ノンパラメトリックベイズの基礎

「続わかりやすい パターン認識」
第11章

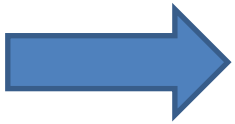
この本の内容をベースに私の理解を説明します

新納浩幸

ノンパラメトリック手法

統計学の手法、
　例）回帰の問題

観測データ　$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$



$y = f(\mathbf{x})$　　f の推定

関数 f に基本の形（2次式など）を仮定し、基本の形に含まれるパラメータを推定・・・パラメトリック手法

基本の形を仮定しない・・・ノンパラメトリック手法

ノンパラメトリックベイズ

ノンパラメトリック手法とは直接の関係はない

LDA は混合分布の（ベイズの）モデル

混合数（トピックの数）をパラメータ化したものが
ノンパラメトリックベイズモデル

深い意味ではつながっている（？）けど、
クラスタリングのクラスタ数を推定できる
モデルと単純に考えてもOK

ホップの壺モデル

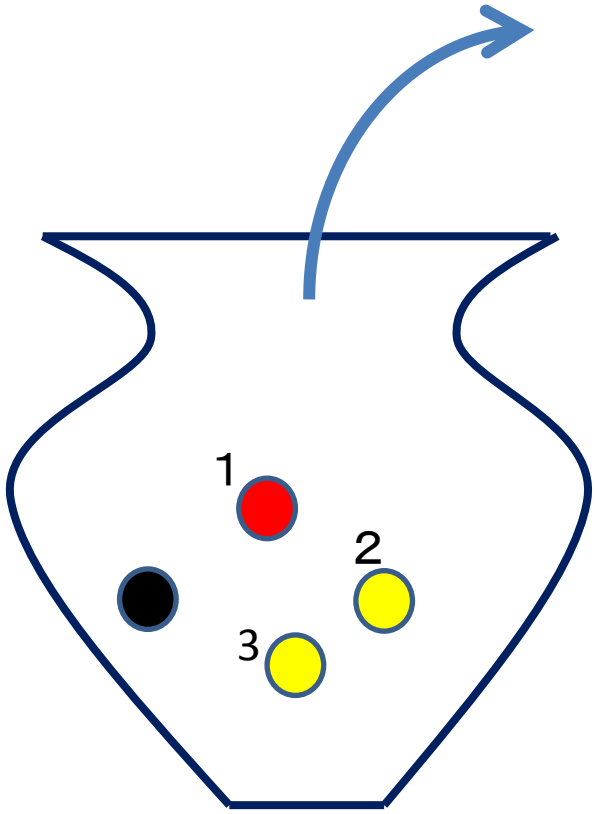
ノンパラメトリックベイズモデルの基本モデル

壺に様々な色の玉が入っている、黒玉の重さは α、
それ以外の玉の重さは 1

(1) 最初、壺に1個の黒玉が入っている
(2) 壺から重さに比例した確率で1つ玉を取り出す
(3) 黒玉だったら、壺にない色の玉を1つ選んで、
黒玉と一緒に戻す、黒玉以外なら、同じ色の玉を
取り出した玉と一緒に戻す・・・(2) に戻る

壺の中の玉の分布はどうなっていくか？

イメージ図



1つ取り出して、取り出した玉と、
もう1つの玉を壺に戻す

追加されていく玉の色に注目

左の状態で次に追加される玉の色が
赤である確率は、
赤玉を取り出す確率と同じ

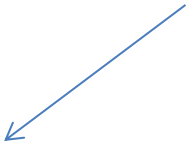
赤玉の個数

$$\frac{1}{3+\alpha}$$

全体の重さ

計算例

赤、黄、黄、赤、赤、緑、っと追加される確率



P(赤、黄、黄、赤、赤、緑)

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

$$= \frac{\alpha^3 \cdot 2!}{AF(\alpha,6)}$$

$$AF(\alpha,n) = \alpha \cdot (\alpha+1) \cdot (\alpha+2) \cdots (\alpha+n-1)$$

交換可能性

先の確率は玉の取り出し順に依存しない

色の違いだけに注目していることも大事

P(赤、黄、黄、赤、赤、緑)

= P(緑、青、白、青、青、白)

