Nonparametric Bayesian Models for Unsupervised Learning

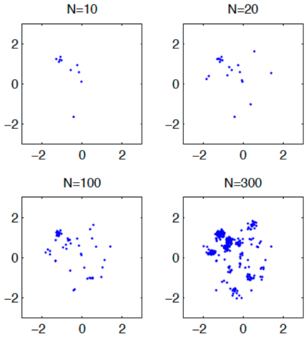
Piyush Rai

Probabilistic Machine Learning (CS772A)

Nov 4, 2017

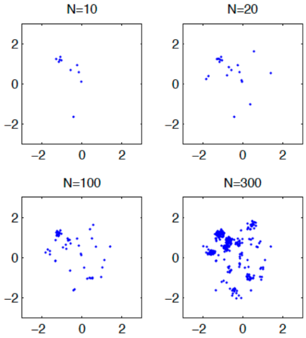
Nonparametric Bayesian Models: Motivation

- Often more/newer structures may appear as we observe more and more data



Nonparametric Bayesian Models: Motivation

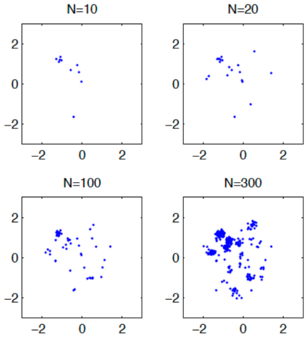
- Often more/newer structures may appear as we observe more and more data



- Would like to have models that don't assume a fixed-sized structure in advance, e.g.,

Nonparametric Bayesian Models: Motivation

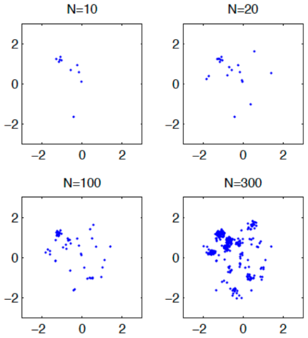
- Often more/newer structures may appear as we observe more and more data



- Would like to have models that don't assume a fixed-sized structure in advance, e.g.,
  - Would like to have a mixture model s.t. number of clusters can grow/adapt with data

Nonparametric Bayesian Models: Motivation

- Often more/newer structures may appear as we observe more and more data



- Would like to have models that don't assume a fixed-sized structure in advance, e.g.,
  - Would like to have a mixture model s.t. number of clusters can grow/adapt with data
  - Would like to have a neural net which can grow/adapt in size (number/width of layers) with data

Nonparametric Bayesian Models for Clustering

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

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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

$$p(\boldsymbol{\pi}|\alpha) = \text{Dirichlet}(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})$$

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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

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

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 $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, $\sum_{k=1}^{K} \pi_k = 1$

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

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 $\mathbf{Z}$

$$
p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi}
$$

Finite Mixture Model

- Since $\boldsymbol{z}_n$'s are i.i.d. given $\boldsymbol{\pi}$, we have

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \int \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi}$$

Finite Mixture Model

- Since $\boldsymbol{z}_n$'s are i.i.d. given $\boldsymbol{\pi}$, we have

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \int \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi}$$

- Substituting $p(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$ and $p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\frac{\alpha}{K})} \prod_{k=1}^{K} \pi_k^{\frac{\alpha}{K}-1}$

Finite Mixture Model

- Since $\boldsymbol{z}_n$'s are i.i.d. given $\boldsymbol{\pi}$, we have

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \int \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi}$$

- Substituting $p(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$ and $p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \prod_{k=1}^{K} \pi_k^{\frac{\alpha}{K}-1}$, we have

$$p(\mathbf{Z}|\alpha) = \frac{1}{B(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})} \int \prod_{k=1}^{K} \pi_k^{m_k + \frac{\alpha}{K} - 1} d\boldsymbol{\pi}$$

Finite Mixture Model

- Since $\boldsymbol{z}_n$'s are i.i.d. given $\boldsymbol{\pi}$, we have

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \int \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi}$$

