This Lecture: Advanced Machine Learning

- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes
- Bayesian Estimation & Bayesian Non-Parametrics
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes
- SVMs and Structured Output Learning
  - SV Regression, SVDD
  - Large-margin Learning

Recap: Bayesian Mixture Models

- Let’s be Bayesian about mixture models
  - Place priors over our parameters
  - Again, introduce variable $z_n$ as indicator which component data point $x_n$ belongs to.
  - $z_n \sim \text{Multinomial}(\pi)$
  - $x_n | z_n = k, \mu_k, \Sigma_k \sim \mathcal{N}(\mu_k, \Sigma_k)$
  - Introduce conjugate priors over parameters
    - $\pi \sim \text{Dirichlet}(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$
    - $\mu_k, \Sigma_k \sim \mathcal{N} - \mathcal{W}(0, s, d, \phi)$

Recap: Mixture Models with Dirichlet Priors

- Integrating out the mixing proportions $\pi$
  - $p(z|\pi) = \int p(z|\pi)p(\pi|\alpha)d\pi$
    - $p(z|\pi) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(N_k + \alpha/K)}$
  - $\Gamma(N + \alpha) = \prod_{k=1}^{K} \Gamma(N_k + \alpha/K)$

- Conditional probabilities
  - Examine the conditional of $z_n$ given all other variables $z_{-n}$
    - $p(z_{nk} = 1 | z_{-n}, \alpha) = \frac{p(z_{nk} = 1, z_{-n} | \alpha)}{p(z_{-n} | \alpha)}$
      - $= \frac{N_{-n,k} + \alpha/K}{N - 1 + \alpha}$
    - The more populous a class is, the more likely it is to be joined!
Recap: Infinite Dirichlet Mixture Models

• Conditional probabilities: Finite $K$
  $$p(z_{nk} = 1|z_{\cdot n}, \alpha) = \frac{N_{n,k} + \alpha/K}{N - 1 + \alpha}, \quad N_{n,k} \text{ def } \sum_{i=1,i\neq n}^{N} z_{ik}$$

• Conditional probabilities: Infinite $K$
  - Taking the limit as $K \to \infty$ yields the conditionals
    $$p(z_{nk} = 1|z_{\cdot n}, \alpha) = \begin{cases} \frac{N_{n,k}}{N - 1 + \alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N - 1 + \alpha} & \text{if all } k \text{ not represented} \end{cases}$$
  - Left-over mass $\alpha \Rightarrow$ countably infinite number of indicator settings

Note

• Why this term if all $k$ are not represented?
  $$p(z_{nk} = 1|z_{\cdot n}, \alpha) = \begin{cases} \frac{N_{n,k}}{N - 1 + \alpha} & \text{if } k \text{ represented} \\ \frac{\alpha}{N - 1 + \alpha} & \text{if all } k \text{ not represented} \end{cases}$$
  - The total probability assigned to all unoccupied clusters is determined by the complement of existing cluster weights:
    $$\lim_{K \to \infty} p(z_{nk} = m|z_{\cdot n}, \alpha) = 1 - \sum_{k=1}^{K} \frac{N_{n,k}}{N - 1 + \alpha} = \frac{N - 1 + \alpha - (N - 1)}{N - 1 + \alpha} = \frac{\alpha}{N - 1 + \alpha}$$

Topics of This Lecture

• Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
• Dirichlet Processes
  - Motivation
  - Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-breaking construction
• Applying DPMMs
  - Efficient sampling
  - Applications

Recap: Gibbs Sampling

• Approach
  - MCMC-algorithm that is simple and widely applicable.
  - May be seen as a special case of Metropolis-Hastings.
• Idea
  - Sample variable-wise: replace $x_i$ by a value drawn from the distribution $p(z_i|x_i)$.
  - This means we update one coordinate at a time.
  - Repeat procedure either by cycling through all variables or by choosing the next variable.
• Properties
  - The algorithm always accepts!
  - Completely parameter free.
  - Can also be applied to subsets of variables.

