a true Pitman-Yor.

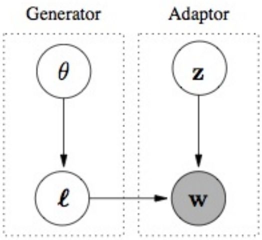
- Prediction rule (Polya Urn)

$$
\begin{aligned}
P(w_i = w \,|\, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) &= \sum_k \sum_{\ell_k} P(w_i = w \,|\, z_i = k, \ell_k) P(\ell_k \,|\, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) P(z_i = k \,|\, \mathbf{z}_{-i}) \\
&= \sum_{k=1}^{K(\mathbf{z}_{-i})} \frac{n_k^{(\mathbf{z}_{-i})} - a}{i - 1 + b} I(\ell_k = w) + \frac{K(\mathbf{z}_{-i})a + b}{i - 1 + b} \theta_w \qquad (3)
\end{aligned}
$$



- Prediction rule (Polya Urn)

$$
\begin{aligned}
P(w_i = w \mid \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) &= \sum_k \sum_{\ell_k} P(w_i = w \mid z_i = k, \ell_k) P(\ell_k \mid \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) P(z_i = k \mid \mathbf{z}_{-i}) \\
&= \sum_{k=1}^{K(\mathbf{z}_{-i})} \frac{n_k^{(\mathbf{z}_{-i})} - a}{i - 1 + b} I(\ell_k = w) + \frac{K(\mathbf{z}_{-i})a + b}{i - 1 + b} \theta_w \qquad (3)
\end{aligned}
$$

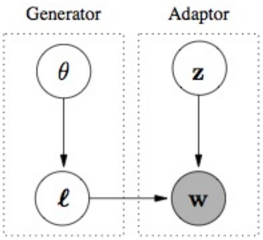


- Compare with DP prediction rule
- The authors set $b = 0$. *Why?*

- Prediction rule (Polya Urn)

$$P(w_i = w \mid \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) = \sum_k \sum_{\ell_k} P(w_i = w \mid z_i = k, \ell_k) P(\ell_k \mid \mathbf{w}_{-i}, \mathbf{z}_{-i}, \theta) P(z_i = k \mid \mathbf{z}_{-i})$$

$$= \sum_{k=1}^{K(\mathbf{z}_{-i})} \frac{n_k^{(\mathbf{z}_{-i})} - a}{i - 1 + b} I(\ell_k = w) + \frac{K(\mathbf{z}_{-i})a + b}{i - 1 + b} \theta_w \quad (3)$$



- Compare with DP prediction rule
- The authors set $b = 0$. *Why?*
    - Maybe because its value makes no difference.

Estimate based on tokens or types?

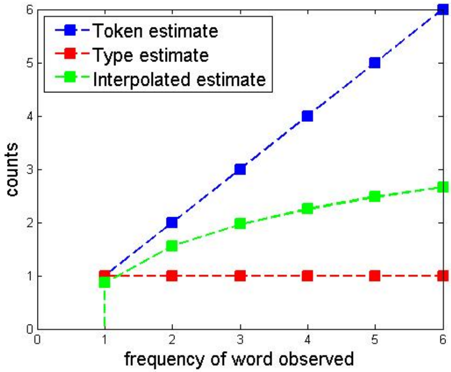
Simplified setting: given observation of a set of N words, derive a distribution over all words.

- Based on tokens: $\hat{\pi}_{w,1} = \frac{n_w}{N}$
- Based on types: $\hat{\pi}_{w,2} \propto I(w \in \mathbf{W})$
- Interpolate between them: $\hat{\pi} = \alpha(n_w)\hat{\pi}_{w,1} + \beta\hat{\pi}_{w,2}$

Estimate based on tokens or types?

Simplified setting: given observation of a set of N words, derive a distribution over all words.

- Based on tokens: $\hat{\pi}_{w,1} = \frac{n_w}{N}$
- Based on types: $\hat{\pi}_{w,2} \propto I(w \in \mathbf{W})$
- Interpolate between them: $\hat{\pi} = \alpha(n_w)\hat{\pi}_{w,1} + \beta\hat{\pi}_{w,2}$



Interpolated Kneser-Ney (IKN)

- Task: estimate distribution of $(w_{N+1}|\mathbf{w_{N-n+2\cdots N}})$;
- Given: a vector of N words $\mathbf{w}$ that share a common history ($n-1$ previous words) and vectors of words with different histories $\mathbf{w^{(1)}}, \cdots \mathbf{w^{(H)}}$

Interpolated Kneser-Ney (IKN)

- Task: estimate distribution of $(w_{N+1} | \mathbf{w_{N-n+2 \cdots N}})$;
- Given: a vector of N words $\mathbf{w}$ that share a common history ($n-1$ previous words) and vectors of words with different histories $\mathbf{w^{(1)}}, \cdots \mathbf{w^{(H)}}$

IKN estimator:

