Dirichlet process mixture models

Theory and clustering applications

Erika Lena, Francesco Ortu

Overview

1. Introduction
2. Dirichlet Process
3. Dirichlet Process Mixture Models
4. Application to image segmentation
5. Conclusions

Introduction

Introduction

Non-parametric Bayesian clustering

The Dirichlet process mixture model is a Bayesian non parametric method, widely used in clustering tasks.

Contrary to Gaussian mixture models, in DPMM we do not specify the number of clusters in advance, and this guarantees more flexibiliy to the model, especially when the geometry of the data is unclear.

Dirichlet Process

Dirichlet Process
Definition

Let be $(\mathcal{X}, \mathcal{B})$ a measurable space, G_0 a (probability) measure, $\alpha \in \mathbb{R}_+$.
A *Dirichlet process* $G \sim DP(G_0, \alpha)$ is a stochastic process indexed by $\mathcal{B}$ such that for any $\{B_i\}_{i=1}^n$ partition of $\mathcal{X}$:

$$\Big(G(B_1), \dots, G(B_n)\Big) \sim Dir\Big(\alpha G_0(B_1), \dots, \alpha G_0(B_n)\Big)$$

In other words, a *Dirichlet process* is a measure over measure parameterized by αG_0. [3] [2]

Dirichlet Process

Polya Urn

Suppose to have an infinite set of colors and an Urn. This algorithm generate samples from a Dirichlet process [1]:

- At the start we have α black ball and one colored (whose "color" is randomly selected according to G_0)
- for $n > 1$ we draw a ball from the urn:
    1. if the ball is black we returned it to the urn along with a ball of a new color randomly selected according to G0;
    2. if the ball is colored we returned it to the urn along with a ball with the same color.

Dirichlet Process

The cluster propriety

The polya urn scheme lead us to the cluster propriety of the Dirichlet process. In fact, let be $\theta_1, \theta_2 ...$ a draw of $\theta \sim G$ where $G \sim DP(G_0, \alpha)$. Then we have

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \sum_{i=1}^{n} \frac{1}{\alpha + n} \delta_{\theta_i} + \frac{\alpha}{\alpha + n} G_0$$

There is a positive probability that each θ_i will take one the value of another θ_j, leading some of the variables to share values. In Addition, greater α means that, for θ_n is more likely to have a newly draw from G_0 rather than take on one of the values from $\theta_{1:n-1}$.

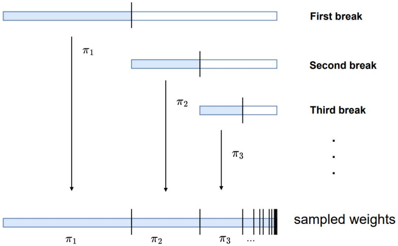
Dirichlet Process

The stick-breaking construction

Another algorithm to generate samples form a Dirichlet process is the *stick-breaking process* [5]. This involves repeatedly breaking off and discarding a random fraction of a "stick" that is initially of length 1. We obtain the probability mass function for $\theta \sim G$:

$$f(\theta) = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}(\theta) \qquad \theta_i \overset{iid}{\sim} G_0$$

$$\pi_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j) \qquad \beta_j \overset{iid}{\sim} Beta(1, \alpha)$$



Dirichlet Process Mixture Models

Dirichlet Process Mixture Models

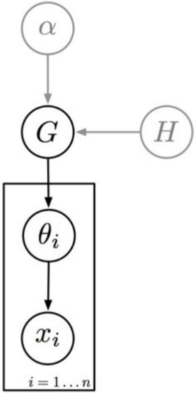
The most common application of the DPs in the using *DP mixture models*, used to do clustering.

$$G|\alpha, G_0 \sim DP(\alpha, G_0)$$

$$\theta_i|G \sim G$$

$$x_i|\theta_i \sim F(\theta_i)$$

When clustering data generate from Gaussian distribution we have a $F = \mathcal{N}(\theta)$.

