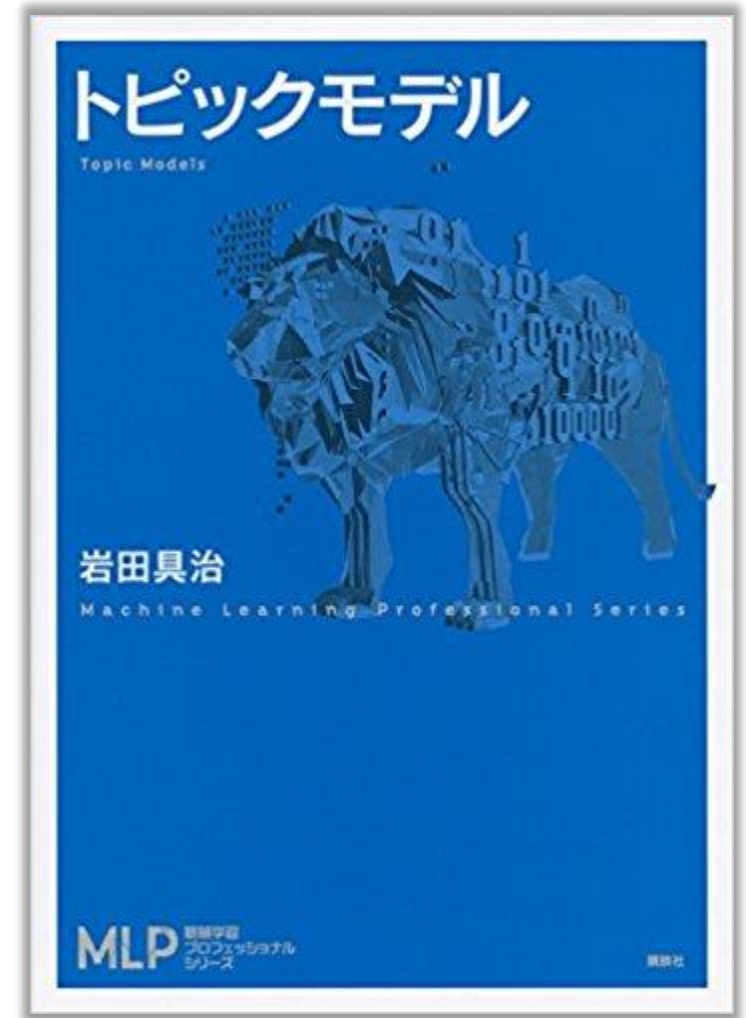


統計的モデリング基礎⑪ ～ベイズモデリング～

鹿島久嗣
(情報学科 計算機科学コース)

目次： ベイズモデリング

- ベイズ統計の基礎
 - ベイズの公式、事前分布、事後分布
 - 事後分布による意思決定
 - ナイーブベイズ予測
- ベイズモデリング
 - 離散分布のベイズ推定
 - 階層ベイズモデリング
 - 経験ベイズ法



ベイズ統計：

ベイズの公式によって事前知識と証拠を組み合わせて推論

- 事前知識と観測された証拠から確率を更新する
 - 事前知識：事前分布によって表される、自分が予めもっている、ある事象がどの程度起こりそうかという信念
- 新たな証拠が得られたとき、信念の更新はベイズの公式に基づいて行われる

はじめの信念／事前知識
(事前分布)

明日は20%の
確率で雨かな



猫が顔を洗った
(証拠)



信念の更新

ベイズの公式

証拠を確認後の信念
(事後分布)

明日は90%の
確率で雨だな



ベイズ統計で中心的役割を果たすベイズの公式： 条件付確率の条件部と帰結部を入れ替える



- 条件付確率 $P(\text{rain} \mid \text{cat})$: 猫が顔を洗うのを目撃したときに、次の日に雨が降る確率

- $$P(\text{rain} \mid \text{cat}) = \frac{P(\text{rain, cat})}{P(\text{cat})}$$

- ベイズの公式：条件付確率の条件部分と帰結部分を入れ替える公式

雨の日の前日に猫が顔を洗っている確率

猫がどうか関係なく、そもそも雨が降る確率
(事前分布)

$$P(\text{rain} \mid \text{cat}) = \frac{P(\text{cat} \mid \text{rain})P(\text{rain})}{P(\text{cat})}$$

