

オンライン学習におけるディリクレ過程

大阪大・基礎工 水間 浩太郎

大阪大・基礎工 濱田 悦生

1 はじめに

混合モデルにおける推測において利用されている代表的なアルゴリズムは、EM アルゴリズムやマルコフ連鎖モンテカルロ法 (MCMC 法) などの計算機を利用した手法である。特にベイズ統計学を利用して用いられる手法には、いくつかの問題点や考慮すべき点が存在する。まず、混合モデルではクラスターに関するパラメータの対称性があるため、MCMC を利用する際に工夫が必要である (Stephens, 2000)。また、混合モデルにおけるクラスター数を予め指定する必要がある (Escobar and West, 1995)。この下で本研究では、クラスター数を自動的に決めることができるノンパラメトリックベイズを利用して、MCMC を利用しないような高速な手法を提案する。

2 提案手法

本研究では、ノンパラメトリックベイズ統計学をパラメトリックな尤度のパラメータに利用するディリクレ過程混合 (DPM) モデルを利用する。サイズが N であるデータ $(X_i)_{i=1}^N$ が $N(\mu_i, \sigma_i^2)$ に従うとし、パラメータ $\theta_i = (\mu_i, \sigma_i^2)$ が $\text{Dir}(\alpha F_0)$ に従うとする。ここで α は事前分布の強度で F_0 は期待値測度とする。なお、ディリクレ過程はランダム測度を周辺化することで次のような表現ができる:

$$\theta_i | \theta_1, \dots, \theta_{i-1} \sim \frac{\alpha}{\alpha + i - 1} F_0 + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}.$$

今回はクラスターに関心があるものとして、 θ_i ではなくクラスター添字 γ_i を利用して推定する:

$$P\{\gamma_i = h | \gamma_1, \dots, \gamma_{i-1}\} = \begin{cases} \frac{\sum_{j=1}^{i-1} 1_{\gamma_j = h}}{\alpha + i - 1}, & h = 1, \dots, k_{i-1}, \\ \frac{\alpha}{\alpha + i - 1}, & h = k_{i-1} + 1. \end{cases}$$

なお、 k_{i-1} は $i-1$ までのデータのクラスター数とする。クラスターによる推定を行うことで、より正確な推定が可能になり、ディリクレ過程を用いることでクラスター数についても自動的に決定することが可能になる。

3 シミュレーション

本モデルのシミュレーション結果は当日報告を行う。

参考文献

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.