多項分布を使った
ベイズ的モデリング

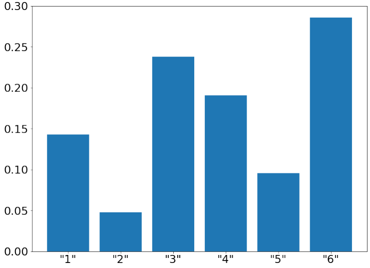
正田 備也

masada@rikkyo.ac.jp

Contents

カテゴリカル分布

- $\mathsf{V} = \{\mathsf{v}_1, \ldots, \mathsf{v}_W\}$ を W 種類のアイテムの集合とする
  例 1. サイコロの目（$W = 6$）
  例 2. 自然言語の語彙 ($W =$ 数千〜数十万?)
- カテゴリカル分布は V 上に定義された離散確率分布
- パラメータは $\phi = (\phi_1, \ldots, \phi_W)$
  - アイテム v_w が出現する確率 ϕ_w
  - $\sum_{w=1}^{W} \phi_w = 1$ を満たす



多項分布 multinomial distribution

- 複数回の独立な試行のモデリングには、多項分布を使う
- 計 n 回の試行のうち各アイテムが何回ずつ出現するか、その可能なすべての場合に確率を割り振る確率分布が多項分布
  - 次のスライド参照
- パラメータは n と $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_W)$
  - 試行の回数 n（観測データから決まる）
  - アイテム v_w の出現確率 ϕ_w（$\sum_w \phi_w = 1$ を満たす）
  - $\sum_w \phi_w = 1$ が満たされる

多項分布はどのような集合の上に定義されるか

- 多項分布は "計 n 回の試行のうち各アイテムが何回ずつ出現するかの、可能な全ての場合の集合" の上に定義される
- W 種類のアイテムから重複を許して n 個を選ぶすべての場合にわたって確率を合計すると、1 になる
  - W 種類のアイテムから重複を許して n 個を選ぶ場合の数は？
  - n 個の「○」と $W-1$ 個の「｜」（仕切り）を並べる場合の数と同じ

例. 「○○｜｜○｜○○○」は、$n=6$ で、v_1 が 2 回、v_2 が 0 回、v_3 が 1 回、v_4 が 3 回、それぞれ出現する場合を表す

多項分布の確率質量関数

- アイテム v_w の出現回数を c_w と書くことにする
- 総試行回数を n とすると、$\sum_w c_w = n$ が成り立つ
- このとき、多項分布の確率質量関数 pmf は以下ように書ける

$$p((c_1, \ldots, c_W); \boldsymbol{\phi}, n) = \frac{n!}{\prod_{w=1}^{W} c_w!} \prod_{w=1}^{W} \phi_w^{c_w} \tag{1}$$

  - $\frac{n!}{\prod_w c_w!} = \frac{n!}{c_1! \cdots c_W!}$ の部分は、n 回の試行のうち、アイテム v_w が c_w 回出現するような試行の列の総数をあらわしている
- 多項分布は、各アイテムの出現回数が同じで、出現順が違うだけの試行列を区別できない

Contents

多項分布によるモデリングに登場する変数

- アイテムの出現列を表す確率変数 $\boldsymbol{x} = \{x_1, \ldots, x_n\}$
  - x_i は、i 番目に出現したアイテムを表す<u>確率変数</u>

    例. $x_i =$ “apple” は「i 番目に出現する単語は “apple”」という意味
  - x_i の値はすでに観測されている（値が既知の変数）
- 多項分布のパラメータ $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_W)$
  - ϕ_w は、アイテム v_w の出現確率を表す<u>パラメータ</u>

    例. $\phi_w = 0.0013$ は「単語 v_w の出現確率が 0.0013」という意味
  - ϕ_w は値が未知の変数
  - この値の推定が、多項分布によるモデリングにおいて解くべき問題

多項分布の最尤推定

- 観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ はアイテムの出現の列
- 多項分布によるモデリングでは、出現順序は無視される
- つまり、各アイテム v$_w$ の出現回数 c_w だけが問題とされる
- このとき、観測データ $\boldsymbol{x}$ の尤度は、ϕ の関数として、以下のように書ける

$$p(\boldsymbol{x}; \boldsymbol{\phi}, n) = \frac{n!}{\prod_{w=1}^{W} c_w!} \prod_{w=1}^{W} \phi_w^{c_w} \tag{2}$$