雨がどうか関係なく、そもそも猫が顔を洗う確率

例：

ベイズの公式に基づく事後確率計算

- 事前確率：これまでの経験から雨が降る確率は20%
 - 雨が降らない確率は80%

- 雨がふる日の前日に猫が顔を洗っている確率は80%とすると：

$$P(\text{rain} \mid \text{cat}) = \frac{P(\text{cat} \mid \text{rain})P(\text{rain})}{P(\text{cat})} = \frac{0.8 \times 0.2}{P(\text{cat})} = \frac{0.16}{P(\text{cat})} \text{ がわかる}$$

- 一方、雨が降らない日の前日に猫が顔を洗っている確率は50%とすると：

$$P(\neg \text{rain} \mid \text{cat}) = \frac{P(\text{cat} \mid \neg \text{rain})P(\neg \text{rain})}{P(\text{cat})} = \frac{0.5 \times 0.8}{P(\text{cat})} = \frac{0.40}{P(\text{cat})}$$

- 両者より $P(\text{rain} \mid \text{cat}) = \frac{0.16}{0.16 + 0.40} = 0.29$ (29%) となる

9ポイント増えたよ！

ベイズ決定： 事後分布をもちいた意思決定

- 予測×効用（結果のうれしさ度合い）によって意思決定

効用（結果のうれしさ度合い）		ビールの仕入れ量	
		多め	少な目
実際の天候	☀️晴れ	+50	0
	☔️雨	-100	-10

- 猫が顔を洗った場合の期待効用 U

0.29

- $U(\text{多めに仕入れ}) = -100 \times P(\text{rain} \mid \text{cat}) + 50 \times P(\neg \text{rain} \mid \text{cat}) = -100 \times 0.29 + 50 \times 0.71 = 6.5$
- $U(\text{少な目に仕入れ}) = -10 \times P(\text{rain} \mid \text{cat}) + 0 \times P(\neg \text{rain} \mid \text{cat}) = -10 \times 0.29 + 0 \times 0.71 = -2.9$
- 多めに仕入れたほうが期待効用が高い

ナイーブベイズ予測： テキスト分類の初等的手法

- ある文書が特定のカテゴリに属する確率を知りたい
 - たとえば、Webページをみて、そのトピックが経済なのか、スポーツなのか、政治なのか、芸能なのか、...を判別するのに使える
 - つまり、事後確率 $P(\text{topic} \mid \text{text})$ を知りたい
- ベイズの公式により、事後確率は：

$$P(\text{topic} \mid \text{text}) = \frac{P(\text{text} \mid \text{topic})P(\text{topic})}{P(\text{text})}$$

- $P(\text{topic})$ ：そのトピックが観測される確率（事前確率）
- $P(\text{text} \mid \text{topic})$ ：あるトピックが決まった時に、そのWebページのテキストが作られる確率

ナイーブベイズ予測：

予測にはテキストの生成確率の計算が必要

- テキストが与えられたときのトピックの事後確率は：

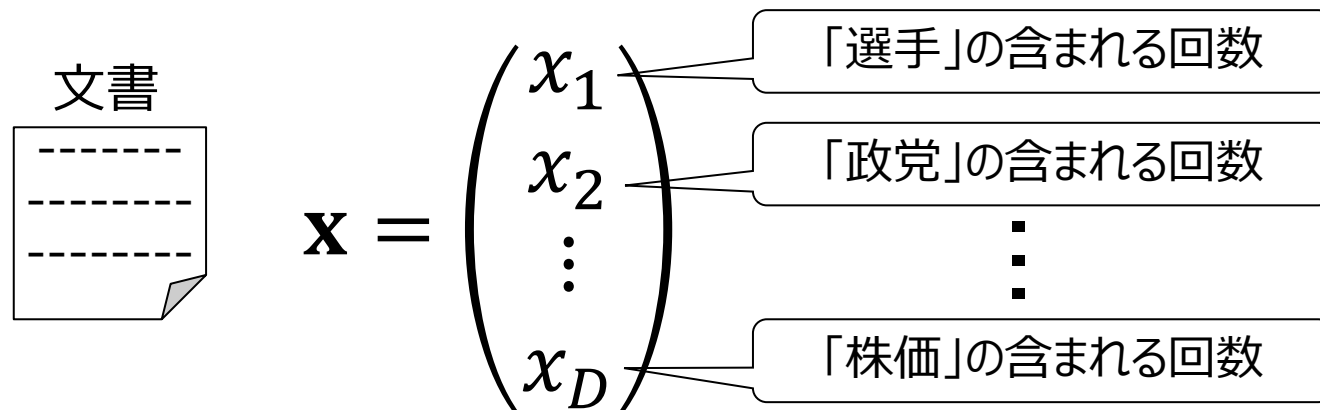
$$P(\text{topic} \mid \text{text}) = \frac{P(\text{text} \mid \text{topic})P(\text{topic})}{\sum_{\text{topic}'} P(\text{text} \mid \text{topic}')P(\text{topic}')}$$

