$p(\text{conclusions}|\text{Skipping } \{*2*\})$

Bayesian Language Modelling with Skipgrams

Louis Onrust

The topics for today

- Bayesian non-parametrics
- Language modelling
- Chinese restaurants
- Results
- Skipgrams
- The future

Motivation for BNP

- As the data grows, more patterns emerge:
  - this implies a growing, unbounded number of degrees of freedom;
  - risk of overfitting.

Non-parametric with parameters

- Non-parametric does not mean there are no parameters:
  - instead, the number of parameters is not fixed.
  - “Allows an infinite number of parameters.”

Clustering and the First Formulae

Dirichlet Process

- Centrepiece of BNP: Introduced by Ferguson
- Infinite-dimensional analog of the Dirichlet distribution.

Clustering

- Each data point x_i is assigned to one of K clusters
  - with probability w_k, for $k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} w_k = 1$ and a Dirichlet prior placed on the probabilities $\{w_k\}$.
- Treat clustering problem as inferring partitions
  - by placing probability distributions on partitions;
  - Chinese Restaurant Process

Clusteling with a Chinese Lestaulant Plocess

Partitions and Clusters

- A partition of N points is denoted as $\pi_{[N]}$
  - $\pi_{[10]} = \{\{3,5\},\{1,2,9,10\},\{4,6,7\},\{8\}\}$
  - $\pi_{[N]}$ is a set of subsets, where subsets are clusters

Chinese Restaurant Process

- CRP is a probability distribution on partitions.
- Restaurant metaphor: points are customers, clusters are tables.
- Sequential process: each point at a time is added to an existing set of clusters.
  - The first customer is seated alone;
  - Each subsequent customer is either:
    - seated at one of the already occupied tables, or
    - starting a new table.

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

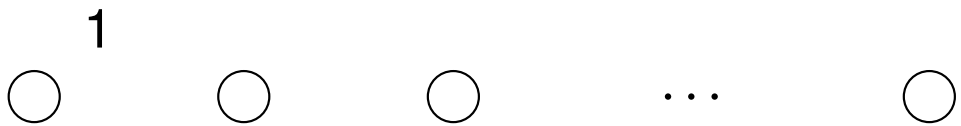
Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$

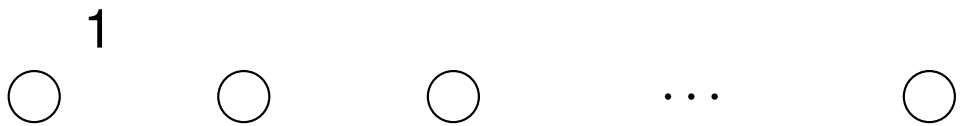


$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



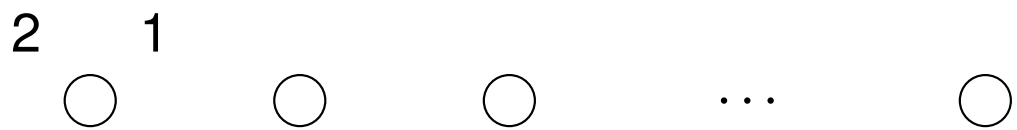
$P = \frac{\alpha}{\alpha}$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



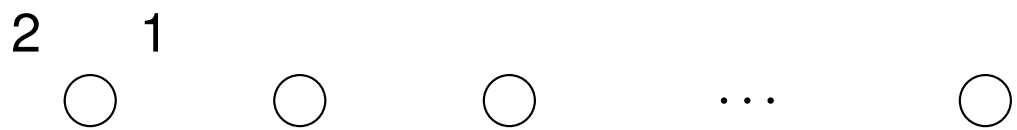
$P = \frac{\alpha}{\alpha}$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



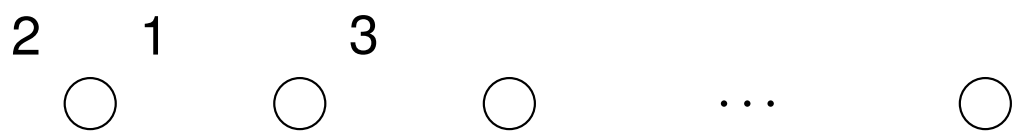
$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1}$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



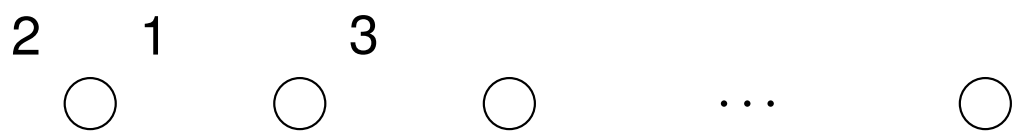
$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1}$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



