HILBERT SPACE EMBEDDING FOR DIRICHLET PROCESS MIXTURES

Krikamol Muandet

Department of Empirical Inference
Max Planck Institute for Intelligent Systems
Tübingen, Germany

MOTIVATIONS

Bayesian nonparametrics

- Pros: Flexible, i.e., the complexity of the model is determined by the data.
- Cons: Exact inference is often intractable.
  - Markov chain Monte Carlo (Neal 2000).
  - variational Bayes (Blei and Jordan 2005, Kurihara 2007).
  - etc.

Kernel methods

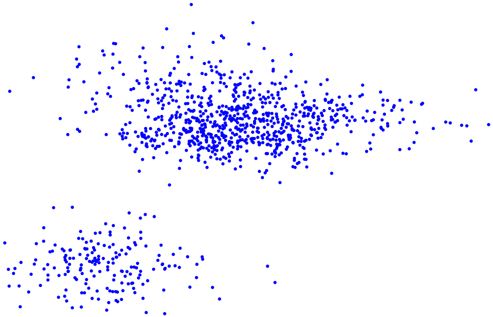
- Pros: The problem usually results in the convex optimization.
- Cons: The algorithm suffers from model selection/comparison, e.g., need cross validation to specify the right model complexity.

MOTIVATIONS

Model prior $\rightarrow$ Bayes' theorem $\rightarrow$ Inference

$$\underbrace{P(Y|X)}_{\text{posterior}} \propto \underbrace{P(X|Y)}_{\text{likelihood}} \underbrace{P(Y)}_{\text{prior}}$$

BAYESIAN CLUSTERING PROBLEM



$$x_i \sim \sum_{j=1}^{K} \pi_i \mathbb{P}_j, \; i = 1, \ldots, n$$

Key question: what is the right number of clusters, i.e., K?

Dilichlet Process
(Ferguson 1973)

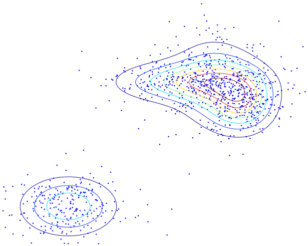
Definition

A Dirichlet Process is a distribution of a random probability measure G over a measurable space $(\Omega, \mathcal{B})$, such that for any finite partition $(A_1, \ldots, A_r)$ of Ω (i.e., $\Omega = \coprod_{i=1}^{r} A_i$, where $\coprod$ means disjoint union and $A_i \in \mathcal{B}$), we have

$$(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha G_0(A_1), \ldots, \alpha G_0(A_r))$$

where $G(A_i) = \int_{A_i} dG$ and $G_0(A_i) = \int_{A_i} dG_0$ for $i = 1, \ldots, r$.

Dirichlet Process Mixtures



The DP mixture model can be summarized as follow:

$$P \sim DP(G_0, \alpha),\ \Theta_i \sim P,\ x_i|\theta_i \sim f(\cdot|\theta_i)$$

where θ_i is a latent variable that parametrizes the distribution of an observed data points. For example,

$$x_i|\theta_i \sim f(\cdot|\theta_i = \{m, \Sigma\}) = \mathcal{N}(\cdot|m, \Sigma)$$

STICK-BREAKING CONSTRUCTION

(SETHURAMAN 1994)

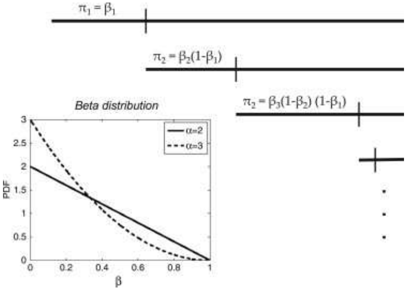
$$\beta_i \sim \text{Beta}(1, \alpha)$$

$$\pi_i = \beta_i \prod_{k=1}^{i-1} (1 - \beta_k)$$

$$\theta_i \sim G_0$$

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

c



Theorem

The stick breaking construction gives the same probability measure over all random measures on the measurable space $(\Omega, \mathcal{B})$ with the Dirichlet Process with same parameter α and G_0.

Hilbert Space Embedding for DPM

The Dirichlet Process Mixture Embedding (DPME) is defined as

$$\begin{aligned} \Upsilon \; : \; \mathfrak{P}_{\alpha,\Theta} &\longrightarrow \mathcal{H}_k \\ \mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta}} &\longmapsto \int k(x,\cdot)\,\mathrm{d}\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta}}(x) \triangleq \sum_{i=1}^{\infty} \pi_i \int k(x,\cdot)\,\mathrm{d}f_{\theta_i}(x) \end{aligned}$$

$\mathfrak{P}_{\alpha,\Theta}$	a space of all Dirichlet Process mixture model.
$\mathcal{H}_k$	a reproducing kernel Hilbert space (RKHS) with reproducing kernel k.
$\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta}}$	a Dirichlet Process mixture model $\sum_{i=1}^{\infty} \pi_i f_{\theta_i}(x)$.
f_{θ_i}	a density function such that $f_{\theta_i}(\cdot) \geq 0$ and $\int \mathrm{d}f_{\theta_i} = 1$.

