Introduction to Nonparametric Bayesian Modeling

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Plan

- Need for nonparametric Bayesian modeling
- Some basic ideas
- Some examples of NPBayes modeling for
  - Mixture Models
  - Latent Feature Models and Matrix Factorization
- Some standard ways of constructing NPBayes models
  - Stick-breaking process, Dirichlet process
  - Some metaphors: Chinese Restaurant Process and Indian Buffet Process

Motivating Problem: Mixture Models

- Suppose each observation is generated from a K component mixture model

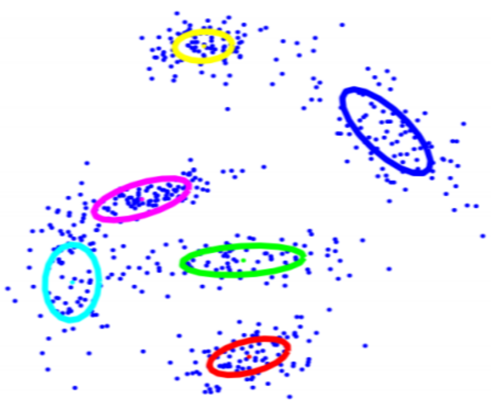
Cluster id of the n^{th} observation

K-dim probability vector of component mixing proportions

n^{th} observation

$$\mathbf{z}_n \sim \text{multinoulli}(\pi)$$

$$\mathbf{x}_n \sim \mathcal{N}(\mu_{\mathbf{z}_n}, \Sigma_{\mathbf{z}_n})$$



- How to learn K, i.e., the number of components (clusters) for such a mixture model?

- Can use marginal-likelihood based model selection but is expensive
  - Need to train the model several times for each possible value of K

- Also difficult if the data is streaming (hard to know beforehand how many clusters)

- How about a prior over $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]$ (or $\boldsymbol{\pi}$) that allows learning the "right" K?

Motivating Problem: Latent Feature Models

- Suppose each observation is a <u>subset sum</u> of K "basis vectors" (or "latent features"*)

Is k^{th} latent feature present in the n^{th} observation?

An example: Each text document in a collection being a subset sum of K "latent" themes or topics present in the collection

$$z_{nk} \sim \text{Bernoulli}(\pi_k) \qquad k = 1, \ldots, K$$

The n^{th} observation $(D \times 1)$ expressed as a subset sum of the K latent features (each $D \times 1$), plus some noise

$$\boldsymbol{x}_n = \sum_{k=1}^{K} z_{nk}\boldsymbol{a}_k + \epsilon_n = \mathbf{A}\boldsymbol{z}_n + \epsilon_n$$

Noise (e.g., zero mean Gaussian)

The k^{th} latent feature $(D \times 1)$

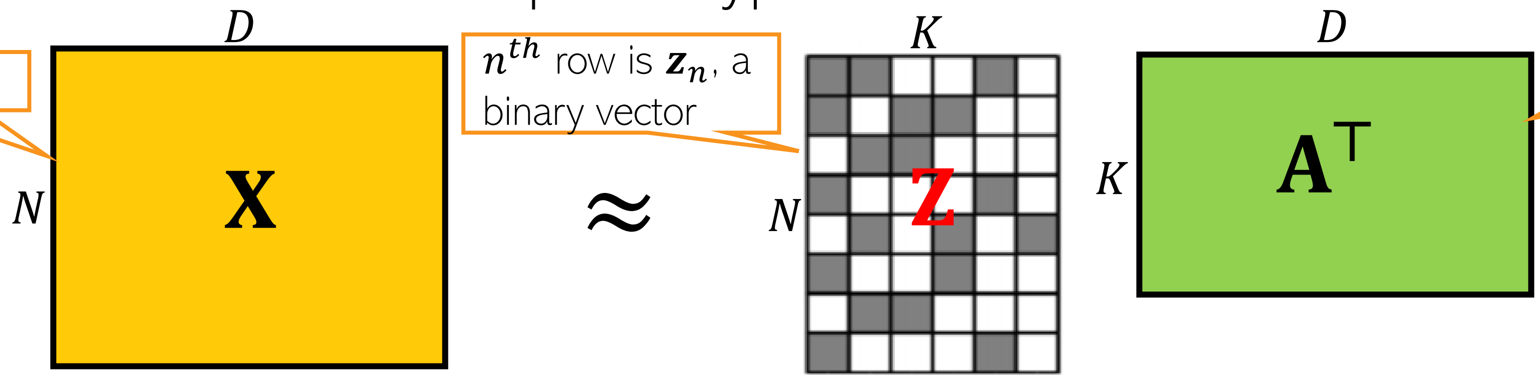
A binary sparse matrix

- This can also be seen as special type of matrix factorization $\mathbf{X} = \mathbf{Z}\mathbf{A}^\top + \mathbf{E}$

n^{th} row is $\boldsymbol{x}_n$

n^{th} row is $\boldsymbol{z}_n$, a binary vector

k^{th} row is $\boldsymbol{a}_k^\top$



Just like mixture models, selecting it based on marg-lik will be expensive

- How about a prior over $\mathbf{Z}$ (or $\mathbf{A}$ or $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$) that allows learning the "right" K?