Gibbs Sampling for Finite Mixtures

• We need approximate inference here
  - Gibbs Sampling: Conditionals are simple to compute
    - $p(z_{nk} = k|\text{other terms}) \propto \sum_{i \neq n} p(z_{nk} = k|z_{\cdot n}, \alpha)$
    - $p(x|z \sim \text{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K))$
    - $\mu_k, \Sigma_k|\text{others} \sim N - \text{IW}(\nu', s', d', d')$
**Gibbs Sampling for Finite Mixtures**

- We need approximate inference here
  - **Gibbs Sampling**: Conditionals are simple to compute
    \[ p(z_{ik} = k | \text{other data}) \propto \sum_{k=1}^{K} p(z_{ik} = k | \mu_k, \Sigma_k) \]
    \[ p(z_{ik} = k | \text{other data}) \propto N(z_i | \mu_k, \Sigma_k) \]
    \[ \pi \sim \text{Dir}(\alpha_1 / K, \ldots, \alpha_K / K) \]
    \[ \mu_k, \Sigma_k | \text{other data} \sim N - IW(\nu', \nu', \nu', \nu') \]
- However, this will be rather inefficient...
  - In each iteration, algorithm can only change the assignment for individual data points.
  - There are often groups of data points that are associated with high probability to the same component. \( \Rightarrow \) Unlikely that group is moved.
  - Better performance by collapsed Gibbs sampling which integrates out the parameters \( \pi, \mu, \Sigma \).

**Collapsed Finite Bayesian Mixture**

- More efficient algorithm
  - Conjugate priors allow analytic integration of some parameters
  - Resulting sampler operates on reduced space of cluster assignments (implicitly considers all possible cluster shapes)

**Procedure**

- The model implies the factorization
  \[ p(x | \alpha) \propto p(\pi | \alpha) p(x | \pi) \]
- Derive
  \[ p(z | \alpha) = \int p(z | \pi) p(\pi | \alpha) d\pi \]
  \[ p(x_n | z_n, H) = \int p(x_n | \theta_n, H) p(\theta_n | z_n, H) d\theta_n \]

\( \Rightarrow \) Conjugate prior, Normal - Inverse Wishart

**Collapsed Finite Mixture Sampler**

- **Algorithm**
  1. Sample a random permutation \( \pi(\cdot) \) of the integers \( \{1, \ldots, N\} \).
  2. Set \( z = z^{(1)} \). For each \( n \in \{1, \ldots, \pi(N)\} \), sequentially resample \( z_n \) as follows
    a) For each of the \( K \) clusters, determine the predictive likelihood (this can be computed from cached sufficient statistics)
    \[ p_i(x_n | z_n, H) = p(x_n | \pi(n) \neq m, H) \]
    b) Sample a new assignment \( z_n \) from the multinomial distribution
    \[ z_n \sim \frac{\sum_{k=1}^{K} \pi_k \sum_{j \neq m} p_i(x_n | z_n, H)}{\sum_{k=1}^{K} \sum_{j \neq m} \pi_k p_i(x_n | z_n, H)} \]
    c) Update cached sufficient statistics to reflect assignment \( z_n \).
  3. Set \( z^{(t)} = z \). Optionally, mixture parameters may be sampled via steps 2-3 of the standard finite mixture sampler.
Standard vs. Collapsed Samplers

- Theorem (Rao-Blackwell)
  "Analytical marginalization of some variables from a joint distribution always reduces the variance of later estimates."

Collapsed sampler converges much more quickly.

Discussion

- Collapsed Gibbs sampling
  Integrates out the parameters $\pi, \mu, \Sigma$.
  $p(z_k = 1 | oth.) \propto \frac{N_{z_k} + \alpha/K}{N - 1 + \alpha} p_k(x_n | \mu, \Sigma)$

- Properties
  - Can change all assignments in each iteration.
  - Able to move entire groups between clusters.
  - Faster convergence.
  - However, similar worst-case performance as standard sampler, may get stuck in local optima for many iterations.

