Chinese Restaurant Process

David Mareček

November 11, 2021

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Text Segmentation

The task

大颇瓦法,为佛教密宗无上瑜伽部的甚深密法.在藏传佛教各大教派中最著名的仍是直贡大颇瓦法. 有许多信徒不远万里来到直贡接受颇瓦法或希望从直贡噶举派上师处得到此法. 享誉海内外的直贡颇瓦法，分为口传颇瓦法和伏藏颇瓦法两类. 口传颇瓦法,由初始佛持金刚传来,又分为语旨觉受传承和证语意义传承. 语旨觉受传承是四大语旨教授由金刚持依次传授因陀罗,龙树,玛当吉,而至底洛巴.其法有父续密集及圆满次第五道直教和颇瓦法. 证悟意义传承是底洛巴亲自从金刚持处聆听，依次传授给那若巴，玛尔巴，米拉日巴，塔布拉吉，帕木珠巴，觉巴大师，止至今天其耳传甚深密法，传承从未间断. 此颇瓦法称之为彩虹颇瓦法，因得此颇瓦法临终出现彩虹迷漫,故此得名.

伏藏颇瓦法，由阿弥陀佛依次传授给莲花生,赤颂的大巨尼玛. 尼玛将其臧在塔拉岗布山后的黑曼陀罗湖中. 后来，由大巨尼玛的转世牧羊人尼达桑杰发掘并依次传授贤士. 囊卡坚赞，多旦.更堆桑布，帕果.智美洛珠，法王桑杰坚赞，门才.坚赞教授，直贡羊日岗.堪布洛.平措朗杰. 此后，由直贡噶举做为甚深密法发扬光大. 此颇瓦法，称之为入草颇瓦法，因 其颇瓦法成功后，在头盖上能插入茅草，故此得名. 颇瓦，意为迁识，使亡者不经过中阴，将灵魂往生净土之法. 据说，接受此甚深密法颇瓦法，从人的头盖骨里出黄水或能插入茅草等现象. 接受直贡颇瓦法,使人立刻感觉头疼，头顶渗出水珠,晕倒，打通梵净穴，流鼻血等现象，故此灵验而闻名于世
(欲探详情，请参阅[捷径颇瓦法彩虹修行次第概要.自显自显虹光]，[捷径颇瓦法红哄教授.利众自焰]等口传颇瓦法典，以及[不修成佛之法甚深密法.入草颇瓦法]等法典).

The task

Let's switch the problem from Chinese to English so that we understand the results.

- We delete all spaces and try to restore them back in an unsupervised manner.
- Suppose that we do not have any rules, any dictionary, etc.

Theresa May launches a frantic two-week campaign today to save her Brexit deal and premiership by telling MPs to do their duty and support her or face going "back to square one". In a high-risk strategy to turn the tide of opposition in Westminster, the prime minister will then embark on a nationwide tour designed to sell her plan directly to the electorate.

theresamaylaunchesafrantictwo-weekcampaigntodaytosaveherbrexitdealand premiershipbytellingmpstodotheirdutyandsupportherorfacegoing"backtosquareone" inahigh-riskstrategytoturnthetideofoppositioninwestminster,theprimeministerwill thenembarkonanationwidetourdesignedtosellherplandirectlytotheelectorate.

Maximum Likelihood Estimation ?

Consider the following generative model:

1. Generate a word according to a finite probability distribution over words $p(w)$.
2. Write down the word chosen.
3. With probability 0.99, go to step 1. With probability 0.01, quit.

$$P_{MLE}(T) = \prod_{i=1}^{N} p(w_i) \cdot 0.99^{N-1} \cdot 0.01$$

What would be the the best maximum likelihood solution?
Note that the vocabulary is not fixed.

Where is the problem?

We need a better model reflecting the **re-use** of words in the text.

Requirements for the model

- We do not want too many words. We would like to see words repeating many times.
  - $\rightarrow$ Dirichlet priors with concentration parameters < 1?
  - But there are possibly infinitely many different words. (The length of a word is not limited.)
- We want to avoid trivial solutions.
  - The whole document is one word.
  - Each character is a word.
- We want to regulate somehow the average length of words.