- 尤度を最大化する ϕ の値を推定値とするのが最尤推定
  - 最尤推定のほかにも ϕ の値を推定する方法はある

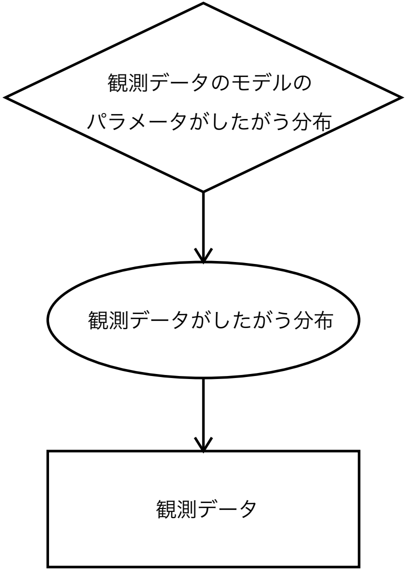
多項分布の最尤推定の問題点

- 観測データに現れるアイテム以外のアイテムについては、出現確率 ϕ_w がゼロと推定される
- よって、最尤推定の結果を使って未知データの確率を計算するとき、ひとつでも観測データに現れないアイテムが含まれていると、確率はゼロと算出されてしまう
  - テキストデータで言えば、out-of-vocabulary (OoV) words の問題
- ベイズ的な考え方を使うと、この問題にひとつの解決を与えることができる

Contents

ベイズ的なモデリングとは

- 統計モデルは観測データの不確かさ uncertainty を表現する
- だが、ベイズ的な統計モデリングでは、観測データをもとにして統計モデルのパラメータを決めること自体にも不確かさ uncertainty があると考える
- そこで、パラメータも確率変数とみなし、パラメータも確率分布にしたがっているものとしてモデリングする
- そこで導入されるのが事前分布である
- 事前分布はパラメータがしたがう確率分布として導入される


観測データのモデルの
パラメータがしたがう分布
観測データがしたがう分布
観測データ

多項分布を使うベイズ的モデルの “部品”

- $p(\boldsymbol{x}|\boldsymbol{\phi})$ 観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ の尤度
  - x_i は i 番目に出現するアイテムを表す確率変数
  - 事前分布を使わないときは $p(\boldsymbol{x};\boldsymbol{\phi})$ と書いていた
  - ベイズ的モデリングでは、$p(\boldsymbol{x}|\boldsymbol{\phi})$ と、条件付き確率として書く
  - これは、観測変数 x_i だけでなく、$\boldsymbol{\phi}$ も確率変数となるからである
- $p(\boldsymbol{\phi};\boldsymbol{\beta})$ 多項分布のパラメータ $\boldsymbol{\phi}$ が従う事前分布
  - $\boldsymbol{\beta}$ は事前分布のパラメータ
  - 事前分布のパラメータを一般にハイパーパラメータと呼ぶ
  - ここにどんな分布を使えばいいか？（以下、説明。）

多項分布の事前分布はどのような分布か

- 多項分布によるモデリングでは、W 種類のアイテムの出現頻度 c_w をモデリングする
  - アイテムの出現順序は無視される
- W 個のパラメータ $\phi_1, \ldots, \phi_W$ は、いずれも非負で、和が 1
- 非負で和が 1 になる実数の組 $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_W)$ は無数にある
- これら無数の組の上に、確率分布を定義したい
- 非負で和が 1 になる実数の組の上に定義される確率分布なら、事前分布として使える

ディリクレ分布 Dirichlet distribution

- 非負で和が 1 になる W 個の実数の組は無数にある
- ディリクレ分布は、それら無数の組の上に定義される確率分布のひとつ
  - つまり、ディリクレ分布の台 support は $W-1$ 次元単体 simplex
- 多項分布のパラメータがしたがう事前分布として利用できる
- ディリクレ分布の確率密度関数は

$$p(\boldsymbol{\phi};\boldsymbol{\beta}) = \frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\prod_{w=1}^{W}\Gamma(\beta_w)}\prod_{w=1}^{W}\phi_w^{\beta_w-1} \tag{3}$$