Topics of This Lecture

- Gaussian Processes
  Gaussian Processes (GP) define a distribution over functions
  $f \sim \text{GP}(\mu, \Sigma)$
  where $\mu$ is the mean function and $\Sigma$ is the covariance function.
  $\Rightarrow$ We can think of GPs as "infinite-dimensional" Gaussians.

- Dirichlet Processes
  Dirichlet Processes (DP) define a distribution over distributions (a measure on measures)
  $G \sim \text{DP}(\nu | G_0, \alpha)$
  Where $\alpha > 0$ is a scaling parameter and $G_0$ is the base measure.
  $\Rightarrow$ We can think of DPs as "infinite-dimensional" Dirichlet distributions.

Sidenote: Bayesian Nonparametric Methods

- Bayesian Nonparametric Methods (BNPs)
  Both Gaussian Processes and Dirichlet Processes are examples of BNPs.

- What does that mean?
  - Nonparametric: does NOT mean there are no parameters!
  - It means (very roughly) that the number of parameters grows with the number of data points.

- Parametric methods:
  - Get data, build model, predict using model

- Nonparametric methods:
  - Get data, predict directly based on data

Dirichlet Processes

- Definition [Ferguson, 1973]
  Let $\Theta$ be a measurable space, $G_0$ be a probability measure on $\Theta$, and $\alpha$ a positive real number.
  For all $(A_1, \ldots, A_N)$ finite partitions of $\Theta$,
  $G \sim \text{DP}(\nu | G_0, \alpha)$
  means that
  $(G(A_1), \ldots, G(A_N)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_N))$

- Translation
  A random probability distribution $G$ on $\Theta$ is drawn from a Dirichlet Process if its measure on every finite partition follows a Dirichlet distribution.
**Definition** [Ferguson, 1973]

- Let \( \Theta \) be a measurable space, \( G_0 \) be a probability measure on \( \Theta \), and \( \alpha \) a positive real number.
- For all \((A_1, \ldots, A_K)\) finite partitions of \( \Theta \),
  \[ G \sim \text{DP}(G_0, \alpha) \]
  means that
  \[ (G(A_1), \ldots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K)) \]

**Important property** [Blackwell]
- Draws from a DP will always place all their mass on a countable set of points, the so-called atoms \( \delta_{\theta_i} \).
  \[ G(\theta) = \sum_{i=1}^{\infty} \frac{1}{\pi_i} \delta_{\theta_i} \]
  \[ \sum_{i=1}^{\infty} \pi_i = 1 \]
  where \( \delta_{\theta_i} \) is a Dirac delta at \( \theta_i \), and \( \theta_i \sim \text{G}(\cdot) \).
  \( \Rightarrow \) Samples from DP are discrete with probability one.

**Moments**
  \[ \mathbb{E}[G(A)] = G_0(A), \quad \text{var}[G(A)] = \frac{G_0(A)(1 - G_0(A))}{\alpha + 1} \]

**Sampling**
- Since \( G \) is a probability measure, we can draw samples from it
  \[ G \sim \text{DP}(G_0, \alpha) \]
  \[ \theta_1, \ldots, \theta_N | G \sim G \]
- Posterior of \( G \) given observations \( \theta_1, \ldots, \theta_N \)?
  - The usual Dirichlet-multinomial conjugacy carries over to the nonparametric DP as well. \( \Rightarrow \) Posterior is again a DP.
  \[ G|\theta_1, \ldots, \theta_N \sim \text{DP} \left( \alpha + N, \frac{\alpha G_0 + \sum_{i=1}^{N} \delta_{\theta_i}}{\alpha + N} \right) \]

