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

\[
P(Y|X) \propto P(X|Y) \cdot P(Y)
\]

- Posterior
- Likelihood
- 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., \( K \)?
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 $\Omega$ (i.e., $\Omega = \bigsqcup_{i=1}^{r} A_i$, where $\bigsqcup$ means disjoint union and $A_i \in \mathcal{B}$), we have

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

where $G(A_i) = \int_{A_i} dG$ and $G_0(A_i) = \int_{A_i} dG_0$ for $i = 1, \ldots, r$. 
The DP mixture model can be summarized as follow:

\[ P \sim DP(G_0, \alpha), \quad \Theta_i \sim P, \quad x_i|\theta_i \sim f(\cdot|\theta_i) \]

where \( \theta_i \) 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)*

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

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

\[ \theta_i \sim G_0 \]

\[ G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \cdot \]

**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 \(\alpha\) and \(G_0\).
**Hilbert Space Embedding for DPM**

The Dirichlet Process Mixture Embedding (DPME) is defined as

\[ \mathcal{Y} : \mathcal{P}_{\alpha, \Theta} \rightarrow \mathcal{H}_k \]

\[ \mathbb{P}_{\pi, \theta} \mapsto \int k(x, \cdot) d\mathbb{P}_{\pi, \theta}(x) \triangleq \sum_{i=1}^{\infty} \pi_i \int k(x, \cdot) df_{\theta_i}(x) \]

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

$$
\mathcal{T} : \mathcal{P}_{\alpha, \Theta, T} \rightarrow \mathcal{H}_k
$$

$$
\mathbb{P}_{\pi, \theta, T} \mapsto \int k(x, \cdot) \, d\mathbb{P}_{\pi, \theta, T}(x) \triangleq \sum_{i=1}^{T} \pi_i \int k(x, \cdot) \, d\theta_i(x)
$$
**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:

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

where \( C \) is an arbitrary constant.

**Proof.**

\[
\| \gamma[\mathbb{P}_{\pi, \theta}] - \gamma[\mathbb{P}_{\pi, \theta, T}] \|_{\mathcal{H}}^2 = \left\| \sum_{i=T+1}^{\infty} \pi_i \int k(x, \cdot)df_{\theta_i}(x) \right\|_{\mathcal{H}}^2
\]

\[
= R \left( 1 - \sum_{i=1}^{T} \pi_i \right) \approx R \cdot \exp \left( -\frac{T}{\alpha} \right)
\]
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 $\delta$, one must have

$$T > -\alpha \log \left( \frac{\delta}{R} \right)$$
Given a truncated DPME $\gamma[\mathbb{P}_{\pi, \theta}, T]$ and observation $x_1, x_2, \ldots, x_n$, we learn $\pi$ and $\theta$ by solving the following optimization problem:

$$\min_{\pi, \theta} \|\hat{\mu}_X - \gamma[\mathbb{P}_{\pi, \theta}, T]\|_H^2 \quad \text{subject to } \pi^T 1 = 1, \pi_i \geq 0$$

To prevent overfitting, we introduce a regularizer $\Omega(\pi) = \frac{1}{2} \|\pi\|_2^2$ with a regularization constant $\varepsilon > 0$. Substituting $\hat{\mu}_X$ and $\gamma[\mathbb{P}_{\pi, \theta}, T]$ back yields a quadratic programming (QP) for $\pi$:

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

where $S_{ij} = \langle \mu[f_{\theta_i}], \mu[f_{\theta_j}] \rangle_H$ and $R_j = \langle \hat{\mu}_X, \mu[f_{\theta_j}] \rangle_H$. 

12
1. Set $\alpha$, $\delta$ and estimate $T$.

2. Do until convergence
   2.1 Optimize the mixing proportion $\pi$ via quadratic programming (QP).
   2.2 Optimize the parameters $\theta$ 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


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