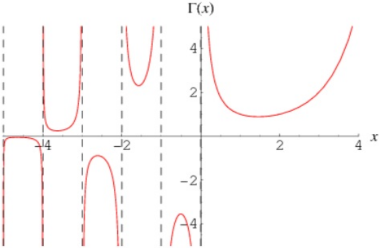
$$P(w_{N+1} = w \,|\, \mathbf{w}) = \frac{n_w^{(\mathbf{w})} - I(n_w^{(\mathbf{w})} > D)D}{N} + \frac{\sum_w I(n_w^{(\mathbf{w})} > D)D}{N} \frac{\sum_h I(w \in \mathbf{w}^{(h)})}{\sum_w \sum_h I(w \in \mathbf{w}^{(h)})} \quad (5)$$

Two stage language model revisited: adjust parameters

Likelihood:

$$P(\mathbf{w} \mid \theta) = \sum_{\mathbf{z},\boldsymbol{\ell}} \left(\prod_{k=1}^{K(\mathbf{z})} \theta_{\ell_k} \right) \cdot \frac{\Gamma(K(\mathbf{z}))}{\Gamma(N)} \cdot a^{K(\mathbf{z})} \cdot \left(\prod_{k=1}^{K(\mathbf{z})} \frac{\Gamma(n_k^{(\mathbf{z})} - a)}{\Gamma(1-a)} \right)$$

$a \in [0, 1)$

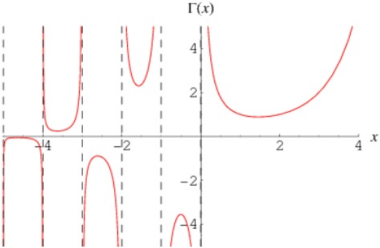


Two stage language model revisited: adjust parameters

Likelihood:

$$P(\mathbf{w} \mid \theta) = \sum_{\mathbf{z}, \boldsymbol{\ell}} \left(\prod_{k=1}^{K(\mathbf{z})} \theta_{\ell_k} \right) \cdot \frac{\Gamma(K(\mathbf{z}))}{\Gamma(N)} \cdot a^{K(\mathbf{z})} \cdot \left(\prod_{k=1}^{K(\mathbf{z})} \frac{\Gamma(n_k^{(\mathbf{z})} - a)}{\Gamma(1 - a)} \right)$$

$a \in [0, 1)$



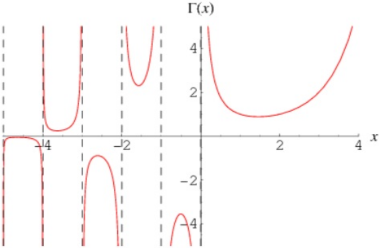
- When $a \longrightarrow^{-} 1$,
  $\Gamma(1 - a) \longrightarrow \infty$, so
  $\frac{\Gamma(n_k^{\mathbf{z}} - a)}{\Gamma(1-a)} \longrightarrow 0$ unless $n_k^{(\mathbf{z})} = 1$

Two stage language model revisited: adjust parameters

Likelihood:

$$P(\mathbf{w} \mid \theta) = \sum_{\mathbf{z}, \boldsymbol{\ell}} \left(\prod_{k=1}^{K(\mathbf{z})} \theta_{\ell_k} \right) \cdot \frac{\Gamma(K(\mathbf{z}))}{\Gamma(N)} \cdot a^{K(\mathbf{z})} \cdot \left(\prod_{k=1}^{K(\mathbf{z})} \frac{\Gamma(n_k^{(\mathbf{z})} - a)}{\Gamma(1-a)} \right)$$

$a \in [0, 1)$



- When $a \longrightarrow^{-} 1$, $\Gamma(1-a) \longrightarrow \infty$, so $\frac{\Gamma(n_k^{\mathbf{z}} - a)}{\Gamma(1-a)} \longrightarrow 0$ unless $n_k^{(\mathbf{z})} = 1$
- When $a \longrightarrow^{+} 0$, $K(\mathbf{z})$ tends to be small small, due to the $a^{K(\mathbf{z})}$ term
    - Actually, $K(\mathbf{z}) \approx$ number of distinct words in $\mathbf{w}$

Correspondence

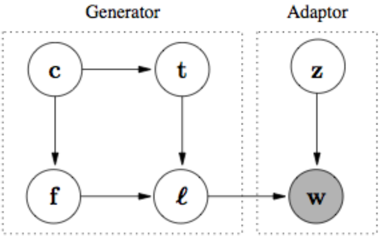
Two-stage model:

$$P(w_{N+1} = w \mid \mathbf{w}, \theta) = \frac{n_w^{\mathbf{w}} - E_{\mathbf{z}}[K_w(\mathbf{z})]\, a}{N} + \frac{\sum_w E_{\mathbf{z}}[K_w(\mathbf{z})]\, a}{N}\, \theta_w \qquad (6)$$