**Existence of Dirichlet Processes**

- **Summary so far**
  - A probability measure is a function from subsets of a space \( \Theta \) to \([0,1]\) satisfying certain properties.
  - A DP is a distribution over probability measures such that marginals on finite partitions are Dirichlet distributed.
- **How do we know that such an object exists?**
  - Kolmogorov Consistency Theorem: If we can prescribe consistent finite dimensional distributions, then a distribution over functions exists.
  - De Finetti’s Theorem: If we have an infinite exchangeable sequence of random variables, then a distribution over measures exists making them independent.
  - Pólya’s urn, Chinese Restaurant Process
  - Stick-Breaking Construction: Just construct it.

**Properties**

- **Summary so far**
  - We have seen some of the formal properties of DPs.
  - But how can we use them?
  - How can we sample from them?
  - In the following, we will characterize DPs through several different constructions in order to highlight key properties...
- **Constructions**
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking Construction
Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

Dirichlet Process Mixture Models

- During this lecture, we will use the following two forms for DPMMs...

- Indicator variable representation
  - Form of an infinite mixture model
  - The DP is implicit through the choice of priors
  - We will use this form whenever we want to make the assignment of points to clusters explicit (⇒ use for clustering).

- Distributional form
  - Explicit representation of the DP through the node $G$.
  - Useful when we want to use the DPMM’s predictive distribution.

Dirichlet Process Mixture Models

- Base distribution $G_0$
- Infinite discrete distribution on $G_0$, defines the clusters
- Parameters of the cluster that generates $x$, given the cluster
- Likelihood of $x$, given the cluster

Pólya’s Urns

[Blackwell & MacQueen, 1973]

- Can we sample observations without constructing $G$?
  - $G \sim \text{DP}(G_0, \alpha)$
  - $\theta_i \sim G$
- Yes, by a variation of the classical balls-in-urns analogy
  - Assume that $G_0$ is a distribution over colors, and that each $\theta_i$ represents the color of a single ball placed in the urn.
  - Start with an empty urn. Repeat for $N$ steps:
    1. With probability proportional to $\alpha$, draw $\theta_i \sim G_0$ and add a ball of that color to the urn.
    2. With probability proportional to $n - 1$ (i.e., the number of balls currently in the urn), pick a ball at random from the urn. Record its color as $\theta_i$ and return the ball into the urn, along with a new one of the same color.
Pólya’s Urns: Discussion

- Pólya Urn scheme
  - Simple generative process for the predictive distribution of a DP
  - Consider a set of $N$ observations $\tilde{x}_i \sim G$ taking $K$ distinct values $(\theta_k)_{k=1}^K$. The predictive distribution of the next observation is then
  $$p(\tilde{x}_{N+1} = \theta | \tilde{x}_1, \ldots, \tilde{x}_N, \alpha, H) = \frac{\alpha H(\theta) + \sum_{k=1}^K N_k \delta(\theta, \theta_k)}{N + \alpha}$$

- Remarks
  - This procedure can be used to sample observations from a DP without explicitly constructing the underlying mixture.
  - DPs lead to simple predictive distributions that can be evaluated by caching the number of previous observations taking each distinct value.

De Finetti’s Theorem

- Theorem
  - For any infinitely exchangeable sequence of random variables $\{x_i\}_{i=1}^\infty$, $x_i \in X$, there exists some space $\Theta$ of probability measures and corresponding distribution $P(\theta)$ such that the joint probability of any $N$ observations has a mixture representation
  $$p(x_1, x_2, \ldots, x_N) = \int \prod_{n=1}^N p(x_n | \theta) dP(\theta)$$

- Interpretation
  - If you assert exchangeability, it is reasonable to act as if there is an underlying parameter, there is a prior on this parameter, and the data are i.i.d. given that parameter.
  - In order for this to work, we need to allow $\theta$ to range over $\Theta$ in which case $P(\theta)$ is a distribution over measures.
  - As we know, the Dirichlet Process is a distribution on measures!

