Bayesian Non Parametric and its Inference

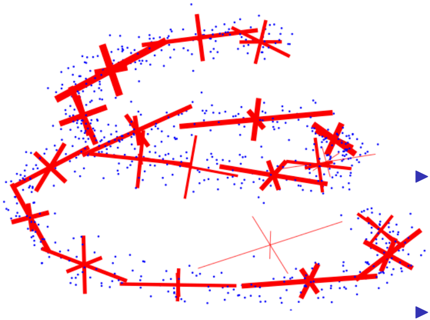
A/Prof Richard Yi Da Xu
Yida.Xu@uts.edu.au
https://github.com/roboticcam/machine-learning-notes

University of Technology Sydney (UTS)

June 8, 2018

Dirichlet Process: A diagrammatic representation

Rasmussen, Infinite Gaussian Mixture Model (1999):



- For a mixture model:
  Let $\mathbf{X} = x_1, \ldots, x_N$:

$$P(\mathbf{X}|\theta_1, \ldots \theta_K, w_1, \ldots w_K) = \sum_{l=1}^{K} w_l f(\mathbf{X}|\theta_l)$$

where $\sum_{l=1}^{K} w_l = 1$

- If we allow K to also vary, what happens if you want to:

$$\underset{\theta_1,\ldots\theta_K, w_1,\ldots,w_K, K}{\arg\max} P(\mathbf{X}|\theta_1, \ldots \theta_K, w_1, \ldots w_K, K)?$$

- K = N for Gaussian case. Of course it's not desirable!

Dirichlet Process: Motivation

- For data $x_1, \ldots, x_N$, each x_i is associating with a parameter θ_i
- We need to a good prior for $\Pr(\theta_1 \ldots \theta_N)$:
- You also want K potentially be infinite
- A "clustering" property, controllable through a single parameter α
- Let's define it using Hierarchical prior, its marginal is:

$$p(\theta_1, \ldots \theta_n) = \int_G \Pr(\theta_1, \ldots, \theta_n | G) \mathbf{p}(\mathbf{G})$$

So, we are interested in the property of G:

- G needs to be **discrete** random distribution.
- Perhaps it should also some resemblence with some basic distribution H.

Dirichlet Process Definition

- We say G is a Dirichlet process, distributed with base distribution H and concentration parameter α:

$$G \sim DP(\alpha, H), \text{if}$$
$$(G(A1), ..., G(Ar)) \sim \text{Dir}(\alpha H(A1), ..., \alpha H(Ar))$$

- for every finite measurable partition $A_1, ..., Ar$ of Θ.
- What does this all mean? Let's visualise it!
- **note** $(A_1 \cup A_2 \cup \cdots \cup A_r) \subseteq \Omega$, this can be seen from the fact that:

$$(x_1, \ldots, x_k, \ldots, x_K) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$$
$$\implies \left(\frac{x_1}{1 - x_k}, \ldots, \frac{x_{k-1}}{1 - x_k}, \frac{x_{k+1}}{1 - x_k}, \ldots, \frac{x_K}{1 - x_k} \right) \sim \text{Dir}(\alpha_1, \alpha_{k-1}, \alpha_{k+1}, \alpha_K)$$

You need both the posterior and predictive distribution of Multinomial-Dirichlet:

Posterior

$$
\begin{aligned}
&P(p_1, \dots, p_k | n_1, \dots, n_k) \\
&\propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}}_{\text{Dir}(p_1,\dots,p_k|\alpha_1,\dots,\alpha_k)} \underbrace{\frac{n!}{n_1! \dots n_k!} \prod_{i=1}^k p_i^{n_i}}_{\text{Mult}(n_1,\dots,n_k|p_1,\dots p_k)} \\
&\propto \prod_{i=1}^k p_i^{\alpha_i - 1} \prod_{i=1}^k p_i^{n_i} = \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \text{Dir}(p_1, \dots p_k | \alpha_i + n_i, \dots \alpha_k + n_k)
\end{aligned}
$$

Marginal

$$
\begin{aligned}
p(n_1, \dots n_k) &= \int_{p_1,\dots,p_k} P(p_1, \dots, p_k, n_1, \dots, n_k) \\
&= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \frac{n!}{n_1! \dots n_k!} \int_{p_1,\dots,p_k} \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
&= \frac{N!}{n_1! \dots n_k!} \times \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^k \Gamma(\alpha_i + n_i)}{\Gamma\left(N + \sum_{i=1}^k \alpha_i\right)}
\end{aligned}
$$

Expectation

- for any measurable set $A_i \in \Omega$: we have $\mathbb{E}[G(A_i)] = H(A_i)$, why?
- for a dirichlet distribution:

$$f(x_1, \dots, x_K | \alpha_1, \dots, \alpha_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

- the expectation: $E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$
- Therefore:

$$\mathbb{E}[G(A_i)] = \frac{\alpha H(A_i)}{\sum_i \alpha H(A_i)} = \frac{\alpha H(A_i)}{\alpha \sum_i H(A_i)} = H(A_i)$$

- note that the expectation is **independent of** α

Variances

- Variances for Dirichlet Distribution:

$$\mathbb{VAR}[X_i] = \frac{\alpha_i \left(\left(\sum_i^K \alpha_{i=1} \right) - \alpha_i \right)}{\left(\sum_i^K \alpha_{i=1} \right)^2 \left(\sum_i^K \alpha_{i=1} + 1 \right)}$$

