An Introduction to the Dirichlet Process and Nonparametric Bayesian Models

David Pfau

Columbia University

26 Apr 2010

Motivation for Nonparametric Bayes

Gaussian Mixture Model

- EM for GMM
- Bayesian GMM
- The Infinite Limit

Dirichlet Process

- Definition
- Stick Breaking Construction
- Pólya Urn Scheme and Chinese Restaurant Process
- Extensions: Pitman-Yor Process and Hierarchical DP

Many successful applications of Bayesian models:

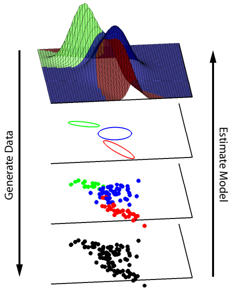
- Machine Learning
- Cognitive Science
- Theoretical Neuroscience?

But complex models have to be specified in advance. Not yet *fully* unsupervised learning.

For models with a fixed number of parameters (e.g. clustering, HMM) many ways to pick the optimal number of parameters:

$$\begin{aligned} AIC &= -2 \ln P(D|\hat{\Theta}_k) + 2k \\ BIC &= -2 \ln P(D|\hat{\Theta}_k) + k \ln |D| \\ P(D|\mathcal{M}_k) &= \int P(D|\Theta_k) P(\Theta_k|\mathcal{M}_k) d\Theta_k \end{aligned}$$

Different methods have different shortcomings.

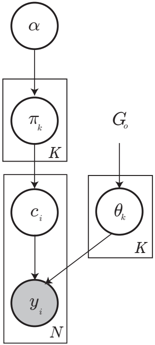


GMM

$$
\begin{aligned}
\theta_k &= \{\vec{\mu}_k, \Sigma_k\} \\
c_i|\vec{\pi} &\sim \text{Discrete}(\pi_1, \ldots, \pi_K) \\
\vec{y}_i|c_i = k, \Theta &\sim \text{Gaussian}(\theta_k)
\end{aligned}
$$

Expectation-Maximization

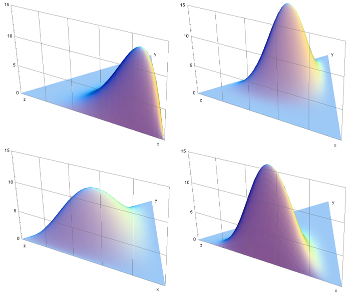
$$
\begin{array}{lll}
\text{E-step} & T_{i,k}^{(t)} = P(c_i = k|\vec{y}_i, \theta_k^{(t)}) \\
 & Q(\Theta, \vec{\pi}|\Theta^{(t)}, \vec{\pi}^{(t)}) = \mathbb{E}[\log L(\Theta, \vec{\pi}|\mathbf{y}, T^{(t)})] \\
\text{M-step} & (\Theta^{(t+1)}, \vec{\pi}^{(t+1)}) = \arg\max_{\Theta, \vec{\pi}} Q(\Theta, \vec{\pi}|\Theta^{(t)}, \vec{\pi}^{(t)})
\end{array}
$$



Bayesian GMM

$$
\begin{aligned}
\Sigma_k &\sim \mathrm{IW}_{\nu_0}(\Lambda_0^{-1}) \\
\vec{\mu}_k &\sim \mathrm{Gaussian}(\vec{\mu}_0, \Sigma_k/\kappa_0) \\
\vec{\pi}|\alpha &\sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \\
\theta_k &= \{\vec{\mu}_k, \Sigma_k\} \\
c_i|\vec{\pi} &\sim \mathrm{Discrete}(\pi_1, \ldots, \pi_K) \\
\vec{y}_i|c_i = k, \Theta &\sim \mathrm{Gaussian}(\theta_k)
\end{aligned}
$$

- Inference: sample posterior of c_i via MCMC
- Applied to spike sorting by Lewicki [1994].



