Algorytm EM

Niech będzie dany zestaw obserwacji x i sparametryzowna rodzina rozkładów prawdopodobieństwa P_θ z gęstością p_θ. Metoda estymacji największej wiarygodności polega na maksymalizacji $p_\theta(x)$, tzn. szukamy θ dla którego $p_\theta(X)$ jest maksymalne. Jednakże bezpośrednia maksymalizacja $p_\theta(X)$ może być bardzo trudna. Algorytm EM jest iteracyjną procedurą która może być łatwiejsza od bezpośredniej maksymalizacji.

Poniżej będziemy zakładać że wszystkie potrzebne nam rozkłady prawdopodobieństwa mają gęstości, co w szczególności jest spełnione w przypadku dyskretnej przestrzeni probabilistycznej (choć dla uproszczenia notacji poniżej używamy symbole tak jakbyśmy mieli przypadek ciągły).

W algorytmie EM zakładamy że istnieje nieobserwowalna wielkość y która w pewnym sensie wyznacza x. Dokładniej, zakładamy że rozkład warunkowy (a właściwie gęstość warunkowa) x pod warunkiem y nie zależy od θ czyli oznaczając przez r_θ gęstość y, przez $q(x|y)$ gęstość warunkową x pod warunkiem y, zaś przez q_θ gęstość łączną mamy

$$q_\theta(x,y) = r_\theta(y)q(x|y).$$

Poniżej będziemy zakładać że zbiór y takich że $r_\theta(y) > 0$ nie zależy od θ. Z tego założenia wynika że zbiór par (x,y) takich że $q_\theta(x,y) > 0$ nie zależy od θ. Niech $q_\theta(y|x)$ będzie gęstością warunkową y pod warunkiem x. Mamy

$$q_\theta(x,y) = p_\theta(x)q_\theta(y|x).$$

Teraz, zakładając że $q_\theta(x,y) > 0$ (czyli również $q_{\theta_n}(x,y) > 0$) mamy $q_{\theta_n}(y|x) > 0$, $q(x|y) > 0$, $p_{\theta_n}(x) > 0$ i

$$\frac{q_\theta(x,y)}{p_{\theta_n}(x)} = \frac{q_\theta(x,y)q_{\theta_n}(y|x)}{p_{\theta_n}(x)q_{\theta_n}(y|x)} = \frac{q_\theta(x,y)}{q_{\theta_n}(x,y)}q_{\theta_n}(y|x) =$$

$$\frac{r_\theta(y)q(x|y)}{r_{\theta_n}(y)q(x|y)}q_{\theta_n}(y|x) = \frac{r_\theta(y)}{r_{\theta_n}(y)}q_{\theta_n}(y|x)$$

Z powyższego wynika że dla $p_\theta(x) > 0$ mamy

$$\frac{p_\theta(x)}{p_{\theta_n}(x)} = \int \frac{q_\theta(x,y)}{p_{\theta_n}(x)}dy = \int_{q(x,y)>0} \frac{q_\theta(x,y)}{p_{\theta_n}(x)}dy = \int \frac{r_\theta(y)}{r_{\theta_n}(y)}q_{\theta_n}(y|x)dy.$$

W szczególności z istnienia (x,y) takiego że $q_\theta(x,y) > 0$ wynika że $p_{\theta_n}(x) > 0$.

Zauważmy teraz że logarytm jest funkcją ściśle wklęsłą. Przy ustalonym x oznaczając $L(\theta) = \log(p_\theta(x))$ z nierówności Jensena mamy

$$L(\theta) - L(\theta_n) = \log(\frac{p_\theta(x)}{p_{\theta_n}(x)}) = \log(\int \frac{r_\theta(y)}{r_{\theta_n}(y)} q_{\theta_n}(y|x)dy)$$

$$\geq \int \log(\frac{r_\theta(y)}{r_{\theta_n}(y)}) q_{\theta_n}(y|x)dy = Q(\theta, \theta_n).$$