分母 = $P(\text{text})$

- $P(\text{topic}) = \frac{\text{そのトピックの文書数}}{\text{全文書数}}$ で計算できる（最尤推定）
- 一方、 $P(\text{text} \mid \text{topic})$ をどのように考えるかは自明ではない
 - ◆ トピックが決まった時の文書生成モデルが必要
 - ◆ 例えば、マルコフモデル？ 隠れマルコフ？ RNN？
 - ◆ もうちょっと簡単な場合を考えてみる...

テキストの発生確率モデル： 単語袋(bag-of-words)モデル

- 文書 \mathbf{x} を、文書中に出現する単語によって表現



bag-of-words 表現

- 単純化のための仮定：各単語の発生は独立とする

$$P(\text{text} \mid \text{topic})$$

$$= P(w_1 \mid \text{topic})^{n_1} P(w_2 \mid \text{topic})^{n_2} \cdots P(w_D \mid \text{topic})^{n_D}$$

辞書に含まれるある単語

ナイーブベイズ予測：

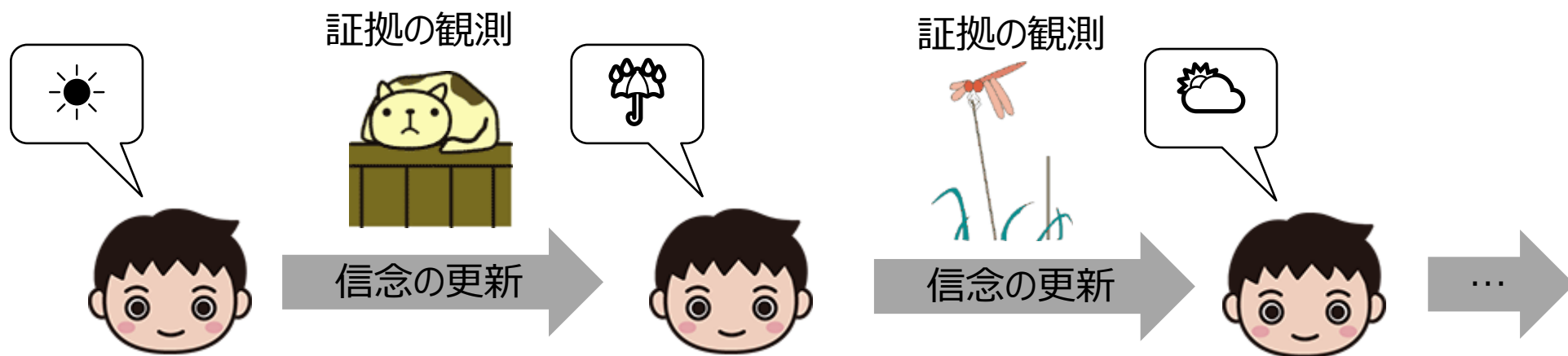
単語袋モデルに基づくテキストの発生確率の計算

- 計算したいのは $P(\text{topic} \mid \text{text}) \propto P(\text{text} \mid \text{topic})P(\text{topic})$
「スポーツ」 「スポーツ」 「スポーツ」
- $P(\text{topic}) = \frac{\text{そのトピックの文書数}}{\text{全文書数}}$ で推定できる（最尤推定）
- $P(\text{text} \mid \text{topic}) =$
 $P(w_1 \mid \text{topic})^{n_1} P(w_2 \mid \text{topic})^{n_2} \cdots P(w_D \mid \text{topic})^{n_D}$
「選手」 「スポーツ」 「スポーツ」 「選手」
 - $P(w_1 \mid \text{topic}) = \frac{\text{そのトピックの文書中で } w_1 \text{ が現れた回数}}{\text{そのトピックの文書中の総単語出現数}}$ と推定（最尤推定）
「選手」 「スポーツ」
- すべてのtopicについて $P(\text{text} \mid \text{topic})P(\text{topic})$ を計算し正規化する