- Substituting $p(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$ and $p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \prod_{k=1}^{K} \pi_k^{\frac{\alpha}{K}-1}$, we have

$$p(\mathbf{Z}|\alpha) = \frac{1}{B(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})} \int \prod_{k=1}^{K} \pi_k^{m_k + \frac{\alpha}{K} - 1} d\boldsymbol{\pi} = \frac{B(m_1 + \frac{\alpha}{K}, m_2 + \frac{\alpha}{K}, \ldots, m_K + \frac{\alpha}{K})}{B(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$, $B(\alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$, Γ is gamma func. $(\Gamma(x+1) = x\Gamma(x))$

Finite Mixture Model

- Since $\boldsymbol{z}_n$'s are i.i.d. given $\boldsymbol{\pi}$, we have

$$p(\mathbf{Z}|\alpha) = \int p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} = \int \prod_{n=1}^{N} p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi}$$

- Substituting $p(\boldsymbol{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$ and $p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\frac{\alpha}{K})} \prod_{k=1}^{K} \pi_k^{\frac{\alpha}{K}-1}$, we have

$$p(\mathbf{Z}|\alpha) = \frac{1}{B(\frac{\alpha}{K}, \frac{\alpha}{K}, \dots, \frac{\alpha}{K})} \int \prod_{k=1}^{K} \pi_k^{m_k+\frac{\alpha}{K}-1} d\boldsymbol{\pi} = \frac{B(m_1 + \frac{\alpha}{K}, m_2 + \frac{\alpha}{K}, \dots, m_K + \frac{\alpha}{K})}{B(\frac{\alpha}{K}, \frac{\alpha}{K}, \dots, \frac{\alpha}{K})}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$, $B(\alpha_1, \alpha_2, \dots, \alpha_K) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$, Γ is gamma func. $(\Gamma(x+1) = x\Gamma(x))$

- Using the above, the $p(\mathbf{Z}|\alpha) = p(\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_N|\alpha)$ can be written as

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

Finite Mixture Model

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

Finite Mixture Model

- What is $p(z_n|\mathbf{Z}_{-n}, \alpha)$, i.e., the distribution of z_n given cluster assignment $\mathbf{Z}_{-n}$ of all other points?

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

Finite Mixture Model

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

Finite Mixture Model

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

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

Finite Mixture Model

- What is $p(z_n|\mathbf{Z}_{-n},\alpha)$, i.e., the distribution of z_n given cluster assignment $\mathbf{Z}_{-n}$ of all other points?

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

- Note that the above is a discrete distribution (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

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

Finite Mixture Model

- What is $p(z_n|\mathbf{Z}_{-n},\alpha)$, i.e., the distribution of z_n given cluster assignment $\mathbf{Z}_{-n}$ of all other points?

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

- Note that the above is a discrete distribution (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

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

Finite Mixture Model

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

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

- Note that the above is a discrete distribution ($\mathbf{z}_n$ 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

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

- Thus prior prob. of $\mathbf{z}_n = j$ 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)

Finite Mixture Model

- Note that with $\boldsymbol{\pi}$ integrated out, the cluster assignments are not i.i.d. anymore

$$p(z_n = j | \mathbf{Z}_{-n}, \alpha) \quad = \quad \frac{m_{-n,j} + \frac{\alpha}{K}}{N - 1 + \alpha}$$

Finite Mixture Model

- Note that with $\boldsymbol{\pi}$ integrated out, the cluster assignments are not i.i.d. anymore

$$p(\boldsymbol{z}_n = j | \mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j} + \frac{\alpha}{K}}{N - 1 + \alpha}$$

.. as now prior prob. of $\boldsymbol{z}_n = j$ depends on how many other examples are assigned to cluster j

Finite Mixture Model

- Note that with $\boldsymbol{\pi}$ integrated out, the cluster assignments are not i.i.d. anymore

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) \quad = \quad \frac{m_{-n,j} + \frac{\alpha}{K}}{N - 1 + \alpha}$$

.. as now prior prob. of $\boldsymbol{z}_n = j$ depends on how many other examples are assigned to cluster j

- Note: We can also derive the above result as

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \int p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{Z}_{-n}, \alpha)d\boldsymbol{\pi} \qquad \text{(complete the exercise!)}$$

Taking the Infinite Limit..

- We saw that $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j} + \frac{\alpha}{K}}{N-1+\alpha}$

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 the number of clusters is very large, i.e., $K \to \infty$?

