Haupthseminar: Machine Learning

Chinese Restaurant Process, Indian Buffet Process

Agenda

- Motivation
- Chinese Restaurant Process- CRP
  - Dirichlet Process
  - Interlude on CRP
  - Infinite and CRP mixture model
  - Estimation of required tables
- Indian Buffet Process- IBP
  - Latent Feature Model - Finite Features
  - Interlude on IBP- Infinite Features

Motivation

Stochastic process : ‘Study of non deterministic systems’

Discrete Probability distribution

Continuous Probability distribution

Non parametric and Unsupervised learning especially clustering, when data points are coming from hidden groups.

Chinese Restaurant Process

Beta Distribution

Beta Distribution is a continuous probability distribution between interval [0,1] which provides a belief that event can happen with certain probability θ.

The distribution is parameterized by α and β defining the shape of distribution.

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Chinese Restaurant Process - CRP

Dirichlet Distribution

Dirichlet Distribution is a generalization of Beta distribution which gives a belief/probability that set of events have corresponding probabilities given that every event is occurred at least α-1 times.

Formal definition,

$$X = \{x_1, x_2, x_3, ...x_K\} \; ; \sum_{i=1}^{K} x_i = 1$$

$$P(X) \; \alpha \; \prod_{i=1}^{K} x_i^{\alpha_i - 1} = (1/Beta(\alpha)) * \prod_{i=i}^{K} x_i^{\alpha_i - 1}$$

$$\alpha = \{\alpha_1, \alpha_2, ..., \alpha_K\} \; ; \; \alpha_i > 0$$

CRP- continued, Sampling from Dirichlet Process

Inputs:

1. { R_1 , R_2 R_3 ,.....} ; Set of random variables in study
2. B ; Base distribution.

Process:

1. Draw R_1 from Base distribution B.
2. for n >1

Draw/Normalize R_n from Base B with P(R_n)= $\alpha/\alpha+n-1$ or

Sample R_n with P(R_n)= $|n_{R_n}|/(\alpha+n-1)$ such that

$n_R=\{R_j\}\, j<n$ where every element of n_{R_n} is a previous observation.

Chinese Restaurant Process - CRP

Dirichlet Process

The Dirichlet Process is a stochastic discrete probability distribution process usually used when modeling data that tends to repeat previous values in a "rich gets richer" fashion.

if R ~ DP(B,α) then,

$$R(\{o_i\}_{i=1}^{n}) \sim Dirichlet(\boldsymbol{\alpha} B(\{o_i\}_{i=1}^{n}))$$

Interlude on CRP

Scenario

Chinese restaurant where the customer enters in to the restaurant having tables with infinite capacity. At a random time the customer can choose to occupy the table which is empty or can choose the table which is occupied before by certain customers.

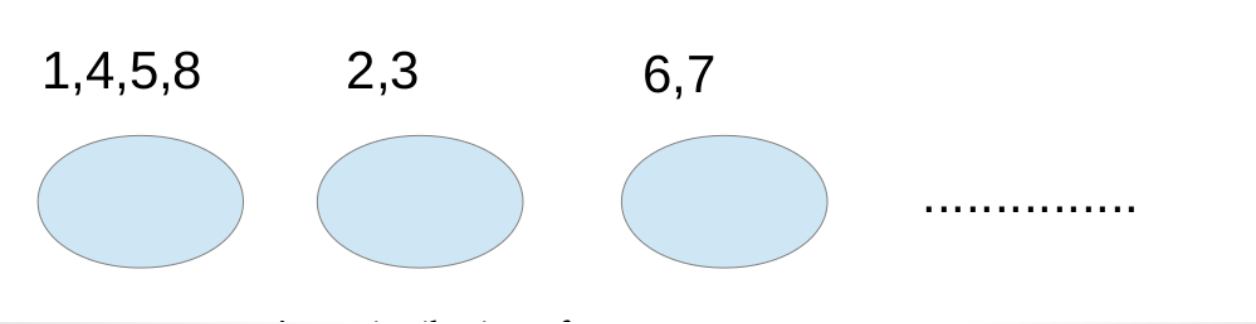
- The important focal point here is to study the probability distribution of such random event.