この確率は何回玉を入れるかの n と色の種類数 c だけで決まる

Ewens の抽出公式

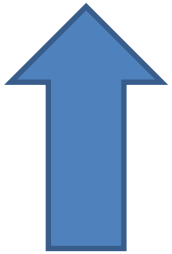
$$P_E(n_1, n_2, \cdots, n_c) = \frac{\alpha^c \cdot \prod_{i=1}^{c} (n_i - 1)!}{AF(\alpha, n)}$$

n_i ：色 i の玉の個数

順番に関係ないので、各色
の玉がいくつ取り出されたか
だけが効く

C について

先の公式は C に依存しているように見えるが、事前に C の値を決めておく必要はない



C の値は取り出しが終了した時点で決まる

α について

α の値が大きいほど、色の種類数が増える。α の値が小さいほど、特定の色の玉が増える。

CRP (Chinese Restaurant Process)

ホップの壺モデルを基礎とした類似のモデル

中華料理店に1人づつ客がくる、どのテーブルに着くか？
最終的なテーブルの様子は？

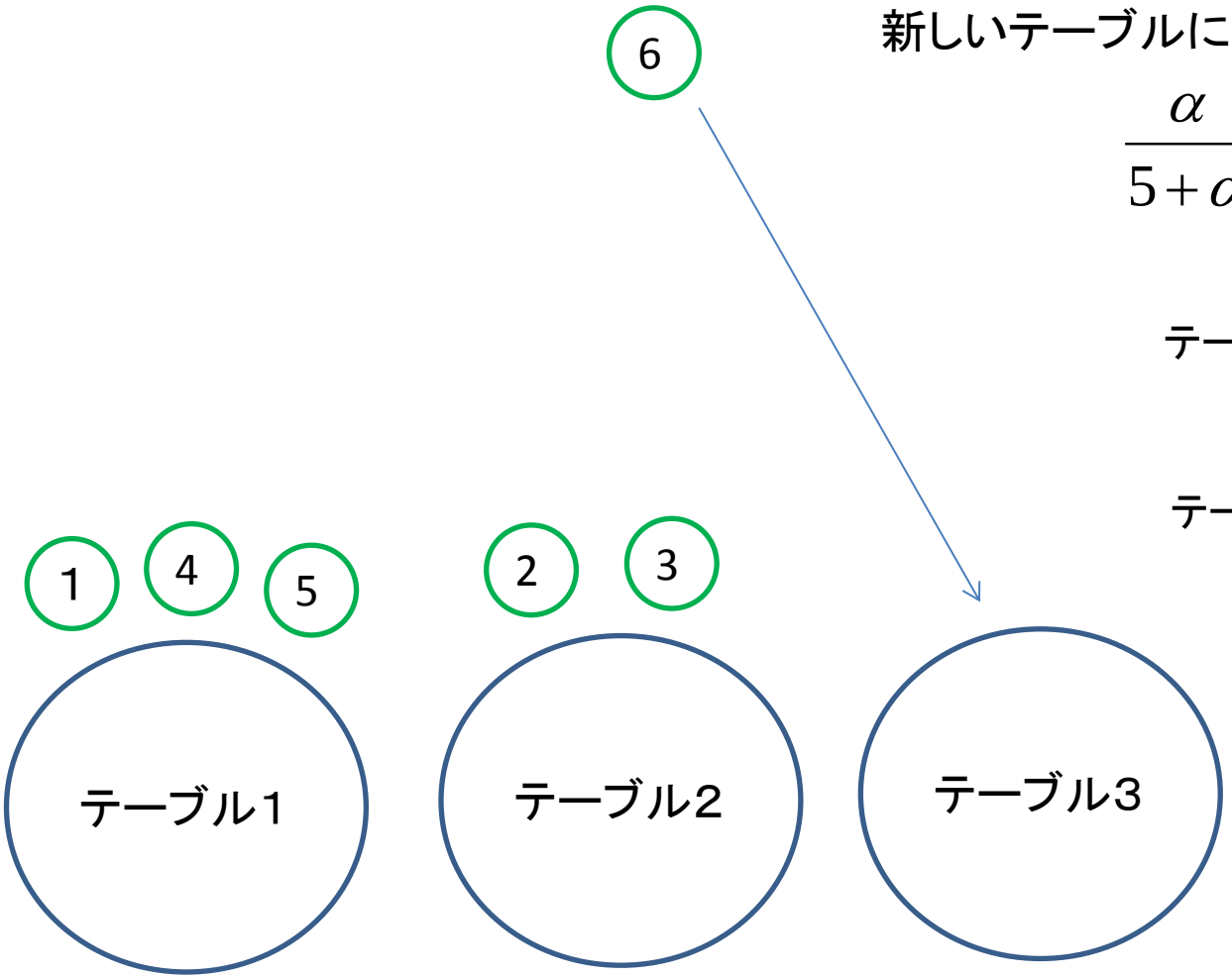
(1) 最初の客は任意のテーブルに着く

(2) n 番目の客が テーブル i に着く確率 → $\frac{n_i}{n-1+\alpha}$

n_i : テーブル i にいる人数

新たなテーブルに着く確率 → $\frac{\alpha}{n-1+\alpha}$

イメージ図



新しいテーブルに着く確率は

$$\frac{\alpha}{5+\alpha}$$

テーブル1に着く確率は $\frac{3}{5+\alpha}$

テーブル2に着く確率は $\frac{2}{5+\alpha}$

計算例

客がどのテーブルに着くかに注目

P(1, 2, 2, 1, 1, 3)

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

$$= \frac{\alpha^3 \cdot 2!}{AF(\alpha,6)}$$

= P(赤、黄、黄、赤、赤、緑)