信念の逐次更新：

証拠が得られるごとに信念が更新される

- ベイズの定理は証拠をもとに信念を更新する
- 証拠が新しく得られるたびに信念が更新される：
 $P(\text{rain}) \rightarrow P(\text{rain} \mid \text{cat}) \rightarrow P(\text{rain} \mid \text{cat}, \text{dragonfly}) \rightarrow \dots$
 - 証拠間が条件付き独立であるなら証拠の観測順序は関係ない
 - 他の例：テキスト分類で、一単語観測されるごとに予測を更新



ベイズ的統計モデリングの考え方： 最尤推定の尤度の代わりに事後分布を考える

- ベイズ統計では事後分布 $P(\text{パラメータ} | \text{データ})$ を考える

- 事後分布ではパラメータを確率変数と考える

- 事後分布：

$$P(\text{パラメータ} | \text{データ}) = \frac{P(\text{データ} | \text{パラメータ})P(\text{パラメータ})}{P(\text{データ})}$$

ベイズの定理

- 対数事後分布：

$$\begin{aligned} \log P(\text{パラメータ} | \text{データ}) \\ = \underbrace{\log P(\text{データ} | \text{パラメータ})}_{\text{尤度}} + \underbrace{\log P(\text{パラメータ})}_{\text{事前分布}} + \underbrace{\log P(\text{データ})}_{\text{定数}} \end{aligned}$$

離散分布のベイズ推定： ディリクレ分布を事前分布とする

- 離散分布 $\mathbf{p} = (p_1, p_2, \dots, p_k)$, $\sum_{i=1}^k p_i = 1, p_i \geq 0$ を考える
- データ : $\mathbf{n} = (n_1, n_2, \dots, n_k)$
 - n_j : 各シンボル $j \in \{1, 2, \dots, k\}$ の観測数
- パラメータの事後分布 $P(\mathbf{p} | \mathbf{n}) = \frac{P(\mathbf{n}|\mathbf{p})P(\mathbf{p})}{P(\mathbf{n})} = \frac{P(\mathbf{n}|\mathbf{p})P(\mathbf{p})}{\int_{\mathbf{p}} P(\mathbf{n}|\mathbf{p})P(\mathbf{p})d\mathbf{p}}$
- 事前分布 $P(\mathbf{p})$ はディリクレ分布とする
 - ディリクレ分布は離散分布の共役事前分布
 - 共役事前分布 : 事後分布と事前分布の形が同じになるような事前分布

ディリクレ分布： 離散分布の（共役）事前分布

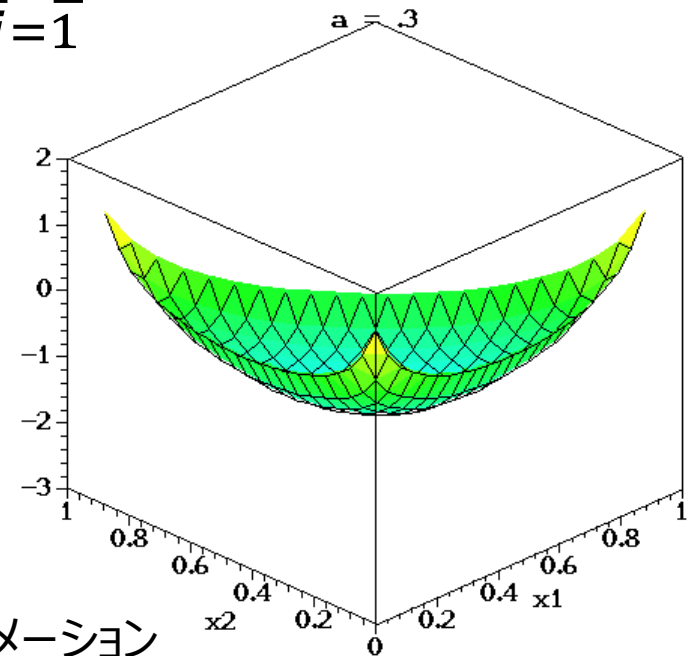
- ディリクレ分布：離散分布 $\mathbf{p} = (p_1, p_2, \dots, p_k), p_j \geq 0, \sum_{j=1}^k p_j = 1$ を生成する確率モデル