- t_i table occupied by customer *i*.

$$P(\{t_i\}_{i=1}^{n}) = \prod_{i=i}^{n} P(t_i|t_1, t_2, t_3 \ldots\ldots\ldots\ldots, t_{i-1})$$

Interlude on CRP continued….

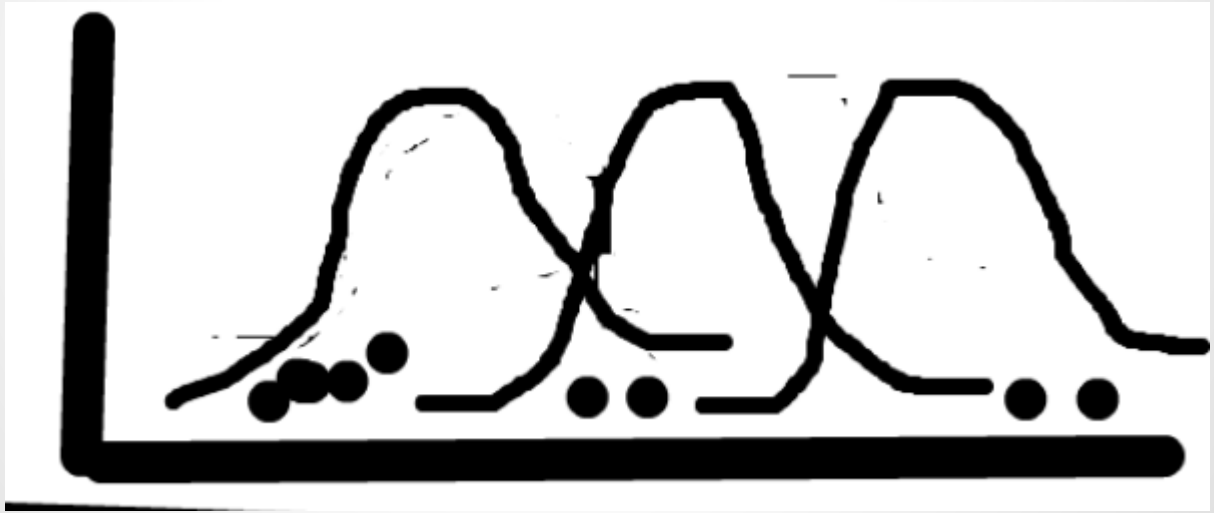
Distribution of 8 customers around 3 tables at time t=8.



$$p(t_1, t_2, t_3 \ldots\ldots, t_n) = \frac{\prod_{k=1}^{K_n}(b_k - 1)!}{n!} \quad Where, \; b_k = \sum_{i=1}^{n} II(t_k = t_i)$$

CRP Continued

Infinite Mixture Model



CRP mixture

- Endow each table with mean $m_k \sim N(0, \sigma_m^2)$

- Choose the cluster assignment $t_n \sim CRP(\alpha; t_1, t_2, \ldots\ldots t_{n-1})$

- Draw a data point $c_n \sim N(m_{t_n}, \sigma_x^2)$

- Posterior distribution of data point given *n* data points

$$p(c_{n+1} | c_{1:n}, m_{1:n}, \sigma_x^2, \sigma_m^2) = \sum_{t=1}^{K_n+1} p(t|t_{1:n}) p(c_{n+1} | t, m_t)$$

It is also observed that all the clusters are Independent and identically Distributed(i.i.d), therefore the random numbers (numerator) can be exchanged without modifying the probabilities(denominator) and this is called Theory of Ex-changeability.

Estimation of Required tables

Its just sufficient to consider the occupied tables for probability calculation and hence the estimation of required tables is vital.

