多項分布の最尤推定と MAP 推定

Yoshihiko Suhara

2011-07-13

普段，無意識のうちに使っている多項分布のパラメータ推定方法の証明と，よく出てくる Dirichlet スムージングが実は多項分布の MAP 推定だったということを解説.

事前知識として，最尤推定，MAP 推定，Lagrange 乗数法くらいの知識が必要かも.

1 多項分布

多項分布の復習．多項分布の確率密度関数は,

$$P(x_1, x_2, \ldots, x_K, n; \theta_1, \theta_2, \ldots, \theta_K) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{j=1}^{K} \theta_j^{x_j}$$

ただし,

$$\sum_{i=1}^{K} \theta_i = 1$$

で与えられる．$\frac{n!}{\prod_{i=1}^{K} x_i!}$ は，多項係数で，長さ n の順列において，それぞれの状態の回数が $x_1, x_2, \ldots, x_K$ という場合の並び方の総数を表している.

2 最尤推定

まず，多項分布のパラメータの最尤推定から.

流れとしては，ふつうの最尤推定パラメータを求める方法と同じ.

- 多項分布の尤度関数を作る
- 尤度関数の対数を取る
- パラメータで偏微分して極値を求める

こんな感じ.

K 種類の単語の出現確率を多項分布でモデル化する場合の例で説明する．L 個のデータ $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$ (ただし，$\mathbf{x}_1 = \{x_{11}, x_{12}, \ldots, x_{1K}\}$) が与えられたとする．ここで，$x_{11}$ は，1 番目の文書における単語 w_1 の出現回数，x_{12} は，同じ文書における単語 w_2 の出現回数とする.

すると，このデータが与えられた場合の尤度は,

$$P(\mathbf{X}|\theta) = \prod_{i=1}^{L} \frac{N_i!}{\prod_{j=1}^{K} x_{ij}!} \prod_{k=1}^{K} \theta^{x_{ik}}$$

となる．ここで N_i は，i 番目の文書に含まれる単語の総数を表している．尤度の対数を取ると，

$$\log P(\mathbf{X}|\theta) = \sum_{i=1}^{L} \left(\log N_i! - \sum_{j=1}^{K} \log x_{ij}! + \sum_{k=1}^{K} x_{ik} \log \theta_k \right)$$

となる．ここで，log は凹関数，かつ $x_{ik} \geq 0 \ \forall i, j$ なので，上述の対数尤度は，θ_k に関して凹関数であることがわかる．したがって，最大値を見つけるためには極値を求めればよい．

多項分布のパラメータの制約 $\sum_{i=1}^{K} \theta_i = 1$ があるため，Lagrange 乗数法を用いて極値を求める．等式制約 $\sum_{i=1}^{K} \theta_i - 1 = 0$ を含んだラグランジュ関数は，

$$L = \log P(\mathbf{X}|\theta) + \lambda \left(\sum_{i=1}^{K} \theta_i - 1 \right)$$

すなわち，

$$L = \sum_{i=1}^{L} \left(\log N_i! - \sum_{j=1}^{K} \log x_{ij}! + \sum_{k=1}^{K} x_{ik} \log \theta_k \right) + \lambda \left(\sum_{i=1}^{K} \theta_i - 1 \right) \tag{1}$$

となる．ここで λ は，等式制約に対する Lagrange 乗数である．あとは，式 (1) をパラメータ θ_k と λ で偏微分して 0 とおいて方程式を解けばよい．

$$\frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{L} x_{ik} \frac{1}{\theta_k} + \lambda = 0$$

$$-\frac{1}{\lambda} \sum_{i=1}^{L} x_{ik} = \theta_k \tag{2}$$

式 (2) のままでは解くことができないので，少し工夫をする．上述の例では θ_k に関する偏微分を求めたが，全ての k について偏微分を求めて 0 とおいた式を足し合わせると，

$$-\frac{1}{\lambda} \sum_{i=1}^{L} \sum_{k=1}^{K} x_{ik} = \sum_{k=1}^{K} \theta_k$$

