

Online Chinese Restaurant Process

Emmanuel Adjei
University of California, Irvine
Statistics Department

June 12, 2024

OUTLINE

- Introduction
- Online Learning Framework for Linear Classification
- Dirichlet Process Mixture Models
- Online Chinese Restaurant Process
- Experimental Results
- Conclusions
- Reference

Introduction



Figure: Source:<https://www.scribbr.com/ai-tools/machine-learning/>

Big data originates from computers, phones, and sensor networks. Processing this vast amount of data has become a significant challenge for many researchers, highlighting the need to develop effective algorithms for data processing. Some methods to tackle this issue include partitioning the data and using online learning approaches.

Online Learning Framework for Linear Classification

Online learning operates on a sequence of data samples with time stamps. At each step t , the learner receives an incoming sample $\mathbf{x}_t \in \mathcal{X}$ in a k -dimensional vector space, that is, $\mathcal{X} = \mathbb{R}^k$ [1].

- Initialize: $\mathbf{w}_1 = 0$ $t = 1, 2, \dots, T$
- The learner receives an incoming instance: $\mathbf{x}_t \in \mathcal{X}$;
- The learner predicts the class label: $\hat{y}_t = \text{sgn}(f(\mathbf{x}_t; \mathbf{w}_t))$;
- The true class label is revealed from the environment: $y_t \in \mathcal{Y}$;
- The learner calculates the loss: $\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t))$; If $\ell(\mathbf{w}_t; (\mathbf{x}_t, y_t)) > 0$
- The learner updates the classification model:
 $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \Delta(\mathbf{w}_t; (\mathbf{x}_t, y_t))$;

Dirichlet Process Mixture Models

Dirichlet Process

A DP is a distribution over probability measures. A DP has two parameters, base measure θ (as the mean), and a concentration parameter α (as precision). Denote it by $\text{DP}(\alpha, \theta)$

Therefore if a probability measure $G \sim \text{DP}(\alpha, \theta)$, then

$$(G(B_1), \dots, G(B_K)) \sim \text{Dir}(\alpha\theta(B_1), \dots, \alpha\theta(B_K))$$

for any finite measurable partition (B_1, \dots, B_K) of Ω .

This is based on the idea behind Dirichlet distribution. Supposed K -dimensional weight vector (π_1, \dots, π_K) satisfying $\sum \pi_i = 1$ and each $\pi_i \geq 0$, we can define a Dirichlet distribution with parameter $(\alpha_1, \dots, \alpha_K)$ with a density [3]

$$\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

Dirichlet Process Mixture Models

DP mixture is essentially a convolution of a continuous kernel $K(y | \theta_i, \phi)$ with a DP discrete realization. By convolution, we gain easy interpretation (clustering) and smooth functions, e.g., densities.

$$Y | \theta_i, \phi \sim K(y | \theta_i, \phi)$$

Consider a DP mixture model:

$$y_i | \theta_i \sim K(y_i | \theta_i), \theta_i | G \sim G, G \sim DP(\alpha, G_0).$$

θ_i 's are subject-level random effects, and there will be duplicates (because G is discrete)[3].

Chinese Restaurant Process

- The Chinese restaurant process (CRP) mixture model is one of the representations of the DP mixture model.
- The Chinese restaurant process (CRP), a discrete-time stochastic process, defines a distribution over partitions that embodies the assumed prior distribution over cluster structures

Imagine a Chinese restaurant with an infinite number of tables each with infinite capacity, and a sequence of n customers who enter the restaurant and sit down. The first customer enters the restaurant and sits at the first table. The i th subsequent customer sits at an occupied table, or at the next unoccupied table as follows:

$$P(z_i = j \mid z_{-i}, \alpha) \propto \begin{cases} \frac{m_j}{i-1+\alpha} & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha} & \text{if } j = k+1 \end{cases} \quad \text{where } m_j \text{ is the number of}$$

people sitting on table j or sits at a new table with a probability that is proportional to α . The tables are analogous to clusters, and customers to observations or data points.

Properties of Chinese Restaurant Process

The Chinese Restaurant as a Distribution

To generate an observation, you first sit down at a table. You sit down at a table proportional to the number of people sitting at the table.

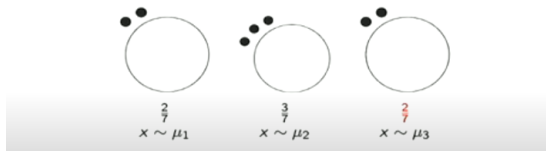


Figure: Source: <https://www.scribbr.com/ai-tools/machine-learning/>

- The CRP exhibits a clustering property due to a rich-gets-richer phenomenon, since it is proportional to m_j .
- Each customer's assignment to a new table is positive.

