A Review of Yee Whye Teh's A Hierarchical Language Model based on the Pitman-Yor Process

Jessica Forde

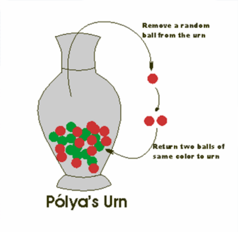
Columbia University

March 26, 2013

Outline

Some Intuition: Polya Urns

- Imagine an urn with balls in k colors, where n_i is the number of balls with color i and $\alpha_i = \frac{n_i}{\sum_{i=1}^{k} n_i}$
- After each draw, the ball drawn is returned with an additional ball of the same color



- Each draw defines a distribution over the set of all unique colors
- As the number of draws approaches infinity, the balls in the urn will be distributed $\mathit{Dirichlet}(\alpha_1, ..., \alpha_K)$
- The limit of the color proportions in the urn defined by these draws can be described as a Dirichlet Process (DP)[3]

Dirichlet Processes

- Θ has measurable partition $A_1, ..., A_k$ if $\cup_{i=1}^{k} A_i = \Theta$ and $A_1, ..., A_k$ is closed under complementation and countable union
- Given event space, Θ with measurable partitions $A_1, ..., A_k$, base distribution H (e.g. $H \sim \mathcal{N}$), and scale parameter α, we say G is distributed DP [3][2] if

$$(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha H(A_1), ..., \alpha H(A_k))$$

- For all $i \in [1, K]$, $E[G(A_i)] = H(A_i)$ and $Var[G(A_i)] = \frac{H(A_i)(1-H(A_i))}{\alpha+1}$
- From an NLP perspective,
    - if Θ is the set of all words, G is a distribution over words where α indicates the similarity between H and G [5]
    - if $\theta_i \in \Theta$ is a word token and x_i is an observed string, a typical mixture model set up states that $\theta_i \sim G$ and $x_i|\theta_i \sim F(\theta_i)$

Blackwell MacQueen Urn Scheme

- Another useful metaphor for a DP marginalizes out G itself [2][3]

$$p(\theta_1, ..., \theta_n) = \int (\prod_{i=1}^{n} p(\theta_i | G)) p(G) \partial G$$

- We now have an urn, G, which is initially empty, and a paintbox H
- To initialize, we first draw color from H and put a ball with that color in G, $\theta_1 \sim H$
- For ball θ_{n+1}, we draw a new color $\theta_{n+1} \sim H$ with probability $\frac{\alpha}{n+\alpha}$ to color the ball, or we draw $\theta_{n+1} \sim G$ like in the Polya Urn setup and return two balls with that same color with probability $\frac{n}{n+\alpha}$

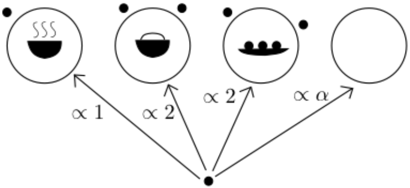
- Ferguson [3] proved that DP's are the infinite sum of discrete distributions; Let δ_{θ_i} be an indicator function, called an atom, equalling 1 if $\theta_i \in A_j$ and let π_i be the probability mass of δ_{θ_i}

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$$

- Because we are working with cojugate distributions, we can describe our intuition from the Blackwell MacQueen urn scheme in the following ways
  - $G \sim DP(\alpha, H)$
  - $\theta_{1:n} | G \sim G$
  - $\theta_i | \theta_{1:n \setminus i}, G \sim G$
  - $G | \theta_{1:n} \sim DP(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n})$
  - $\theta_i \sim H$
  - $\theta_{n+1} | \theta_{1:n} \sim \frac{aH + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$

Chinese Restaurant Process

- We can observe that draws from a Blackwell MacQueen urn define a random partition
- Imagine now there are k colors drawn from H in the urn after n draws
- This distribtuion over the partition from $[1:n]$ into these k clusters is a Chinese Restaurant Process[1], $\theta_{n+1}|\theta_{1:n} \sim CRP(H)$



$$P(\theta_{n+1} \in [1:k]) = \sum_{j=1}^{k} \frac{n_j}{n+\alpha}$$

Pitman-Yor Processes

- In a typical CRP setup, the probability of adding a additional component to a mixture model given n observations is $\frac{\alpha}{\alpha+n}$
- Pitman-Yor (PY) Processes add a rate parameter d to control the addition of components
- Instead, the probability of an additional table at given k components is $\frac{\alpha+dk}{n+\alpha}$
- The number of unique words in an NLP set up is therefore $O(\alpha n^d)$ instead of $O(\alpha \log n)$
- Goldwater et al. [4] observe that PYs are better suited to linguistic other DPs because they mimic the power law distributions seen in natural languages
    - if $t(c)$ is the expected number of PY components with c observations, $t(c+1) = (1 + \frac{d}{\alpha+c})t(c) + \frac{\alpha}{\alpha+c}$