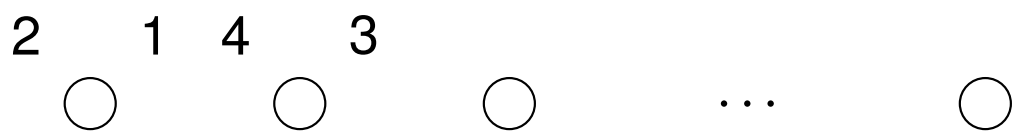
$$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1} \frac{\alpha}{\alpha+2}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



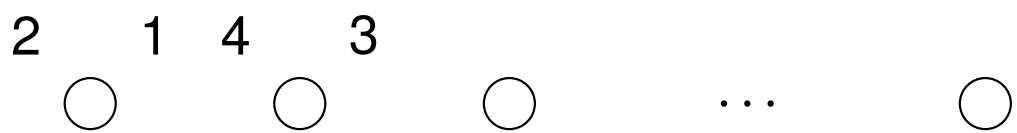
$$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1} \frac{\alpha}{\alpha+2}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



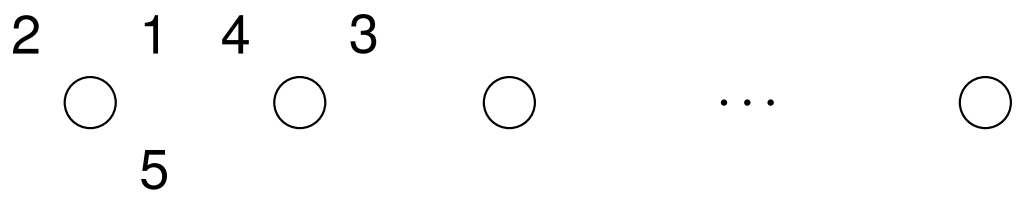
$$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1} \frac{\alpha}{\alpha+2} \frac{1}{\alpha+3}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



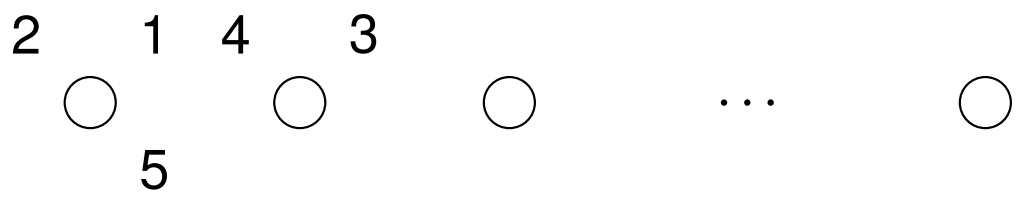
$$P = \frac{\alpha}{\alpha}\frac{1}{\alpha+1}\frac{\alpha}{\alpha+2}\frac{1}{\alpha+3}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



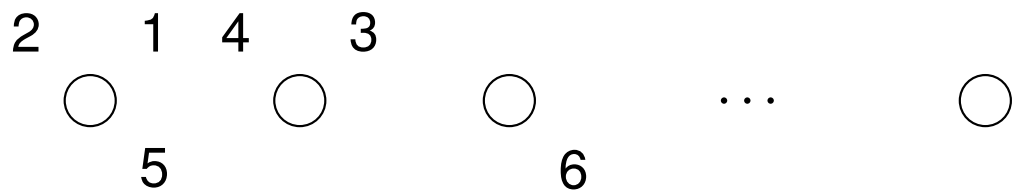
$$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1} \frac{\alpha}{\alpha+2} \frac{1}{\alpha+3} \frac{2}{\alpha+4}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



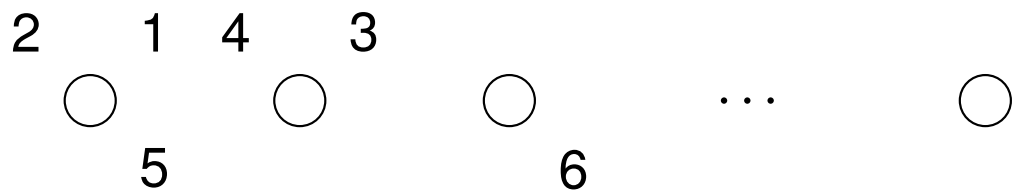
$$P = \frac{\alpha}{\alpha} \frac{1}{\alpha+1} \frac{\alpha}{\alpha+2} \frac{1}{\alpha+3} \frac{2}{\alpha+4}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