Chinese Restaurant Process

- Each sample from such process produce a distribution over possibly infinite number of classes.

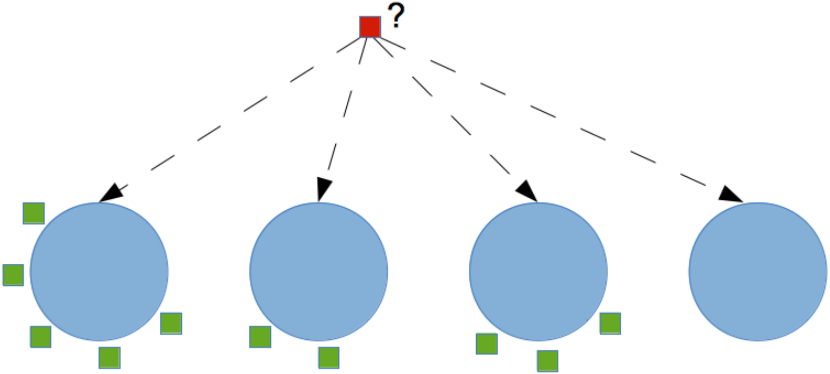
Chinese Restaurant Process

Chinese restaurant process (CRP)

A random process, where task is analogous of seating customers in Chinese Restaurant.

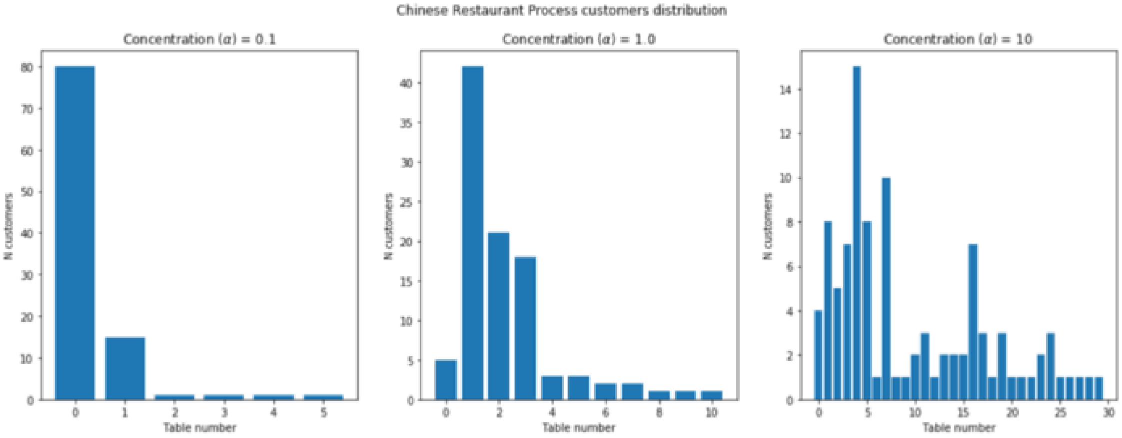
- Imagine a restaurant has infinite number of round tables.
- Each table accomodates an infinite number of customers.
- The first customer walks in, sits down at the first table and order a meal from the base probability distribution P_0.
- Suppose there are $n-1$ customers already sitting down at various tables and a new nth customer walks in.
- With probability $\alpha/(\alpha+n-1)$, he starts a new table and order a meal from the base probability distribution P_0.
- With probability $(n-1)/(\alpha+n-1)$, he randomly picks already-seated customer and sits down at his table with already ordered meal.
- Parameter α is a scalar concentration parameter of the process.
- CRP generates a probability distribution.

Chinese restaurant process (CRP)



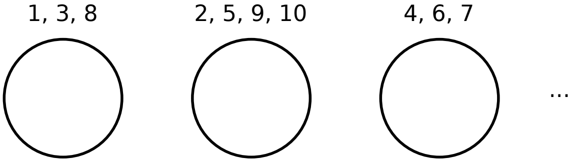
Demo:
http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html?parameters=2&dp=1#