  - $\frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\prod_{w=1}^{W}\Gamma(\beta_w)}$ は規格化定数で、$\Gamma(\cdot)$ はガンマ関数

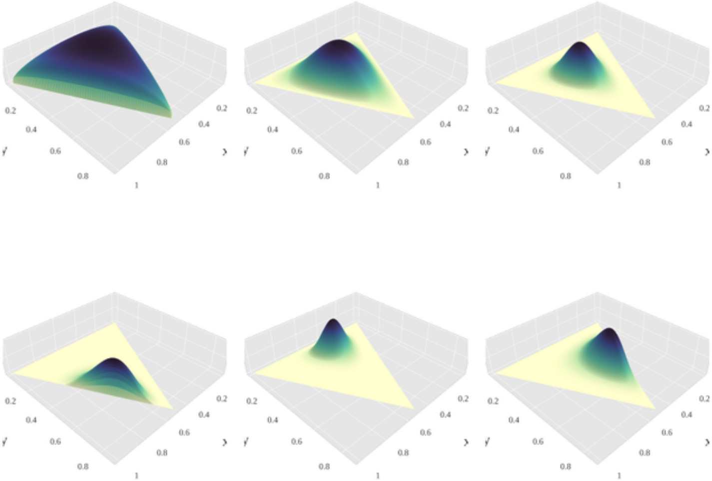


Figure: ディリクレ分布の確率密度関数の例 ($W = 3$)

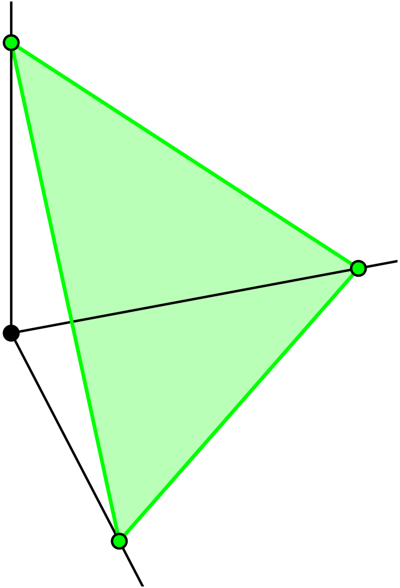
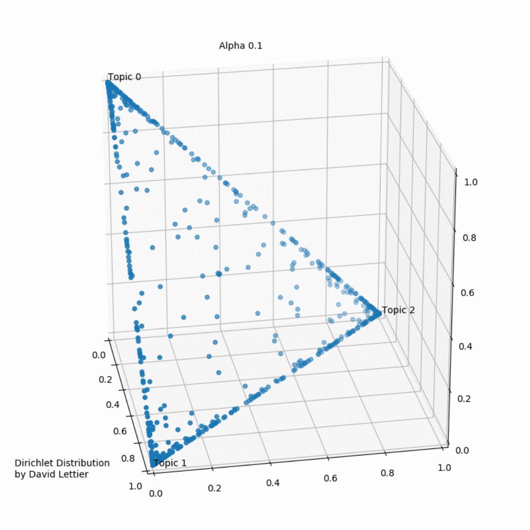


Figure: https://en.wikipedia.org/wiki/Simplex より standard 2-simplex の図



https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

ガンマ関数

- ガンマ関数について次の等式が成り立つ

$$\Gamma(x+1) = x\Gamma(x) \tag{4}$$

- この授業でガンマ関数について把握すべきことは式 (4) だけ
- 上式より、自然数 n について、$\Gamma(n+1) = n!$
  - 式 (4) はガンマ関数の定義から導かれるが、定義は知らなくてよい
  - ガンマ関数は実際は複素数について定義されるが、ここでは実数、しかも正の実数を引数とする場合しか扱わない
  - ディリクレ分布の規格化定数が $\frac{\Gamma(\sum_{w=1}^{W}\beta_w)}{\prod_{w=1}^{W}\Gamma(\beta_w)}$ である証明は割愛する

    cf. https://math.stackexchange.com/questions/207073/definite-integral-over-a-simplex