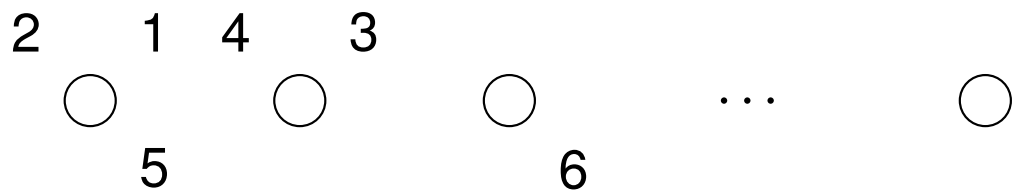
$$P = \frac{\alpha}{\alpha}\frac{1}{\alpha+1}\frac{\alpha}{\alpha+2}\frac{1}{\alpha+3}\frac{2}{\alpha+4}\frac{\alpha}{\alpha+5}$$

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha}{\alpha+n} & \text{otherwise.} \end{cases}$$

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{CRP}(\alpha, N)$$

Example for $\{\{1,2,5\},\{3,4\},\{6\}\}$



$$P = \frac{\alpha}{\alpha}\frac{1}{\alpha+1}\frac{\alpha}{\alpha+2}\frac{1}{\alpha+3}\frac{2}{\alpha+4}\frac{\alpha}{\alpha+5} \rightarrow P(\pi_{[N]}) = \frac{\alpha^K}{\alpha^{(N)}}\prod_{c\in\pi_{[N]}}(|c|-1)!$$

Two-Parameter Poisson-Dirichlet Processes

Why Pitman-Yor Processes?

- DP generates an infinite number of atoms, with a relatively slow rate.
- Many real-world phenomena have a power-law growth.
- The DP cannot generate such power-laws, hence we use PYP.

PYP as generalisation of the CRP

- With DP the rate for selecting a new table is $\frac{\alpha}{\alpha+N}$, and choosing an occupied table goes to $\frac{N}{\alpha+N}$:
  - $\sum_{n=1}^{N} \frac{\alpha}{\alpha+n} \asymp \log(N)$
- PYP allows α to grow, with the rate of a discount parameter σ
- PYP reduces to DP with $\sigma = 0$.

PYCRP

$$P(\text{customer } n+1 \text{ joins table } c|\pi_{[n]}) = \begin{cases} \frac{|c|-\sigma}{\alpha+n} & \text{if } c \in \pi_{[n]}, \\ \frac{\alpha+\sigma K_n}{\alpha+n} & \text{otherwise.} \end{cases}$$

- The probability of joining an existing table is reduced by an amount proportional to σ relative to the CRP.
- Reductions are added to the probability of starting a new table.

The seating pattern after N customers defines a set of clusters:

$$\pi_{[N]} \sim \text{PYP}(\alpha, \sigma, N)$$

$$P(\pi_{[N]}) = \frac{\alpha(\alpha+\sigma)\cdots(\alpha+\sigma(K_N-1))}{\alpha^{(N)}} \prod_{c\in\pi_{[N]}} (1-\sigma)(2-\sigma)\cdots(|c|-1-\sigma).$$

Finally, Language Models

Unigram LM with PYP

W is a fixed vocabulary of V words. For each $w \in W$ let $G(w)$ be the probability of w, and $G = [G(w)]_{w \in W}$ the vector of word probabilities.

$$G \sim \text{PYP}(\alpha, \sigma, G_0).$$

Inference for our Unigram LM

- Training data $\mathscr{D}$ consists of occurrence counts c_w.
- We are interested in the posterior distribution $P(G, \Theta | \mathscr{D}) = P(G, \Theta, \mathscr{D}) / P(\mathscr{D})$.
  - The CRP marginalises out G, replacing it with the seating arrangement S
  - The new posterior is then: $P(S, \Theta | \mathscr{D}) = P(S, \Theta, \mathscr{D}) / P(\mathscr{D})$.
  - Predictive probability: $p(w | \mathscr{D}) = \int P(w | S, \Theta) P(S, \Theta | \mathscr{D}) \mathrm{d}(S, \sigma)$.
    - $P(w | S, \Theta) = \frac{c_w - \sigma t_w}{\alpha + c_\cdot} + \frac{\alpha + \sigma t_\cdot}{\alpha + c_\cdot}$.

