EM アルゴリズム

jojonki

Mar. 2020

1 はじめに

北先生の言語と計算 (4) 確率的言語モデルや高村先生の言語処理のための機械学習入門を読んで EM アルゴリズムを学習したので自分用にメモする．どちらも初学者に良い本なので超オススメである．

基本的の今回のメモは北先生の本を辿っていくのだが，式を若干であるが丁寧に展開して説明する．

誤り箇所があれば Twitter 等で@jojonki までメンションを投げてもらえると助かる．

2 EM アルゴリズム

単純なモデルである場合は，観測データに対してモデルをできるだけフィットさせる最尤推定法で解ける．例えば観測値 $X_1, ..., X_N$ が未知パラメタ θ を含む確率分布 $P_\theta(X)$ から抽出された標本だとする．各 X_i は独立に抽出された標本だとすると，最尤推定では尤度関数 $L(\theta) = \prod_{i=1}^{N} P_\theta(X_i)$ を最大化するように解く．

ただし HMM などでは，観測データを生成したモデルの内部状態を一意的に決定することはできないため，直接最尤推定法を適用することができない．そんなときは EM アルゴリズムの登場である．EM アルゴリズムは観測した不完全データから，未知パラメタ θ を推測するアルゴリズムである．北先生の本に従い，EM アルゴリズムの考え方を最初に説明する．

2.1 EM アルゴリズムの考え方

簡単な混合モデル（Mixture Model）で説明する．例えばある言語モデル $P(x)$ を考えたとき，このモデルは，２つの言語モデル $P_A(x)$ と $P_B(x)$ から確率的に選ばれて構成されるとする．

$$P(x) = \lambda P_A(x) + (1 - \lambda) P_B(x) \tag{1}$$

このとき，$x_1, ..., x_N$ の生成において，P_A の選ばれた回数が分かるとすると，$\lambda = P_A$が選ばれた回数$/N$ と単純である．ただしどちらが選ばれたかなどはモデルの内部状態に隠れており，観測データ $x_1, ..., x_N$ からは決定できない．このような状態の時の観測データを不完全データ (incomplete data) と呼ぶ．

EM アルゴリズムでは，このようなときに未知パラメタ λ にまずは適当な値を設定して，最尤推定により各観測データ x_i の生成にモデル P_A が使われた回数の期待値 $E_\lambda[A|x_i]$ を求める．

$$\begin{aligned} E_\lambda[A|x_i] &= \frac{\text{モデル } P_A\text{により } x_i\text{が生成された確率}}{\text{いずれかのモデルにより } x_i\text{が生成された確率}} \\ &= \frac{\lambda P_A(x_i)}{\lambda P_A(x_i) + (1 - \lambda) P_B(x_i)} \end{aligned} \tag{2}$$

よって観測データ全体から P_A が選ばれた期待値は，この式の和で表されるので，パラメタ λ を下記のように更新できる．

$$\bar{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda P_A(x_i)}{\lambda P_A(x_i) + (1 - \lambda) P_B(x_i)} \tag{3}$$

そして更新された $\bar{\lambda}$ で，式 2 を計算し（Expectation ステップ），式 3 でまた λ を更新する（Maximization ステップ）．これをパラメタが収束するまで繰り返す．

つまり，E ステップでは，各観測事例がそれぞれのクラス（今回は A か B）にどの程度属しているかを計算している．M ステップでは，得られた期待値をもとにパラメタを更新している．

EM アルゴリズムでは，この E ステップと M ステップを繰り返し行い，不完全データに対して尤度が大きくなるようにパラメタを決定する枠組みである

2.2 EM アルゴリズム

先程は 2 つの言語モデルが未知の割合で混合された言語モデルを考えたが，より一般的な状況について述べる．再び不完全データの観測データ $x_1, ..., x_N$ を考え，未知のパラメタ θ を求めるとする．EM アルゴリズムではこのような状況下においても，モデルの対数尤度を増加させる方向にパラメタを推定していく方法である．すでに計算したパラメタ θ (θ の初期値は適当) からモデルの対数尤度を増加させるパラメタ $\bar{\theta}$ を考えたときの，対数尤度の差を求めてみる．