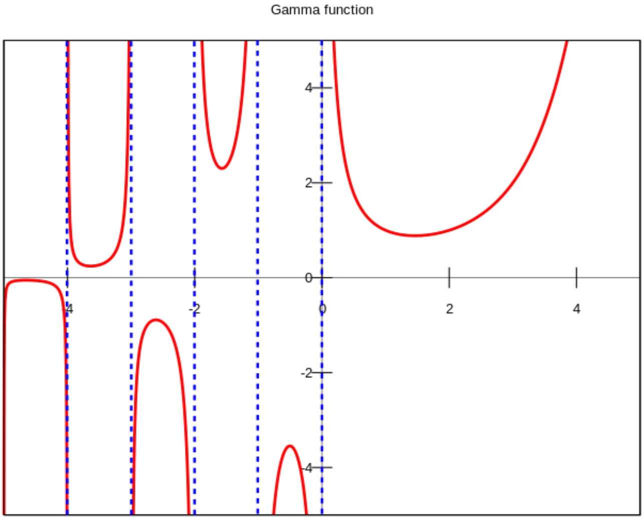


Figure: ガンマ関数

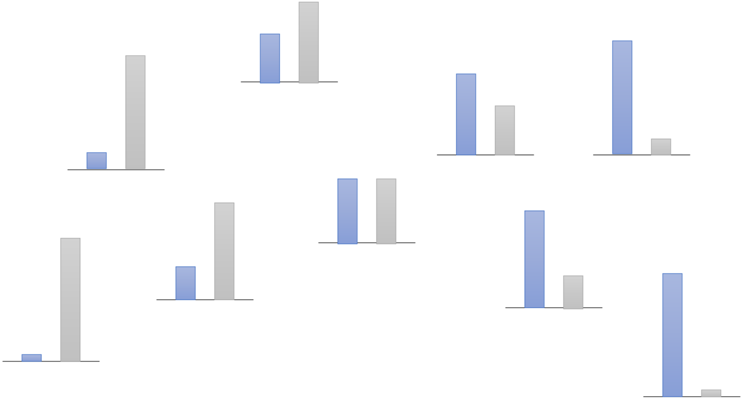
ベータ分布とディリクレ分布

- ベータ分布は、ディリクレ分布で $W = 2$ の場合に対応する
- アイテムの種類が 2 種類の場合と 3 種類以上の場合との対応関係は、以下の表のとおり

アイテムが 2 種類の場合	アイテムが 3 種類以上の場合
ベルヌーイ分布	カテゴリカル分布
二項分布	多項分布
ベータ分布	ディリクレ分布

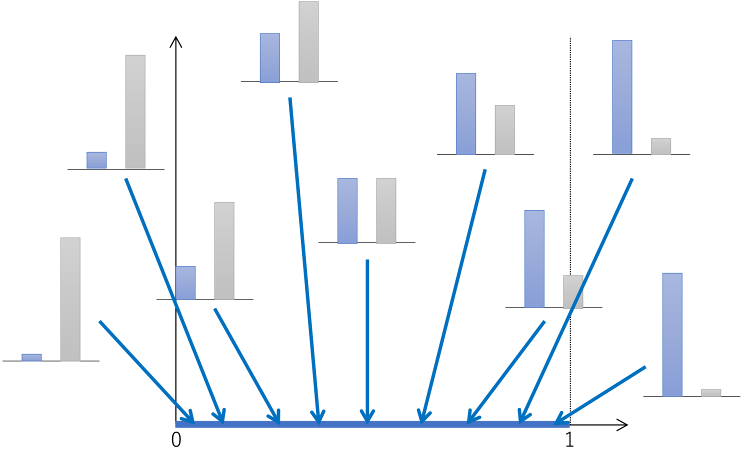
ベータ分布の二項分布に対する関係 (1/2)

- 「二項分布をひとつ選ぶこと」＝「ϕ_1 の値を選ぶこと」



ベータ分布の二項分布に対する関係 (2/2)

