

講習会資料 Sample

Column ～ 対数尤度関数の最大化と KL-divergence ～

EM アルゴリズムで対数尤度を最大化する際、各ステップで対数尤度関数が増加 (もしくは不変) することを EM アルゴリズムを一般的に取り扱うことで示すことができる。パラメータをまとめて θ として表すと、最大化する対数尤度関数は、

$$\begin{aligned}\ln p(X|\theta) &= \ln \sum_Z p(X, Z|\theta) \\ &= \ln \sum_Z q(Z) \frac{p(X, Z|\theta)}{q(Z)} \\ &\geq \sum_Z q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)}\end{aligned}$$

ここで $q(Z)$ は潜在変数 Z についての分布である。なお、3 行目の変形ではイェンセン不等式、

$$\ln E[x] \geq E[\ln x]$$

を用いた。最終的に求められた式、

$$L(q, \theta) = \sum_Z q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)}$$

には変分下界という名前がつけられている。 $(L(q, \theta))$ は θ の関数、 $q(Z)$ の汎関数となっている。汎関数については次コラム参照。) 次に対数尤度関数 $\ln p(X|\theta)$ とこの変分下界 $L(q, \theta)$ の差について考えると、

$$\begin{aligned}\ln p(X|\theta) - L(q, \theta) &= \dots\dots \\ &= \dots\dots \\ &= - \sum_Z q(Z) \ln \frac{p(Z|X, \theta)}{q(Z)} \\ &= KL[q(Z)||p(Z|X, \theta)]\end{aligned}$$

となる。よって対数尤度関数は、

$$\ln p(X|\theta) = L(q, \theta) + KL[q(Z)||p(Z|X, \theta)]$$

と変形することができることがわかる。これを図式化したものを 図 1 に示す。

この対数尤度関数の最尤解を求めるためには以下のような方法をとると、常に $\ln p(X|\theta)$ が増加するように最適化を行うことができる。

1. $L(q, \theta)$ を $q(Z)$ について最大化する。(E ステップ)
2. $L(q, \theta)$ を θ について最大化する。(M ステップ)
3. 1 に戻る

現在のパラメータ θ の値を $\theta = \theta^{old}$ としてそれぞれのステップについて詳しく見ていく。

まず 1 のステップ (E ステップ) では $q(Z)$ についての最大化を行う。その際、左辺の $\ln p(X|\theta)$ は $q(Z)$ に依存しないので、このステップにおいて不変である。このことから $L(q, \theta)$ の最大化は、 $KL[q(Z)||p(Z|X, \theta)]$ の最小化と等価になり、 $L(q, \theta)$ の最大値は KL ダイバージェンスが 0 になるとき、つまり、

$$q(Z) = p(Z|X, \theta^{old})$$

であることがわかる。またこのとき当然ながら変分下界は対数尤度と一致する。(図 2)

次に 2 のステップ (M ステップ) について考える。1 のステップが終わった段階で、 $KL[q(Z)||p(Z|X, \theta)]$ 、 $L(q, \theta)$ の値は、

$$\begin{aligned} KL[q(Z)||p(Z|X, \theta)] &= 0 \\ L(q, \theta) &= \sum_Z q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)} \\ &= \sum_Z p(Z|X, \theta^{old}) \ln \frac{p(X, Z|\theta)}{p(Z|X, \theta^{old})} \\ &= \sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{old}) \ln p(Z|X, \theta^{old}) \\ &= Q(\theta, \theta^{old}) + \text{const.} \end{aligned}$$

となる。なお、第 2 項は θ には依存しないので定数扱いしている。よって $L(q, \theta)$ を最大化するには、 $Q(\theta, \theta^{old})$ を最大化すればよい。また、KL ダイバージェンスについても、KL ダイバージェンス自体が正の値しか取り得ないことと、E ステップが終わった時点で 0 となっていることを考えれば、M ステップにおいてどんな $\theta = \theta^{new}$ をとったところで必ず増加することがわかる。(図 3)

以上より、このような方法で対数尤度 $\ln p(X|\theta)$ を最大化すれば、パラメータ更新手続きは常に対数尤度を増加させることがわかる。(この更新手続きを視覚化したのが図 4 である。)

■補足

M ステップにおいて、 $L(q, \theta)$ を最大化させるのに $Q(\theta, \theta^{old})$ を最大化させたが、これを実際に行う際には、そもそも $Q(\theta, \theta^{old})$ が最適化可能であるという前提が必要となる。 $Q(\theta, \theta^{old})$ に現れる同時分布 $p(X, Z|\theta)$ の尤度は、「完全データ対数尤度関数」と呼ばれており、上記の式変形は、これが十分に簡単に最大化することができることを仮定したものである。(例えば、 $p(X, Z|\theta)$ が指数型分布族である場合は、対数と指数がキャンセルして $\ln p(X, Z|\theta)$ の最大化は用意になる。)