Nonparametric Bayesian Modeling

Piyush Rai

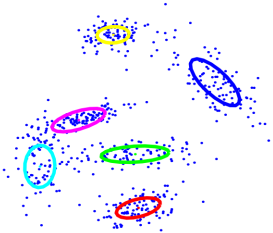
Topics in Probabilistic Modeling and Inference (CS698X)

April 10, 2018

Motivation: Mixture Models

- Suppose each observation is generated from a mixture model

$$\begin{aligned} \boldsymbol{z}_n &\sim \text{multinoulli}(\boldsymbol{\pi}) \\ \boldsymbol{x}_n &\sim \mathcal{N}(\mu_{\boldsymbol{z}_n}, \Sigma_{\boldsymbol{z}_n}) \end{aligned}$$



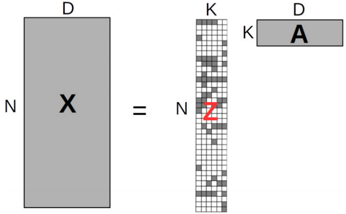
- How to learn K, i.e., the number of clusters for such a mixture model?
- Need to model/prior for $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$ that allows number of clusters to be unbounded

Motivation: Latent Feature Models

- Suppose each observation is a subset sum of K latent features

$$\begin{aligned} z_{nk} &\sim \text{Bernoulli}(\pi_k) \qquad k = 1, \ldots, K \\ \boldsymbol{x}_n &= \sum_{k=1}^{K} z_{nk}\boldsymbol{a}_k + \epsilon_n = \mathbf{A}\boldsymbol{z}_n + \epsilon_n \end{aligned}$$

- This can also be seen as a form of matrix factorization with one matrices being binary



- How to learn K, i.e., the number of latent features (i.e., number of columns/rows in $\mathbf{Z}/\mathbf{A}$)?
- Need to model/prior for $\mathbf{Z}$ that allows number of columns in $\mathbf{Z}$ to remain unbounded

Motivation: Matrix Factorization

- Consider the following singular value decomposition model for an $N \times M$ matrix $\mathbf{X}$

$$\mathbf{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top + \mathbf{E} = \mathbf{U}\Lambda\mathbf{V}^\top + \mathbf{E}$$

  where $\boldsymbol{u}_k \in \mathbb{R}^N$ and $\boldsymbol{v}_k \in \mathbb{R}^M$, Λ is a $K \times K$ diagonal matrix

- This is basically a weighted sum of rank-1 matrices
- How to learn K, i.e., the "rank" of the above factorization?
- Need a model/prior for Λ that allows the size of Λ to be unbounded

Nonparametric Bayesian Modeling

- Enables constructing models that do not have an a priori fixed size
- Nonparametric does not mean no parameters
  - Instead, have a possibly infinite (unbounded) number of parameters
  - Note: We've already seen Gaussian Processes which is a nonparametric Bayesian model
- Usually constructed via one of the following ways
  - Take a finite model (e.g., a finite mixture model) and consider its "infinite limit"
  - Have a model that allows very large number of parameters but has a "shrinkage" effect, e.g.,

$$\mathbf{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top + \mathbf{E} \qquad \lambda_k \to 0 \quad \text{as} \quad k \to \infty$$

- We will look at some examples of both these approaches

Being Nonparametric by taking Infinite Limit of Finite Models

Finite Mixture Model

- Data $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$, cluster assignments $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$, K clusters
- Denote the mixing proportion by a vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

$$\begin{aligned}
p(\boldsymbol{\pi}|\alpha) &= \text{Dirichlet}(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \\
p(\boldsymbol{z}_n|\pi) &= \prod_{k=1}^{K} \pi_k^{z_{nk}} \\
p(\mathbf{X}|\boldsymbol{\pi}) &= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\boldsymbol{z}_n = k)
\end{aligned}$$

- Integrating out $\boldsymbol{\pi}$, the marginal prior probability of cluster assignments $\mathbf{Z}$

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \qquad \text{(verify)}$$

where m_k = no. of points with $\boldsymbol{z}_n = k$

Finite Mixture Model

- What is the prior $p(\boldsymbol{z}_n|\mathbf{Z}_{-n},\alpha)$, i.e., the dist. of $\boldsymbol{z}_n$ given cluster assignment $\mathbf{Z}_{-n}$ of other points?

$$p(\boldsymbol{z}_n|\mathbf{Z}_{-n},\alpha) = \frac{p(\boldsymbol{z}_n,\mathbf{Z}_{-n}|\alpha)}{p(\mathbf{Z}_{-n}|\alpha)} = \frac{p(\mathbf{Z}|\alpha)}{p(\mathbf{Z}_{-n}|\alpha)}$$