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 the number of clusters is very large, i.e., $K \to \infty$?
- In that case, we will have $p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha}$

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 the number of clusters is very large, i.e., $K \to \infty$?
- 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 example)

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 the number of clusters is very large, i.e., $K \rightarrow \infty$?
- 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 example)
- Total probability of example n going to any of these K_+ clusters $= \sum_{j=1}^{K_+} \frac{m_{-n,j}}{N-1+\alpha}$

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 the number of clusters is very large, i.e., $K \to \infty$?
- 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 example)
- Total probability of example n going to any of these K_+ clusters $= \sum_{j=1}^{K_+} \frac{m_{-n,j}}{N-1+\alpha} = \frac{N-1}{N-1+\alpha}$

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 the number of clusters is very large, i.e., $K \to \infty$?
- 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 example)
- Total probability of example 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 example n going to a new (i.e., so far unoccupied) cluster $= \frac{\alpha}{N-1+\alpha}$

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 the number of clusters is very large, i.e., $K \to \infty$?
- 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 example)
- Total probability of example 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 example 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

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

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 the number of clusters is very large, i.e., $K \rightarrow \infty$?
- 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 example)
- Total probability of example 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 example 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

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 the number of clusters is very large, i.e., $K \to \infty$?
- 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 example)
- Total probability of example 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 example 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

A Metaphor: Chinese Restaurant Process (CRP)

- Consider a restaurant with infinite number of tables (each table denotes a cluster)

A Metaphor: 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)

A Metaphor: 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

A Metaphor: 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 $\frac{m_k}{n-1+\alpha}$

A Metaphor: Chinese Restaurant Process (CRP)

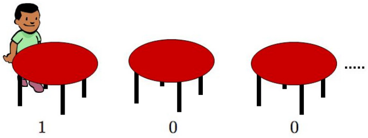
- Consider a restaurant with infinite number of tables (each table denotes a cluster)
- Customer 1 sits at a randomly closen table (all tables are equivalent to begin with)
- Each subsequent customer $n > 1$ sits using the following scheme
  - Sits at an already occupied table k with probability $\frac{m_k}{n-1+\alpha}$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

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

A Metaphor: 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 $\frac{\alpha}{n-1+\alpha}$

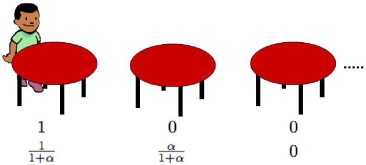
$$p(\boldsymbol{z}_n = j | \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n-1}, \alpha) = \begin{cases} \frac{m_{-n,j}}{n-1+\alpha} & \text{(for } j = 1, \ldots, K_+\text{)} \\ \frac{\alpha}{n-1+\alpha} & \text{(create a new cluster } K_+ + 1\text{)} \end{cases}$$



A Metaphor: 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 $\frac{\alpha}{n-1+\alpha}$

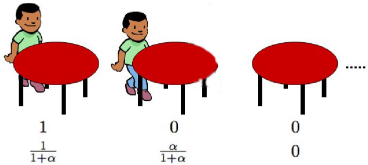
$$p(\boldsymbol{z}_n = j | \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n-1}, \alpha) = \begin{cases} \frac{m_{-n,j}}{n-1+\alpha} & \text{(for } j = 1, \ldots, K_+\text{)} \\ \frac{\alpha}{n-1+\alpha} & \text{(create a new cluster } K_+ + 1\text{)} \end{cases}$$



A Metaphor: 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 $\frac{\alpha}{n-1+\alpha}$

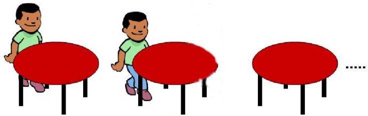
$$p(\boldsymbol{z}_n = j | \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n-1}, \alpha) = \begin{cases} \frac{m_{-n,j}}{n-1+\alpha} & \text{(for } j = 1, \ldots, K_+) \\ \frac{\alpha}{n-1+\alpha} & \text{(create a new cluster } K_+ + 1) \end{cases}$$



A Metaphor: 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 $\frac{\alpha}{n-1+\alpha}$