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

Motivating Problem: SVD-style Matrix Factorization

- Consider the following SVD-style decomposition for an $N \times M$ matrix $\mathbf{X}$

Rank 1 matrix

Zero mean Gaussian noise

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

- Each $\boldsymbol{u}_k \in \mathbb{R}^N$, $\boldsymbol{v}_k \in \mathbb{R}^M$, $\boldsymbol{\lambda}_k \in \mathbb{R}$, and $\boldsymbol{\Lambda}$ is a $K \times K$ diag matrix with $\boldsymbol{\lambda}_k$'s on diags
- This is basically a <u>weighted</u> sum of K rank-1 matrices
  - λ_k's are the weights
  - λ_k's are akin to the singular values in SVD
- How to learn K, i.e., the "rank" of the above factorization?
- How about a prior on $\boldsymbol{\Lambda}$, or $\mathbf{U}$ or $\mathbf{V}$, that allows us to learn the "right" K ?

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

Nonparametric Bayesian Modeling

A vast area of research in ML and statistics. We will only be looking at a basic flavor of some approaches

- 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

And can potentially grow as we see more and more data (actual number will depend on the amount/properties of data)

  - 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 params 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

A tutorial on Bayesian nonparametric models (Gershman and Blei, 2012)

Being Nonparametric by Taking Infinite Limit of Finite Models

A Finite Mixture Model

- Data $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, cluster assignments $\mathbf{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N]$, K clusters
- Denote the mixing proportion by a vector $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

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

a.k.a. “collapsing” a variable; one less variable to infer now

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

Number of points with $\boldsymbol{z}_n = k$

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

A Finite Mixture Model

- The prior distribution of $\boldsymbol{z}_n$ given cluster assignment $\mathbf{Z}_{-n}$ of other points?

A discrete distribution (multinoulli) since $\boldsymbol{z}_n$ can take one of K possibilities

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

This "conditional" prior is needed when computing the posterior of $\boldsymbol{z}_n$ since we have integrated out $\boldsymbol{\pi}$

- Using $p(\mathbf{Z}|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\prod_{k=1}^{K}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}$ we have

Number of points in cluster $\boldsymbol{j}$, not counting $\boldsymbol{x}_n$

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n},\alpha) = \frac{p(\boldsymbol{z}_n = j,\mathbf{Z}_{-n}|\alpha)}{p(\mathbf{Z}_{-n}|\alpha)} = \frac{\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\frac{\Gamma(m_j+\frac{\alpha}{K})\prod_{k\neq j}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}}{\frac{\Gamma(\alpha)}{\Gamma(N-1+\alpha)}\frac{\Gamma(m_j-1+\frac{\alpha}{K})\prod_{k\neq j}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}} = \frac{m_{-n,j}+\frac{\alpha}{K}}{N-1+\alpha}$$

- Note: Can also get this result using $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n},\alpha) = \int p(\boldsymbol{z}_n = j|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{Z}_{-n},\alpha)d\boldsymbol{\pi}$

- Thus prior prob. of $\boldsymbol{z}_n = j$ is proportional to how many other points are in cluster j

- Note that it also implies that mixture models have a rich-gets-richer property
  - Meaning: *a priori*, a cluster with more points is likely to attract more points

Taking the Infinite Limit..

- Since $p(z_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j} + \frac{\alpha}{K}}{N-1+\alpha}$, as $K \to \infty$, $p(z_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha}$
- Suppose only K_+ clusters are currently occupied (i.e., have at least one data point)
- Total prob. of $\boldsymbol{x}_n$ 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 $\boldsymbol{x}_n$ 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

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

- The above gives us a prior distribution for mixture models with unbounded K
  - Can combine it now with the suitable likelihood to infer the posterior* of $\boldsymbol{Z}$
- Note: Prob. of starting a new cluster is prop. to Dirichlet hyperparam α (can learn it)

*Markov chain sampling methods for Dirichlet process mixture models, (Neal, 2000), Variational inference for Dirichlet process mixtures (Blei and Jordan, 2006)

A Metaphor: Chinese Restaurant Process (CRP)