ガンマ関数

$$P(p_1, p_2, \dots, p_k) = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k (p_j)^{\alpha_j - 1}$$

- $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k) \geq 0$ は（超）パラメータ

- $\int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{\alpha_j - 1} d\mathbf{p} = \frac{\prod_{j=1}^k \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^k \alpha_j)}$



0.3 ≤ α₁ = α₂ = α₃ ≤ 2.0 のアニメーション

https://en.wikipedia.org/wiki/Dirichlet_distribution#/media/File:LogDirichletDensity-alpha_0.3_to_alpha_2.0.gif

事後分布の計算：

事前分布と事後分布がともにディリクレ分布になる

- 事後分布：
$$P(\mathbf{p} \mid \mathbf{n}) = \frac{P(\mathbf{n} \mid \mathbf{p})P(\mathbf{p})}{P(\mathbf{n})} = \frac{P(\mathbf{n} \mid \mathbf{p})P(\mathbf{p})}{\int_{\mathbf{p}} P(\mathbf{n} \mid \mathbf{p})P(\mathbf{p})d\mathbf{p}}$$

- $P(\mathbf{n} \mid \mathbf{p})P(\mathbf{p}) = \prod_{j=1}^k (p_j)^{n_j} \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k (p_j)^{\alpha_j - 1}$

事前分布がディリクレ分布

- 以上を用いて
$$P(\mathbf{p} \mid \mathbf{n}) = \frac{\prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1}}{\int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1} d\mathbf{p}}$$
$$= \frac{\Gamma(\sum_{j=1}^k n_j + \alpha_j)}{\prod_{j=1}^k \Gamma(n_j + \alpha_j)} \prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1}$$

事後分布もディリクレ分布

ベイズ予測分布： 推定のばらつきを考慮した予測

- MAP推定では事後分布が最大となるパラメータを**点推定**する

$$\hat{\mathbf{p}} = \operatorname{argmax}_{\mathbf{p}} P(\mathbf{p} \mid \mathbf{n})$$

- 得られたパラメータを次のシンボル x の予測に用いる

$$P(x \mid \hat{\mathbf{p}}) = \hat{p}_x, x \in \{1, 2, \dots, k\}$$

- ベイズ予測では「事後分布そのもの」を用いて予測する

$$P(x \mid \mathbf{n}) = \int_{\mathbf{p}} P(x \mid \mathbf{p}) P(\mathbf{p} \mid \mathbf{n}) d\mathbf{p}$$

- あらゆるパラメータのモデルの予測を事後確率で重みづけて予測
- パラメータを点推定するのではなく**事後分布ごっそり使う**

離散分布のベイズ予測分布：
MAP推定同様に加算平滑化が導かれる

$$\int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{\alpha_j-1} d\mathbf{p} = \frac{\prod_{j=1}^k \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^k \alpha_k)}$$

$$\begin{aligned} P(x \mid \mathbf{n}) &= \int_{\mathbf{p}} P(x \mid \mathbf{p}) P(\mathbf{p} \mid \mathbf{n}) d\mathbf{p} \\ &= \int_{\mathbf{p}} p_x \frac{\Gamma(\sum_{j=1}^k n_j + \alpha_j)}{\prod_{j=1}^k \Gamma(n_j + \alpha_j)} \prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1} d\mathbf{p} \\ &= \frac{\Gamma(\sum_{j=1}^k n_j + \alpha_j)}{\prod_{j=1}^k \Gamma(n_j + \alpha_j)} \int_{\mathbf{p}} (p_x)^{n_x + 1 + \alpha_x - 1} \prod_{j \neq x} (p_j)^{n_j + \alpha_j - 1} d\mathbf{p} \\ &= \frac{\Gamma(\sum_{j=1}^k n_j + \alpha_j)}{\prod_{j=1}^k \Gamma(n_j + \alpha_j)} \frac{\Gamma(n_x + 1 + \alpha_x) \prod_{j \neq x} \Gamma(n_j + \alpha_j)}{\Gamma(1 + \sum_{j=1}^k n_j + \alpha_j)} \\ &= \frac{n_x + \alpha_x}{\sum_{j=1}^k n_j + \alpha_j} \end{aligned}$$