- Recall that n-gram models use the conditional distribution of a word given its $n-1$ predecessors to approximate a sentence
  - $P(sentence) \approx \prod_{i=1}^{T} P(word_i | word_{i-n+1}^{i-1})$
- Teh [7] places a prior on this model based on the Hierarchical Pitman-Yor (HPY)
  - Given the context $\mathbf{u} = \{u_1, ..., u_m\}, m \leq n-1$:

$$G_{\mathbf{u}} \sim PY(d_{|\mathbf{u}|}, \theta_{|\mathbf{u}|}, G_{\pi(\mathbf{u})})$$

  - $G_{\pi(\mathbf{u})}$ is the base distribution of the observed word given the suffix $\pi(\mathbf{u}) = \{u_1, ..., u_{m-1}\}$
  - $G_{\pi(\mathbf{u})}$ is drawn recursively until we reach $G_{\emptyset} \sim PY(d_0.\theta_0, G_0)$, the probability of the current word given the empty set
  - This prior takes the structure of a suffix tree of depth n

Inference in the HPY Model via HCRP

- For inference, this model is reframed in the context of a Hierarchical Chinese Restaurant Process (HCRP) [6]
- Teh [7] uses Gibbs sampling to approximate the posterior over the seating arrangements and the model parameters
- Like in the Blackwell MacQueen example, $G_{\mathbf{u}}$ is marginalized out and instead replaced with $S_{\mathbf{u}}$, which corresponds to a seating arrangement
- The probability of a word given the context and the data is approximately

$$P(w|u, \mathcal{D}) \approx \sum_{i=1}^{I} p(w|\mathbf{u}, S^{(i)}, \Theta^{(i)})$$

- Sampling takes $O(nT)$ time and requires $O(M)$ space
- Teh [7] notes that interpolated Kneser-Ney (IKN) smoothing approximates this model by assuming each cluster has a unique token
- HPY outperforms IKN on the APNews corpus

Gibbs Sampling in the HCRP

- Let $\mathbf{u}$ be a restaurant with $c_{\mathbf{u}wk}$ customers sitting at table k and eating dish w and $t_{\mathbf{u}w}$ be the number of tables serving w
- To draw a new word given context $\mathbf{u}$
  - If $\mathbf{u} == 0$, return $w \in W$ with probability $G_0(w)$
  - else sit customer at table k with probability $\propto c_{\mathbf{u}wk} - d_{|\mathbf{u}|}$
  - or sit customer at a new table serving dish w with probability $\propto \theta_{|\mathbf{u}|} + t_{|\mathbf{u}|} d_{|\mathbf{u}|}$
- The probability of the next word after context $\mathbf{u} = 0$ is $G_0(w)$ else it is

$$P^{HPY}_{\mathbf{u}}(w|S_{\mathbf{u}}) = \frac{c_{\mathbf{u}w\cdot} - d_{|\mathbf{u}|} t_{|\mathbf{u}w|}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}}} + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|} t_{|\mathbf{u}|}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}}} P^{HPY}_{\pi(\mathbf{u})}(w|S_{\mathbf{u}})$$

- Note that this equation is similar to IKN by setting $t_{|\mathbf{u}w|} = 1$

References I

[1] D. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer, 1985.

[2] D. Blackwell and J. B. Macqueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[3] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[4] S. Goldwater, T. L. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18*, pages 459–466, Cambridge, MA, 2006. MIT Press.

[5] N. Sharif-razavian and A. Zollmann. An overview of nonparametric bayesian models and applications to natural language processing. *Science*, pages 71–93, 2008. URL `http://www.dcs.shef.ac.uk/intranet/teaching/projects/archive/msc2001/pdf/m0sk.pdf`.

[6] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.

[7] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006. URL `http://www.aclweb.org/anthology/P/P06/P06-1124`.