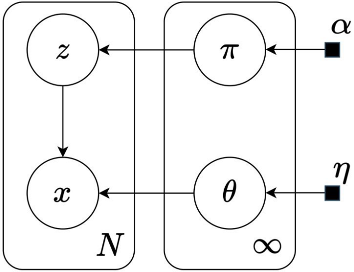


Dirichlet Process Mixture Models

From the equivalence of the stick breaking construction for Dirichlet process we used the following model:

$$\pi|\alpha \sim GEM(\alpha) \qquad z_i|\pi \sim Categorical(\pi)$$

$$\theta_k|\eta \sim G_0 = \mathcal{N}(\eta) \qquad x_i|z_i, \theta_{z_i} \sim \mathcal{N}(\theta_{z_i}, 1)$$



Application to image segmentation

Application to image segmentation

Pyro implementation of previous DPMM:

```
def model(data):
    with pyro.plate("beta_plate", K-1):
        beta = pyro.sample("beta", Beta(1, alpha))

    with pyro.plate("mu_plate", K):
        mu = pyro.sample("mu", Normal(0., 5.))

    with pyro.plate("data", N):
        z = pyro.sample("z", Categorical(mix_weights(beta)))
        pyro.sample("obs", Normal(mu[z], 1.), obs=data)
```

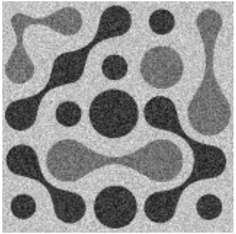
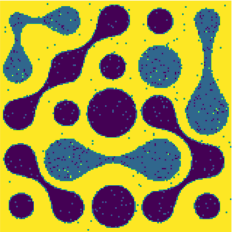
Application to image segmentation

```
Sample: 100%|██████████| 300/300 [14:48,  2.96s/it, step size=1.24e-01, acc. prob=0.833]
```

	mean	std	median	5.0%	95.0%	n_eff	r_hat
beta[0]	0.32	0.00	0.32	0.31	0.32	194.80	1.01
beta[1]	1.00	0.00	1.00	1.00	1.00	102.30	1.00
beta[2]	0.89	0.22	1.00	0.53	1.00	101.13	0.99
beta[3]	0.92	0.19	1.00	0.79	1.00	89.15	0.99
beta[4]	0.89	0.23	1.00	0.49	1.00	145.30	0.99
beta[5]	0.90	0.20	1.00	0.58	1.00	120.45	0.99
beta[6]	0.93	0.15	0.99	0.76	1.00	86.05	0.99
mu[0]	6.36	0.02	6.36	6.33	6.39	111.38	0.99
mu[1]	1.98	0.01	1.98	1.96	1.99	175.31	0.99
mu[2]	0.71	4.81	0.54	-6.39	9.08	239.16	0.99
mu[3]	-0.57	5.10	-0.06	-8.24	7.35	53.18	1.03
mu[4]	0.45	5.38	-0.13	-7.58	8.40	84.13	1.02
mu[5]	-0.11	6.85	-0.96	-11.67	8.61	102.27	0.99
mu[6]	-0.27	5.41	-0.10	-9.23	8.69	80.17	1.01
mu[7]	0.24	5.34	0.15	-8.98	8.53	194.19	0.99

```
Number of divergences: 0
```

Application to image segmentation

Application to image segmentation

Conclusions

Conclusions and possible improvements

- It is possible to extend the model using Markov random fields in order to exploit spatial information. [4]
- We could use Variational Inference instead MCMC to reduce the computation time.
- It is possible to better investigate the role of best hyperparameter α. []

References I

D. Blackwell and J. B. MacQueen.
Ferguson Distributions Via Polya Urn Schemes.
The Annals of Statistics, 1(2):353 – 355, 1973.

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

B. A. Frigyik, A. Kapila, and M. R. Gupta.
Introduction to the dirichlet distribution and related processes.
Technical Report UWEETR-2010-0006, Department of Electrical Engineering, University of Washington, 2010.

References II

P. Orbanz and J. Buhmann.
Nonparametric bayesian image segmentation.
International Journal of Computer Vision, 77:25–45, 05 2008.

J. Sethuraman.
A constructive definition of dirichlet priors.
Statistica sinica, pages 639–650, 1994.