This results in the O(α log(α+ n−1)).

$$K_n = \int_i \frac{\alpha}{\alpha+n-i} di = \alpha H_n$$

Latent Feature Model

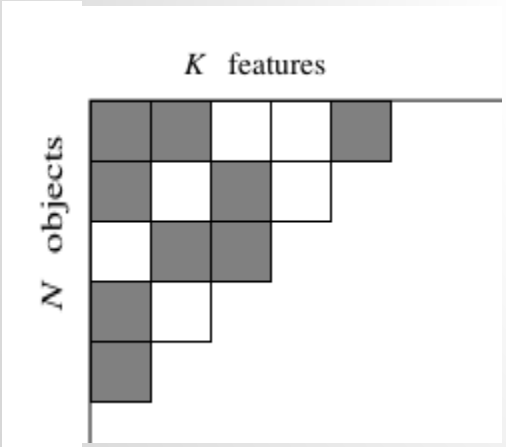
The Latent feature model aims to study the relationship of the features associated with the heterogeneous objects. The relationship here refers to the probability study of features that object can possess.

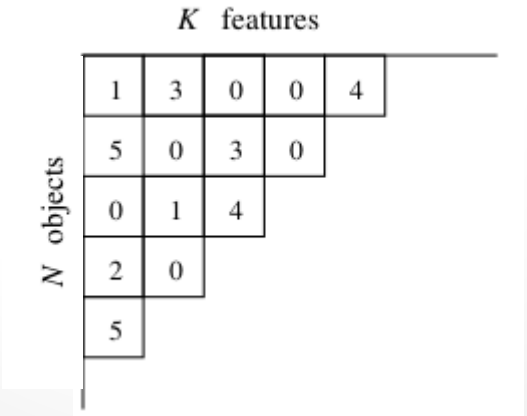
The latent feature model is represented using matrix M= $[f_1, f_2, \ldots, f_n]$ here every is a column vector having *k* dimension representing *k* feature values for an object and hence it forms k X n matrix.

Latent Feature model

-- M= B ⊗ V

-- Latent Feature Matrix M is generated by the elementary multiplication of corresponding feature vector and Binary Vector





Latent Feature Model

Consider the random variable for every feature defines a probability with which the particular object can take a feature i.e. $\theta=\theta_1, \theta_2, \ldots\ldots\ldots, \theta_K$

$$P(\theta_k) = (\Gamma(a+b)/\Gamma(a)\Gamma(b)) * Beta(\theta_k; a, b)$$

$$Beta(\theta_k; a, b) = (1/Beta(a,b)) * {\theta_k}^{a-1}(1-\theta_k)^{b-1} = P(\theta_k)$$

Latent Feature Model

The probability of producing such Binary matrix can be analyzed using marginal probability distribution.

In latent feature model we have two random variables(B,Θ) with collection of discrete values. The probability of producing such B with 1's and 0's with the condition where the probability that object can take feature *k* is θ_k

$$P(B) = \prod_{k=1}^{K} \int \left(\prod_{i=1}^{N} P(B_{ik} = 1 | \theta_k) \right) P(\theta_k)\, d\theta_k$$

Latent Feature Model

The marginal distribution can be represented as product of Beta distribution and by applying gamma transformation $\text{Beta(a,b)} = \Gamma(a)\,\Gamma(b)/\Gamma(a+b)$:he probability of producing such binary matrix is given by following equation, $n_k = \sum_{i=1}^{N} B_{ik}$

$$P(B) = \prod_{k=1}^{K} Beta(n_k + \alpha/K, N - n_k + 1)/Beta(\alpha/K, 1)$$

$$\prod_{k=1}^{K} \frac{(\alpha/K)(n_k + \alpha/K - 1)! * (N - n_k)!}{(N + \alpha/K)!}$$

Latent Feature Model, Taking Infinite Limit

-- The Equivalent class represents the set of matrices having identical lof, i.e [B] implies that set of binary matrices having lof equivalent to B.

-- The probability of such equivalence class helps in analysis when number of features tends to infinity.

$$P([B]) = K!/\prod_{i=1}^{2^N-1} K_h! \prod_{k=1}^{K} \frac{\alpha/K\,\Gamma(n_k+\alpha/K)*\Gamma(N-n_k+1)!}{\Gamma(N+1+\alpha/K)}$$

$$P([B]) = \frac{\alpha^{K^+} e^{-\alpha H n}}{\prod_{i=1}^{2^N-1} K_h!} \prod_{k=1}^{K^+} \frac{(n_k-1)!*(N-n_k)!}{N!}$$

Indian Buffet Process (IBP): Latent Feature Model- Infinite features

- In India, Buffet is a common style of offering variety of dishes to the incoming customers.The customers enters in can sample dishes in random order which grabs the interest in analyzing the probability distribution of such random events.
- Assume the scenario where infinite number of dishes are available and N customers enters in to the restaurant and start sampling the dishes.
- We can define a distribution over infinite binary matrices by specifying a procedure by which customers (objects) choose dishes (features).