- substitute $\alpha \to \alpha H(A_i)$:

$$\begin{aligned} \mathbb{VAR}(G(A_i)) &= \frac{\alpha H(A_i) \left(\alpha - \alpha H(A_i) \right)}{\alpha^2 \left(\alpha + 1 \right)} \\ &= \frac{H(A_i) \left(1 - H(A_i) \right)}{\left(\alpha + 1 \right)} \end{aligned}$$

- when $\alpha = 0$:

$$\mathbb{VAR}(G(A_i)_{\alpha=0}) = H(A_i)(1 - H(A_i))$$

Posterior

- from **multinomial-dirichlet conjugacy**, we have:

$$G' = G(A_1), \dots, G(A_r) | \theta_1, \dots, \theta_n \sim \mathrm{Dir}(\alpha H(A_1) + n_1, \dots, \alpha H(A_k) + n_k)$$

- DP provides a conjugate family of priors over distributions that is **closed** under posterior updates given observations:

$$G' \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right), \text{ or}$$

$$G' \sim \mathrm{DP}\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}\right)$$

- another way of specifying this is:

$$G_u \sim \mathrm{DP}(\alpha, H) \qquad G' = \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i} + \frac{\alpha}{\alpha + n} G_u$$

In words: posterior of $\mathrm{DP}(\alpha, H)$ is to **squash** $\mathrm{DP}(\alpha, H)$ to a total mass of $\frac{\alpha}{\alpha+n}$ remaining mass was assigned to discrete points $\sum_{i=1}^{n} \delta_{\theta_i}$.

Predictive

- Let $P(\theta_{n+1} \in A|G) = G(A)$:

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, \dots, \theta_n) &= \int_G P(\theta_{n+1} \in A|G)P(G|\theta_1, \dots, \theta_n)dG \\ &= \mathbb{E}(G(A)|\theta_1, \dots, \theta_n) \\ &= \mathbb{E}(G'(A)) \end{aligned}$$

- We know that:

$$\mathbb{E}(G(A)) = H(A) \implies \mathbb{E}(G'(A)) = \frac{\alpha}{\alpha + n} H(A) + \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$$

Stick-Breaking construction

- $v_k \sim \text{Beta}(1, \alpha)$
- $\pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l)$
- $\theta_k \sim H$
- $G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$

posterior sampling of π

- $v_k \sim \text{Beta}(1, \alpha)$
- $\pi_k = v_k \prod_{l=1}^{k-1} (1 - v_l)$
- given samples $\theta_1, \dots, \theta_N$ with k distinct values having $n_1, \dots, n_K$ counts

$$
\begin{aligned}
G' &= G(A_1), \dots, G(A_K) | \theta_1, \dots, \theta_n \\
&\sim \text{Dir}(\alpha H(A_1) + n_1, \dots, \alpha H(A_K) + n_k) \\
&\sim \text{Dir}\Big(\delta_{\theta_1 \in B_1} n_1, \dots, \delta_{\theta_K \in B_K} n_K, \alpha H(\Omega \setminus \{\mathrm{d}B_1, \dots \mathrm{d}B_K\}_{\|\mathrm{d}B_k\| \to 0 \; \forall k}
\end{aligned}
$$

$$
\implies (\pi_1, \dots, \pi_k, \pi_u) \sim \text{Dir}(n_1, n_2, \dots n_K, \alpha)
$$

- where π_u are all the probability mass assign to $\theta_{K+1}, \dots, \theta\infty$

Predictive

Let $\alpha_i = \frac{\alpha}{k}$: compute the density of i^{th} data belonging to existing component m.

$$
\begin{aligned}
\Pr(z_i = m|\mathbf{z}_{-1}) &= \int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_k) P(p_1,\dots,p_k|n_{1,-i},\dots,n_{k,-i}) \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K) P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K) P(p_1,\dots,p_K)}{P(n_{1,-i},\dots,n_{k,-i})} \\
&= \frac{\int_{p_1,\dots,p_k} P(z_i = m|p_1,\dots,p_K) P(n_{1,-i},\dots,n_{k,-i}|p_1,\dots,p_K) P(p_1,\dots,p_K)}{\int_{p_1,\dots,p_K} P(n_1^{-i},\dots,n_K^{-i}|p_1,\dots,p_K) P(p_1,\dots,p_K)} \\
&= \frac{\Gamma(\frac{\alpha}{k} + n_{m,-i} + 1) \prod_{l=1, l\neq m}^{k} \Gamma(\frac{\alpha}{k} + n_{l,-i})}{\Gamma(N+\alpha)} \times \frac{\Gamma(N-1+\alpha)}{\prod_{l=1}^{k} \Gamma(\frac{\alpha}{k} + n_{l,-1})} \\
&= \frac{\frac{\alpha}{k} + n_{m,-i}}{N+\alpha-1} \qquad \text{Let } k \to \infty = \frac{n_{m,-i}}{N+\alpha-1}
\end{aligned}
\tag{1}
$$

$\Pr(z_i = \text{new}) = \frac{\alpha}{N+\alpha-1}$.