Pólya Urn Scheme

- Existence proof for DP
  - Starting with a DP, we constructed Pólya’s urn scheme.
  - The reverse is possible using De Finetti’s theorem:
  - Since the $\theta_i$ are i.i.d. $\sim G$, their joint distribution is invariant to permutations, thus $\theta_1, \theta_2, \ldots$ are exchangeable.
  - Thus a distribution over measures must exist making them i.i.d.
  - This is the DP.

- We have just proven that DPs exist
  - Hooray!
  - Now, let’s move on to see how we can use them...

Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Dirichlet Process Mixture Models
  - Pólya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction
- Applying DPMs
  - Efficient sampling
  - Applications

Sidenote on Partitions

- Problem with partitions
  - If our goal is clustering, the output grouping is defined by an assignment of indicator variables
  $$z_n \sim \text{Mult}(\pi) \quad \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$$
  - The number of ways of assigning $N$ data points to $K$ mixtures is $K^N$.
  - If $K \geq N$, this is much larger than the number of ways of partitioning the data!
  - Example: $N = 5$: 52 partitions vs. $5^5 = 3125$

\Rightarrow Need representation that is invariant to relabeling!
Chinese Restaurant Process (CRP)

- How can DPs support clustering?
- Chinese Restaurant Process
  - Visualize clustering as a sequential process of customers sitting at tables in an (infinitely large) restaurant.
    - Customers ⇔ observed data to be clustered
    - Tables ⇔ distinct blocks of partition, or clusters
      - This will help us see the clustering effect of DPs explicitly
- Relation to the clustering problem
  - We typically don’t know the number of clusters and want to learn it from data
  - CRPs address this problem by assuming that there is an infinite number of latent clusters, but that only a finite number of them is used to generate the observed data.

**Procedure**

- Imagine a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers.
- The 1st customer enters and sits at the first table.
- The Nth customer enters and sits at table k with prob $\frac{N_k}{N}$ for $k = 1, \ldots, K$
- New table with prob $\frac{N + 1}{N + 1 + \alpha}$

where $N_k$ is the number of customers already sitting at table $k$.

**Remark**

- Metaphor was motivated by the seemingly infinite seating capability of Chinese restaurants in San Francisco...

**Resulting conditional distribution**

$p(z_N = z_1, \ldots, z_{N-1}, z_N = \alpha) = \frac{1}{N - 1 + \alpha} \left( \sum_{k=1}^{K} N_k \delta(z_k, k) + \alpha \delta(z_N, K + 1) \right)$

**Relationship between CRPs and DPs**

- Discussion
  - DP is a distribution over distributions.
  - DP results in discrete distributions, so if you draw N points, you are likely to get repeated values.
  - A DP therefore induces a partitioning of the N points.
  - The CRP is the corresponding distribution over partitions.
  - We can easily get back from the CRP to the Polya urn scheme by the following extension:
    - When the first customer sits down at an empty table, he independently chooses a dish $\theta_i$ for the entire table from a prior distribution $G_0$.
    - Dish $\rightarrow$ parameters of the cluster

**The CRP exhibits the clustering property of the DP.**

- Rich-gets-richer effect implies small number of large clusters.
- Expected number of clusters is $K = O(\alpha \log N)$. 
CRPs & Exchangeable Partitions

\[ p(z_N = z_{1}, ... , z_{N-1}, \cdot | \alpha) = \frac{1}{N-1 + \alpha} \sum_{k=1}^{N} N_k \delta(z, k) + \alpha \delta(z, k) \]