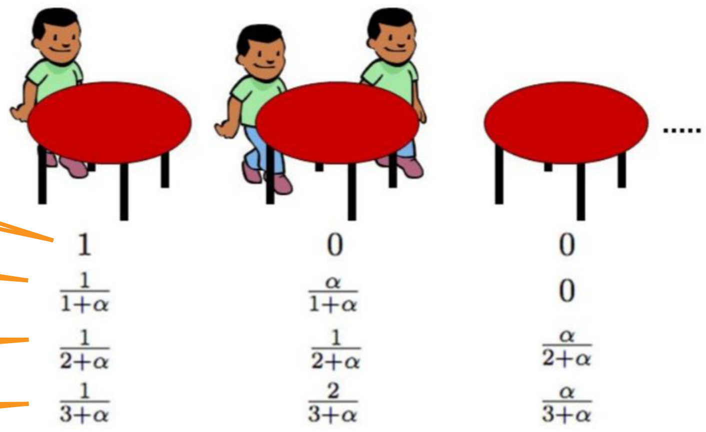
- Assume a restaurant with infinite number of tables (each table denotes a cluster)
- Customer 1 sits at a randomly chosen table (all tables are equivalent to begin with)
- Each subsequent customer $n > 1$ sits using the following scheme
  - Sits at an already occupied table k with probability $\frac{m_k}{n-1+\alpha}$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

When customer 1 comes

When customer 2 comes after customer 1 chose table

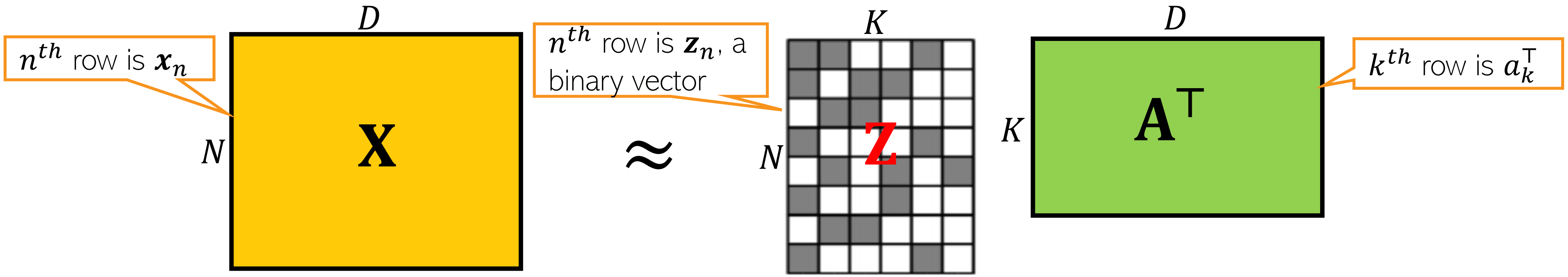
When customer 3 comes after customer 1 and 2 chose table

When customer 4 comes after customer 1,2 and 3 chose table



Nonparametric Bayesian Latent Feature Model

- Recall the subset-sum problem: $\boldsymbol{x}_n = \mathbf{A}\boldsymbol{z}_n + \epsilon_n$



- To learn K, one option is to model $\mathbf{Z}$ such that number of columns can be unbounded
- For the finite case of $N \times K$ matrix $\mathbf{Z}$, assume the following generative process

Prob. of an entry of column k being 1

Number of other entries in column k that are 1

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

Prob that this entry is 1 given values of other entries in column k

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

As $K \to \infty$, $\quad p(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$

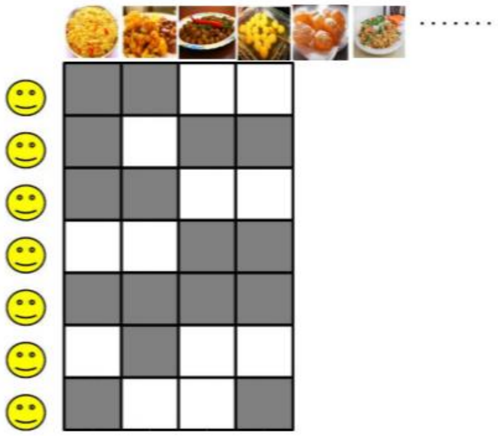
Proportional to how many other entries in column k are 1

Rich-gets-richer just like mixture models

A Metaphor: Indian Buffet Process (IBP)

- IBP is a metaphor for latent feature model similar to CRP for mixture model

- Assume a buffet with infinite dishes
  - Customer 1 selects $\mathbf{Poisson}(\alpha)$ dishes
  - Customer n makes selection as follows
    - Select each already selected dish k with prob. $\frac{m_{-n,k}}{N}$
    - Select $\mathbf{Poisson}(\alpha/n)$ new dishes



- Customer-dish assignment matrix is a binary matrix
  - Thus this "process" defines a prior over binary matrices without a pre-defined number of columns

Since new dishes are added as well

- Note that as n grows, number of new dishes goes to zero (and the number of columns K converges to some finite number)

- Refer to (Griffiths and Ghahramani, 2011) for examples and other theoretical details. Also has connections to Beta Process (just like CRP has with Dirichlet Process)

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