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

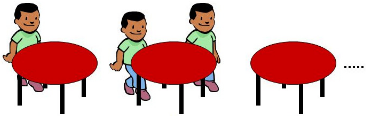


1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$

A Metaphor: 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 $\frac{\alpha}{n-1+\alpha}$

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

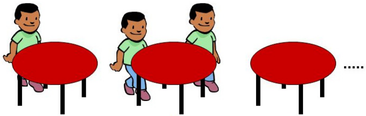


1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$

A Metaphor: 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
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$
- Pitman-Yor Process: A variant of CRP for which K has a power-law growth $\mathcal{O}(N^d)$, where $0 \leq d < 1$ is an additional "discount" parameter and $\alpha > -d$

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$
- Pitman-Yor Process: A variant of CRP for which K has a power-law growth $\mathcal{O}(N^d)$, where $0 \leq d < 1$ is an additional "discount" parameter and $\alpha > -d$
- For the n-th customer, the probabilities are

$$p(\text{table} = k) \quad \propto \quad \frac{n_k - d}{n - 1 + \alpha} \qquad k = 1, \ldots, K$$

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$
- Pitman-Yor Process: A variant of CRP for which K has a power-law growth $\mathcal{O}(N^d)$, where $0 \leq d < 1$ is an additional "discount" parameter and $\alpha > -d$
- For the n-th customer, the probabilities are

$$
\begin{aligned}
p(\text{table} = k) &\propto \frac{n_k - d}{n - 1 + \alpha} \qquad k = 1, \ldots, K \\
p(\text{new table}) &\propto \frac{\alpha + dK}{n - 1 + \alpha}
\end{aligned}
$$

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$
- Pitman-Yor Process: A variant of CRP for which K has a power-law growth $\mathcal{O}(N^d)$, where $0 \leq d < 1$ is an additional "discount" parameter and $\alpha > -d$
- For the n-th customer, the probabilities are

$$
\begin{aligned}
p(\text{table} = k) &\propto \frac{n_k - d}{n - 1 + \alpha} \qquad k = 1, \ldots, K \\
p(\text{new table}) &\propto \frac{\alpha + dK}{n - 1 + \alpha}
\end{aligned}
$$

- For PY process, probability of occupying existing tables with discounted by d
  - Suppresses the "rich-gets-richer" property to some extent

Some Other Properties/Extensions

- The *a priori* expected number of clusters (when using the CRP prior): $K = \mathcal{O}(\alpha \log N)$
- Pitman-Yor Process: A variant of CRP for which K has a power-law growth $\mathcal{O}(N^d)$, where $0 \leq d < 1$ is an additional "discount" parameter and $\alpha > -d$
- For the n-th customer, the probabilities are

$$
\begin{aligned}
p(\text{table} = k) &\propto \frac{n_k - d}{n - 1 + \alpha} \qquad k = 1, \ldots, K \\
p(\text{new table}) &\propto \frac{\alpha + dK}{n - 1 + \alpha}
\end{aligned}
$$

- For PY process, probability of occupying existing tables with discounted by d
  - Suppresses the "rich-gets-richer" property to some extent
- In PYP, the creation of new tables is encouraged more and more as K grows

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each z_n as follows (K_+ = current number of clusters)

$$p(z_n = j|\mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ & \end{cases}$$

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each z_n as follows (K_+ = current number of clusters)

$$p(z_n = j|\mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\mathbf{x}_n|\mu_*, \Sigma_*) p(\mu_*, \Sigma_*) d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each $\boldsymbol{z}_n$ as follows (K_+ = current number of clusters)

$$p(\boldsymbol{z}_n = j | \mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\boldsymbol{x}_n | \mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\boldsymbol{x}_n | \mu_*, \Sigma_*) p(\mu_*, \Sigma_*) d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

- Just like MCMC for GMM, except that we consider $K_+ + 1$ possibilities for each $\boldsymbol{z}_n$

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each $\boldsymbol{z}_n$ as follows (K_+ = current number of clusters)

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\boldsymbol{x}_n|\mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\boldsymbol{x}_n|\mu_*, \Sigma_*)p(\mu_*, \Sigma_*)d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

- Just like MCMC for GMM, except that we consider $K_+ + 1$ possibilities for each $\boldsymbol{z}_n$
- In addition to sampling $\mathbf{Z}$, the sampler needs to sample $\{\mu_j, \Sigma_j\}_{j=1}^{K_+}$ (and α hyperparam)