Sampling from CRP



Chinese restaurant Process - Properties

- Denote z_i the table occupied by the i-th customer.
- A possible arrangement of 10 customers:



$$P(z_1, \ldots, z_{10}) = P(z_1)P(z_2|z_1) \cdots P(z_{10}|z_1, \ldots, z_9) =$$

$$= \frac{\alpha}{\alpha} \frac{\alpha}{1+\alpha} \frac{1}{2+\alpha} \frac{\alpha}{3+\alpha} \frac{1}{4+\alpha} \frac{1}{5+\alpha} \frac{2}{6+\alpha} \frac{2}{7+\alpha} \frac{2}{8+\alpha} \frac{3}{9+\alpha}$$

Exchangeability

$$P(z_1, \dots, z_{10}) = P(z_1)P(z_2|z_1) \cdots P(z_{10}|z_1, \dots, z_9) =$$

$$= \frac{\alpha}{\alpha} \frac{\alpha}{1+\alpha} \frac{1}{2+\alpha} \frac{\alpha}{3+\alpha} \frac{1}{4+\alpha} \frac{1}{5+\alpha} \frac{2}{6+\alpha} \frac{2}{7+\alpha} \frac{2}{8+\alpha} \frac{3}{9+\alpha}$$

- The probability of a seating is invariant under permutations.
- Permuting the customers permutes the numerators in the above computation, while the denominators remains the same.
- This property is known as **exchangeability**.

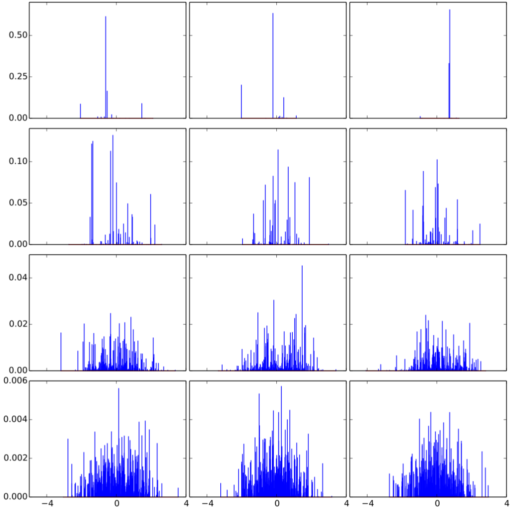
Dirichlet Processes (DP)

A family of stochastic processes $G \sim \mathsf{DP}(P_0, \alpha)$.

- P_0 is a base distribution
- α is a concentration parameter

Samples form DP generates probability distributions.

Figure: samples from DP with Gaussian base distribution and with parameter $\alpha = 1$, 10, 100, and 1000.



Dirichlet Processes (DP)

There are several equivalent views of the Dirichlet process:

- **Chinese Restaurant Process**
- **Stick Breaking Process**: We start with a unit-length stick and in each step we break off a portion of the remaining stick according to $\text{Beta}(1, \alpha)$.
- **Pólya urn scheme**: Imagine that we start with an urn filled with α black balls. Then we proceed as follows:
    1. Each time we need an observation, we draw a ball from the urn.
    2. If the ball is black, we generate a new (non-black) colour uniformly, label a new ball this colour, drop the new ball into the urn along with the ball we drew, and return the colour we generated.
    3. Otherwise, label a new ball with the colour of the ball we drew, drop the new ball into the urn along with the ball we drew, and return the colour we observed.

Text Segmentation

Chinese Restaurant Process for Text Segmentation

Assume the following generative model:

1. $i = 0$ number of words generated.
2. $i = i + 1$
3. With probability $\frac{\alpha}{\alpha+i-1}$, generate a word according to base probability P_0.
4. With probability $\frac{i-1}{\alpha+i-1}$, repeat one word that was already generated before.
5. Write down the word chosen.
6. With probability p_{cont}, go to step 2. With probability $1 - p_{cont}$, quit. ($p_{cont} = 0.99$)

Probability of the a given observed text is:

$$P(T) = \prod_{i=1}^{n} \frac{\alpha P_0(w_i) + \text{count}(w_i \text{ in previous words selections})}{\alpha + i - 1} \cdot p_{cont}^{n-1} \cdot (1 - p_{cont})$$

Base distribution

How to set the base distribution for words?

