

<https://goo.gl/tYmyZL>

KYOTO UNIVERSITY

統計的モデリング基礎⑫ ～ベイズモデリング～

鹿島久嗣
(情報学科 計算機科学コース)

DEPARTMENT OF INTELLIGENCE SCIENCE
AND TECHNOLOGY

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
- 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



3

KYOTO UNIVERSITY

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 $P(\text{rain} | \text{cat})$: 猫が顔を洗うのを目撃したときに、明日雨が降る確率

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

- 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

当該日の前日に猫が
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)

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

雨がどうか関係なく、
そもそも猫が顔を洗う
確率

4

KYOTO UNIVERSITY

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明日雨が降る確率は20%
- 雨がふる日の前日に猫が顔を洗っている確率は80%
- $P(\text{rain} | \text{cat}) = \frac{P(\text{cat}|\text{rain})P(\text{rain})}{P(\text{cat})} = \frac{0.8 \times 0.2}{P(\text{cat})} = \frac{0.16}{P(\text{cat})}$ がわかる
- 一方、これまでの経験から雨が降らない確率は80%であることと、雨が降らない日の前日に猫が顔を洗っている確率は50%
- $P(\neg\text{rain} | \text{cat}) = \frac{P(\text{cat}|\neg\text{rain})P(\neg\text{rain})}{P(\text{cat})} = \frac{0.5 \times 0.8}{P(\text{cat})} = \frac{0.40}{P(\text{cat})}$
- 両者より $P(\text{rain} | \text{cat}) = \frac{0.16}{0.16+0.40} = 0.29$ (29%) となる

9ポイント増えたよ！

5

ベイズ決定：

事後分布をもちいた意思決定

- 予測×効用によって意思決定

		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

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

6

KYOTO UNIVERSITY

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
 - 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
 - つまり、事後確率 $P(\text{topic} | \text{text})$ を知りたい
- ベイズの公式により、事後確率は：

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

- $P(\text{topic})$ ：そのトピックが観測される確率
- $P(\text{text} | \text{topic})$ ：あるトピックが決まった時に、そのWebページ（のようなテキスト）がつくられる確率

7

KYOTO UNIVERSITY

ナイーブベイズ予測： テキストの生成確率が必要

- テキストが与えられたときのトピックの事後確率：

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

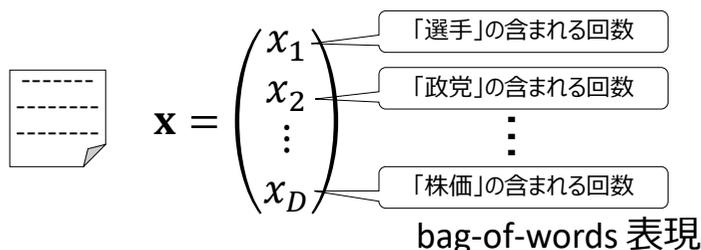
- $P(\text{topic}) = \frac{\text{そのトピックの文書数}}{\text{全文書数}}$ で計算できる（最尤推定）
- $P(\text{text} | \text{topic})$ ：これをどのように考えるかは自明ではない
 - ◆ トピックが決まった時の文書生成モデルが必要
 - ◆ マルコフモデル？

8

KYOTO UNIVERSITY

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、出現する単語によって表現



- 単純化のための仮定：各単語の発生は独立とする

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

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

辞書に含まれるある単語

9

KYOTO UNIVERSITY

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

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

- $P(\text{topic}) = \frac{\text{そのトピックの文書数}}{\text{全文書数}}$ で計算できる（最尤推定）

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

- $P(w_1 \mid \text{topic}) = \frac{\text{そのトピックの文書中で } w_1 \text{ が現れた回数}}{\text{そのトピックの文書中の総単語出現数}}$ で計算（最尤推定）