- **Closer analysis**
  - Consider the probability of a certain seating arrangement:
  \[ p(z_1, ... , z_N | \alpha) = p(z_1 | \alpha)p(z_2 | z_1, \alpha) \cdots p(z_N | z_{N-1}, ... , z_1, \alpha) \]
  \[ = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k) \]
  - Derivation of the terms

  \[
  \frac{1 \cdot 2 \cdots (N-1)!}{1 + \alpha \cdot 2 + \alpha \cdots N - 1 + \alpha} = \frac{\Gamma(N)}{\Gamma(N + \alpha)}
  \]

- **Probability of a seating arrangement**

\[ p(z_1, ... , z_N | \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \alpha^K \prod_{k=1}^{K} \Gamma(N_k) \]

- **Exchangeability property**
  - The probability of a seating arrangement of \( N \) customers is **independent** of the order they enter the restaurant!
  - The CRP is thus a prior on infinitely exchangeable partitions.
  - **(Definition exchangeability):** The joint probability underlying the data is invariant to permutation.

- **Why is this of importance?**
  - Two reasons...

Reason 1: De Finetti’s Theorem

- **Putting all of this together…**
  - De Finetti’s theorem tells us that the CRP has an underlying mixture distribution with a prior distribution over measures.
  - The Dirichlet Process is the De Finetti mixing distribution for the CRP.

- **Graphical model visualization**
  - This means, when we integrate out \( G \), we get the CRP:
  \[ p(\theta_1, \ldots, \theta_N) = \prod_{n=1}^{N} p(\theta_n | G) dP(G) \]

\[ \Rightarrow \text{If the DP is the prior on } G, \text{ then the CRP defines how points are assigned to clusters when we integrate out } G. \]

Reason 2: Efficient Inference

- **Taking advantage of exchangeability…**
  - In clustering applications, we are ultimately interested in the cluster assignments \( z_1, \ldots, z_N \).
  - Equivalent question in the CRP: Where should customer \( n \) sit, conditioned on the seating choices of all the other customers?
  - This is easy when customer \( n \) is the last customer to arrive:
  - (Seemingly) hard otherwise...

\[ \Rightarrow \text{Because of exchangeability, we can always swap customer } n \text{ with the final customer and use the above formula!} \]
\[ \Rightarrow \text{We’ll use this for efficient Gibbs sampling later on...} \]

Big Picture: CRPs and the DP

\[ G \sim \text{DP} (\alpha, \psi) \]

The CRP describes the partitions of \( \theta \) when \( G \) is marginalized out

Topics of This Lecture

- **Finite Bayesian Mixture Models**
  - Recap
  - Approximate inference

- **Dirichlet Processes**
  - Motivation
  - Definition
  - Dirichlet Process Mixture Models
  - Polya Urn scheme
  - Chinese Restaurant Process
  - Stick-Breaking construction

- **Applying DPMMs**
  - Efficient sampling
  - Applications
**Stick-Breaking Construction** [Sethuraman, 1994]

- Explicit construction for the weights in DP realizations
  - Define an infinite sequence of random variables
    \[ \beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, 2, \ldots \]
  - Then define an infinite sequence of mixing proportions as
    \[ \pi_1 = \beta_1 \]
    \[ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \quad k = 2, 3, \ldots \]
  - This can be viewed as breaking off portions of a stick
    \[ \beta_1 \quad \beta_2 \quad (1 - \beta_1) \quad \ldots \]
  - When the \( \pi_k \) are drawn this way, we can write \( \pi \sim \text{GEM}(\alpha) \).
    (where GEM stands for Griffiths, Engen, McCloskey)

**Stick-Breaking Example**

- Interpretation
  - Mixture weights \( \pi_k \) partition a unit-length “stick” of probability mass among an infinite set of random parameters.
  - Note: The weights do not decrease monotonically!