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each $\boldsymbol{z}_n$ as follows (K_+ = current number of clusters)

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\boldsymbol{x}_n|\mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\boldsymbol{x}_n|\mu_*, \Sigma_*) p(\mu_*, \Sigma_*) d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

- Just like MCMC for GMM, except that we consider $K_+ + 1$ possibilities for each $\boldsymbol{z}_n$
- In addition to sampling $\mathbf{Z}$, the sampler needs to sample $\{\mu_j, \Sigma_j\}_{j=1}^{K_+}$ (and α hyperparam)
  - No need to store/sample the parameter of "unused" clusters as we are integrating those out

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each $\boldsymbol{z}_n$ as follows (K_+ = current number of clusters)

$$p(\boldsymbol{z}_n = j | \mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\boldsymbol{x}_n | \mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\boldsymbol{x}_n | \mu_*, \Sigma_*) p(\mu_*, \Sigma_*) d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

- Just like MCMC for GMM, except that we consider $K_+ + 1$ possibilities for each $\boldsymbol{z}_n$
- In addition to sampling $\mathbf{Z}$, the sampler needs to sample $\{\mu_j, \Sigma_j\}_{j=1}^{K_+}$ (and α hyperparam)
  - No need to store/sample the parameter of "unused" clusters as we are integrating those out
  - If $\boldsymbol{z}_n$ is assigned to a new cluster, we create new cluster's parameters μ_*, Σ_* by sampling from posterior

An Infinite Gaussian Mixture Model

- Let's use our infinite prior (CRP) in a Gaussian mixture model. The CRP prior as we saw is

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

- Using this prior on $\mathbf{Z}$, we can derive an MCMC sampler. Many ways do do this (Neal, 2000)
- One option is to sample for each $\boldsymbol{z}_n$ as follows (K_+ = current number of clusters)

$$p(\boldsymbol{z}_n = j|\mathbf{Z}_{-n}, \alpha, \mathbf{X}) \propto \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} \times \mathcal{N}(\boldsymbol{x}_n|\mu_j, \Sigma_j) & (j = 1, \ldots, K_+) \\ \frac{\alpha}{N-1+\alpha} \int \mathcal{N}(\boldsymbol{x}_n|\mu_*, \Sigma_*)p(\mu_*, \Sigma_*)d\mu_* d\Sigma_* & \text{(new cluster)} \end{cases}$$

- Just like MCMC for GMM, except that we consider $K_+ + 1$ possibilities for each $\boldsymbol{z}_n$
- In addition to sampling $\mathbf{Z}$, the sampler needs to sample $\{\mu_j, \Sigma_j\}_{j=1}^{K_+}$ (and α hyperparam)
  - No need to store/sample the parameter of "unused" clusters as we are integrating those out
  - If $\boldsymbol{z}_n$ is assigned to a new cluster, we create new cluster's parameters μ_*, Σ_* by sampling from posterior
- Also possible to do collaposed Gibbs sampling by integrating out $\{\mu_j, \Sigma_j\}_{j=1}^{K_+}$ (Neal, 2000)

Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models

Nonparametric Bayesian Clustering: Other Perspectives

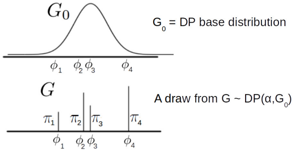
- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via

Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via
  - Dirichlet Process (DP): Informally, an infinite dimensional version of Dirichlet distribution

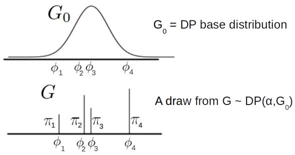
Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via
  - Dirichlet Process (DP): Informally, an infinite dimensional version of Dirichlet distribution
    - DP is a distribution over discrete distributions (thus can be used in clustering problems)



Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via
  - Dirichlet Process (DP): Informally, an infinite dimensional version of Dirichlet distribution
    - DP is a distribution over discrete distributions (thus can be used in clustering problems)