$$
\begin{aligned}
\log P_{\bar{\theta}}(x_i) - \log P_{\theta}(x_i) &= \log \frac{P_{\bar{\theta}(x_i)}}{P_{\theta}(x_i)} \\
&\quad y \text{ の周辺化} \\
&= \sum_y P_{\theta}(y|x_i) \log \frac{P_{\bar{\theta}(x_i)}}{P_{\theta}(x_i)} \\
&\quad P_{\bar{\theta}} \text{ と } P_{\theta} \text{ をベイズの定理で展開} \\
&= \sum_y P_{\theta}(y|x_i) \log \left[\frac{P_{\bar{\theta}}(x_i, y)}{P_{\bar{\theta}}(y|x_i)} \frac{P_{\theta}(y|x_i)}{P_{\theta}(x_i, y)} \right] \\
&= \sum_y P_{\theta}(y|x_i) \log \left[\frac{P_{\bar{\theta}}(x_i, y)}{P_{\theta}(x_i, y)} \frac{P_{\theta}(y|x_i)}{P_{\bar{\theta}}(y|x_i)} \right] \\
&= \sum_y P_{\theta}(y|x_i) \log \frac{P_{\bar{\theta}}(x_i, y)}{P_{\theta}(x_i, y)} + \sum_y P_{\theta}(y|x_i) \log \frac{P_{\theta}(y|x_i)}{P_{\bar{\theta}}(y|x_i)} \\
&\quad \text{ここで } Q(\theta, \bar{\theta}) = \sum_y P_{\theta}(y|x_i) \log P_{\bar{\theta}}(x_i, y) \text{ とおくと，} \\
&= Q(\theta, \bar{\theta}) - Q(\theta, \theta) + \sum_y P_{\theta}(y|x_i) \log \frac{P_{\theta}(y|x_i)}{P_{\bar{\theta}}(y|x_i)} \\
&\quad \text{第 2 項は非負であるので（後述），} \\
&\geq Q(\theta, \bar{\theta}) - Q(\theta, \theta)
\end{aligned}
\tag{4}
$$

右辺は対数尤度の差に対する下限となっており，$Q(\theta, \bar{\theta}) > Q(\theta, \theta)$ となるような $\bar{\theta}$ を見つけることで，x_i に対する対数尤度を増加させることができる．つまり $Q(\theta, \bar{\theta})$ を最大にするような $\bar{\theta}$ を求めれば良い．

整理すると，EM をアルゴリズムはまず θ に適当なパラメタを与えたあとに，E ステップで $Q(\theta, \bar{\theta})$ を計算し，M ステップで Q 関数を最大化する θ を新しいパラメタ $\bar{\theta}$ として更新．この E ステップと M ステップをパラメタ θ が収束するまで更新する．E ステップを Q 関数の計算と書いたが，式 2 で示したような事例のクラスに属する割合の期待値計算をすることで Q 関数の値が求まるので，Q 関数の計算を E ステップと言っても問題はない．

ところで，先程の事例は 1 事例 x_i に対する Q 関数であったが，複数の観測データに適用するためには，下記のように定義するだけで良い．

$$
Q(\theta, \bar{\theta}) = \sum_{i=1}^{N} \sum_y P_{\theta}(y|x_i) \log P_{\bar{\theta}}(x_i, y) \tag{5}
$$

ところで式 4 で，第 2 項は非負といった証明をする．そのためには確率分布 $P(x), Q(x)$ を考え，$\log x \leq x - 1$

の性質を利用する．

$$\begin{aligned}\sum_x P(x)\log\frac{Q(x)}{P(x)} &\le \sum_x P(x)\Big(\frac{Q(x)}{P(x)}-1\Big)\\ &= \sum_x Q(x) - \sum_x P(x)\\ &= 1-1=0\end{aligned}$$

この式を ≥ 0 の形にするためマイナスをかけると対数の分母分子が入れ替わるため，

$\sum_x P(x)\log\frac{P(x)}{Q(x)} \ge 0$ が導ける．（証明終わり）

(6)