となる．ここで，左辺の $\sum_{i=1}^{L} \sum_{k=1}^{K} x_{ik}$ は，全文書に含まれる単語の総数を表している．これを N とおく．右辺については，$\sum_{k=1}^{K} \theta_k = 1$ より，1 となるため，

$$-\frac{1}{\lambda} N = 1$$

$$\lambda = -N$$

となる．

さて，これを式 (2) に代入すると，

$$\sum_{i=1}^{L} x_{ik} \frac{1}{\theta_k} = N$$

したがって，多項分布のパラメータの最尤推定値は，

$$\theta_k = \frac{\sum_{i=1}^{L} x_{ik}}{N}$$

であることがわかる．

3 MAP 推定

MAP 推定値も最尤推定の証明と同じ流れで導出できる．

まず，パラメータ θ の MAP 推定は，

$$\underset{\theta}{\operatorname{argmax}}\, P(\theta|\mathbf{X}) = \underset{\theta}{\operatorname{argmax}}\, P(\mathbf{X}|\theta)P(\theta)$$

で求めることができる．$P(\mathbf{X}|\theta)$ は，さきほど説明した尤度，$P(\theta)$ は，パラメータの事前確率である．多項分布の共役事前分布は Dirichlet 分布なので，$P(\theta)$ を Dirichlet 分布で

$$P(\theta) = \frac{1}{Z}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

と表現できる．ここで Z は

$$Z = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}$$

で表現されるベータ関数であるが，θ_k に依存しておらず，後ほど行う偏微分にも影響を与えないため，Z のまま記述する．

したがって，$P(\mathbf{X}|\theta)P(\theta)$ は，

$$P(\mathbf{X}|\theta)P(\theta) = \prod_{i=1}^{L}\frac{N_i!}{\prod_{j=1}^{K}x_{ij}!}\prod_{k=1}^{K}\theta^{x_{ik}}\frac{1}{Z}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

となり，これの対数を取ると，

$$\log P(\mathbf{X}|\theta)P(\theta) = \sum_{i=1}^{L}\left(\log N_i! - \sum_{j=1}^{K}\log x_{ij}! + \sum_{k=1}^{K}x_{ik}\log\theta_k\right) + \left(-\log Z + \sum_{k=1}^{K}(\alpha_k-1)\log\theta_k\right)$$

となる[*1]

さきほどと同様に，等式制約 $\sum_{i=1}^{K}\theta_i - 1 = 0$ の下での極値を求める．等式制約と Lagrange 乗数を加えた Lagrange 関数を θ_k で偏微分すると，

$$\sum_{i=1}^{L}\left(x_{ik}\frac{1}{\theta_k}\right) + \frac{\alpha_k-1}{\theta_k} + \lambda = 0$$

$$-\frac{1}{\lambda}\left\{\sum_{i=1}^{L}x_{ik} + (\alpha_k-1)\right\} = \theta_k \tag{3}$$

ここで，先ほどと同じ方法で全ての k について式 (3) を足し合わせると，

$$-\frac{1}{\lambda}\left\{\sum_{i=1}^{L}\sum_{k=1}^{K}x_{ik} + \sum_{k=1}^{K}(\alpha_k-1)\right\} = \sum_{k=1}^{K}\theta_k$$

[*1] (疑問) 凹関数であるためには，$alpha_k - 1 \geq 0\ \forall k$ の必要がある．Dirichlet 分布的には $\alpha_k > 0$ なので，$1 > \alpha_k > 0$ のとき，凹関数でなくなってしまうので，$\alpha_k \geq 1$ という制約にする必要がある．

左辺の $\sum_{i=1}^{L}\sum_{k=1}^{K} x_{ik}$ は，先ほどと同様に全文書の単語総数 N となり，右辺 $\sum_{k=1}^{K}\theta_k$ は 1 となる．これより，

$$-\frac{1}{\lambda}\left\{N+\sum_{k=1}^{K}(\alpha_k-1)\right\}=1$$

$$N+\sum_{k=1}^{K}(\alpha_k-1)=-\lambda$$

が得られる．

これを式 (3) に代入し，多項分布のパラメータの MAP 推定値

$$\theta_k=\frac{\sum_{i=1}^{L} x_{ik}+(\alpha_k-1)}{N+\sum_{j=1}^{K}(\alpha_j-1)} \tag{4}$$

を得ることができる．

$\alpha_k=2\ \forall k$ とした場合には，

$$\theta_k=\frac{\sum_{i=1}^{L} x_{ik}+1}{N+K} \tag{5}$$

となり，Laplace スムージングと等しくなることがわかる．この場合には，それぞれのパラメータが α_k-1 回 (すなわち 1 回) 観測されたとする Dirichlet 分布によって事前知識を与えていると解釈することができる．