- conditional density, i.e., to determine which table customer i sit based on seating of the previous customers $1, \dots, i-1$: 假若我们知道第 1 到第 $n-1$ 的人坐的桌子是什么，那第 n 个人坐的桌子的条件概率是:

$$\Pr(z_i = m | \mathbf{z}_{-i}, \alpha) \propto \begin{cases} \frac{n_m^{-i}}{N+\alpha-1} & m^{\text{th}} \text{ existing table (坐第 } m \text{ 张有人坐的桌子)} \\ \frac{\alpha}{N+\alpha-1} & \text{new table (坐新的桌子)} \end{cases}$$

where n_m^{-i} is number of customers **exclude** customer i, sit in table m

n_m^{-i}指的是第 m 张桌子坐了几个人 (除了第 i 个人以外)

- using above conditional density $\Pr(z_i = m | \mathbf{z}_{-i}, \alpha)$, **sampling** of joint density $\Pr(z_1, \dots z_n)$ may be:

1. sample from joint density **directly**:
   直接从联合分布中采样

$$(z_1, z_2, \dots, z_n) \sim \left(\Pr(z_1) \Pr(z_2 | z_1) \dots \Pr(z_n | z_1, \dots z_{n-1}) \equiv \Pr(z_1, \dots z_n) \right)$$

2. sample from joint density: $\Pr(z_n, z_1, \dots z_n)$ using **Gibbs sampling**
   用吉布斯采样从联合分布中采样

Chinese Restaurant Sampling algorithm 中国餐馆过程的采样算法 (2)

Algorithm 1 direct sampling

Require: N = number of customer
Require: T = number of iterations
Require: α = DP concentration

for $t = 1, \ldots, T$ **do**
  for $n = 1, \ldots, N$ **do**
    $z_1^{(t)} \sim \Pr(z_1 | \alpha)$
    $z_2^{(t)} \sim \Pr(z_2 | z_1^{(t)}, \alpha)$
    ...
    $z_i^{(t)} \sim \Pr(z_i | z_1^{(t)}, \ldots, z_{i-1}^{(t)}, \alpha)$
    ...
    $z_N^{(t)} \sim \Pr(z_N | z_1^{(t)}, \ldots, z_{N-1}^{(t)}, \alpha)$
  end for
  generated a sample: $\mathbf{Z}^{(t)} \equiv (z_1^t, \ldots, z_N^t)$
end for
return all samples: $(\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(T)})$

Algorithm 2 Gibbs sampling

Require: N = number of customer
Require: T = number of iterations
Require: α = DP concentration
Require: b = burn-in

start with one joint sample, e.g., $(z_1^{(0)} = 1, \ldots z_N^{(0)} = 1)$
for $t = 1, \ldots, T + b$ **do**
  for $n = 1, \ldots, N$ **do**
    $z_1^{(t)} \sim \Pr\left(z_1 | z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_N^{(t-1)}, \alpha\right)$
    $z_2^{(t)} \sim \Pr\left(z_2 | z_1^{(t)}, z_3^{(t-1)}, \ldots, z_N^{(t-1)}, \alpha\right)$
    ...
    $z_i^{(t)} \sim \Pr\left(z_i | z_1^{(t)}, \ldots, z_{i-1}^{(t)}, z_{i+1}^{(t-1)}, \ldots, z_N^{(t-1)}, \alpha\right)$
    ...
    $z_N^{(t)} \sim \Pr\left(z_N | z_1^{(t)}, z_2^{(t)}, \ldots, z_{N-1}^{(t)}, \alpha\right)$
  end for
  generated a sample: $\mathbf{Z}^{(t)} \equiv (z_1^{(t)}, \ldots, z_N^{(t)})$
end for
return the last T samples after discard burn-in: $(\mathbf{Z}^{(1+b)}, \ldots, \mathbf{Z}^{(T+b)})$

how can we know samples are drawn correctly?, we can check **theoretical** vs **empirical** for:

- Expected number of tables, K 被占桌子个数 K 的期望
- probability over the numbers of occupied tables $\Pr(K = k)$ 被占桌子个数 $\Pr(K = k)$ 的概率

Two Chinese Restaurant Process theoretical properties: $\mathbb{E}(K)$ and $\Pr(K)$ 两个中国餐馆过程的理论特性

if customers all sit according to a Chinese Restaurant Process (CRP) then,

- **Property 1**: what is the **expected number of** occupied tables?
  假若在餐馆中的人按照中国餐馆过程坐的话，那被占的桌子个数的期望是什么呢？
- **Property 2**: what is the **probability on number** of occupied tables?
  假若在餐馆中的人按照中国餐馆过程坐的话，那被占的桌子个数的概率是什么呢？

expected number of occupied tables $\mathbb{E}(K)$ (被占桌子个数的期望) (1)

- We know for expectation, $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ for any two random variables X and Y regardless of whether they are independent or not
  我们都知对于期望来说 $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ 不管随机变量 X 和 Y 是否独立
- **proof**

$$
\begin{aligned}
\mathbb{E}(X+Y) &= \sum_x \sum_y (x+y) P_{XY}(x,y) \\
&= \sum_x \sum_y x P_{XY}(x,y) + \sum_y \sum_x y P_{XY}(x,y) \\
&= \sum_x x \sum_y P_{XY}(x,y) + \sum_y y \sum_x P_{XY}(x,y) \\
&= \sum_x x P_X(x) + \sum_y y P_Y(y) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}
$$