Online Chinese Restaurant Process

Online learning properties

- True label information is available after the predictions are made.
- The model can use the true label and cost label to refine the predictive model.

Several differences exist between the two processes:

- The CRP is an unsupervised learning method, and it is usually used in clustering applications. Conversely, the online CRP is an online algorithm, in which the label information y_i is available after the prediction of x_i is made.
- Second, the prior of the online CRP differs from that of the CRP.
- Third, the movement of the datum x_i between classes simultaneously alters the parameters of the classes that are indexed as z_i and y_i in the online CRP.

Online Chinese Restaurant Process

Relaxing Function

It is noted that table assignments for i th customer comprise two results, one is assigned by waiter, denoted as z_i , and the other one is the final sitting table, represented as y_i . For table j and customer i , f_j and e_j track the misassignment of the previous $i - 1$ customers as shown in Equation (2), where \mathbb{I} is an indicator function.

$$f_j = \sum_{a=1}^{i-1} \mathbb{I} \{y_a = j \wedge z_a \neq j\}$$

$$e_j = \sum_{a=1}^{i-1} \mathbb{I} \{z_a = j \wedge y_a \neq j\}$$

$$g(\gamma_1, \gamma_2, e_j, f_j) = (1 + \gamma_1)^{f_j} (1 - \gamma_2)^{e_j} - - - - - (1)$$

This work proposes a relaxing function, Equation (1), where γ_1 and γ_2 are regret rates, and the information about misassignment that is carried by f_j and e_j are viewed as prior knowledge in determining the distribution of table assignment probabilities.

$$P(z_i = j \mid z_{-i}, \mathbf{x}_i, y_i, \theta, G_0, \alpha) \\ \propto \begin{cases} g(\gamma_1, \gamma_2, e_j, f_j) \frac{m_j}{i-1+\alpha} H(\mathbf{x}_i, \theta_j) & \text{if } j \leq k \\ \frac{\alpha}{i-1+\alpha} \int H(\mathbf{x}_i, \theta_j) dG_0(\theta_j) & \text{if } j = k+1 \end{cases} \quad \text{---(2)}$$

Equation (2) presents the posterior distribution estimation of online CRP, in which the prior comprises the relaxing function and the prior of the CRP, and $H(\mathbf{x}_i, \theta_j)$ denotes the likelihood that datum \mathbf{x}_i is a member of class j . Notably, the online CRP reduces to the CRP when $f_j = 0$ and $e_j = 0$ for all j .

From here, we update class parameters, to check overestimation and under estimation.

Graphical Model of Online Chinese Restaurant Process

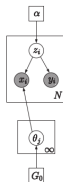


Figure: Graphical Model of Online Chinese Restaurant Process

$$z_i | \alpha \sim \text{Mult}(\text{Online CRP}(\alpha))$$

$$\theta_{z_i} | G_0 \sim G_0$$

$$x_i | \theta_{z_i} \sim F(\theta_{z_i}) \text{ --- (3)}$$

Equation (3) specifies that z_i is predicted class of datum x_i . For each datum x_i , the predicted label z_i can be sampled using the online CRP. The class parameter θ_{z_i} is drawn from the base distribution, G_0 , and each data point x_i is generated by a distribution F associated with parameter θ_{z_i} .

Experimental Results

In this presentation, two data sets (Reuters Corpus and Volume I (RCV1) and Wikipedia data set) are used to assess system performance and several methods are compared with the proposed algorithm.

Table: 1: Experimental Results on RCV1 Data Set

	Error Rate	Cluster F_1	Execution Time (sec)
Online CRP	0.1477	0.8517	54,958
Last Perceptron	0.1176	0.8905	259,512

Table: 2: Experimental Results on Wikipedia Data Set

	Error Rate	Cluster F_1	Execution Time (sec)
Online CRP	0.1776	0.7910	102,549




Conclusions

Table 2 shows the online CRP outperforms the variants of Perceptron on Wikipedia data sets.

Last Perceptron outperforms the proposed method on the RCV1 data set as shown in Table 1, but its execution time is much longer.

- This work combines Bayesian nonparametric learning with online to adapt model complexity and parameters to data.
- The experimental results indicate that the proposed method works well and efficiently on massive data sets.
- The future work is to extend the proposed approach to exploratory learning.

Reference

-  [1] Online learning: A comprehensive survey, Hoi, Steven CH and Sahoo, Doyen and Lu, Jing and Zhao, Peilin, Neurocomputing, 459, 249–289, 2021, Elsevier
-  [2] Liu, Chien-Liang, Tsung-Hsun Tsai, and Chia-Hoang Lee. "Online chinese restaurant process." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
-  [3] Weining Shen. " Bayesian Statistical Analysis." Stats 225, University of California Irvine, 2024. Lecture slides.