Hilbert Space Embedding for DPM

- Ishwaran and James (2001) made an important observation that a truncation of the stick-breaking representation at a sufficiently large T already provides an excellent approximation to the full DPMM model.
- As a result, we propose the *truncated Dirichlet Process Mixture Embedding* (tDPME):

$$\begin{aligned} \Upsilon \; : \; & \mathfrak{P}_{\alpha,\Theta,T} \longrightarrow \mathcal{H}_k \\ & \mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T} \longmapsto \int k(x,\cdot)\,\mathrm{d}\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}(x) \triangleq \sum_{i=1}^{T} \pi_i \int k(x,\cdot)\,\mathrm{d}f_{\theta_i}(x) \end{aligned}$$

Almost-sure Truncation

Theorem

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) with a reproducing kernel k. Assume that $\|k(x,\cdot)\|_{\mathcal{H}}^2 \leq R$ for all x. The following inequality holds:

$$\left\|\Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta}}] - \Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}]\right\|_{\mathcal{H}}^2 \leq R \cdot \exp\left(-T/\alpha\right)$$

where C is an arbitrary constant.

Proof.

$$\begin{aligned}\|\Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta}}] - \Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}]\|_{\mathcal{H}}^2 &= \left\|\sum_{i=T+1}^{\infty} \pi_i \int k(x,\cdot)\mathrm{d}f_{\theta_i}(x)\right\|_{\mathcal{H}}^2 \\ &= R\left(1 - \sum_{i=1}^{T}\pi_i\right) \approx R \cdot \exp\left(-\frac{T}{\alpha}\right)\end{aligned}$$

ALMOST-SURE TRUNCATION

- With sufficiently large truncation level T, the error is small.
- The truncated DPME can be used as a surrogate to the true DPME.
- The bound also suggests how to choose the truncation level T. That is, for an error to be smaller than δ, one must have

$$T > -\alpha \log \left(\frac{\delta}{R} \right)$$

Optimization
(Song et al. 2008)

Given a truncated DPME $\Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}]$ and observation $x_1, x_2, \ldots, x_n$, we learn $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ by solving the following optimization problem:

$$\min_{\boldsymbol{\pi},\boldsymbol{\theta}} \|\widehat{\mu}_X - \Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}]\|_{\mathcal{H}}^2 \quad \text{subject to } \boldsymbol{\pi}^\top \mathbf{1} = 1, \pi_i \geq 0$$

To prevent overfitting, we introduce a regularizer $\Omega(\boldsymbol{\pi}) = \frac{1}{2}\|\boldsymbol{\pi}\|^2$ with a regularization constant $\varepsilon > 0$. Substituting $\widehat{\mu}_X$ and $\Upsilon[\mathbb{P}_{\boldsymbol{\pi},\boldsymbol{\theta},T}]$ back yields a quadratic programming (QP) for $\boldsymbol{\pi}$:

$$\min_{\boldsymbol{\pi}} \frac{1}{2}\boldsymbol{\pi}^\top(\mathbf{S} + \varepsilon \mathbf{I})\boldsymbol{\pi} - \mathbf{R}^\top \boldsymbol{\pi} \quad \text{subject to } \boldsymbol{\pi}^\top \mathbf{1} = 1, \pi_i \geq 0$$

where $\mathbf{S}_{ij} = \langle \mu[f_{\theta_i}], \mu[f_{\theta_j}]\rangle_{\mathcal{H}}$ and $\mathbf{R}_j = \langle \widehat{\mu}_X, \mu[f_{\theta_j}]\rangle_{\mathcal{H}}$.

OPTIMIZATION
(SONG ET AL. 2008)

1. Set α, δ and estimate T.
2. Do until convergence
   2.1 Optimize the mixing proportion $\boldsymbol{\pi}$ via quadratic programming (QP).
   2.2 Optimize the parameters $\boldsymbol{\theta}$ via constraint optimization.
3. Cluster the data points according to the resulting mixture model.

Demo

CONCLUSIONS & DISCUSSIONS

Conclusion

- the conjunction between Bayesian nonparametrics and kernel methods.
  - the Hilbert space embedding of the Dirichlet Process mixtures.

Open questions

- How to avoid truncation?
- Is the solution of DPME related to ML/MAP solutions?
- How to choose the kernel k?
- Kernel methods and random measures.

Acknowledgement

Philipp Hennig and Francis Bach

Many thanks!

krikamol@tuebingen.mpg.de

Bibliography I

D. M. Blei and M. I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):pp. 209–230, 1973. ISSN 00905364.

H. Ishwaran and James. Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, pages 161–173, Mar. 2001. ISSN 0162-1459.

K. Kurihara. Collapsed variational dirichlet process mixture models. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI07*, 2007.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, 9(2): 249–265, 2000.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th international conference on Machine learning*, ICML ’08, pages 992–999, New York, NY, USA, 2008. ACM.