Non-Parametric Bayes

Mark Schmidt

UBC Machine Learning Reading Group

January 2016
Bayesian learning includes:

- Gaussian processes.
- Approximate inference.
- Bayesian nonparametrics.
Motivation: Choosing Number of Mixture Components

Consider density estimation with mixture of Gaussians:

How many clusters should we use?
Motivation: Choosing Number of Mixture Components

Consider density estimation with mixture of Gaussians:

How many clusters should we use?

Standard approach:

1. Try out a bunch of different values for number of clusters.
2. Use a model selection criterion to decide (BIC, cross-validation, etc.).
Motivation: Choosing Number of Mixture Components

Consider density estimation with mixture of Gaussians:

How many clusters should we use?

Bayesian non-parametric approach:

- Fit a single model where number of clusters adapts to data.
Motivation: Choosing Number of Mixture Components

Consider density estimation with mixture of Gaussians:

How many clusters should we use?

Bayesian non-parametric approach:
- Fit a single model where number of clusters adapts to data.
- Number of clusters increases with dataset size.
- Standard **Gaussian mixture model** with \( k \) mixtures.

\[
x^i | z^i = c, \theta_c \sim \mathcal{N}(\mu_c, \Sigma_c), \quad z^i \sim \text{Cat}(\theta_1, \theta_2, \ldots, \theta_k),
\]
Finite Mixture Models

- Standard **Gaussian mixture model** with $k$ mixtures.

$$x^i | z^i = c, \theta_c \sim N(\mu_c, \Sigma_c), \quad z^i \sim \text{Cat}(\theta_1, \theta_2, \ldots, \theta_k),$$

- The conjugate prior to the categorical distribution

$$p(z^i = c | \theta) = \theta_c,$$

is the **Dirichlet distribution**,

$$p(\theta | \alpha) \propto \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\ldots\theta_k^{\alpha_k-1}.$$

- We can think of Dirichlet as **distribution over probabilities** of $k$ variables.
Finite Mixture Models

- **Standard Gaussian mixture model** with $k$ mixtures.

$$x^i | z^i = c, \theta_c \sim \mathcal{N}(\mu_c, \Sigma_c), \quad z^i \sim \text{Cat}(\theta_1, \theta_2, \ldots, \theta_k),$$

- The conjugate prior to the categorical distribution

$$p(z^i = c | \theta) = \theta_c,$$

is the **Dirichlet distribution**,

$$p(\theta | \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

- We can think of Dirichlet as distribution over probabilities of $k$ variables.
- With this and MCMC/variational inference, we can do the usual Bayesian stuff.
- However, this model requires us to **pre-specify** $k$. \/
We don’t want to pre-specify \( k \).

Naive approach:

- Put a prior over \( k \).
- Work with posterior over \( k, \theta \), and mixture parameters.
Infinite Mixture Models

- We don’t want to pre-specify $k$.
- Naive approach:
  - Put a prior over $k$.
  - Work with posterior over $k$, $\theta$, and mixture parameters.
- Challenges:
  - Do we have to fit a model for every $k$?
  - For $k' < k$, posterior are defined over different spaces (needs reversible-jump MCMC).
We don’t want to pre-specify $k$.

Naive approach:
- Put a prior over $k$.
- Work with posterior over $k$, $\theta$, and mixture parameters.

Challenges:
- Do we have to fit a model for every $k$?
- For $k' < k$, posterior are defined over different spaces (needs reversible-jump MCMC).

Non-parametric Bayesian approach:
- Assume $k = \infty$, but only a finite number were used to generate data.
Infinite Mixture Models

- We don’t want to pre-specify $k$.

- Naive approach:
  - Put a prior over $k$.
  - Work with posterior over $k$, $\theta$, and mixture parameters.

- Challenges:
  - Do we have to fit a model for every $k$?
  - For $k' < k$, posterior are defined over different spaces (needs reversible-jump MCMC).

- Non-parametric Bayesian approach:
  - Assume $k = \infty$, but only a finite number were used to generate data.
  - Posterior will contain assignments of points to these clusters.
  - Posterior predictive can assign point to new cluster.
Recall that **stochastic process** is an infinite collection of random variables.

**Gaussian process**: “infinite-dimensional” Gaussian.
- Process is defined by mean function and covariance function.
- Useful non-parametric prior for continuous distributions.

**Dirichlet process**: “infinite-dimensional” Dirichlet.
- Process defined by concentration parameter $\alpha$.
- Useful non-parametric prior for categorical distributions.