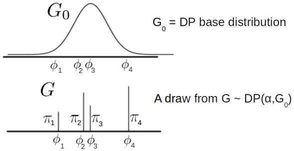
  - Stick-breaking process: Assume a distribution $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ and generate mixing probabilities

$$\beta_k \sim \text{Beta}(1, \alpha) \qquad k = 1, \ldots, \infty$$

$$\pi_1 = \beta_1, \quad \pi_k = \beta_k \prod_{\ell=1}^{k-1}(1 - \beta_\ell) \qquad k = 2, \ldots, \infty$$

Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via
  - Dirichlet Process (DP): Informally, an infinite dimensional version of Dirichlet distribution
    - DP is a distribution over discrete distributions (thus can be used in clustering problems)



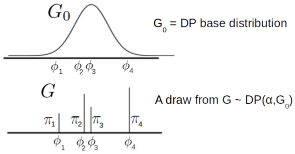
  - Stick-breaking process: Assume a distribution $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ and generate mixing probabilities

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

  - Pólya-Urn scheme: A metaphor to color a set of balls using infinite many possible colors

Nonparametric Bayesian Clustering: Other Perspectives

- The infinite limit and CRP based constructions are not the only way to define such models
- Equivalent constructions for Nonparametric Bayesian clustering can also be obtained via
  - Dirichlet Process (DP): Informally, an infinite dimensional version of Dirichlet distribution
    - DP is a distribution over discrete distributions (thus can be used in clustering problems)



  - Stick-breaking process: Assume a distribution $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ and generate mixing probabilities

$$\beta_k \sim \text{Beta}(1, \alpha) \qquad k = 1, \ldots, \infty$$

$$\pi_1 = \beta_1, \quad \pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \qquad k = 2, \ldots, \infty$$

  - Pólya-Urn scheme: A metaphor to color a set of balls using infinite many possible colors

May refer to "Dirichlet Process" by Teh (2010) for some of these connections

Nonparametric Bayesian Models for Latent Feature Learning

Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

- Note that each $\boldsymbol{x}_n$ is a binary weighted sum of K basis vectors $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$

Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

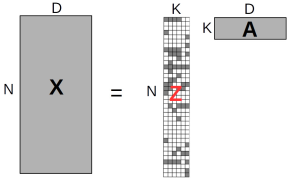
- Note that each $\boldsymbol{x}_n$ is a binary weighted sum of K basis vectors $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$
- We can think of $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]$ as K binary latent features representing $\boldsymbol{x}_n$

Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

- Note that each $\boldsymbol{x}_n$ is a binary weighted sum of K basis vectors $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$
- We can think of $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]$ as K binary latent features representing $\boldsymbol{x}_n$
- Note that $\mathbf{X} \approx \mathbf{ZA}$ with $\mathbf{Z}$ being $N \times K$ (binary), and $\mathbf{A}$ being $K \times D$. Also, K usually unknown

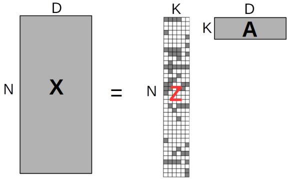


Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

- Note that each $\boldsymbol{x}_n$ is a binary weighted sum of K basis vectors $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$
- We can think of $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]$ as K binary latent features representing $\boldsymbol{x}_n$
- Note that $\mathbf{X} \approx \mathbf{ZA}$ with $\mathbf{Z}$ being $N \times K$ (binary), and $\mathbf{A}$ being $K \times D$. Also, K usually unknown



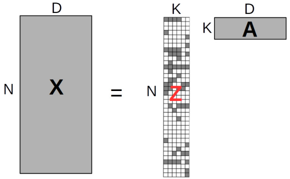
- Different from a mixture model: Here $\boldsymbol{z}_n$ is a binary vector, rather than a one-hot vector

Learning Binary Latent Features

- Suppose we have N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and each $\boldsymbol{x}_n$ can be written as

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

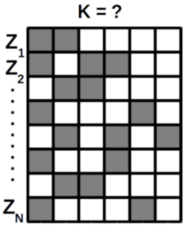
- Note that each $\boldsymbol{x}_n$ is a binary weighted sum of K basis vectors $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$
- We can think of $\boldsymbol{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]$ as K binary latent features representing $\boldsymbol{x}_n$
- Note that $\mathbf{X} \approx \mathbf{ZA}$ with $\mathbf{Z}$ being $N \times K$ (binary), and $\mathbf{A}$ being $K \times D$. Also, K usually unknown



- Different from a mixture model: Here $\boldsymbol{z}_n$ is a binary vector, rather than a one-hot vector
- How do we learn K in this case (i.e., number of columns in $\mathbf{Z}$)?