**Stick-Breaking Construction**

- We now have an explicit formula for each \( \pi_k \):
  \[ \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l) \]

**Stick-Breaking and the DP**

- Graphical Model representation
  - The Stick-Breaking representation provides another interpretation of the concentration parameter \( \alpha \).
  - Since \( \beta_k \sim \text{Beta}(1, \alpha) \), we can apply standard moment formulas and find
    \[ \text{E}[\beta_k] = \frac{1}{1 + \alpha} \]
    \[ \Rightarrow \text{For small } \alpha, \text{ the first few mixture components are typically assigned the majority of the probability mass.} \]
    \[ \Rightarrow \text{For } \alpha \to \infty, \text{ samples } G \sim \text{DP}(\alpha, G_0) \text{ approach the base measure } G_0 \text{ by assigning small, roughly uniform weights to a densely sampled set of discrete parameters.} \]

**Summary: Pólya Urns, CRPs, and Stick-Breaking**

- \( G \sim \text{DP}(\alpha, G_0) \)
  - The Pólya urn describes the predictive distribution of \( \theta \) when \( G \) is marginalized out
  - The CRP describes the partitions of \( \theta \) when \( G \) is marginalized out

- **Big Picture:** Stick-Breaking allows us to sample directly from the weights
Topics of This Lecture

- Finite Bayesian Mixture Models
  - Recap
  - Approximate inference
- Dirichlet Processes
  - Motivation
  - Definition
  - Dirichlet Process Mixture Models
  - Polya Urn scheme
  - Chinese Restaurant Process
  - Stick-breaking construction
- Applying DPMMs
  - Efficient sampling
  - Applications

Back to the clustering problem...

Motivation

Stick

Sample a random permutation

Algorithm is valid if

The model implies the factorization

\[ E \]

Chinese Restaurant Process

If any current clusters are empty (\[ z_{nk} = 0 \]), remove them and decrement \( K \) accordingly.

Update cached sufficient statistics to reflect assignment \( z_{nk} \). If \( z_{nk} = k \), create a new cluster and increment \( K \).

Applications

Exchangeability

Remarks

Recap

Definition

Indicator variable representation

Distributional form

Derived the collapsed (Rao-Blackwellized) Gibbs sampler we derived for finite mixtures.

As before, sample the indicator variables \( z_k \) assigning observations to latent clusters, marginalizing mixture weights \( \pi_k \) and parameters \( \theta_k \).

Assume the cluster priors \( H(\lambda) \) are conjugate.

The model implies the factorization

\[ p(z_k | z_{-k}, X; \pi, \lambda) \propto p(z_k | z_{-k}, \pi) p(X | z_k, X_{-k}, \lambda) \]

Prior on partitions expressed by the CRP!

Efficient algorithm

- Generalize the collapsed (Rao-Blackwellized) Gibbs sampler we derived for finite mixtures
- As before, sample the indicator variables \( z_k \) assigning observations to latent clusters, marginalizing mixture weights \( \pi_k \) and parameters \( \theta_k \)
- Assume the cluster priors \( H(\lambda) \) are conjugate

Derivation

- The model implies the factorization

\[ p(z_k | z_{-k}, X; \pi, \lambda) \propto p(z_k | z_{-k}, \pi) p(X | z_k, X_{-k}, \lambda) \]

\( \text{Prior on partitions expressed by the CRP!} \)

Good collapse DP Mixture Sampler

- Efficient algorithm

  - Generalize the collapsed (Rao-Blackwellized) Gibbs sampler we derived for finite mixtures
  - As before, sample the indicator variables \( z_k \) assigning observations to latent clusters, marginalizing mixture weights \( \pi_k \) and parameters \( \theta_k \)
  - Assume the cluster priors \( H(\lambda) \) are conjugate

Derivation

- The model implies the factorization

\[ p(z_k | z_{-k}, X; \pi, \lambda) \propto p(z_k | z_{-k}, \pi) p(X | z_k, X_{-k}, \lambda) \]

\( \text{Prior on partitions expressed by the CRP!} \)