ホップの壺モデルと同じ

交換可能性

CRP でもホップの壺モデルと同様、
交換可能性が成立

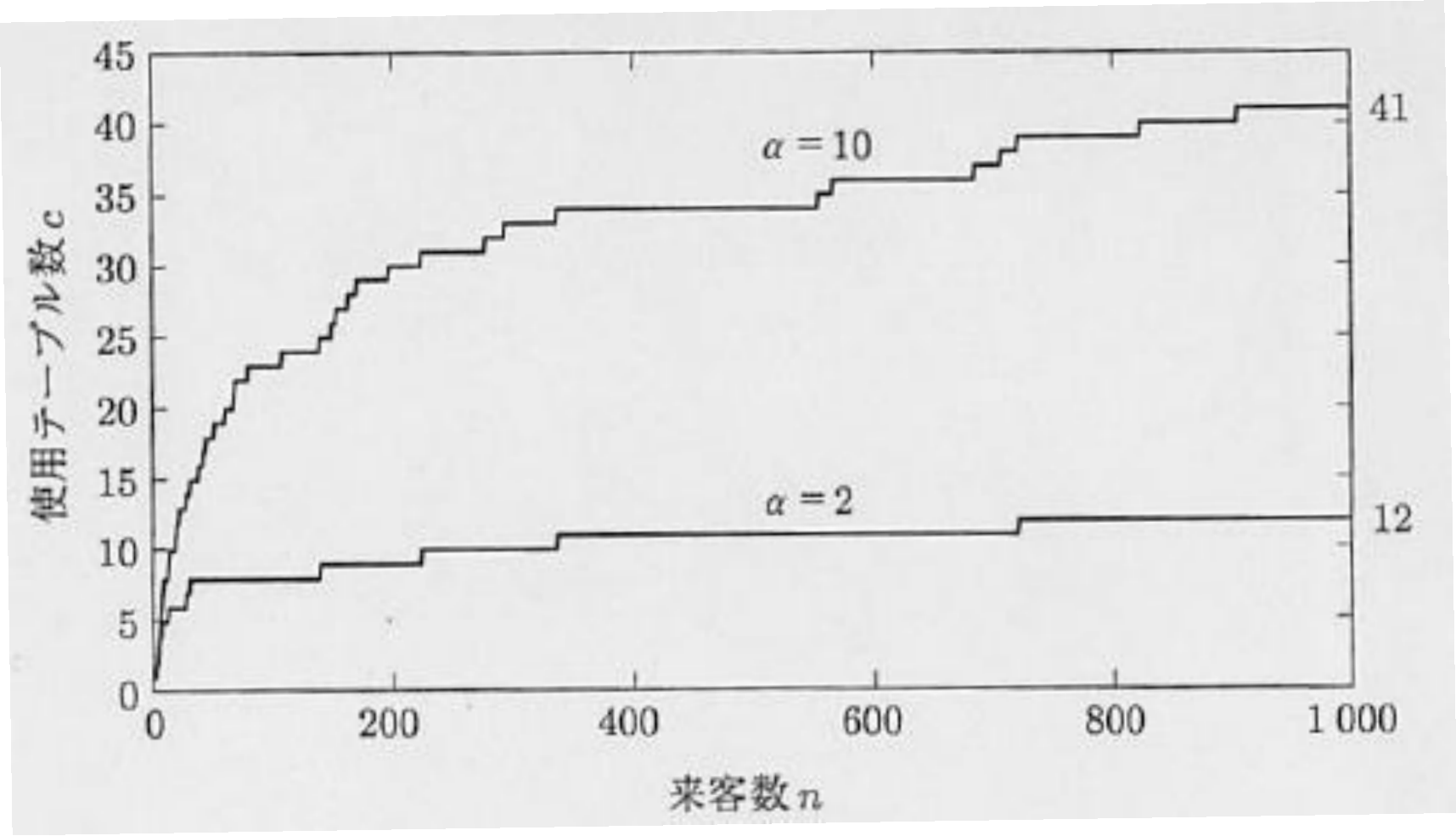
着席の確率は客の来る順序に依存しない

$$P(1, 2, 2, 1, 1, 3) = P(1, 1, 2, 3, 2, 1)$$

テーブルの分割とその確率

(a) テーブル数 c	(b) テーブルの客数 $(n_1, \ldots, n_c)$	(c) 客の分割方法	(d) 生起確率 $P_E(n_1, \ldots, n_c)$
1	(4)	(① ② ③ ④)	1/10
2	(2, 2)	(① ②) (③ ④)	1/30
		(① ③) (② ④)	1/30
		(① ④) (② ③)	1/30
	(3, 1)	(① ② ③) (④)	1/15
		(① ② ④) (③)	1/15
		(① ③ ④) (②)	1/15
		(② ③ ④) (①)	1/15
3	(2, 1, 1)	(① ②) (③) (④)	1/15
		(① ③) (②) (④)	1/15
		(① ④) (②) (③)	1/15
		(② ③) (①) (④)	1/15
		(② ④) (①) (③)	1/15
		(③ ④) (①) (②)	1/15
4	(1, 1, 1, 1)	(①) (②) (③) (④)	2/15

CRP の実験例


45
40
35
30
25
20
15
10
5
0
使用テーブル数 c
$\alpha = 10$
41
$\alpha = 2$
12
0
200
400
600
800
1 000
来客数 n