10

KYOTO UNIVERSITY

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：
 $P(\text{rain}) \rightarrow P(\text{rain} \mid \text{cat}) \rightarrow P(\text{rain} \mid \text{cat}, \text{dragonfly}) \rightarrow \dots$
 - 証拠間が独立であるならば証拠の得られる順番は関係ない
 - 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



11

KYOTO UNIVERSITY

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 $P(\text{パラメータ} \mid \text{データ})$ 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：

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

ベイズの定理

- 対数事後分布：
 $\log P(\text{パラメータ} \mid \text{データ})$
 $= \underbrace{\log P(\text{データ} \mid \text{パラメータ})}_{\text{尤度}} + \underbrace{\log P(\text{パラメータ})}_{\text{事前分布}} + \underbrace{\log P(\text{データ})}_{\text{定数}}$

12

KYOTO UNIVERSITY

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
- 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
 - 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：
ディリクレ分布を事前分布とする

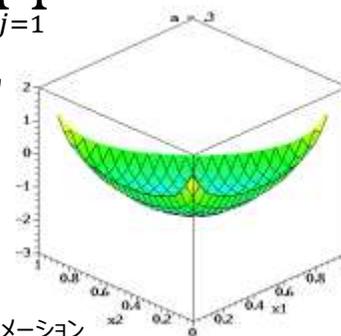
- ディリクレ分布：離散分布 $\mathbf{p} = (p_1, p_2, \dots, p_k)$, $p_j \geq 0, \sum_{j=1}^k p_j = 1$ を生成する確率モデル

ガンマ関数

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

- $\alpha = (\alpha_1, \dots, \alpha_k) \geq 0$ 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：
事後分布もディリクレ分布になる

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

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

事前分布がディリクレ分布

■
$$P(\mathbf{p} | \mathbf{n}) = \frac{\prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1}}{\int_{\mathbf{p}} \prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1} d\mathbf{p}}$$

$$= \frac{\Gamma(\sum_{j=1}^k n_j + \alpha_j)}{\prod_{j=1}^k \Gamma(n_j + \alpha_j)} \prod_{j=1}^k (p_j)^{n_j + \alpha_j - 1}$$

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

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

- 得られたパラメータを次のシンボル x の予測に用いる

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

- ベイズ予測では事後分布そのものを用いて予測する

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

- 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
- 最適化問題を解いてパラメータを点推定するのではなく、全部使う

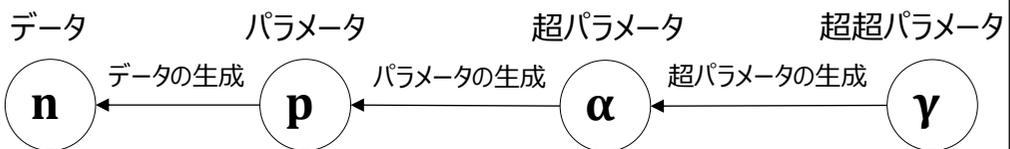
離散分布のベイズ予測分布：
やはり加算平滑化になる

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

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

階層ベイズモデル：
事前分布の事前分布

- 事前分布 $P(\mathbf{p} | \boldsymbol{\alpha})$ もパラメータ（超パラメータ） $\boldsymbol{\alpha}$ 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 $P(\boldsymbol{\alpha})$ 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



経験ベイズ推定： 周辺尤度の最大化

- 周辺尤度： $P(\mathbf{n} | \boldsymbol{\alpha}) = \int_{\mathbf{p}} P(\mathbf{n} | \mathbf{p}) P(\mathbf{p} | \boldsymbol{\alpha}) d\mathbf{p}$
- 周辺尤度を最大化する超パラメータ $\boldsymbol{\alpha}$ を求める：
 $\hat{\boldsymbol{\alpha}} = \operatorname{argmax}_{\boldsymbol{\alpha}} P(\mathbf{n} | \boldsymbol{\alpha})$
- 離散分布の周辺尤度（数值的に最大化する）：

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