Hierarchical PYP

HPYLM

Given a context $\mathbf{u}$, let $G_{\mathbf{u}}(w)$ be the probability of the current word taking on value w. The HPYP has a prior for $G_{\mathbf{u}} \sim \text{PYP}(\alpha_{|\mathbf{u}|}, \sigma_{|\mathbf{u}|}, G_{\pi(\mathbf{u})})$, with $\pi(\mathbf{u})$ being the suffix of $\mathbf{u}$ of all but the first word. $G_{\pi(\mathbf{u})}$ is also unknown, so we place a recursive prior over it, with parameters $\alpha_{|\pi(\mathbf{u})|}, \sigma_{|\pi(\mathbf{u})|}$ and mean vector $G_{\pi(\pi(\mathbf{u}))}$; $G_{\varnothing}$ being the empty context, which is the same as for the unigram LM.

Interpolated Kneser-Ney can be considered to be an approximate inference of HPYLM.

Chinese Restaurant Franchise

- A Chinese Restaurant Franchise consists of Chinese Restaurants
- There is a global menu with an unbounded number of dishes

Adding Bayesiawesomeness to Colibri

Existing Bayesian language models. . .

- are merely an algorithmic showcase without real LM aspirations;
- cannot handle really big datasets.

Colibri

- C++ and Python library for basic linguistic constructions
- Generates *n*-grams, skipgrams, and flexgrams
- http://proycon.github.io/colibri-core/

Bayesian Colibri

- We extend the C++ library for BNP with PYP (https://github.com/redpony/cpyp)
- BaCo is available at https://github.com/naiaden/cococpyp

From *n*-gram to *Y*-gram

We want to beat a MKN approach with skipgrams: Pickhardt et al., *A Generalized Language Model as the Combination of Skipped n-grams and Modified Kneser Ney Smoothing*, ACL'14.

Comparison

- Compare with existing systems
- Compare the performance of backoff methods
- Compare the within and cross domain performance
- Compare the ngram and skipgram performance

Data

- JRC-Acquis
- Google 1 Billion Word Corpus
- EMEA (European Medicines Agency)

Comparison with existing systems

n-grams

	baco			srilm			cpyp		
	jrc	1bw	emea	jrc	1bw	emea	jrc	1bw	emea
jrc	13	1195	961	22	1664	1310	13	1536	1098
1bws	768	158	946	1460	211	1516	785	155	987
emea	600	1143	4	1115	1745	10	794	1597	4

skipgrams

	baco		
	jrc	1bw	emea
jrc	13	1162	939
1bws	751	162	921
emea	581	1155	4

Relative change in perplexity

	jrc	1bw	emea
jrc	2.03	2.80	2.30
1bws	2.23	2.38	2.63
emea	3.20	1.09	0.67

Influence of backoff method

n-grams

		jrc	1bw	emea
jrc	ngram	13	1510	1081
	bobaco	14	1477	1122
	glm	69	1195	961
1bws	ngram	768	158	946
	bobaco	815	185	1025
	glm	801	264	1039
emea	ngram	769	1552	4
	bobaco	779	1385	4
	glm	600	1143	32

skipgrams

		jrc	1bw	emea
jrc	ngram	13	1843	1295
	bobaco	13	1542	1149
	glm	65	1195	939
1bws	ngram	879	163	1105
	bobaco	751	162	921
	glm	768	252	988
emea	ngram	969	2089	4
	bobaco	838	1655	4
	glm	581	1155	32

ngram only backoff to shorter *n*-grams

bobaco backoff to all patterns $\leq n$, until match

glm backoff to all patterns $\leq n$

Future work

Monday

- Further investigate clues cross domain language modelling
- Further investigate comparison to state-of-the-art skipgram language modelling
- Influence of size training data
- Test hypotheses on more corpora
- Get up to date on (Bayesian) statistical significance tests

Someday

- Create language models of multiple domains: DHPYPLM
- Skipgrams versus flexgrams

Dirichlet Distribution

Informal introduction

- Dirichlet distribution can model the randomness of pmfs
  - We have a dictionary of k possible words
  - Each document can be represented by a pmf of length k by normalising the empirical frequency of its words
  - A group of documents produces a collections of pmfs, and we use the Dirichlet distribution to capture the variability

Formal introduction

Let $Q = [Q_1, Q_2, \ldots, Q_k]$ be a random pmf with k components, that is $Q_i \geq 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} Q_i = 1$

$$f(q; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1}.$$

Back to the Dirichlet Process.

This is a dice