Binary Matrices with Unbounded Number of Columns

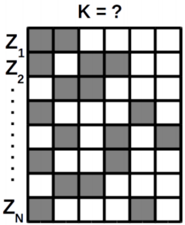
- So how do we model binary matrices for which number of columns K is *a priori* unknown?



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

Binary Matrices with Unbounded Number of Columns

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

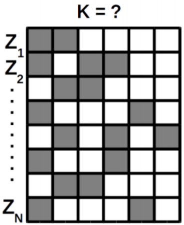


- A nonparam. Bayesian model called "Indian Buffet Process" (IBP) defines a prior for such matrices

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

Binary Matrices with Unbounded Number of Columns

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

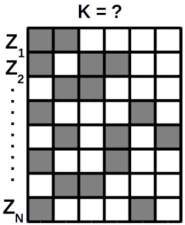


- A nonparam. Bayesian model called "Indian Buffet Process" (IBP) defines a prior for such matrices
- Just like CRP, the IBP is a metaphor to describe the process that generates such matrices

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

Binary Matrices with Unbounded Number of Columns

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

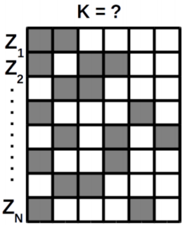


- A nonparam. Bayesian model called "Indian Buffet Process" (IBP) defines a prior for such matrices
- Just like CRP, the IBP is a metaphor to describe the process that generates such matrices
- Note: In some models, the number of rows can be unknown (and number of columns fixed)

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

Binary Matrices with Unbounded Number of Columns

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

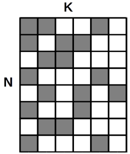


- A nonparam. Bayesian model called "Indian Buffet Process" (IBP) defines a prior for such matrices
- Just like CRP, the IBP is a metaphor to describe the process that generates such matrices
- Note: In some models, the number of rows can be unknown (and number of columns fixed)
  - The IBP model still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

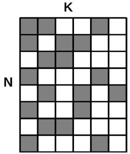
- Consider the generative process of an $N \times K$ binary matrix Z



- Rows denote the N examples, columns denote the K latent features

Modeling Binary Matrices with Finite Many Columns

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

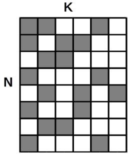


- Rows denote the N examples, columns denote the K latent features
- Assume $\pi_k \in (0, 1)$ to be probabiliy of latent feature k being 1

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

Modeling Binary Matrices with Finite Many Columns

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



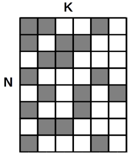
- Rows denote the N examples, columns denote the K latent features
- Assume $\pi_k \in (0,1)$ to be probabiliy of latent feature k being 1

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

- Note: All z_{nk}'s are i.i.d. given π_k

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

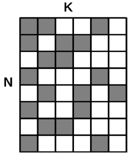
$$z_{nk} \sim \text{Bernoulli}(\pi_k), \qquad \pi_k \sim \text{Beta}(\alpha/K, 1)$$

- 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 $\mathbf{Z}$

$$p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \int p(z_{nk} = 1 | \pi_k) p(\pi_k | \boldsymbol{z}_{-n,k})$$

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 $\mathbf{Z}$

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

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

Modeling Binary Matrices with Infinite Many Columns

- For the finite K case, we saw that $p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$

[a]Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)

Modeling Binary Matrices with Infinite Many Columns

- 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 $p(z_{nk} = 0|\boldsymbol{z}_{-n,k}) = \frac{N-m_{-n,k}}{N}$

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

Modeling Binary Matrices with Infinite Many Columns

- 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 $K \rightarrow \infty$, 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)

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

Modeling Binary Matrices with Infinite Many Columns

- 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

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

Modeling Binary Matrices with Infinite Many Columns

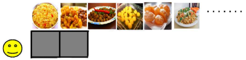
- For the finite K case, we saw that $p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0 | \boldsymbol{z}_{-n,k}) = \frac{N - m_{-n,k}}{N}$
- 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

 ....