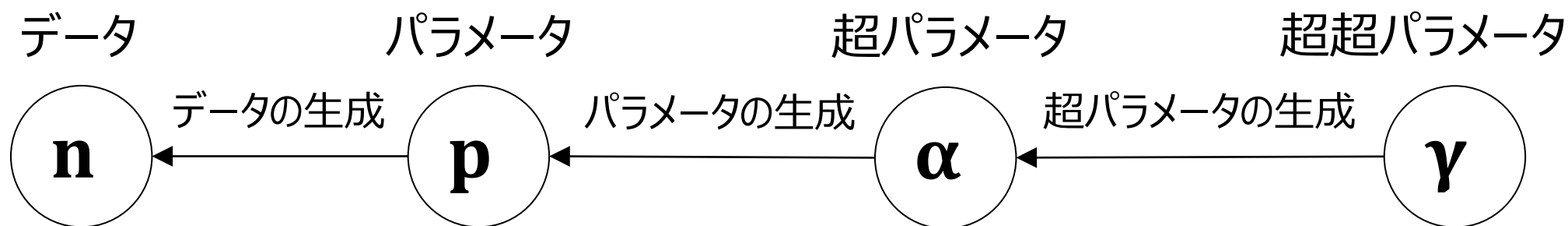
$$\Gamma(x + 1) = x\Gamma(x)$$

階層ベイズモデル：

事前分布の事前分布（の事前分布の・・・）を考える

- 事前分布 $P(\mathbf{p} \mid \alpha)$ も（超）パラメータ α をもつ
→ どのように決めたらよいか
- 超パラメータの事前分布 $P(\alpha)$ を考える
 - あるいは、データにあわせてチューニングする（後述）
- 事前分布のパラメータもまた事前分布をもつとするとモデルが階層化される

この辺になると
なんかもうどうでも
よくなってくる



経験ベイズ推定：

周辺尤度の最大化による超パラメータの推定

- 超パラメータを推定したい
- 周辺尤度： $P(\mathbf{n} \mid \boldsymbol{\alpha}) = \int_{\mathbf{p}} P(\mathbf{n} \mid \mathbf{p}) P(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p}$ を考える
 - パラメータ \mathbf{p} が積分消去（周辺化）されていることに注意
- 経験ベイズ法：
周辺尤度を最大化する超パラメータ $\boldsymbol{\alpha}$ を推定値とする
$$\hat{\boldsymbol{\alpha}} = \operatorname{argmax}_{\boldsymbol{\alpha}} P(\mathbf{n} \mid \boldsymbol{\alpha})$$
- MAP推定では超パラメータは推定できないことに注意する

離散分布の経験ベイズ推定：

対数周辺尤度を勾配法によって最大化する

- 周辺尤度： $P(\mathbf{n} \mid \boldsymbol{\alpha}) = \int_{\mathbf{p}} P(\mathbf{n} \mid \mathbf{p}) P(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p}$

- 離散分布の周辺尤度（数値的に最大化する）：

$$\begin{aligned} P(\mathbf{n} \mid \boldsymbol{\alpha}) &= \int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{n_j} \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k (p_j)^{\alpha_j-1} d\mathbf{p} \\ &= \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{n_j+\alpha_j-1} d\mathbf{p} = \frac{\Gamma(\sum_{j=1}^k \alpha_j)}{\prod_{j=1}^k \Gamma(\alpha_j)} \frac{\prod_{j=1}^k \Gamma(n_j + \alpha_j)}{\Gamma(\sum_{j=1}^k n_j + \alpha_j)} \end{aligned}$$

- 勾配法による $\boldsymbol{\alpha}$ の最適化：

$$\boldsymbol{\alpha}^{\text{NEW}} = \boldsymbol{\alpha} + \eta \nabla_{\boldsymbol{\alpha}} \log P(\mathbf{n} \mid \boldsymbol{\alpha})$$

- ディガンマ関数（対数ガンマ関数の微分）が必要でやや面倒

まとめ： ベイズモデリング

■ ベイズ統計の基礎

- ベイズ統計ではパラメータの事後分布を考える
 - ◆ ベイズの公式、事前分布、事後分布
- 事後分布による意思決定：証拠（データ）をもとに信念を更新
 - ◆ ナイーブベイズ予測

■ ベイズモデリング

- ディリクレ分布を共役事前分布とした離散分布のベイズ推定
- 階層ベイズモデリングによるモデルの抽象化
- 経験ベイズ法による超パラメータの推定