- BTW, what happens to variance of sum of random variables?
  顺便说一下，方差的值可是和变量之间是否独立有关哟！

$$\begin{aligned}
\mathbb{VAR}\left(\sum_{i=1}^n a_i X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^n a_i X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right)^2 = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n a_i a_j X_i X_j\right] - \left(\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right)^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{E}[X_i X_j] - \left(\sum_{i=1}^n a_i \mathbb{E}[X_i]\right)^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{E}[X_i X_j] - \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{E}[X_i]\mathbb{E}[X_j] \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \left(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathbb{COV}(X_i, X_j) = \sum_{i=1}^n a_i^2 \mathbb{VAR}(X_i) + 2 \sum_{i=1}^n \sum_{j:\ j>i} a_i a_j \mathbb{COV}(X_i, X_j)
\end{aligned}$$

expected number of occupied tables $\mathbb{E}(K)$ (被占桌子个数的期望) (3)

- we think the problem as:
  - if N customers are occupying K random number of total tables:
    如果有 N 个顾客霸占了 K 个的桌子 - K 是个随机变量
  - and we have a set of N **binary** random variables $\{k_i^{\text{new}}\}$ each indicate if or not a customer sits on a **new** table
    而且我们有 N 个 0/1 随机变量 $\{k_i^{\text{new}}\}$, 告诉我们第 i 个人是否坐新的桌子
  - then, the random variable K is:

$$K = \sum_{i=1}^{N} k_i^{\text{new}}$$

- then for $\mathbb{E}[K]$:

$$\mathbb{E}[K] = \mathbb{E}(\#\text{ of occupied tables}) = \mathbb{E}\left(\sum_{i=1}^{N} k_i^{\text{new}}\right) = \sum_{i=1}^{N} \mathbb{E}(k_i^{\text{new}})$$

expected number of occupied tables $\mathbb{E}(K)$ (被占桌子个数的期望) (4)

- if every customer has **same** probability t to sit at a **new** table, then K is sum of N **i.i.d Bernoulli** random variables:
  如果每个人都是以 t 概率坐新的桌子

$$\Pr\left(k_i^{\text{new}} = 1(\text{occupied})\right) = t \implies \mathbb{E}(k_i^{\text{new}}) = t$$

- $\Pr(k_i^{\text{new}} | k_1^{\text{new}}, \ldots, k_{i-1}^{\text{new}}) = P(k_i^{\text{new}}) = t$
- for Bernoulli distribution that $\mathbb{E}(X)$ and $P(X = 1)$ has the same value
  在伯努利分布上，期望 $\mathbb{E}(X)$ 和概率 $P(X = 1)$ 值都是一样的

$$\mathbb{E}(\# \text{ of occupied tables}) = \sum_{i=1}^{N} \mathbb{E}(k_i^{\text{new}}) = N \times t$$

- in **CRP**: $\Pr(k_i^{\text{new}} | k_1^{\text{new}}, \ldots, k_{i-1}^{\text{new}}) \neq P(k_i^{\text{new}})$
  - **however** in CRP: $P(k_i^{\text{new}})$ is **independent** of the actual value of previous $\{k_t^{\text{new}}\}_{t=1}^{i-1}$, i.e., **independent** of existing seating arrangement
    在中国餐馆过程中，$\Pr(k_i^{\text{new}})$ 和之前的人到底坐那张桌子无关
  - However, it does depend on how many people are in the restaurant, i.e., value of $i-1$
    但在中国餐馆过程中，$\Pr(k_i^{\text{new}})$ 和之前一共来了几个人有关

- we know each i^{th} **new** person has $\frac{\alpha}{\alpha+i-1}$ probability of occupying a new table:
  我们知道第 i 个新进来的人有 $\frac{1}{\alpha+i-1}$ 概率坐新的桌子

$$\begin{aligned}
\mathbb{E}(\#\text{ of occupied tables}) = \mathbb{E}(K) &= \sum_{i=1}^{N} \mathbb{E}(k_i^{\text{new}}) \\
&= \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1} = \sum_{i=0}^{N-1} \frac{\alpha}{\alpha + i} \\
\text{using following relations:} \qquad & \psi(x+N) - \psi(x) = \sum_{k=0}^{N-1} \frac{1}{x+k} \\
&= \alpha\big(\psi(\alpha+N) - \psi(\alpha)\big) \\
&\approx \alpha \log\left(1 + \frac{n}{\alpha}\right)
\end{aligned}$$

- where

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx \qquad\qquad \psi(x) = \frac{d}{dx} \ln\big(\Gamma(x)\big) = \frac{\Gamma'(x)}{\Gamma(x)}$$

- **Homework** to also prove:

$$\mathbb{VAR}(\#\text{ of occupied tables}) = \alpha\Big(\psi(\alpha+n) - \psi(\alpha)\Big) + \alpha^2\big(\psi'(\alpha+n) - \psi'(\alpha)\big)$$

probability of the number of occupied tables $\Pr(K = k)$

- we already know how to compute the expectation of occupied table, but what about the probability of a particular number of occupied tables?
  我们已然知道怎样计算被占桌子的个数的均值，那么，被占桌子的概率是什么呢?
- K is a random variable indicating "number of times customers sit at **new** tables"
  我们让 K 作为被占桌子的随机变量
- say we are interested in $\Pr(K = 3)$, then:
  假设我们想要知道 $K = 3$ 的概率
  - $1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}$ customer sit at **new tables**, or
  - $1^{\text{st}}, 6^{\text{th}}, 9^{\text{th}}$ customer sit at **new tables**

  can both contribute to $\Pr(K = 3)$
  以上的两种情况都会成为 $\Pr(K = 3)$ 的一部分
- the question is what are the combinations (i.e, coefficient) for each $\Pr(K)$?
  所以每个 $\Pr(K = k)$ 应该有个概率集的组合，也就是系数。我们看下如何得到它？