gdzie ostatnia równość jest definicją $Q(\theta, \theta_n)$. W algorytmie EM wybieramy jako θ_{n+1} takie θ które zmaksymalizuje $Q(\theta, \theta_n)$, albo przynajmniej takie by $Q(\theta_{n+1}, \theta_n) > 0$. Przy takim wyborze θ_n wartości $L(\theta_n)$ tworzą ciąg ściśle rosnący. Jeśli dla dowolnego ustalonego x gestość $p_\theta(x)$ jest ograniczona, to również $L(\theta)$ jest ograniczona i ciąg $L(\theta_n)$ jest ograniczony, a wiec jest ciągiem zbieżnym. Przy rozsądnych założeniach, np. że dla dowolnego a zbiór θ takich że $L(\theta) \geq a$ jest zwarty, zaś Q ma ciągłą pochodną wynika stąd istnienie podciągu ciągu θ_n zbiegającego do θ_∞ takiego że $Q(\theta_\infty, \theta_\infty) = 0$ daje punkt stacjonarny $Q(\theta, \theta_\infty)$ (tzn. $\nabla_\theta Q(\theta, \theta_\infty) = 0$).

Zauważmy że $Q(\theta, \theta) = L(\theta) - L(\theta) = 0$, czyli jeśli $Q(\theta, \theta_n)$ nie osiąga maksimum dla $\theta = \theta_n$ to istnieje θ takie że $Q(\theta, \theta_n) > 0$. Jak zauważyliśmy wyżej przy słabych założeniach o Q istnieje podciąg zbieżny do θ_n i pochodna $\nabla_\theta Q(\theta, \theta_n) = 0$. Czyli jeśli również L jest różniczkowalna to pochodna L w θ_∞ jest równa 0. Czyli θ_∞ jest punktem stacjonarnym L. W specjalnych przypadkach może się zdarzyć że θ_∞ to punkt siodłowy czy nawet lokalne minimum L, ale zwykle jest to maksimum lokalne.

Otrzymany wzór na $Q(\theta, \theta_n)$ można nieco przekształcić:

$$Q(\theta, \theta_n) = \int \log(\frac{r_\theta(y)}{r_{\theta_n}(y)}) q_{\theta_n}(y|x)dy =$$

$$\int \log(r_\theta(y)) q_{\theta_n}(y|x)dy - \int \log(r_{\theta_n}(y)) q_{\theta_n}(y|x)dy$$

A więc maksymalizacja $Q(\theta, \theta_n)$ jest równoważna maksymalizacji

$$\int \log(r_\theta(y)) q_{\theta_n}(y|x)dy = E_{q_{\theta_n}}(\log(r_\theta(y))|x)$$

bo drugi człon w $Q(\theta, \theta_n)$ nie zależy od θ. Ostatnia równość wyjaśnia nazwę EM: expectation maximization. Mianowicie, w najpierw wyliczamy wyrażenie na $E_{q_{\theta_n}}(\log(r_\theta(y))|x)$, jest to krok E (expectation). Następnie maksymalizujemy otrzymane wyrażenie względem θ, czyli robimy krok M (maximization).

Powyżej przedstawiliśmy algorytm EM w bardzo ogólnej postaci. Zwykle zakłada się że $y = (x, z)$ dla pewnego dyskretnego z. Oczywiście wtedy x jest jednoznacznie wyznaczone przez y, a więc nasze założenie o tym że rozkład x pod warunkiem y nie zależy od θ jest automatycznie spełnione.

W ogólnej sytuacji nie jest jasne czy maksymalizacja $Q(\theta, \theta_n)$ jest łatwiejsza od maksymalizacji $L(\theta)$. Ale w wielu przypadkach, np. modelu IBM 1 czy przy szacowaniu ukrytych modeli Markowa maksymalizacja $Q(\theta, \theta_n)$ może być łatwo przeprowadzona dokładnie. W niektórych przypadkach dokładne wzory na θ maksymalizujące $Q(\theta, \theta_n)$ są niepraktyczne, ale można je zadowalająco przybliżać.