Bayesian Analysis of the Poisson-Dirichlet Process

Wray Buntine and Marcus Hutter
National ICT Australia (NICTA)

16th October 2008

Pitman-Yor and Dirichlet Processes

- Two-Parameter Poisson-Dirichlet Distribution proposed by Pitman and Yor in 1997, denoted $\mathrm{PD}(a, b)$.
- Usually embedded in the Two-Parameter Poisson-Dirichlet Process, denoted $\mathrm{PDP}(a, b)$.
- Subsequently called Pitman-Yor Process.
- Parameters usually are α, θ, but we use a, b.
- When first parameter (a or α) is zero, reduces to the Dirichlet Process (DP).
- Used for so-called non-parametric Bayesian analysis, since they are infinite dimensional distributions.
- At NIPS is considered double-plus good.

Typical Application: Clustering

- Data is sequence $\vec{y}_1, \vec{y}_2, \ldots$.
- Assume clusters have means distributed apriori as Gaussian$(0, \vec{I})$.
- Mixture model places $\vec{y}_i$ in cluster k_i: Can make the mixture probabilities PD(a, b), thus allowing infinite many clusters.
- Or place $\vec{y}_i$ with mean $\vec{\mu}_i$, where $\vec{\mu}_i$ are PDP $\left(a, b, \text{Gaussian}(0, \vec{I})\right)$.
    - The PDP will intrinsically do the mixture model for you.

Typical Application: Language Model

Word probability models:

- Data is sequence of English words, "from", "apple", "to", "from", "from", "cat", "to", ...
- Model the word probabilities as PD(a, b), thus allowing potentially infinite many words.

Variable n-gram models:

- Data is English text tokenised at spaces, "The quick brown fox jumped over the lazy dog ..."
- Model the word context (previous n words) be a tree with context frequency at each node. Place a single word probability model (as above) at each node as well.
- Thus combined model gives word probabilities for each context.

Our Motivation

- PDP's and DPs are usually defined procedurally (with sampling or sorting rules) or axiomatically.
- We would like to better understand what sort of prior they are when used for non-parametric Bayesian inference, *e.g.* for "infinite" dimensional clustering.
- When used in language modelling, we would like more efficient inference algorithms.

Model Family

Consider the distributional formula:

$$\begin{aligned} p(k) &= p_k \qquad \text{where } \sum_{k=1}^{\infty} p_k = 1, \\ p(X|k) &= \delta_{X_k}(\cdot) \end{aligned}$$

where $\delta_{X_k}(\cdot)$ is a discrete measure concentrated at X_k. We assume the values $X_k \in \mathcal{X}$ are independently and identically distributed according to same base measure $H(\cdot)$.

- We might model a sequence of X values generated.
- We might model just a sequence of k values.
- We might model the equivalence classes in a sequence of k's.
- We might model the p_k themselves.

Model Family, cont.

Assume $H(\cdot)$ is the uniform distribution on the unit interval $[0,1]$.

- We might model a sequence of X values generated.

  0.4674, 0.3925, 0.1937, 0.4674, 0.4674, 0.3947, 0.1937, ...,

- We might model just a sequence of k values.

  12, 435, 7198, 12, 12, 35, 7198, ...,

- We might model the equivalence classes in a sequence of k's.

  1, 2, 3, 1, 1, 4, 3, ...,

- We might model the p_k themselves (assuming $a=0$ and $b=2$),

  $p_1 = 3/9,\ p_2 = 1/9,\ p_3 = 2/9,\ p_4 = 1/9,$

Model Family, cont.