“binomial coefficient” vs “Stirling number of the first kind”
“二项式系数” 和 “第一种 Stirling 数值”

- in binomial expansion, fixed y:在二项式展开中,y 值是固定的

$$(x+y)^n = \underbrace{(x+y)(x+y)\dots(x+y)}_{n \text{ identical terms}} = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

in terms of probability distribution:将总数为 1 的概率分散在值 $= K$ 的各项当中

$$\mathbf{1} = (\underbrace{p}_{x} + \underbrace{(1-p)}_{y})^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

- now, instead of fix y, increase y by one in each term: $\prod_{y=0}^{n}(x+y)$ 现在我们需要 y 值每次都增加 1 所以肯定不能用二项式展开:

$$\prod_{y=0}^{n}(x+y) = (x+0)(x+1)(x+2)\dots(x+n) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} x^k$$

$\begin{bmatrix} n \\ k \end{bmatrix}$ is called **Stirling number of the first kind**: $\begin{bmatrix} n \\ k \end{bmatrix}$ 系数叫做 “第一种 Stirling 数值”

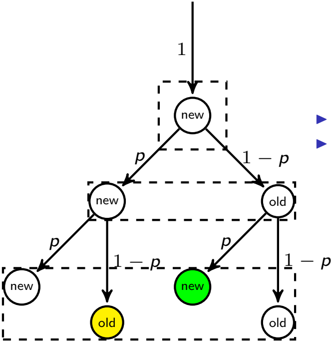
in terms of probability distribution: 将总数为 1 的概率分散在值等于 k 的各项当中:

$$\mathbf{1} = \frac{(x+0)(x+1)(x+2)\dots(x+n)}{(x+0)(x+1)(x+2)\dots(x+n)} = \frac{\sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} x^k}{(x+0)(x+1)(x+2)\dots(x+n)}$$

- let's see which of these above probability distribution used in Chinese Restaurant Process: 我们看下 CRP 到底需要什么样的概率展开

probability on "number of tables" $\Pr(K)$ - independent probabilities

- **case 1** every person sits on a new table with probablity p: 在此情况，每个人不管啥时候进入餐馆，都是以 p 概率坐新桌子:



- first person always occupies the first table with probability of 1; therefore, when computing $\Pr(K = k|N)$, we can only assign $N - 1$ to $k - 1$ tables: 第一个人总是坐一张新桌子，所以当我们计算 $\Pr(K = k|N)$ 时，我们只能对剩下 $N - 1$ 人进行 $k - 1$ 新桌子的分配
- for example, $N = 3$, $\Pr(K = 2|N = 3)$ requires total of $\binom{N-1}{2-1} = 2$ paths, and the sum of these two paths are: 以上概率会有 2 条路径，他们的概率总和是:

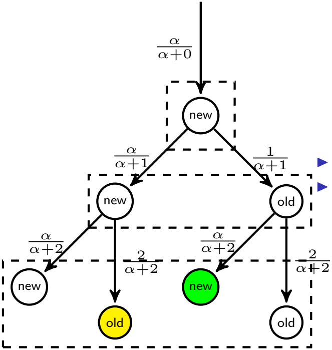
$$\Pr(K = 2|N = 3) = \underbrace{1}_{\text{first}} \times \Big[p(1-p) + (1-p)p\Big]$$
$$= 2(1-p)p$$

- sum of probabilities of all paths must equal 1 所有路径概率的总和加起来等于 1
- probability distribution for each $K = k$ 用独立坐位子概率分配是:

$$\underbrace{1}_{\text{first}} \times \underbrace{(p + (1-p))^{N-1}}_{\text{subsequent}} = 1 \times \left[\sum_{i=0}^{2} \binom{2}{i} p^i (1-p)^{2-i}\right]$$
$$= 1 \times \left[\underbrace{\binom{2}{2} p^2}_{K=3} + \underbrace{\binom{2}{1} p(1-p)}_{K=2} + \underbrace{\binom{2}{0} (1-p)^2}_{K=1}\right]$$
$$= 1 \times \left[p^2 + 2p(1-p) + (1-p)^2\right]$$

probability of the number of tables $\Pr(K)$ - Chinese Restaurant Process

- **case 2** every person sits on a new table with probablity govern by CRP
在此情况，每个人进入餐馆时都是以中国餐馆过程概率坐新桌子:



- first person always occupies the first table with probability of 1; therefore, when computing $\Pr(K = k|N, \alpha)$, we can only assign $N - 1$ to $k - 1$ tables:
第一个人总是坐一张新桌子，所以当我们计算 $\Pr(K = k|N)$ 时，我们只能对剩下 $N - 1$ 人进行 $k - 1$ 新桌子的分配
- for example, $N = 3$, $\Pr(K = 2|N = 3, \alpha)$ requires total of $\binom{N-1}{2-1} = 2$ paths, and the sum of these two paths are: 以上概率会有 2 条路径，他们的概率总和是:

$$\Pr(K = 2|N = 3, \alpha) = \underbrace{\frac{\alpha}{(\alpha + 0)}}_{\text{fixed}} \frac{\alpha}{(\alpha + 1)} \frac{\mathbf{2}}{(\alpha + 2)} + \frac{\alpha}{(\alpha + 0)} \frac{\mathbf{1}}{(\alpha + 1)} \frac{\alpha}{(\alpha + 2)}$$