テーブル数の期待値

$$E(c) = O(\alpha \log n)$$

べき乗則に従っていない、不自然

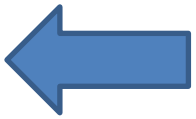
Pitman-Yor Process　（省略）

CRP にパラメータ β を追加、
テーブル数がべき乗則に従う

D. Mochihashi, T. Yamada, N. Ueda,
Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling.
ACL 2009, 100-108 (2009).

ディリクレ分布から分割ルールへ

ディリクレ分布を無限次元に拡張することで、
先の分割ルールが導出できる



ノンパラメトリックベイズの名前の由来

本発表はこれを示して終わり・・・

皆さんの勉強のために、この部分の基礎である
ベイズの考え方、ベータ分布、ディリクレ分布
の基礎を、最初に解説します

2項分布

当たる確率が p のクジがある。
10 本引いて X 本あたる確率は？

X は 0 から 10 の値をとりえる確率変数

$$P(X = x) = {}_{10}C_x \, p^x (1-p)^{10-x}$$

2項分布

最尤推定

当たる確率が p のクジがある。
10 本引いて 2 本当たった。p は？

$$P(X=2)={}_{10}C_2\,p^2(1-p)^8$$

尤度、
p の関数とみて最大化

対数は単調増加関数なので最大化は対数をとっても同じ

$$\log_{10}C_2\,p^2(1-p)^8=\log_{10}C_2+2\log p+8\log(1-p)$$

定数

ここを最大にする p は微分して求まる

$$p=\frac{1}{5}$$

最尤推定値はサンプルに依存

10 本引いて 2 本当たった。p は？ ➡ $p = \frac{1}{5}$

次の日、10 本引いて 1 本当たった。p は？ ➡ $p = \frac{1}{10}$

次の日、9 本引いて 2 本当たった。p は？ ➡ $p = \frac{2}{9}$

p は確率変数とみなせる！（ベイズの考え方）

取り得る値は 0 から 1 の実数値

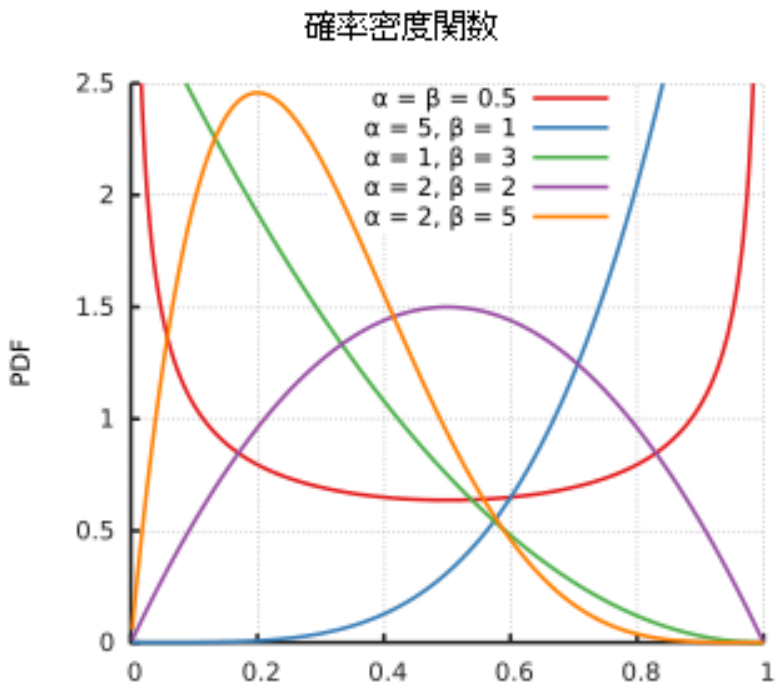
その分布は P(X < x)、連続型なので密度関数で分布を表現

p

ベータ分布

2項分布の p に対する分布として一般に利用

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$



ベータ関数

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$

多項分布

1等、2等、ハズレの結果をもつクジがある。
1等の確率が p1、2等の確率が p2、ハズレの確率が p3
10本引いて、1等が X1 本、2等が X2本、ハズレが X3本となる確率は？

X1,X2,X3 は 0 から 10 の値をとりえる確率変数
X1 + X2 + X3 = 10、p1+p2+p3 = 1 が成立

$$P(X1 = x_1, X2 = x_2, X3 = x_3) = \frac{10!}{x_1! x_2! x_3!} {p_1}^{x_1} {p_2}^{x_2} {p_3}^{x_3}$$

多項分布

最尤推定

先のクジ、10 本引いて1等1本、2等2本、ハズレ7本
