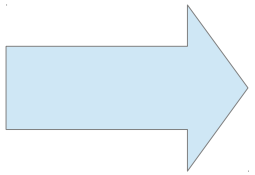
最尤法の簡単な説明と単語の出現確率の算出

新納浩幸

問題

袋の中に赤、青、黄、緑の4つボールが無数に入っている。
今、袋から10個ボールを取り出したら、
赤が4つ、青が3つ、黄が2つ、緑が1つだった。

袋の中の赤、青、黄、緑のボールのそれぞれの割合は？



赤： 0.4
青： 0.3
黄： 0.2
緑： 0.1

Why？

多項分布(1)

P_r ：赤の割合　　P_b ：青の割合

P_y ：黄の割合　　P_g ：緑の割合

$$P_r + P_b + P_y + P_g = 1$$

R ：赤の個数　　B ：青の個数

Y ：黄の個数　　G ：緑の個数

$$R + B + Y + G = 10$$

多項分布(2)

$$P(R=r,B=b,Y=y,G=g)=\frac{(r+b+y+g)!}{r!b!y!g!}P_r^r P_b^b P_y^y P_g^g$$

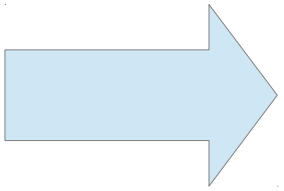
「赤が4つ、青が3つ、黄が2つ、緑が1つ」
となる確率は、

$$P(R=4,B=3,Y=2,G=1)=\frac{10!}{4!3!2!1!}P_r^4 P_b^3 P_y^2 P_g^1$$

最尤法のアイデア

実験の結果、
D ＝「赤が4つ、青が3つ、黄が2つ、緑が1つ」
が起きた。なぜ、D が起きたのか？

D が起こる確率が最も高かったからである



P(D) を最大にするような

$P_r \quad P_b \quad P_y \quad P_g$

をそれぞれの推定値とする

最尤法

最大最小の問題(1)

$$P(R=4,B=3,Y=2,G=1)=\frac{10!}{4!3!2!1!}P_r^4P_b^3P_y^2P_g^1$$

$$r=P_r,b=P_b,y=P_y,g=P_g$$

$$f(r,b,y,g)=Cr^4b^3y^2g^1$$

これを最大にする r,b,y,g を求める、多変数関数の最大最小問題

対数は単調増加なので、対数をとっても同じ

$$f(r,b,y,g)=\log C+4\log r+3\log b+2\log y+\log g$$

最大最小の問題(2)

$$f(r,b,y,g) = \log C + 4\log r + 3\log b + 2\log y + \log g$$

制約がある　　$r + b + y + g = 1$

制約がある場合の最大最小問題は、ラグランジュ乗数法

$$f(r,b,y,g) = \log C + 4\log r + 3\log b + 2\log y + \log g$$
$$-\lambda(r + b + y + g - 1)$$

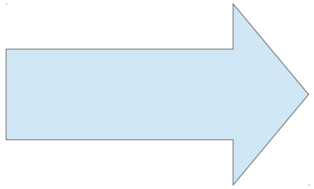
これを最大にする r, b, y, g を求める

最大最小の問題(3)

極値問題として解ける

$$\frac{\partial f}{\partial r} = \frac{4}{r} - \lambda = 0 \qquad \frac{\partial f}{\partial b} = \frac{3}{b} - \lambda = 0$$

$$\frac{\partial f}{\partial y} = \frac{2}{y} - \lambda = 0 \qquad \frac{\partial f}{\partial g} = \frac{1}{g} - \lambda = 0$$

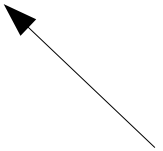


$$\lambda = 10$$

$$r = 0.4, g = 0.3, y = 0.2, g = 0.1$$

最尤法のアプローチ

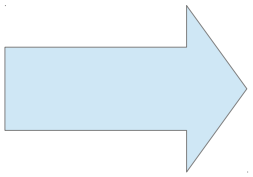
（1）解析対象の現象を確率モデル　P で表現する



パラメトリックモデル
でないとダメ

（2）サンプル D を集める

（3）P(D) を最大にするパラメータを求める



P が完成
解析可能になる

多項分布の最尤推定

起こり得る結果は $Y_1, Y_2, \cdots, Y_V$ のどれか

Y_i が起こる確率 θ_i ← こいつを知りたい

N 回の実験で、それぞれの起こった回数は y_i

$$\sum_{i=1}^{V} y_i = N$$

$$\theta_i = \frac{y_i}{N}$$

N（可変）回の実験を M回

N 回の実験を1セットで、この実験を M 回行う
各実験での N は可変

M 回の実験中の j 回目の実験を D_j

その時のN を n_j

その時のy_i を $y_{j,i}$

最尤法

$$P(allD) = \prod_{j=1}^{M} P(D_j)$$

$$P(allD) = \prod_{j=1}^{M} P(D_j)$$

$$= \prod_{j=1}^{M} P(N = n_j) P(D_j | N = n_j)$$

$$f(\theta) = \sum_{j=1}^{M} \{\log P(N = n_j) + \log P(D_j | N = n_j)\}$$

Θには無関係

クラスには関係あるかもしれないので注意

$$\propto \sum_{j=1}^{M} \log P(D_j | N = n_j)$$

$$f(\theta) = \sum_{j=1}^{M} \log P(D_j | N = n_j)$$

$$= \sum_{j=1}^{M} \log \left(\frac{n_j!}{y_{j,1}! y_{j,2}! \cdots y_{j,V}!} \theta_1^{y_{j,1}} \theta_2^{y_{j,2}} \cdots \theta_V^{y_{j,V}} \right)$$