- 「二項分布をひとつ選ぶこと」＝「$[0,1]$ 上の 1 点を選ぶこと」



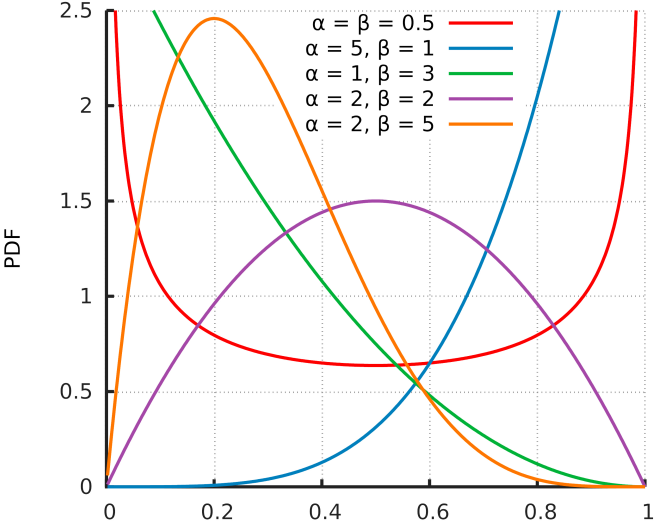


Figure: https://en.wikipedia.org/wiki/Beta_distribution

分布の分布としてのベータ分布

- ベータ分布は $[0, 1]$（1 次元単体）の上に定義される
- $[0, 1]$ 上の一点一点が、別々の二項分布に対応している

  例. $0.6 \in [0, 1]$ は $\phi = (0.6, 0.4)$ をパラメータとする二項分布に対応
  - 2 つのパラメータのうち一方を決めると他方は自動的に決まる
  - ただし、試行の総数 n はあらかじめ固定されているとする
- ということは、ベータ分布は二項分布の集合上に定義された分布とみなせる
  - つまり、ベータ分布は、分布の分布とみなせる

分布の分布としてのディリクレ分布

- ディリクレ分布は $W-1$ 次元単体の上に定義される
  - 1 次元単体は線分、2 次元単体は正三角形、3 次元単体は正四面体
- $W-1$ 次元単体の一点一点が別々の多項分布に対応している
  - 多項分布の W 個のパラメータ $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_W)$ のうち $W-1$ 個、例えば ϕ_1 から ϕ_{W-1} を決めると、残り 1 個は決まる
  - よって多項分布のパラメータ $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_W)$ の取りうる値の組のひとつひとつが、$W-1$ 次元単体に含まれる 1 点 1 点に対応
- ということは、ディリクレ分布は多項分布の集合上に定義された分布とみなせる
  - つまり、ディリクレ分布は、分布の分布とみなせる

多項分布を使うベイズ的モデルの “部品”

- $p(\boldsymbol{x}|\boldsymbol{\phi})$ 観測データ $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ の尤度
  - x_i は i 番目に出現するアイテムを表す確率変数
  - 事前分布を使わないときは $p(\boldsymbol{x}; \phi)$ と書いていた
  - ベイズ的モデリングでは、$p(\boldsymbol{x}|\phi)$ と、条件付き確率として書く
  - これは、観測変数 x_i だけでなく、ϕ も確率変数となるからである
- $p(\boldsymbol{\phi}; \boldsymbol{\beta})$ 多項分布のパラメータ ϕ が従う事前分布
  - $\boldsymbol{\beta}$ は事前分布のパラメータ
  - 事前分布のパラメータを一般にハイパーパラメータと呼ぶ
  - ・・・というわけで、ここにディリクレ分布を使うことにする

事後分布 posterior distribution

$$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \propto p(\boldsymbol{x}|\boldsymbol{\phi})p(\boldsymbol{\phi};\boldsymbol{\beta}) \tag{5}$$

- ベイズ的モデリングは、事後分布を求めることを課題とする
- 事後分布はモデルパラメータ $\boldsymbol{\phi}$ が従う確率分布で、観測データ $\boldsymbol{x}$ が所与の条件付き確率分布
- 事後分布は、式 (5) のように、ベイズ則によって観測データの尤度 $p(\boldsymbol{x}|\boldsymbol{\phi})$ と事前分布 $p(\boldsymbol{\phi};\boldsymbol{\beta})$ とから導き出される
  - 事後分布 $p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta})$ は、パラメータが取りうる値の全てについて、それぞれがどのくらいありえそうかを表している