p1, p2, p3 は？

$$P(X1 = x_1, X2 = x_2, X3 = x_3) = \frac{10!}{1!2!7!} p_1^{\,1} p_2^{\,2} p_3^{\,7}$$ 尤度

$p_1 + p_2 + p_3 = 1$ の条件下での最大化

$$p_1 = \frac{1}{10}, p_2 = \frac{2}{10}, p_3 = \frac{7}{10}$$

多項分布の各確率を確率変数と見なす

$X = (p_1, p_2, p_3)$ ← $p_1 + p_2 + p_3 = 1$

の条件があるので2次元の確率変数

それぞれの次元の値は 0 から 1 の実数値をとる

この確率変数の分布として一般に使われるのが

ディリクレ分布

ディリクレ分布

$$\mathrm{x} = (x_1, x_2, \cdots, x_K) \qquad \sum_{\mathrm{i}=1}^{\mathrm{K}} x_i = 1$$

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$$

$$f(\mathrm{x}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

$$B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$$

ガンマ関数

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

K=2 のディリクレ分布がベータ分布

ディリクレ分布の重要公式

$$f(\mathrm{x}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

密度関数なので

$$\int f(\mathrm{x}) d\mathrm{x} = 1$$

$$\int \prod_{i=1}^{K} x_i^{\alpha_i - 1} d\mathrm{x} = B(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$$

示すのは
結構大変

CRP

ここが天下り的

(1) 最初の客は任意のテーブルに着く

(2) n 番目の客が テーブル i に着く確率 → $\frac{n_i}{n-1+\alpha}$

n_i ：テーブル i にいる人数

新たなテーブルに着く確率 → $\frac{\alpha}{n-1+\alpha}$

π_i ：客がテーブル i に着く確率　　$\sum_{i=1}^{c} \pi_i = 1$

こいつを確率変数だと考える、
その分布をディリクレ分布と考える

設定

π_i ：客がテーブル i に着く確率　　$\sum_{i=1}^{c} \pi_i = 1$

s_k ：k 番目の客があるテーブルに着く事象の略記

ω_i ：テーブル i　　テーブルの数は c 個

$s_k = \omega_i$ ：k 番目の客がテーブル i に着く事象

ディリクレ分布の α_i は全て等しいと仮定 ⇒ $\alpha_i = \frac{\alpha}{c}$

⇒ $f(\pi) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/c)^c} \prod_{i=1}^{c} \pi_i^{\alpha/c-1}$