2.3 混合モデルの推定

さて最初に 2 つの言語モデルを例に EM アルゴリズムの考え方を説明し，次に EM アルゴリズムの一般的な定式化を述べた．ここで再び M 個の混合モデル $P = P_1, ..., P_M$ を考え，EM アルゴリズムを適用して，不完全データ $x_1, ..., x_N$ から，M 個の混合モデルがどのような割合 λ で混ぜ合わさっているか推定しよう．

$$P(x) = \sum_{j=1}^{M} \lambda_j P_j(x),\ (\text{ただし} \sum_{j=1}^{M} \lambda_j = 1) \tag{7}$$

ここで，$P_{\bar{\lambda}}(x_i, j) = \lambda_j P_j(x)$ とおき，Q 関数を下記のように定義する．

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^{N}\sum_{j=1}^{M} P_\lambda(j|x_i)\log P_{\bar{\lambda}}(x_i, j) \tag{8}$$

ここでラグランジュの未定乗数法を用いて Q 関数を最大化するような $\bar{\lambda}$ を求める．制約条件として $\sum_j \bar{\lambda} = 1$ があり，ラグランジュ定数を γ として，ラグランジュ関数は下記になる．

$$\begin{aligned}\mathcal{L}(\bar{\lambda}, \gamma) &= Q(\lambda,) - \gamma\left[\sum_{j=1}^{M}\bar{\lambda}_j - 1\right]\\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} P_\lambda(j|x_i)\log P_{\bar{\lambda}}(x_i, j) - \gamma\left[\sum_{j=1}^{M}\bar{\lambda}_j - 1\right]\end{aligned} \tag{9}$$

これを $\bar{\lambda}_j$ で偏微分して 0 とおいて解いていく，

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\bar{\lambda}_j} &= \sum_{i=1}^{N} P_\lambda(j|x_i)\frac{1}{\bar{\lambda}_j} - \gamma = 0\\ \bar{\lambda}_j &= \frac{1}{\gamma}\sum_{i=1}^{N} P_\lambda(j|x_i)\end{aligned} \tag{10}$$

制約条件を思い出して，

$$\begin{aligned}
\sum_{j=1}^{M} \bar{\lambda}_j = 1 &= \sum_{j=1}^{M} \frac{1}{\gamma} \sum_{i=1}^{N} P_\lambda(j|x_i) \\
&= \frac{1}{\gamma} \sum_{j=1}^{M} \sum_{i=1}^{N} P_\lambda(j|x_i) \\
&= \frac{1}{\gamma} \sum_{i=1}^{N} \sum_{j=1}^{M} P_\lambda(j|x_i) \\
&\quad \text{周辺化} \\
&= \frac{1}{\gamma} \sum_{i=1}^{N} 1 \\
&= \frac{1}{\gamma} N \\
&\quad \text{よって} \\
\gamma &= N
\end{aligned} \tag{11}$$

これを $\bar{\lambda}_j$ の式 10 に戻せば，

$$\bar{\lambda}_j = \frac{1}{N} \sum_{i=1}^{N} P_\lambda(j|x_i) \tag{12}$$

ところで $P_\lambda(j|x_i)$ は下記のように計算できる．これは E ステップに該当する．

$$\begin{aligned}
P_\lambda(j|x_i) &= \frac{P_\lambda(x_i, j)}{P_\lambda(x_i)} \\
&= \frac{\lambda_j P_j(x_i)}{\sum_{j=1}^{M} \lambda_j P_j(x_i)}
\end{aligned} \tag{13}$$

これを先程の式 12 に代入すれば，λ_j の更新式が完成する．これは M ステップに該当する．

$$\bar{\lambda}_j = \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda_j P_j(x_i)}{\sum_{j=1}^{M} \lambda_j P_j(x_i)} \tag{14}$$

3 まとめ

複雑なモデルが確率分布の混合モデルで表せるとすると，EM アルゴリズムでは観測データから，その混合度合いを，尤度を上げる方向で見つけることができる．何となくイメージできただろうか．いずれにせよ北先生の本を見て，この資料はその補助資料として利用していただければ幸いである．

4 参考情報

- 言語と計算 (4) 確率的言語モデル, 北 研二, 辻井 潤一.
- 言語処理のための機械学習入門, 高村 大也, 奥村 学.