- Assume $H(\cdot)$ is the non-atomic distribution uniform on the unit interval $[0,1]$. Given a sequence of X values generated

  0.4674, 0.3925, 0.1937, 0.4674, 0.4674, 0.3947, 0.1937, ...,

  we can be sure the equivalence classes in a sequence of k's are

  1, 2, 3, 1, 1, 4, 3, ...,

- Assume $H(\cdot)$ is a discrete distribution on English words (so different X_k could be the same) given a sequence

  "from", "apple", "to", "from", "from", "cat", "to",...

  it could have any one of the following equivalence classes

  1, 2, 3, 1, 1, 4, 3, ...,
  1, 2, 3, 1, 4, 5, 3, ...,
  1, 2, 3, 1, 1, 4, 5, ...,
  1, 2, 3, 4, 5, 6, 3, ...,

Definitions

Definition. (Pitman and Yor, 1997) For $0 \leq a < 1$ and $b > -a$, suppose that a probability $P_{a,b}$ governs independent random variables $\tilde{Y}_k$ such that $\tilde{Y}_k$ has $\text{Beta}(1-a, b+k\,a)$ distribution. Let

$$\tilde{V}_1 \;=\; \tilde{Y}_1, \qquad \tilde{V}_k \;=\; (1-\tilde{Y}_1)\cdots(1-\tilde{Y}_{k-1})\tilde{Y}_k \qquad k \geq 2$$

and let $V_1 \geq V_2 \geq \cdots$ be the ranked (sorted) values of the $\tilde{V}_k$. Define the Poisson-Dirichlet distribution with parameters a, b, abbreviated $\text{PD}(a, b)$ to be the $P_{a,b}$ distribution of V_n.

Definition. (Ishwaran and James, 2001) Given a base probability (density) function $H(\cdot)$ on a measurable space $\mathcal{X}$, the Poisson-Dirichlet process with parameters a, b, abbreviated $\text{PDP}(a, b, H)$, is given by the discrete probability

$$\sum_{k=1}^{\infty} V_k \delta_{X_k}(\cdot)$$

where $\delta_{X_k}(\cdot)$ is a discrete measure concentrated at X_k and $\vec{V}$ is $\text{PD}(a, b)$.

References

- Pitman and Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 1997. $\longrightarrow$ Basic theory.
- Ishwaran and James, "Gibbs Sampling Methods for Stick Breaking Priors", *JASA*, 2001. $\longrightarrow$ General overview and history of PDP, and sampling methods. *De rigueur* citation at NIPS.
- Y.W. Teh, "A Hierarchical Bayesian Language Model based on Pitman-Yor Processes," *CCL and ACL*, Sydney 2006. $\longrightarrow$ Example of use in the discrete language domain.
- James, "Large sample asymptotics for the two-parameter Poisson-Dirichlet process," *IMS Collections*, 2008. $\longrightarrow$ Recent asymptotic theory.
- Mochihashi and Sumita, "The Infinite Markov Model", *NIPS 20*, 2008. $\longrightarrow$ Application to variable length *n*-gram models.

NB. There are also loads of good tutorials from the SML crowd.

Definition: Partition Model

Definition. A partition model is defined by a countably infinite sequence of probabilities $p_1, p_2, \ldots$ from an infinite-dimensional probability vector $\vec{p}$, where $\sum_{i=1}^{\infty} p_k = 1$. A sample of length N from the model is a sequence of indices $k_1, \ldots, k_N$ drawn i.i.d. according to probability $\vec{p}$.

Definition. The sample from the partition model induces a sample from non-atomic base distribution $H(\cdot)$ as follows. Let $X_k \sim H(\cdot)$ for $k = 1, ..., \infty$. Given the partition represented by indices $I_N = k_1, ..., k_N$, return the sequence $S_N = X_{k_1}, ..., X_{k_N}$.

Definition: Improper Dirichlet Prior

Definition. Given parameters (a, b), where $0 \leq a < 1$ and $b > -a$, define an improper prior (or unnormalised measure) on the parameters $\vec{p}$ of a partition model as follows. For any sub-vector $p_{k_1}, p_{k_2}, ..., p_{k_M}$, use the following measure:

$$p(p_{k_1}, p_{k_2}, ..., p_{k_M}, p_M^+) \;:=\; \left(p_M^+\right)^{b+M\,a} \prod_{m=1}^{M} p_{k_m}^{-a-1} \;,$$

where $p_M^+ = 1 - \sum_{m=1}^{M} p_{k_m}$.

NB. informally, this is a $\text{Dirichlet}_{M+1}(-a, -a, ..., -a, b + M\,a)$,
NB. the definition is consistent across change of variables to sub-vectors.