積分消去

πが与えられている仮定なので、ここは求まる

$$P(s_1 s_2 \cdots s_{n-1}) = \int P(s_1 s_2 \cdots s_{n-1} \mid \pi) f(\pi) d\pi$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha/c)^c} \cdot \int \prod_{i=1}^{c} \pi_i^{\,n_i + \alpha/c - 1} d\pi$$

重要公式

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha/c)^c} \cdot \frac{\prod_{i=1}^{c} \Gamma(n_i + \alpha/c)}{\Gamma(n - 1 + \alpha)}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(n-1+\alpha)} \cdot \prod_{i=1}^{c} \frac{\Gamma(n_i + \alpha/c)}{\Gamma(\alpha/c)}$$

$$P(s_n = \omega_i \mid s_1 s_2 \cdots s_{n-1}) = \frac{P(s_1 s_2 \cdots s_{n-1}, s_n = \omega_i)}{P(s_1 s_2 \cdots s_{n-1})}$$

$$= \frac{P(s_1 s_2 \cdots s_{n-1} s_n)}{P(s_1 s_2 \cdots s_{n-1})}$$

分母は先に求まっている、分子は分母の式の n を n+1 に変更すればよい

$$= \frac{\Gamma(n-1+\alpha)}{\Gamma(n+\alpha)} \cdot \frac{\Gamma(n_i+1+\alpha/c)}{\Gamma(n_i+\alpha/c)}$$

$$= \frac{n_i + \alpha/c}{n-1+\alpha}$$

とりあえず、結論の式

$$P(s_n = \omega_i \mid s_1 s_2 \cdots s_{n-1}) = \frac{n_i + \alpha / c}{n - 1 + \alpha}$$

ここからは簡単、まず

$$C \to \infty$$

$$P(s_n = \omega_i \mid s_1 s_2 \cdots s_{n-1}) = \frac{n_i}{n - 1 + \alpha}$$

設定

$$\Omega_0 = \{\omega_i \mid n_i = 0\}$$

着席者 0 人のテーブルの集合

$$\Omega_1 = \{\omega_i \mid n_i \neq 0\}$$

着席者がいるテーブルの集合

$$\Omega = \Omega_0 \cup \Omega_1 \Rightarrow |\Omega| = |\Omega_0| + |\Omega_1| = c$$

$$P(s_n \in \Omega_0 \mid s_1 s_2 \cdots s_{n-1}) = \sum_{\omega_i \in \Omega_0} P(s_n = \omega_i \mid s_1 s_2 \cdots s_{n-1})$$

$$= \sum_{\omega_i \in \Omega_0} \frac{\alpha / c}{n-1+\alpha}$$

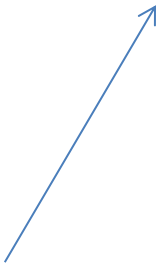
$$= \frac{c - |\Omega_0|}{c} \cdot \frac{\alpha}{n-1+\alpha}$$

$$C \rightarrow \infty$$

$$= \frac{\alpha}{n-1+\alpha}$$

まとめると

$$P(s_n = \omega_i \mid s_1 s_2 \cdots s_{n-1}) = \begin{cases} \dfrac{n_i}{n-1+\alpha} & \omega_i \in \Omega_1 \\ \dfrac{\alpha}{n-1+\alpha} & \omega_i \in \Omega_0 \end{cases}$$



ホップの壺やCRP の分割ルールと同じ

平成24年度茨城大学工学部情報工学科卒業研究論文

ディリクレ混合過程を利用した単語用例の語義別クラスタリング

平成24年度2月8日

工学部情報工学科

執筆者：倉持辰洋(07T4024L)

指導教員：新納浩幸 准教授

國井君、菊池君と同期の
倉持君の卒論

CRP を利用した単語用例の
クラスタリングです

結果、めちゃくちゃでした・・・・、
私の感覚では使えません