IKN model:

$$P(w_{N+1} = w \mid \mathbf{w}) = \frac{n_w^{(\mathbf{w})} - I(n_w^{(\mathbf{w})} > D)D}{N} + \frac{\sum_w I(n_w^{(\mathbf{w})} > D)D}{N} \frac{\sum_h I(w \in \mathbf{w}^{(h)})}{\sum_w \sum_h I(w \in \mathbf{w}^{(h)})} \qquad (5)$$

Application: morphology



- Generator:
  - inflection class: $c_k \sim mult(\kappa)$
  - stem: $t_k | c_k \sim mult(\tau)$
  - suffix: $f_k | c_k \sim mult(\phi)$
  - $l_k = t_k \cdot f_k$
- Adaptor: $z_k \sim \mathcal{PY}(a, 0)$
- Output: $w_k = l_{z_k}$

Gibbs Sampling

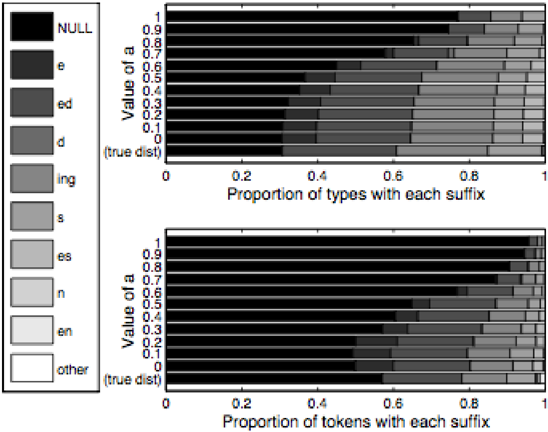
Update $\theta = (c, t, f)$:

$$\begin{aligned}
&P(c_k = c, t_k = t, f_k = f \mid \mathbf{c}_{-k}, \mathbf{t}_{-k}, \mathbf{f}_{-k}, \boldsymbol{\ell}) \\
&\propto \quad I(\ell_k = t_k.f_k) \quad P(c_k = c \mid \mathbf{c}_{-k}) \quad P(t_k = t \mid \mathbf{t}_{-k}, \mathbf{c}) \quad P(f_k = f \mid \mathbf{f}_{-k}, \mathbf{c}) \\
&= \quad I(\ell_k = t_k.f_k) \cdot \frac{n_c + \kappa}{K(\mathbf{z}) - 1 + \kappa C} \cdot \frac{n_{c,t} + \tau}{n_c + \tau T} \cdot \frac{n_{c,f} + \phi}{n_c + \phi F}
\end{aligned}$$

Update z:

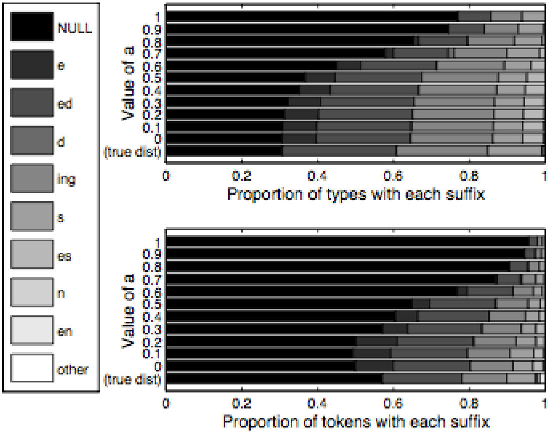
$$P(z_i = k \mid \mathbf{z}_{-i}, \mathbf{w}, \mathbf{c}, \mathbf{t}, \mathbf{f}) \propto \begin{cases} I(\ell_k = w_i)(n_k^{(\mathbf{z}_{-i})} - a) & n_k^{(\mathbf{z}_{-i})} > 0 \\ P(\ell_k = w_i)(K(\mathbf{z}_{-i})a + b) & n_k^{(\mathbf{z}_{-i})} = 0 \end{cases}$$

Evaluation



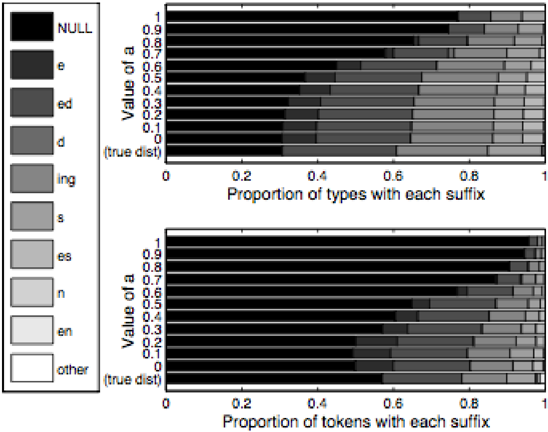
- $a = 0$ works the best, therefore *estimation by type is justified*

Evaluation



- $a = 0$ works the best, therefore *estimation by type is justified*
  - Wait a minute, doesn't $a = 0$ correspond to DP?

Evaluation



- $a = 0$ works the best, therefore *estimation by type is justified*
  - Wait a minute, doesn't $a = 0$ correspond to DP?
- author claims the value of model lies in *flexibility*

Discussion

Questions?