- sum of probabilities of all paths must equal 1 所有路径概率的总和加起来等于 1
- probability distribution for each $K = k$ 用 CRP 的概率分配是:

$$
\begin{aligned}
1 = \frac{(\alpha + 0)(\alpha + 1)(\alpha + 2)}{(\alpha + 0)(\alpha + 1)(\alpha + 2)} &= \frac{\alpha^3 + \left[{3 \atop 2}\right]\alpha^2 + \left[{3 \atop 1}\right]\alpha}{(\alpha + 0)(\alpha + 1)(\alpha + 2)} \\
&= \frac{\alpha^3 + \mathbf{3}\alpha^2 + 2\alpha}{(\alpha + 0)(\alpha + 1)(\alpha + 2)} \\
&= \frac{\Gamma(\alpha)[\alpha^3 + \mathbf{3}\alpha^2 + 2\alpha]}{\Gamma(\alpha + 3)}
\end{aligned}
$$

- for an individual probability: $\Pr(K = k|N, \alpha) = \left[{\mathbf{N} \atop \mathbf{k}}\right]\alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)}$:

Partition Model (划分模型)

- **problem** when there are N customers, one may potentially assign them to $k \in \{1, \dots, N\}$ tables; what is probability of a partition having $\{n_1, n_2, \dots, n_K\}$ customers:
  划分模型的问题是，当有 N 个人的时候，可以把他们划分在 $k \in \{1, \dots, N\}$ 的桌子中。假若某种划分结果是有 K 个桌子，每个桌子有 $\{n_1, n_2, \dots, n_K\}$ 个顾客，那这样划分的概率是什么?
- let $K \equiv n(\Pi)$, number of tables for a particular partition:
- one may have the following two"partitions":
  以下的两种划分都是等价的

$$\{3, 1, 2, 3, 2, 3, 2, 3\} \implies \{n_1 = 1, n_2 = 3, n_3 = 4\}$$
$$\{3, 3, 3, 2, 1, 1, 1, 1\} \implies \{n_1 = 4, n_2 = 1, n_3 = 3\}$$

  they are **equivalent**:
- **in words**: for all partitions of 8 customers having:
  "4 *customers sit one of the table,* 3 *sit one of the table and* 1 *customer sit on one table*"
  then these partitions should be treated **equivalently**, i.e., it does **not** matter which particular table has 4 customers
- obviously, different **process** in generate customers seating result in different probabilities of partitions

Partition Model using CRP (用中国餐馆过程做划分模型)

- we know the conditional density $\Pr(z_i = m | z_{i-1}, \dots z_1) \equiv \Pr(z_i = m | \mathbf{z}_{-i}, \alpha)$

$$\Pr(z_i = m | \mathbf{z}_{-i}, \alpha) \propto \begin{cases} \frac{n_{m,-i}}{N+\alpha-1} & \text{for existing cluster } m \\ \frac{\alpha}{N+\alpha-1} & \text{for new cluster} \end{cases}$$

- using DP, the probability on a partition is:

$$\pi(\Pi_N) = \frac{\alpha^k \prod_{l=1}^{k} \Gamma(n_l)}{\prod_{i=1}^{n} (\alpha + i - 1)}$$

- $k \equiv n(\Pi)$: number of tables (clusters)
- n_l: number of customers in a table (number of data in a cluster)

Partition Model uisng CRP

- probability of a cluster C_j taking value m: $\Pr(C_j = m|\Pi_{-n_j}, \alpha)$ for example, let $j = 2$:

$$\{3, 1, ?, 1, ?, 3, 3, 3\} \implies \underbrace{\{n_1 = 2, \mathbf{n_2} = ?, n_3 = 4\}}_{\Pi_{-n_2}}$$

- knowing **joint** probability of a partition is $\pi(\Pi_N) = \frac{\alpha^k \prod_{l=1}^{k} \Gamma(n_l)}{\prod_{i=1}^{n}(\alpha+i-1)}$:
- sampling of a **conditional** can be achieved in the following fashion: (canceling denominator):

$$\Pr(C_2 = m|\Pi_{-n_2}, \alpha) \propto \begin{cases} \underbrace{\begin{matrix}\left[\alpha\Gamma(n_1 + \mathbf{n_2})\right]\left[\alpha\Gamma(n_3)\right] \\ \left[\alpha\Gamma(n_1)\right]\left[\alpha\Gamma(n_3 + \mathbf{n_2})\right]\end{matrix}}_{} & \begin{matrix} m = 1, \text{ i.e., existing} \\ m = 3, \text{ i.e., existing}\end{matrix} \\ \left[\alpha\Gamma(n_1)\right]\left[\alpha\Gamma(\mathbf{n_2})\right)\left[\alpha\Gamma(n_3)\right] & m = 2, \text{ i.e., new} \end{cases}$$

omit the case of $m = 1$, as it's in the same form as any other existing component indices $m = 3$:

$$\Pr(\mathbf{C_2} = m|\Pi_{-n_2}, \alpha) \propto \begin{cases} \underbrace{\alpha^2\Gamma(n_1)\Gamma(n_3 + \mathbf{n_2})}_{} & m = 3, \text{ i.e., existing} \\ \alpha^3\Gamma(n_1)\Gamma(\mathbf{n_2})\Gamma(n_3) & m = 2, \text{ i.e., new} \end{cases}$$

after cancellation and write it out generally:

$$\Pr(\mathbf{C_j} = m|Pi_{-n_j}, \alpha) \propto \begin{cases} \frac{\Gamma(n_t + \mathbf{n_j})}{\Gamma(n_t)} & m = t, \text{ i.e., existing} \\ \alpha\Gamma(\mathbf{n_j}) & m = \text{new} \end{cases}$$