- Note that the above is a discrete distribution ($\boldsymbol{z}_n$ takes one of K possible values)
- Using $p(\mathbf{Z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\frac{\prod_{k=1}^{K}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}$ (and applying the same to get $p(\mathbf{Z}_{-n}|\alpha)$) we will have

$$\begin{aligned} p(\boldsymbol{z}_n=j|\mathbf{Z}_{-n},\alpha) &= \frac{p(\boldsymbol{z}_n=j,\mathbf{Z}_{-n}|\alpha)}{p(\mathbf{Z}_{-n}|\alpha)} = \frac{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\frac{\Gamma(m_j+\frac{\alpha}{K})\prod_{k\neq j}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}}{\frac{\Gamma(\alpha)}{\Gamma(N-1+\alpha)}\frac{\Gamma(m_j-1+\frac{\alpha}{K})\prod_{k\neq j}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}} \\ &= \frac{m_{-n,j}+\frac{\alpha}{K}}{N-1+\alpha} \qquad (m_{-n,j}=m_j-1 \text{ denotes no. of other examples in cluster } j) \end{aligned}$$

- Thus prior prob. of $\boldsymbol{z}_n=j$ is prop. to $m_{-n,j}$, i.e., number of other examples assigned to cluster j (this is like a "rich gets richer" phenomenon; a popular cluster will attract more examples)

Taking the Infinite Limit..

- We saw that $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}+\frac{\alpha}{K}}{N-1+\alpha}$. What if $K \to \infty$?
- In that case, we will have $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha}$
- Suppose only K_+ clusters are currently occupied (that have at least one data point)
- Total probability of data point n going to any of these K_+ clusters $= \sum_{j=1}^{K_+} \frac{m_{-n,j}}{N-1+\alpha} = \frac{N-1}{N-1+\alpha}$
- Probability of data point n going to a new (i.e., so far unoccupied) cluster $= \frac{\alpha}{N-1+\alpha}$
- Therefore in the limit of an unbounded number of clusters, we have

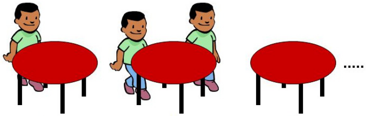
$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} & \text{(prob. of going to } j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} & \text{(prob. of creating a new cluster } K_+ + 1) \end{cases}$$

- The above gives us a prior distribution for clustering with unbounded K
- Note that the probability of starting a new cluster is proportional to Dirichlet hyperparam. α
  - The hyperparam α can also be learned

A metaphor for the same: Chinese Restaurant Process (CRP)

- Consider a restaurant with infinite number of tables (each table denotes a cluster)
- Customer 1 sits at a randomly closen table (all tables are equivalent to begin with)
- Each subsequent customer $n > 1$ sits using the following scheme
  - Sits at an already occupied table k with probability $\frac{m_k}{n-1+\alpha}$
  - Sits at a new table with probability $\frac{\alpha}{n-1+\alpha}$

$$p(\boldsymbol{z}_n = j | \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n-1}, \alpha) = \begin{cases} \frac{m_{-n,j}}{n-1+\alpha} & \text{(for } j = 1, \ldots, K_+\text{)} \\ \frac{\alpha}{n-1+\alpha} & \text{(create a new cluster } K_+ + 1\text{)} \end{cases}$$

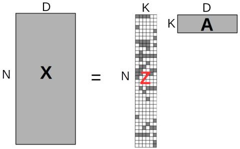


1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$

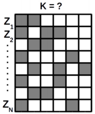
- The total number of occupied tables isn't fixed beforehand (connections to Dirichlet Processes)

Modeling Binary Matrices with Unbounded Number of Columns

- Recall the subset-sum problem



- So how do we model binary matrices for which number of columns K is *a priori* unknown?

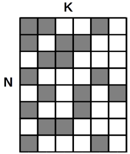


- A nonparam. Bayesian model called "Indian Buffet Process" (IBP) defines a prior for such matrices
- Just like CRP, the IBP is a metaphor to describe the process that generates such matrices
- Note: In some models, the number of rows can be unknown (and number of columns fixed)
  - The IBP model still applies (use IBP as the prior on the transpose of the matrix)

"Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)

Modeling Binary Matrices with Finite Many Columns

- Consider the generative process of an $N \times K$ binary matrix Z



- Rows denote the N examples, columns denote the K latent features
- Assume $\pi_k \in (0, 1)$ to be probabiliy of latent feature k being 1

$$z_{nk} \sim \text{Bernoulli}(\pi_k), \qquad \pi_k \sim \text{Beta}(\alpha/K, 1)$$