Algorithm

1. Sample a random permutation \( \tau(\cdot) \) of the integers \( \{1, \ldots, N\} \).
2. Set \( \alpha = \alpha^{(1)} \) and \( z = z^{(1)} \). For each \( n \in \{\tau(1), \ldots, \tau(N)\} \), sequentially resample \( z_n \) as follows
   a) For each of the \( K \) existing clusters, determine the predictive likelihood

\[ p_k(z_n | z_{-n}, X; \pi, \lambda) \propto p(z_n | z_{-n}, \pi) p(X | z_n, X_{-n}, \lambda) \]

   Also determine the likelihood \( p_k(\cdot | z_{-n}) \) of a potential new cluster \( k' \)

\[ p_k(z_n | z_{-n}, X; \pi, \lambda) = p(z_n | z_{-n}, \pi) \frac{\int p(z_n | \theta) p(\theta | \lambda) d\theta}{\int p(z_n | \theta) p(\theta | \lambda) d\theta} \]

   b) Sample a new assignment \( z_n \) from the multinomial distribution

\[ \pi_n \sim \pi(z_n | z_{-n}, X_{-n}, \lambda) + \sum_{k=1}^K \frac{N_{-nk}}{N} p_k(z_n | z_{-n}, X_{-n}, \lambda) \]

   c) Update cached sufficient statistics to reflect assignment \( z_{nk} \).

Derivation (cont’d)

- Exchangeability: Think of \( z_n \) as the last observation in sequence \( X_n \), as in step 3 of the standard finite mixture sampler.

- The predictive likelihood of \( z_n \) is computed as for finite mixtures:

\[ p(z_n | z_{-n}, X_{-n}, \lambda) = \int p(z_n | \theta) p(\theta | \lambda) d\theta \]

- New clusters \( k' \) are based on the predictive likelihood implied by the hyperparameters \( \lambda \)

\[ p(z_n | z_{-n}, X_{-n}, \lambda) = \int p(z_n | \theta) p(\theta | \lambda) d\theta \]

- Optionally, mixture parameters for the \( K \) currently instantiated clusters may be sampled as in step 3 of the standard finite mixture sampler.

Algorithm (cont’d)

3. Set \( z^{(1)} = z \). Optionally, mixture parameters for the \( K \) currently instantiated clusters may be sampled as in step 3 of the standard finite mixture sampler.
4. If any current clusters are empty (\( N_k = 0 \)), remove them and decrement \( K \) accordingly.

Remarks

- Algorithm is valid if the cluster priors \( H(\lambda) \) are conjugate.
- Cluster assignments \( z^{(1)} \) produced by Gibbs sampler provide estimates \( \hat{K} \) of the number of clusters underlying the observations \( X \), as well as their associated parameters.
- Predictions based on samples average over mixtures of varying size, avoiding difficulties in selecting a single model.
Collapsed DP Sampler: 2 Iterations

Collapsed DP Sampler: 10 Iterations

Collapsed DP Sampler: 50 Iterations

DPMMs vs. Finite Mixture Samplers

• Observations
  - Despite having to search over mixtures of varying order, the DP sampler typically converges faster.
  - Avoids local optima by creating redundant clusters at beginning.

DP Posterior Number of Clusters

Summary: Nonparametric Bayesian Clustering

• DPMMs for Clustering
  - First specify the likelihood. This is application dependent.
  - Next, specify a prior on all parameters - the Dirichlet Process!
  - Exact posterior inference is intractable. But we can use a Gibbs sampler for approximate inference. This is based on the CRP representation.
DPMM Software Packages

- Matlab packages for CRP mixture models

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Author</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variational</td>
<td>K. Kurihara</td>
<td><a href="http://sites.google.com/site/kenichikurihara/academic-software">http://sites.google.com/site/kenichikurihara/academic-software</a></td>
</tr>
</tbody>
</table>

References and Further Reading

- Unfortunately, there are currently no good introductory textbooks on the Dirichlet Process. We will therefore post a number of tutorial papers on their different aspects.