事後分布を求めることと最尤推定との一つの違い

- 最尤推定は、データ尤度 $p(\boldsymbol{x}|\phi)$ をパラメータ ϕ の関数とみなして最大化することで、ϕ の値をひとつに決める

$$\arg\max_{\phi} p(\boldsymbol{x}|\phi)$$

- 一方、ベイズ的なモデリングでは、パラメータ ϕ が取りうる全ての値について、各々どのくらいありえそうかを表している事後分布 $p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta})$ を、求める

$$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \propto p(\boldsymbol{x}|\boldsymbol{\phi})p(\boldsymbol{\phi};\boldsymbol{\beta})$$

事後分布の直感的な意味

$$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \propto p(\boldsymbol{x}|\boldsymbol{\phi})p(\boldsymbol{\phi};\boldsymbol{\beta}) \tag{6}$$

- 上の式は、事前分布 $p(\phi;\boldsymbol{\beta})$ が尤度 $p(\boldsymbol{x}|\phi)$ によって重み付けし直されて事後分布になる、という式
- ϕ が尤度 $p(\boldsymbol{x}|\phi)$ を大きくするような値だと、右辺においてそれだけ大きな値が掛け算される
- よって、左辺の事後分布で ϕ がそのような値を取る確率は大

共役事前分布 conjugate prior distribution

- 共役事前分布とは、事後分布を事前分布と同じ種類の分布にするような事前分布のことをいう
- 例えば、データ尤度 $p(\boldsymbol{x}|\phi)$ が多項分布で表されているとき、事前分布としてディリクレ分布を用いると、事後分布もディリクレ分布となる
- つまり、ディリクレ分布は共役事前分布である
- このため、多項分布を使ってベイズ的なモデリングをするとき、ディリクレ分布を事前分布に使うことが多い
  - ディリクレ分布以外の分布を事前分布として使うこともある
  - 例えば、logit-normal distribution を使うことがある

問題5-1

- ディリクレ分布が共役事前分布であることを示せ
- ヒント：尤度が多項分布を使って表されるとき、事前分布をディリクレ分布にすると、事後分布もディリクレ分布になることを示せばよい

$$p(\boldsymbol{x}|\boldsymbol{\phi})p(\boldsymbol{\phi};\boldsymbol{\beta}) = \frac{n!}{\prod_w c_w!}\prod_w \phi_w^{c_w} \times \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)}\prod_w \phi_w^{\beta_w - 1} \propto \prod_w \phi_w^{c_w + \beta_w - 1} \tag{7}$$

$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \propto p(\boldsymbol{x}|\boldsymbol{\phi})p(\boldsymbol{\phi};\boldsymbol{\beta})$ より、

$$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \propto \prod_w \phi_w^{c_w + \beta_w - 1} \tag{8}$$

あとは、規格化定数 normalizing constant を求めればよい。

$$\int_{\{\boldsymbol{\phi}:\sum_w \phi_w = 1\}} \prod_w \phi_w^{c_w + \beta_w - 1} d\boldsymbol{\phi} = \frac{\prod_w \Gamma(c_w + \beta_w)}{\Gamma(n + \sum_w \beta_w)} \tag{9}$$

よって、

$$p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) = \frac{\Gamma(n + \sum_w \beta_w)}{\prod_w \Gamma(c_w + \beta_w)}\prod_w \phi_w^{c_w + \beta_w - 1} \tag{10}$$

最大事後確率推定（MAP 推定）

- 事後確率を最大化する ϕ の値をモデルパラメータの推定値とする推定方法を、最大事後確率推定という

$$\hat{\phi}_{\text{MAP}} = \arg\max_{\phi} p(\boldsymbol{\phi}|\boldsymbol{x};\boldsymbol{\beta}) \tag{11}$$