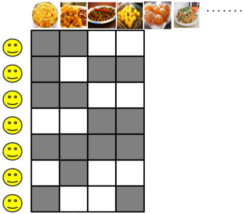
- Note: All z_{nk}'s are i.i.d. given π_k
- For this model, the conditional probability of $z_{nk} = 1$, given other entries in column k of **Z**

$$p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \int p(z_{nk} = 1|\pi_k)p(\pi_k|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}} \quad \text{(verify)}$$

where $m_{-n,k} = \sum_{i \neq n} z_{ik}$ denotes how many other entries in column k are equal to 1

Another Metaphor: Indian Buffet Process

- For the finite K case, we saw that $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}+\frac{\alpha}{K}}{N+\frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\boldsymbol{z}_{-n,k}) = \frac{N-m_{-n,k}}{N}$
- Note that this too exhibits a "rich-gets-richer" phenomenon (just like CRP)
- The Indian Buffet Process is a metaphor for this model. Assume a buffet with infinite dishes
  - Customer 1 selects Poisson(α) dishes
  - The n-th customer selects:
    - Each already selected dish k with probability $m_{-n,k}/n$ (m_k : how many previous customers before n selected dish k)
    - Poisson(α/n) new dishes (this can create new columns in $\mathbf{Z}$)
  - Note that as n grows, number of new dishes goes to zero (and the number of columns K converges to some finite number)



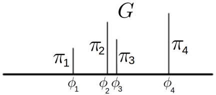
- The above can be used as a prior for $\mathbf{Z}$. Refer to (Griffiths and Ghahramani, 2011) for examples and other theoretical details of the model. Also has connections to Beta Processes

"Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)

Being Nonparametric using Models that have a Shrinkage Effect

Nonparametric Mixture Models: Another Construction

- Consider a finite mixture model with K components with params $(\mu_k, \Sigma_k)_{k=1}^K$



- Denoting $(\mu_k, \Sigma_k) = \phi_k$, this mixture model can be expressed as

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

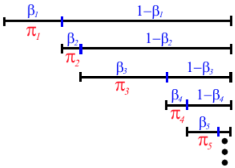
where π_k is the mixing proportion of component k and $\sum_{k=1}^{K} \pi_k = 1$

- In a Bayesian setting, we usually assume $[\pi_1, \ldots, \pi_K] \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- We can make it a nonparametric model by constructing an infinite length probability vector

$$\pi_1, \pi_2, \pi_3, \ldots, \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

- How to construct such an "infinite" length vector? Using an infinite dimensional Dirichlet?

Nonparametric Mixture Models via Stick-Breaking Process



- Assume a mixture distribution $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ and generate mixing proportions as

$$
\begin{aligned}
\beta_k &\sim \text{Beta}(1, \alpha) \qquad k = 1, \ldots, \infty \\
\pi_1 &= \beta_1, \quad \pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \qquad k = 2, \ldots, \infty
\end{aligned}
$$

- Can show that $\sum_{k=1}^{\infty} \pi_k = 1$
- Each "location" ϕ_k can be drawn from an appropriate distribution $\phi_k \sim G_0$
- Therefore, a finite mixture model can be made nonparametric by replacing the Dirichlet prior on π in a mixture model by a stick-breaking process prior as defined above

Another Example: Multiplicative Gamma Process

- Consider the following probabilistic version of SVD

$$\mathbf{X} = \sum_{k=1}^{K} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top + \mathbf{E}$$

- Consider the following prior on the "singular values" λ_k

$$\begin{aligned} \lambda_k &\sim \mathcal{N}(0, \tau_k^{-1}) \\ \tau_k &= \prod_{\ell=1}^{k} \delta_\ell \\ \delta_\ell &\sim \text{Gamma}(\alpha, 1) \quad \text{where } \alpha > 1 \end{aligned}$$

- Note that as k becomes large, τ_k gets larger and larger and λ_k shrinks to zero

---

"Sparse Bayesian infinite factor models (Bhattacharya and Dunson, 2011)

Some Comments

- Nonparametric Bayesian models have been widely used in several applications
  - Clustering, dim-red, regression/classification, time-series models such as HMM, and many others
- Nonparametric Bayesian models are not the only way to learn the right model size
- Marginal likelihood $p(\mathcal{D}|\mathcal{M})$ can be used for model selection from a set of models $\{\mathcal{M}_i\}_{i=1}^{L}$
- Other criteria such as Akaike or Bayesian Information Criteria are also commonly used
  - Usually defined as a sum of negative log-lik. and model size (models with smaller values preferred)

$$\begin{aligned} AIC &= 2k - 2 \times \text{log-lik} \\ BIC &= k \log N - 2 \times \text{log-lik} \end{aligned}$$

  where k denotes the number of parameters of the model, N denotes number of data points
- However, marginal likelihood, AIC/BIC, etc. try multiple models and then choose the best
- In contrast, NPBayes models learn a single model having an unbounded complexity
  - Also natural for streaming data where model selection is difficult/impractical to perform