[Source: Wikimedia Commons]

Dirichlet Distribution

$$\vec{\pi} \sim Dir(\vec{\alpha})$$

$$\vec{\alpha} = \alpha\vec{H}, \sum_{i=1}^{K} H_i = 1$$

$$\mathbb{E}[\vec{\pi}] = \vec{H}$$

$$\alpha \to \infty \Rightarrow \vec{\pi} \to \vec{H}$$

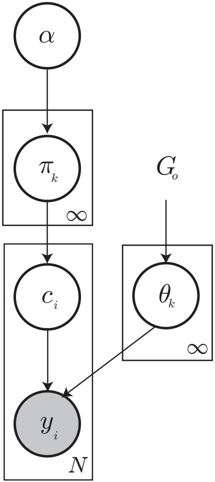
$$\alpha \to 0 \Rightarrow \vec{\pi} \text{ becomes sparse}$$

$$
\begin{aligned}
\vec{\pi} &\sim \mathrm{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \\
P(c_{i+1} = k | c_1, \ldots, c_i, \alpha) &= \int P(c_i + 1 = k | \vec{\pi}) p(\vec{\pi} | c_1, \ldots, c_i, \alpha) d\vec{\pi} \\
= \frac{\Gamma(\alpha + i)}{\prod_{j=1}^{K} \Gamma(\frac{\alpha}{K} + n_j)} & \quad \int \pi_1^{\frac{\alpha}{K} + n_1 - 1} \ldots \pi_k^{\frac{\alpha}{K} + n_k} \ldots \pi_K^{\frac{\alpha}{K} + n_K - 1} d\vec{\pi} \\
&= \frac{n_k + \frac{\alpha}{K}}{\alpha + i}
\end{aligned}
$$

Where n_k is the number of c_j, $j = 1, \ldots, i$ such that $c_j = k$. Order the clusters so $n_k > 0$ if $k \leq K_+$ and $n_k = 0$ if $k > K_+$. Then as $K \rightarrow \infty$

$$
P(c_{i+1} = k | c_1, \ldots, c_i, \alpha) = \begin{cases} \frac{n_k}{\alpha + i} & k \leq K_+ \\ \frac{\alpha}{\alpha + i} & k > K_+ \end{cases} .
$$

This is the *Chinese Restaurant Process*, $CRP(\alpha)$.



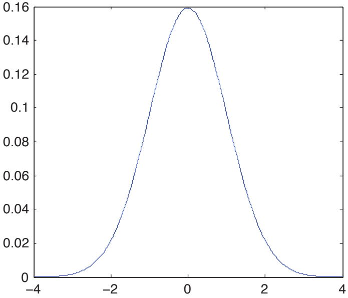
Infinite GMM

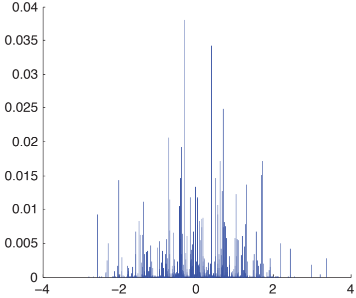
$$
\begin{aligned}
c_i | c_{1:i-1} &\sim CRP(\alpha) \\
\Sigma_k &\sim \mathrm{IW}_{\nu_0}(\Lambda_0^{-1}) \\
\vec{\mu}_k &\sim \mathrm{Gaussian}(\vec{\mu}_0, \Sigma_k / \kappa_0) \\
\theta_k &= \{\vec{\mu}_k, \Sigma_k\} \\
\vec{y}_i | c_i = k, \Theta &\sim \mathrm{Gaussian}(\theta_k)
\end{aligned}
$$

Special case of the Dirichlet Process Mixture Model, due to Rasmussen [2000]

