1 EM Algorithm

EM アルゴリズムは，潜在変数を持つ確率モデルの最尤解を求めるための一般的手法である．変分推論の基礎となる部分でも用いられる．

パラメータ最尤推定の目的は，尤度関数の最大化である．

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{1.1}$$

ここで，完全データ対数尤度 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を定義する．EM アルゴリズムを用いるのは $p(\boldsymbol{X}|\boldsymbol{\theta})$ を直接最大化するのが難しいが，完全データ大数尤度を最大化するのは容易な場合である．ここで，潜在変数 Z についての分布 $q(z)$ を導入する．これを用いて，以下の分解が成り立つ

$$\ln p(\boldsymbol{X}|\theta) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p)$$

$$\mathcal{L}(q, \theta) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{1.2}$$

$$KL(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{1.3}$$

このとき，$KL \geq 0$ なので，$\mathcal{L}(q, \theta)$ は常に尤度関数の下界となる．EM アルゴリズムでは，パラメータ $\boldsymbol{\theta}$ を固定して KL 距離を 0 にする $q(\boldsymbol{Z})$ を選ぶ E-step, $q(\boldsymbol{Z})$ を固定して $\mathcal{L}$ を最大化する M-step を繰り返す．

E-step において，KL 距離が 0 になる $q(\boldsymbol{Z})$ を選ぶが，これは $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ のときに成り立つ．このときの $q(\boldsymbol{Z})$ を $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ とする．M-step ではこの分布を用いて $\mathcal{L}$ を最大化し，新たなパラメータ $\boldsymbol{\theta}$ を求める．

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old}) \tag{1.4}$$

$$= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) + const \tag{1.5}$$

ここで，定数項を除いた Q 関数を定義した．この Q 関数を最大化する $\boldsymbol{\theta}$ が θ^{new} となる．

E-step は Expectation-step の略であるが，これは Q 関数をの定義が完全データ対数尤度の期待値となっているためである．また，M-step は maximization-step であり，その期待値をパラメータ $\boldsymbol{\theta}$ に関して最大化する．これらをまとめると，いか EM アルゴリズムは以下のように表せる．

$$E : Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = E_Z[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})] \tag{1.6}$$

$$M : \quad \boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{1.7}$$

2 Variational Bayes

Variational Bayes(VB，変分ベイズ) は，正確に求めることが難しい潜在変数の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を近似する枠組みである．同様の目的の手法として Sampling 法があるが，計算量の点で VB が有利である．

EM アルゴリズムと同様に，我々の目的は事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ およびモデルのエビデンス $p(X)$ を求めることである．なお，EM アルゴリズムではパラメータ集合 $\boldsymbol{\theta}$ を定義したが，ここでは確率変数として $\boldsymbol{Z}$ に含まれていることにする．

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p) \tag{2.1}$$

$$\mathcal{L}(q) = \int_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} \tag{2.2}$$

$$KL(q||p) = -\int_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} \tag{2.3}$$

EM アルゴリズムの場合と同様に，KL 距離を最小化する事により最適化を行いたいが，ここでは真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を求めるのが不可能であると仮定する．したがって，代わりにある制限したクラスの $q(\boldsymbol{Z})$ を考え，この中での最適化を考える事になる．

$\boldsymbol{Z}$ に関して，$\boldsymbol{Z}$ の要素をいくつかのグループに分解し，それらが独立であることを仮定する．

$$q(\boldsymbol{Z}) = \prod_{i}^{M} \boldsymbol{q}_i(Z_i) \tag{2.4}$$

この仮定のもとで，エビデンスの下界 $\mathcal{L}$ を最大化することを考える．したがって，$q(\boldsymbol{Z})$ の各因子それぞれに関して，順番に最適化を行っていく．記法を簡単にするために $q_i(Z_i)$ を q_i とし，$\mathcal{L}$ を q_i の関数とみなす．

$$\begin{aligned}
\mathcal{L}(q_j) &= \int_{\boldsymbol{Z}} \prod_{i}^{M} q_i \left\{ \ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_{i}^{M} \ln q_i \right\} d\boldsymbol{Z} \\
&= \int q_j \left\{ \int \prod_{i \neq j} q_i \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}_i \right\} d\boldsymbol{Z}_j - \int \prod_i q_i \sum_i \ln q_i d\boldsymbol{Z} + const \\
&= \int q_j \left\{ \int \prod_{i \neq j} q_i \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}_i \right\} d\boldsymbol{Z}_j - \sum_i \int \prod_i q_i \ln q_i d\boldsymbol{Z} + const \\
&= \int q_j \left\{ \int \prod_{i \neq j} q_i \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}_i \right\} d\boldsymbol{Z}_j - \int q_j \ln q_j d\boldsymbol{Z}_j + const \\
&= \int q_j \ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j - \int q_j \ln q_j d\boldsymbol{Z}_j + const
\end{aligned} \tag{2.5}$$

ここで，新しい分布

$$\tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = E_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + const \tag{2.6}$$

を定義した．(2.5) の最大化を考えるが，これは q_i と $\tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ の負の KL 距離となっていることが分かる．よって，

$$\ln q_i^* = E_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + const \tag{2.7}$$

のときに $\mathcal{L}$ は最大になる．なお，定数項を得るためには分布を正規化すれば良い．