IBP

- The first customer starts at the left of the buffet and takes a serving from each dish, stopping after a Poisson(α) number of dishes as his plate becomes overburdened.
- The ith customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself with probability n_k/i, where n_k is the number of dishes sampled before by i previous customers. Having reached the end of all previous sampled dishes, the ith customer then tries a Poisson(α/i) number of new dishes.
- The binary matrix B with N rows and infinitely many columns, where $B_{ik}=1$ if ith customer sampled the kth dish

IBP

The poisson distribution deals with the probability that maximum number of events happening with in certain time interval or space.

In IBP, sampling the new dishes from the right of previously sampled dishes is given by poisson distribution. Where K={1,2,…….} features.

$$P(X = k) = \alpha^k e^{-\alpha}/k!$$

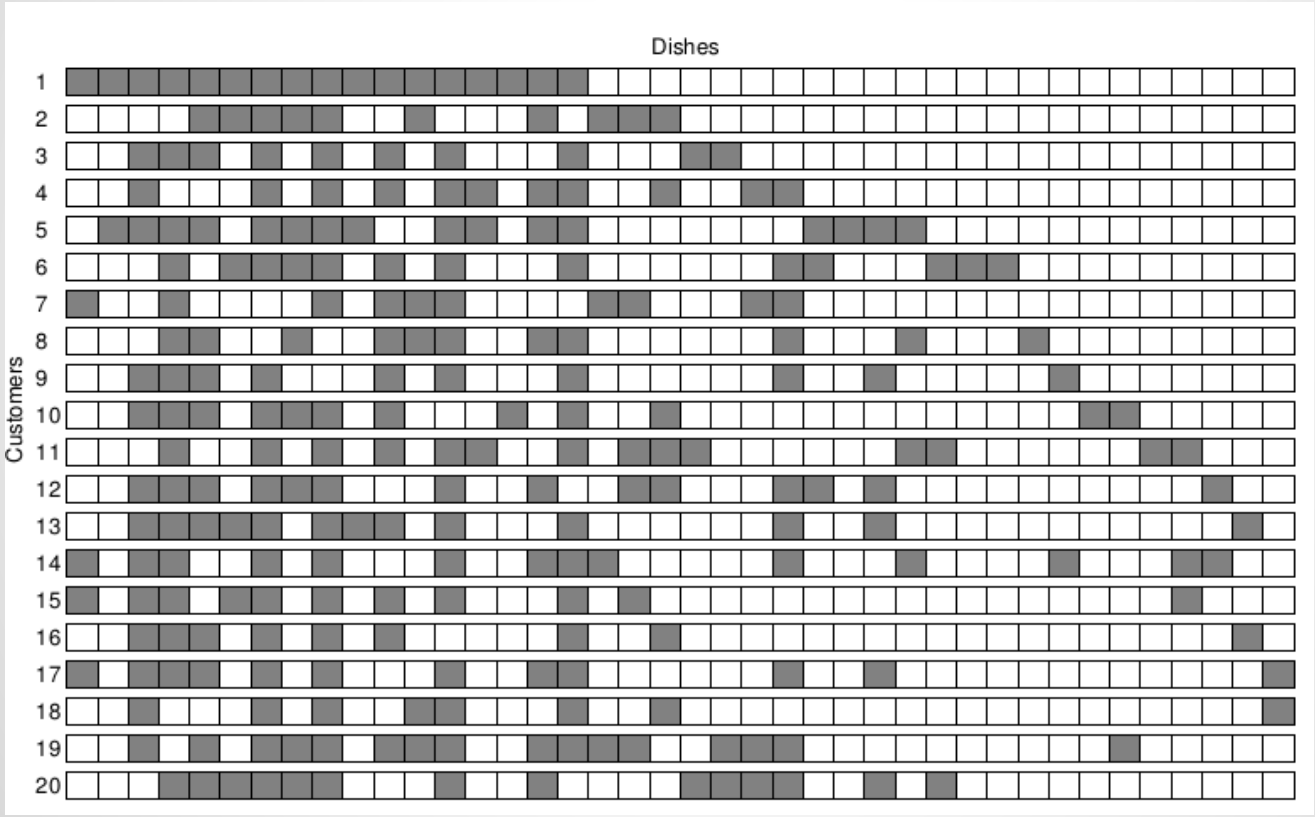
Every Binary matrix at time *i* enable provides the behavior of *i* customers sampling the dishes.