  - MAP 推定と略される（MAP; maximum a posteriori）
- 下記の最尤推定と同様、モデルパラメータ ϕ の値をひとつ選ぶ推定方法

$$\hat{\phi}_{\text{ML}} = \arg\max_{\phi} p(\boldsymbol{x};\boldsymbol{\phi}) \tag{12}$$

多項分布のMAP推定

- 観測データのモデルが多項分布で、ディリクレ分布が事前分布のとき、MAP 推定が与える解は

$$\hat{\phi}_w = \frac{c_w + \beta_w - 1}{\sum_w (c_w + \beta_w - 1)} \tag{13}$$

  - 問：なぜこうなるか、示せ
- アイテムの実際の出現頻度ではなく、$\beta_w - 1$ 回だけ下駄を履かせた出現頻度で確率を計算していることになる
  - 単語の出現確率を求めるとき、このように下駄を履かせた回数を代わりに使うことを、スムージング smoothing という

Contents

情報検索 information retrieval

- たくさんの文書を持っている
- それらの文書をクエリに適合する（relevant な）順にソート
  - 情報検索とは、このようなことをすること
- どう実装すればいい？
- 実装例
  - ひとつひとつの文書について別々に単語出現確率 ϕ をMAP 推定
  - 推定された ϕ を使って、クエリの生成確率を計算
  - この生成確率を高くする順に文書をソート

文書をランキングするための計算

- 上述の MAP 推定は、検索対象の文書群のうち d 番目の文書について単語 v_w の出現確率を $\hat{\phi}_{d,w} = \frac{c_{d,w}+\beta_w-1}{\sum_w (c_{d,w}+\beta_w-1)}$ と与える
  - たとえ v_w が文書 d に現れない単語であっても、つまり $c_{d,w}=0$ であっても、$\beta_w > 1$ ならば確率がゼロにならないことに注意
- この単語確率によってクエリ $\boldsymbol{x}_q$ が生成される確率は：

$$p(\boldsymbol{x}_q|\hat{\phi}_d) = \frac{n_q!}{\prod_w c_{q,w}!} \prod_w \left(\frac{c_{d,w}+\beta_w-1}{\sum_w (c_{d,w}+\beta_w-1)} \right)^{c_{q,w}} \tag{14}$$

  - $c_{q,w}$ はクエリにおける単語 v_w の出現頻度
- $p(\boldsymbol{x}_q|\hat{\phi}_d)$ の降順に、文書をソートすればよい

$$p(\boldsymbol{x}_q; \boldsymbol{\phi}_d) = 2 \cdot {\phi_{d,\mathrm{apple}}}^1 \cdot {\phi_{d,\mathrm{pie}}}^1$$

各文書のパラメータの推定値を使ってクエリの確率を求める

$$\boldsymbol{\phi}_1 = (\phi_{1,1}, \dots, \phi_{1,W})$$

$$\boldsymbol{\phi}_2 = (\phi_{2,1}, \dots, \phi_{2,W})$$

$$\boldsymbol{\phi}_3 = (\phi_{3,1}, \dots, \phi_{3,W})$$

$$\boldsymbol{\phi}_4 = (\phi_{4,1}, \dots, \phi_{4,W})$$

$$\boldsymbol{\phi}_5 = (\phi_{5,1}, \dots, \phi_{5,W})$$

文書ごとに別々のパラメータ集合を用意し文書ごとにMAP推定する。

背景確率を使ったスムージング

- MAP 推定によると d 番目の文書における単語 v$_w$ の出現確率は $\hat{\phi}_{d,w} = \frac{c_{d,w}+\beta_w-1}{\sum_w (c_{d,w}+\beta_w-1)}$ である
- 実際には、コーパス全体における単語 v$_w$ の出現確率 p_w を使って、$\beta_w - 1$ の部分を λp_w で置き換えることが多い
  - p_w のことを背景確率 background probability と呼んだりする
- つまり、$\hat{\phi}_{d,w} = \frac{c_{d,w}+\lambda p_w}{c_d+\lambda}$ とする
  - $c_d \equiv \sum_w c_{d,w}$ は d 番目の文書の長さ
- λ は検証用クエリの検索性能を見ながらチューニングする

cf. https://nlp.stanford.edu/IR-book/html/htmledition/estimating-the-query-generation-probability-1.html