$$= C + \sum_{j=1}^{M} \sum_{i=1}^{V} y_{j,i} \log \theta_i$$

結局、以下の最大化

$$f(\theta) = \sum_{j=1}^{M} \sum_{i=1}^{V} y_{j,i} \log \theta_i - \lambda (\sum_{i=1}^{V} \theta_i - 1)$$

$$\frac{\partial f}{\partial \theta_i} = \sum_{j=1}^{M} \frac{y_{j,i}}{\theta_i} - \lambda = 0$$

$$\lambda \theta_i = \sum_{j=1}^{M} y_{j,i}$$ 両辺を i で総和を取る

$$\lambda = \sum_{j=1}^{M} \sum_{i=1}^{V} y_{j,i} = \sum_{j=1}^{M} n_j = N$$ 総実験回数（総頻度）

$$\theta_i = \frac{\sum_{j=1}^{M} y_{j,i}}{N}$$ ← 各実験セット j での y_i の起こった頻度の総和

文書セットのモデル

先ほどのモデルが文書セットのモデル

文書セット $D = \{D_1, D_2, \cdots, C_M\}$

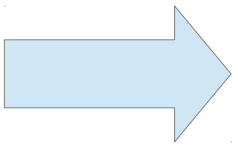
単語セット $W = \{w_1, w_2, \cdots, w_V\}$

D_j 内の w_i の頻度が $y_{j,i}$

θ_i は単語 w_i が現れる確率

スムージング

文書セットが小さいと θ_i の推定が不確か

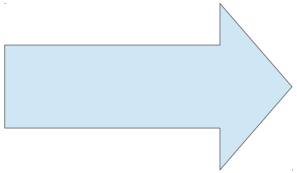


スムージングというテクニックがある

現在はベイズ流で推定するので、スムージングが
自然に実現される

ベイズのアプローチ

$$P(allD) = \prod_{j=1}^{M} P(D_j)$$



$$P(allD) = P(\theta) \prod_{j=1}^{M} P(D_j|\theta)$$

事前分布を導入

$$\log P(allD) = \log P(\theta) + \sum_{j=1}^{M} \log P(D_j|\theta)$$

ディリクレ分布

多項分布の共役事前分布

$$f(\theta;\alpha) = \frac{\prod_{i=1}^{V} \theta_i^{\alpha_i - 1}}{Z(\alpha)}$$

これは密度関数、確率ではないことに注意

$$\log P(\theta) = \log \left(\prod_{i=1}^{V} \theta_i^{\alpha_i - 1} \delta \right) - \log Z_\alpha$$

$$= \log \left(\prod_{i=1}^{V} \theta_i^{\alpha_i - 1} \right) + V \log \delta - \log Z_\alpha$$

$$= \sum_{i=1}^{V} (\alpha_i - 1) \log \theta_i + C$$

$$f(\theta) = \sum_{i=1}^{V} (\alpha_i - 1) \log \theta_i + \sum_{j=1}^{M} \sum_{i=1}^{V} y_{j,i} \log \theta_i$$

$$-\lambda(\sum_{i=1}^{V} \theta_i - 1)$$

こいつの最大化が答え

$$\frac{\partial f}{\partial \theta_i} = \frac{\alpha_i - 1}{\theta_i} + \sum_{j=1}^{M} \frac{y_{j,i}}{\theta_i} - \lambda = 0$$

$$\lambda \theta_i = \alpha_i - 1 + \sum_{j=1}^{M} y_{j,i}$$

$$\lambda = \sum_{i=1}^{V} \alpha_i - V + \sum_{i=1}^{V} \sum_{j=1}^{M} y_{j,i} = \sum_{i=1}^{V} \alpha_i - V + N$$

$\alpha_i = 2$ とすると $\lambda = V + N$

$$\theta_i = \frac{1 + \sum_{j=1}^{M} y_{j,i}}{V + N}$$

お馴染みの式

クラスのある問題

文書のトピック t

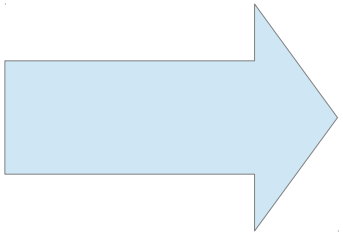
$$\theta = (\theta_{t,1}, \theta_{t,2}, \cdots, \theta_{t,V})$$

解き方は同じ、トピックが t の文書を集めて、
それを D とすればよい
つまり各トピックごとに推定すればよい

トピックモデル

解き方は同じ、トピックが t の文書を集めて、
それを D とすればよい
つまり各トピックごとに推定すればよい

トピック t が不明の場合は？



トピックモデル　PLSI と LDA

文書クラスタリングも可