IBP

The new customer takes previously sampled dishes with the probability $\theta_k = n_k / i$ and from finite latent feature model we know that such probability is beta distributed $\text{Beta}(\alpha/k, 1)$ as we know the number of dishes are infinite and hence applying by limit $K = \infty$ the marginal distribution of B for Indian Buffet process is given below, where K^+ is the number of features possessed by objects.

$$P(B) = \frac{\alpha^{K^+} e^{-\alpha H_N}}{\prod_{i=1}^{N} K_1^{(i)}!} \prod_{k=1}^{K^+} \frac{(n_k - 1)! * (N - n_k)!}{N!}$$

IBP example, α=10 from [9]



IBP example

Previous example shows a matrix generated using the IBP with $\alpha = 10$. The first customer tried 17 dishes. The second customer tried 7 of those dishes, and then tried 3 new dishes.

The third customer tried 3 dishes tried by both previous customers, 5 dishes tried by only the first customer, and 2 new dishes. Vertically concatenating the choices of the customers produces the binary matrix.

IBP Applications

By combining the IBP with different likelihood functions we can get different kinds of models:

•Models for graph structures	(w/ Wood, Griffiths, 2006)
•Models for protein complexe	(w/ Chu, Wild, 2006)
•Models for overlapping clusters	(w/ Heller, 2007)
•Models for choice behaviour	(G¨or¨ur, J¨akel & Rasmussen, 2006)
•Models for users in collaborative filtering	(w/ Meeds, Roweis, Neal, 2006)
•Sparse latent factor models	(w/ Knowles, 2007)

Conclusions

- The CRP provides an evident that dynamic creation of clusters/categories where there is no bound on the number of such clusters.
- The CRP represents Dirichlet process and customers are exchangeable.
- The IBP is a representation of Latent feature model and emphasize on infinite sampling. this distribution can be used to automatically infer the number of features required to account for observed data.

References

[1] Wikipedia, *http://en.wikipedia.org/wiki/Stochastic_process*.

[2]Wikipedia, *http://en.wikipedia.org/wiki/Dirichlet_process*.

[3] Yee Whye. “Teh Bayesian Nonparametric Modelling: Dirichlet Processes, Hierarchical Dirichlet Processes, Indian Buffet Process”. *Gatsby Computational Neuroscience Unit University College London* .April 17, April 25, 2008 / MLII .

[4] David Blei,Peter Frazier and Indraneel Mukherjee. “COS 597C: Bayesian nonparametrics ”, Lecture Notes, September 21 2007.

[5] Brian Kulis . “CSE 788.04: Topics in Machine Learning ”, Lecture Notes, May 7 2012.

[6] Wikipedia, http://en.wikipedia.org/wiki/Beta_distribution.

[7] http://www3.nd.edu/~nancy/Math30530/ Info/beta.pdf, “Relationship between Beta - and Gamma”

[8] http://homepage.tudelft.nl/11r49/documents/wi4006/gammabeta.pdf,“The gamma -and beta function”.

[9] Thomas L. Griffiths and Zoubin Ghahramani. “The Indian Buffet Process: An Introduc--tion and Review”. *Journal of Machine Learning Research, 12 (2011) 1185-1224* .

[10] Wikipedia, http://en.wikipedia.org/wiki/Poisson_distribution.