[a]Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)

Modeling Binary Matrices with Infinite Many Columns

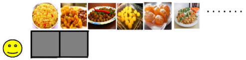
- For the finite K case, we saw that $p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1 | \boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0 | \boldsymbol{z}_{-n,k}) = \frac{N - m_{-n,k}}{N}$
- 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



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

Modeling Binary Matrices with Infinite Many Columns

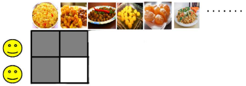
- For the finite K case, we saw that $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}+\frac{\alpha}{K}}{N+\frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\boldsymbol{z}_{-n,k}) = \frac{N-m_{-n,k}}{N}$
- Note that this 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:



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

Modeling Binary Matrices with Infinite Many Columns

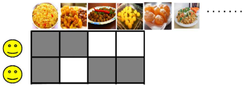
- For the finite K case, we saw that $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}+\frac{\alpha}{K}}{N+\frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\boldsymbol{z}_{-n,k}) = \frac{N-m_{-n,k}}{N}$
- Note that this 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)



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

Modeling Binary Matrices with Infinite Many Columns

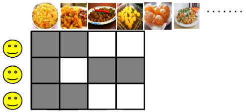
- For the finite K case, we saw that $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$
- As $K \rightarrow \infty$, we will have $p(z_{nk} = 1|\boldsymbol{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\boldsymbol{z}_{-n,k}) = \frac{N - m_{-n,k}}{N}$
- 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 $\mathbf{Z}$)



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

Modeling Binary Matrices with Infinite Many Columns

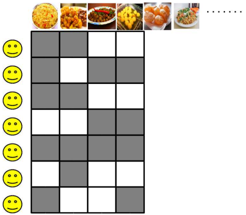
- For the finite K case, we saw that $p(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}+\frac{\alpha}{K}}{N+\frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\mathbf{z}_{-n,k}) = \frac{N-m_{-n,k}}{N}$
- Note that this too exhibits 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)



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

Modeling Binary Matrices with Infinite Many Columns

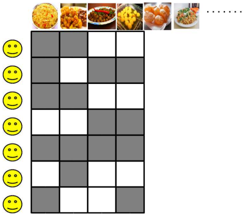
- For the finite K case, we saw that $p(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k} + \frac{\alpha}{K}}{N + \frac{\alpha}{K}}$
- As $K \to \infty$, we will have $p(z_{nk} = 1|\mathbf{z}_{-n,k}) = \frac{m_{-n,k}}{N}$ and $p(z_{nk} = 0|\mathbf{z}_{-n,k}) = \frac{N - m_{-n,k}}{N}$
- Note that this too 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)



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

Modeling Binary Matrices with Infinite Many Columns

- 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)

Some Comments

- Nonparametric Bayesian models are not the only way to learn the right model size

Some Comments

- 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 $\{\mathcal{M}_i\}_{i=1}^{L}$

Some Comments

- 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

Some Comments

- 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)

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

Some Comments

- 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

Some Comments

- 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

Some Comments

- 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

Some Comments

- 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

Some Comments

- 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

Some Comments

- 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
- Again, remember that "nonparametric" doesn't mean no params but unbounded number of params

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems
- We had also seen GP which is another example of a nonparametric Bayesian model

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems
- We had also seen GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (used in both supervised and unsup. learning)

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems
- We had also seen GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (used in both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems
- We had also seen GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (used in both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models
  - Many other such nonparametric Bayesian models for other problems in machine learning
  - For a nice survey, see "A tutorial on Bayesian nonparametric models" by Gershman and Blei (2011)

Summary

- We saw two nonparametric Bayesian models (mainly used in unsupervised learning)
  - CRP/Dirichlet Process: For clustering problems
  - IBP/Beta Process: For latent feature learning problems (also does dimensionality reduction)
- Many applications of these models to solve a wide range of problems
- We had also seen GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (used in both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models
  - Many other such nonparametric Bayesian models for other problems in machine learning
  - For a nice survey, see "A tutorial on Bayesian nonparametric models" by Gershman and Blei (2011)
- Very elegant and rich theory based on stocahstic processes (beyond the score of this course)