- in the case of $n_j = 1$, we get **Chinese Restaurant Process**:

$$\Pr(\mathbf{z_j} = m|\Pi_{-n_2}, \alpha) \propto \begin{cases} \frac{\Gamma(n_t + \mathbf{1})}{\Gamma(n_t)} \\ \alpha\Gamma(\mathbf{1}) \end{cases} = \begin{cases} n_t & m = t, \text{ i.e., existing} \\ \alpha & m = \text{new} \end{cases}$$

sampling for Dirichlet Process with data points

- an infinite mixture density (e.g. Gaussian) can be written as:

$$\Pr(z_i|\mathbf{z}_{-i}, y_i, \Theta) = \frac{\alpha}{n-1+\alpha} \int F(y_i|\theta)H(\theta)\mathrm{d}\theta$$

Slice sampling for Dirichlet Process

- an infinite mixture density (e.g. Gaussian) can be written as:

$$f_{\pi,\theta}(y) = \sum_{j=1}^{\infty} \pi_{j=1} \mathcal{N}(y|\theta_j) \qquad \text{where } \theta = (\mu, \sigma^2)$$

- adding slice variable u:

$$f_{\pi,\theta}(y, u) = \sum_{j=1}^{\infty} \mathbf{1}(u < \pi_j) \mathcal{N}(y|\theta_j)$$

- to ensure **marginal is invariant**:

$$\begin{aligned}
\int f_{\pi,\theta}(y, u) \mathrm{d}u &= \int_0^{\pi_j} \sum_{j=1} \mathbf{1}(u < \pi_j) \mathcal{N}(y|\theta_j) \mathrm{d}u \\
&= \sum_{j=1}^{\infty} \mathcal{N}(y|\theta_j) \int_0^{\pi_j} \mathbf{1}(u < \pi_j) \mathrm{d}u \\
&= \sum_{j=1}^{\infty} \mathcal{N}(y|\theta_j) \times \pi_j \\
&= f_{\pi,\theta}(y)
\end{aligned}$$

- note this is in the **absence of latent variable** z_i (later slides)

Slice variable u

finite model: $$P(y|\pi,\theta) = \frac{1}{K}\sum_{j\in\{1\dots K\}} \mathcal{N}(y|\theta_j)$$

infinite model: $$P(y|\pi,\theta,u) \equiv f_{\pi,\theta}(y|u) = \frac{1}{\underbrace{\#\{A_\pi(u)\}}_{f_\pi(u)}}\sum_{j\in A_\pi(u)} \mathcal{N}(y|\theta_j) = \frac{1}{f_\pi(u)}\sum_{j\in A_\pi(u)} \mathcal{N}(y|\theta_j)$$

- $f_\pi(u)$ is a **a random integer**

$$\begin{aligned} f_\pi(u) &= \sum_{j=0}^{\infty} \mathbf{1}(u < \pi_j) \\ &= \sum_{j=0}^{\infty} \pi_j \mathcal{U}(u|0,\pi_j) \qquad \text{where } \mathcal{U}(u|0,\pi_j) = \begin{cases} \frac{1}{\pi_j}, & u < \pi_j \\ 0, & u > \pi_j \end{cases} \end{aligned}$$

Latent variable z_i

- latent variable z identify the component which y is to be taken:

$$f_{\pi,\theta}(u, z, y) = \mathcal{N}(y|\theta_z)\mathbf{1}(z \in A(u))$$

- for example, $u_6 = 0.15$ and $A(u_6) = \{2, 4, 5, 6\}, k_6 = 4 \in A(u_6) \implies \pi_4 > 0.15$
- If there are n samples, complete data likelihood:

$$\mathcal{L}_{\pi,\theta}(\{y_i, u_i, z_i\}_{i=1}^{n}) = \prod_{i=1}^{n} \mathcal{N}(y_i|\theta_{z_i})\mathbf{1}(u_i < \pi_{z_i})$$

sampling algorithm

1. $u_i \sim U(0, \pi_{z_i})$
2. $f(\theta_j | \cdots) \propto H(\theta_j) \prod_{z_i = j} \mathcal{N}(y_i | \theta_j)$
   If there are no $z_i = j$, then $f(\theta_j | \cdots) = H(\theta_j)$
3. $f(v | \cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}(\pi_{z_i} > u_i)$

$$f(v|\cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}(\pi_{z_i} > u_i) = \pi(v) \prod_{i=1}^{n} \mathbf{1}\Big(\underbrace{v_{z_i} \prod_{l<z_i} (1 - v_l)}_{\pi_{z_i}} > u_i \Big)$$

$$= \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1}\Big(\underbrace{v_{z_i} \prod_{l<z_i} (1 - v_l) > u_i}_{\gamma_j < v_j < \beta_j} \Big)$$

- the above only applies when $j \leq z^*$, where z^* is the maximum of $\{z_1, \ldots, z_n\}$
- for γ_j and β_j must be a function of u_i and α
- for $j > z^*$, $f(v_j | \cdots) = \text{beta}(1, \alpha)$

lower bound γ_j

$$f(v|\cdots) = \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1}\left(\underbrace{v_{k_i} \prod_{l<z_i}(1-v_l) > u_i}_{\gamma_j < v_j < \beta_j} \right)$$