Dirichlet Process: $\mathcal{G} \sim DP(\alpha, H)$

- α - concentration parameter
- H - base distribution
- $\mathcal{G}$ is *atomic:* $p(\theta|\mathcal{G}) = \sum_{k=1}^{\infty} \pi_k \delta(\theta - \theta_k)$





$H = \text{Gaussian}(0, 1)$ $\mathcal{G} \sim DP(100, H)$

Stick Breaking Construction

$$
\begin{aligned}
\pi'_k &\sim Beta(1, \alpha) \\
\pi_k &= \pi'_k \prod_{i=1}^{k-1} (1 - \pi_i) \\
\theta_k &\sim H \\
\mathcal{G} &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}
\end{aligned}
$$

Draws $x_{1:i} \sim \mathcal{G}$ cluster together. Let K_+ be the number of distinct values of $x_{1:i}$, n_k be the number of draws with value θ_k.

$$\mathcal{G}|x_{1:i} \sim DP\left(\alpha + i, \sum_{k=1}^{K_+} \frac{n_k}{\alpha + i}\delta_{\theta_k} + \frac{\alpha}{\alpha + i}H\right)$$

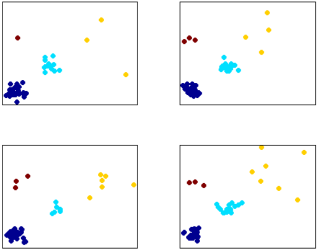
$$x_{i+1}|x_{1:i} \sim \sum_{k=1}^{K_+} \frac{n_k}{\alpha + i}\delta_{\theta_k} + \frac{\alpha}{\alpha + i}H$$

Pitman-Yor Process

$\mathcal{G} \sim PY(\alpha, d, H)$, $PY(\alpha, 0, H) \Leftrightarrow DP(\alpha, H)$

- $d \in [0, 1]$: discount
- Stick breaking construction: $\pi'_k \sim \text{Beta}(1 - d, c + kd)$
- CRP construction:

$$P(c_{i+1} = k | c_1, \ldots, c_i, \alpha, d) = \begin{cases} \frac{n_k - d}{\alpha + i} & k \leq K_+ \\ \frac{\alpha + kd}{\alpha + i} & k > K_+ \end{cases}.$$



Hierarchical Dirichlet Process

Share clusters across groups of data

$$
\begin{aligned}
\mathcal{G}_0 &\sim DP(\alpha, H)\\
\mathcal{G}_j &\sim DP(\alpha, \mathcal{G}_0)\\
\theta_{ji} &\sim \mathcal{G}_j
\end{aligned}
$$

Applications

- Infinite HMM - each $\mathcal{G}_j$ is the transition probability given a state [Teh et al., 2006].
- Variable-length Markov models for language data [Teh, 2006].

- Discrete time, discrete alphabet sequence learning
- Learn probabilistic deterministic finite automata
  - Subclass of HMMs
  - Intermediate between variable-length Markov models and full HMM
  - Use HDP as prior for transition matrix, similar to infinite HMM
  - Inference via Metropolis-Hastings
  - Works on small regular grammars ($\sim$7 states), currently extending to richer data

Nonparametric Bayesian models sidestep the model selection problem, combining model estimation and model selection into one. We define the Dirichlet Process and use it as a prior over parameters that controls the clustering of data. We show that the DP emerges in the limit of certain parametric models as the number of parameters goes to infinity. Draws from a DP can be marginalized out, yielding a tractable model that can be estimated by standard Bayesian methods. The DP can be extended in various ways, and we are applying these tools to discrete alphabet sequence learning with as few free parameters as possible.

Many thanks to:

- Frank Wood
- Liam Paninski
- Nick Bartlett

With support provided by NSF GRFP.

References:

M. S. Lewicki. Bayesian modeling and classification of neural signals. *Neural Computation*, 6:1005–1030, 1994.

C. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, Cambridge, MA, 2000.

Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the Association for Computational Linguistics*, pages 985–992, 2006.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.