Assume the following micro-generative model for generating words:

1. Word $=$ empty
2. Pick a character from the uniform distribution (1/C)
3. Add the chosen character to the end of Word.
4. With probability p_c, go to 2, with probability $1 - p_c$, output the Word. ($p_c = 0.5$)

This base distribution prefers short words to long words, though it assigns positive probability to an infinitely many arbitrarily-long words.

Excercices

1. Imagine there are only three characters a, b, and c and $p_c = 0.5$. What is the base probability of words a, aa, aaa, bcb?
2. We observe a character sequence ab. What is more probable segmentation? $a\ b$ or ab?
3. We observe a character sequence aa. What is more probable segmentation? $a\ a$, or aa?
4. We observe a character sequence $abab$. What is the most probable segmentation?

Inference

- There is 2^{n-1} possible segmentations for n-characters long data.
- $n-1$ latent binary variables s_i: denoting whether there is of isn't a separator between two characters.
- Collapsed Gibbs sampling. Sample one variable conditioned by all the others.
- Exchangeability: if we reorder the words in the sequence, overall probability is the same.
- We can virtually move the changed words at the end of the sequence, compute the overal probablility of the two possibilities and then move the words virtually back.

Algorithm

$\forall i \in \{1 \dots n-1\}$: initialize s_i randomly;
compute initial counts of words $count[word]$ and total word-count t;
for *iter* $\leftarrow$ *1* **to** *numiter* **do**
    for $i \leftarrow$ *randomPermutation*$(1$ **to** $n-1)$ **do**
        $prev =$ previous word; $next =$ next word; $joined = prev + next$;
        if $s_i == 0$ **then** $count[joined]$−−; t−−;
        else $count[prev]$−−; $count[next]$−−; t−=2;
        $p[0] = \frac{\alpha P_0(joined)+count[joined]}{\alpha+t}$;
        $p[1] = \frac{\alpha P_0(prev)+count[prev]}{\alpha+t} \frac{\alpha P_0(next)+count[next]}{\alpha+t+1} p_{cont}$;
        $s_i =$ sample 0 or 1 with weights $p[0]$ and $p[1]$;
        if $s_i == 0$ **then** $count[joined]$++; t++;
        else $count[prev]$++; $count[next]$++; t+=2;
    end
end

Annealing

- When sampling, we can regulate the speed of convergence using so-called temperature $T \in (0, \infty)$
- Instead of sampling with weights $(w_1, w_2, w_3, \ldots, w_n)$ we can use $(w_1^{\frac{1}{T}}, w_2^{\frac{1}{T}}, w_3^{\frac{1}{T}}, \ldots, w_n^{\frac{1}{T}})$.
- The higher T, the larger searching area and the slower convergence.
- For $T \to \infty$, the sampling is uniform.
- For $T \to 0$, we always choose the best option.

Analogy with the temperature T in physics (oscillation of particles).

Pitman-Yor Process

Pitman-Yor process

Chinese Restaurant process

- approaches zero very fast
- after a while, almost no new items are generated

Pitman-Yor process

- generalization of the Chinese Restaurant process
- the generated distributions has a longer tail
- two hyperparameters: α and d

Pitman-Yor process

1. Imagine a restaurant has infinite number of round tables.
2. Each table accomodates an infinite number of customers.
3. The first customer walks in, sits down at the first table and order a meal from the base probability distribution P_0.
4. Suppose there are $n-1$ customers already sitting down at K different tables and a new customer walks in.
5. With probability $(\alpha + dK)/(\alpha + n - 1)$, he starts a new table and order a meal from P_0.
6. With probability $(n-1-dK)/(\alpha+n-1)$, he randomly picks already-seated customer and sits down at his table with already ordered meal. He picks a table with k customers with probability $(k-d)/(\alpha+n-1)$

Pitman-Yor process

$$P(w_i) = \frac{\text{count}(w_i) - d}{\alpha + i - 1} \quad \text{if} \quad \text{count}(w_i) > 0$$

$$P(w_i) = \frac{\alpha + dK}{\alpha + i - 1} P_0(w_i) \quad \text{if} \quad \text{count}(w_i) = 0$$

- $0 \leq d < 1; \qquad \alpha > 0$
- $d = 0$ ... Chinese Restaurant Process
- Is it exchangeable?
- In literature, the two hyperparameters are often called a and b ($b = \alpha$, $a = d$)