- **lower bound** means how **low** you can reduce v_j to
- **reduce** $v_j \implies$ **reduce** π_j
- therefore, one needs to ensure all: $\{\pi_{z_i=j}\} > u_i$:

$$v_{z_i} \prod_{l<z_i}(1-v_l) > \max_{\{i:z_i=j\}}(u_i)$$

$$\implies v_{z_i} > \frac{\max_{\{z_i=j\}}(u_i)}{\prod_{l<z_i}(1-v_l)}$$

$$\implies v_{z_i} > \underbrace{\max_{\{z_i=j\}}\left(\frac{u_i}{\prod_{l<z_i}(1-v_l)}\right)}_{\gamma_j}$$

- $\pi_{j+1}, \pi_{j+2}, \ldots$ will **increase**: there is more to share now - but not affected by lower bound
- $\pi_1, \ldots, \pi_{j-1}$ will **not** be affected

upper bound β_j

$$f(v|\cdots) = \underbrace{\pi(v)}_{\text{beta}(1,\alpha)} \prod_{i=1}^{n} \mathbf{1}\underbrace{\left(v_{z_i} \prod_{l<z_i} (1 - v_l) > u_i \right)}_{\gamma_j < v_j < \beta_j}$$

- **increase** $v_j \implies$ **increase** $\pi_j \implies$ **reduce** $\pi_{j+1}, \pi_{j+2}, \ldots$
- therefore, one needs to ensure all: $\{\pi_{k_i > j}\} > u_i$
- as an **illustrative example**, we let $(j = 3)$ and a particular $(z_i = 5)$:

$$\begin{aligned}
&\pi_{z_i=5} > u_i \\
&\qquad \implies (1 - v_1)(1 - v_2)(\mathbf{1 - v_3})(1 - v_4)v_5 > u_i \\
&\qquad \implies (1 - v_1)(1 - v_2)(1 - v_4)v_5 - \mathbf{v_3}(1 - v_1)(1 - v_2)(1 - v_4)v_5 > u_i \\
&\qquad \implies v_3(1 - v_1)(1 - v_2)(1 - v_4)v_5 < (1 - v_1)(1 - v_2)(1 - v_4)v_5 - u_i \\
&\qquad \implies v_3 < 1 - \frac{u_i}{(1 - v_1)(1 - v_2)(1 - v_4)v_5}
\end{aligned}$$

- however, one needs to ensure v_3 (or v_j in general) satisfies: $\{\forall\ z_i > j\}$, write it generally:

$$\begin{aligned}
v_j &< \min_{\{z_i > j\}} \left(1 - \frac{u_i}{v_{z_j} \prod_{l<z_i, l\neq j} (1 - v_l)} \right) \\
\implies v_j &< \underbrace{1 - \max_{\{z_i > j\}} \left(\frac{u_i}{v_{z_j} \prod_{l<z_i, l\neq j} (1 - v_l)} \right)}_{\beta_i}
\end{aligned}$$

- $\pi_1, \ldots, \pi_{j-1}$ and π_j will **not** be affected

sampling via inverse CDF of v_j

- We can define the **truncated** CDF distribuiton of v:

$$F(v) = \frac{1}{C}\int_{\gamma_j}^{v} f(v|\cdots)\mathrm{d}v$$
$$= \frac{\int_0^v \text{beta}(v|1,\alpha)\mathbf{1}(\gamma_j < v < \beta_j)\mathrm{d}v}{\int_0^1 \text{beta}(v|1,\alpha)\mathbf{1}(\gamma_j < v < \beta_j)\mathrm{d}v} = \frac{\int_{\gamma_j}^{v} \text{beta}(v|1,\alpha)\mathrm{d}v}{\int_{\gamma_j}^{\beta_j} \text{beta}(v|1,\alpha)\mathrm{d}v}$$

- looking at the property of beta distribution:

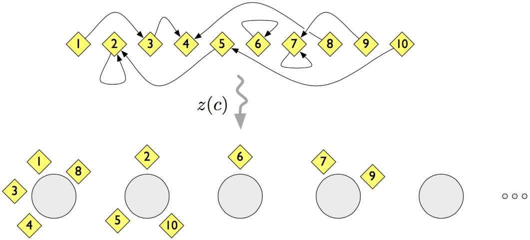
$$\int_{\gamma_j}^{v_j} \text{beta}(v|1,\alpha)\mathrm{d}v = \int_{\gamma_j}^{v_j} \frac{\Gamma(1+\alpha)}{\Gamma(1)\Gamma(\alpha)} v^{1-1}(1-v)^{\alpha-1}\mathrm{d}v$$
$$= \alpha \int_{\gamma_j}^{v_j} (1-v)^{\alpha-1}\mathrm{d}v$$
$$= (1-\gamma_j)^{\alpha} - (1-v_j)^{\alpha}$$

- So, we can prove that:

$$F(v_j) = \frac{(1-\gamma_j)^{\alpha} - (1-v_j)^{\alpha}}{(1-\gamma_j)^{\alpha} - (1-\beta_j)^{\alpha}}$$

- this is where **inverse CDF** becomes useful

dd-CRP



- instead of sample class variable for nodes, it samples links:

$$\Pr(c_i = j | D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{else} \end{cases}$$

- MATLAB code download:
`http://www-staff.it.uts.edu.au/~ydxu/software1.htm`