Also called the *Chinese restaurant process*. 
Recall that **stochastic process** is an infinite collection of random variables.

**Gaussian process**: “infinite-dimensional” Gaussian.
- Process is defined by mean function and covariance function.
- Useful non-parametric prior for continuous distributions.

**Dirichlet process**: “infinite-dimensional” Dirichlet.
- Process defined by concentration parameter $\alpha$.
- Useful non-parametric prior for categorical distributions.
- Also called the **Chinese restaurant process**.
The first customer sits at their own table.
The first customer sits at their own table.

The second customer:
- Sits at a new table with probability \( \frac{\alpha}{1+\alpha} \).
- Sits at first table with probability \( \frac{1}{1+\alpha} \).
The first customer sits at their own table.

The second customer:
- Sits at a new table with probability $\frac{\alpha}{1+\alpha}$.
- Sits at first table with probability $\frac{1}{1+\alpha}$.

The $(n+1)$ customer:
- Sits at a new table with probability $\frac{\alpha}{n+\alpha}$.
- Sits at table $c$ with probability $\frac{n_c}{n+\alpha}$.
Chinese Restaurant Process

- At time $n$, defines probabilities over $k$ “tables” and all others,

$$\left( \frac{n_1}{n + \alpha}, \frac{n_2}{n + \alpha}, \ldots, \frac{n_k}{n + \alpha}, \frac{\alpha}{n + \alpha} \right).$$

- Higher concentration $\alpha$ means more occupied tables.
  - For large $n$ number of tables is $O(\alpha \log n)$.
  - We can put a hyper-prior on $\alpha$.

- A subtle issue is that the CRP is exchangeable:
  - Up to label switching, probabilities are unchanged if order of customers is changed.

- An equivalent view of Dirichlet/Chinese-restaurant process is the “stick-breaking” process.
Dirichlet Process Mixture Models

- Standard finite Gaussian mixture likelihood (fixed variance $\Sigma$)

$$p(x|\Sigma, \theta, \mu_1, \mu_2, \ldots, \mu_k) = \sum_{c=1}^{k} \theta_c p(x|\mu_c, \Sigma),$$

where we might assume $\theta$ comes from a Dirichlet distribution.
Dirichlet Process Mixture Models

- Standard finite Gaussian mixture likelihood (fixed variance $\Sigma$)

$$p(x|\Sigma, \theta, \mu_1, \mu_2, \ldots, \mu_k) = \sum_{c=1}^{k} \theta_c p(x|\mu_c, \Sigma),$$

where we might assume $\theta$ comes from a Dirichlet distribution.

- Infinite Gaussian mixture likelihood,

$$p(x|\Sigma, \theta, \mu_1, \mu_2, \ldots) = \sum_{c=1}^{\infty} \theta_c p(x|\mu_c, \Sigma),$$

where we might assume $\theta$ comes from a Dirichlet process.

- So the DP gives us the non-zero $\theta_c$ values.

- In practice, variational/MCMC inference methods used.

https://www.youtube.com/watch?v=0Vh7qZY9sPs
Non-parametric Bayes place priors over infinite-dimensional objects.

- Complexity of model grows with data.

Various extensions exist (some will be discussed next time):
- Latent Dirichlet allocation (topic models).
- Beta (Indian buffet) process (PCA and factor analysis).
- Hierarchical Dirichlet process.
- Pólya trees (generating trees).
- Infinite hidden Markov models (infinite number of hidden states).
Non-parametric Bayes place priors over infinite-dimensional objects.
  - Complexity of model grows with data.

Gaussian processes define prior over infinite-dimensional functions.

Dirichlet processes define prior over infinite-dimensional probabilities.
  - Interpretation in terms of Chinese restaurant process.

Allows us to fit mixture models without pre-specifying number of mixtures.
Summary

- Non-parametric Bayes place priors over infinite-dimensional objects.
  - Complexity of model grows with data.
- Gaussian processes define prior over infinite-dimensional functions.
- Dirichlet processes define prior over infinite-dimensional probabilities.
  - Interpretation in terms of Chinese restaurant process.
- Allows us to fit mixture models without pre-specifying number of mixtures.
- Various extensions exist (some will be discussed next time):
  - Latent Dirichlet allocation (topic models).
  - Beta (Indian buffet) process (PCA and factor analysis).
  - Hierarchical Dirichlet process.
  - Pólya trees (generating trees).
  - Infinite hidden Markov models (infinite number of hidden states).