مثالهایی از نحوهٔ استفاده از الگوریتم EM در محاسبه برآوردهای درستنمایی ماکزیمم

سید محمد ابراهیم حسینی نسب [۱]   دکتر ناصر رضا ارقامی [۲]

چکیده

گاهی اوقات، در بدست آوردن برآوردهای درستنمایی ماکزیمم، پس از مشتق‌گیری از تابع درستنمایی به معادلاتی می‌رسیم که نمی‌توان از آنها فرم بسته‌ای را برای برآوردهای درستنمایی ماکزیمم نتیجه گرفت. در این گونه موارد الگوریتم EM که یک فن تکراریست و برای اولین بار توسط Dempster و سایرین در سال ۱۹۹۷ ارایه گردید، یک راه عملی برای یافتن جوابهای درستنمایی ماکزیمم پیش پای ما می‌گذارد.

معمولاً، در منابع درسی هنگام معرفی الگوریتم EM کوششی در جهت روشن کردن منطقی که این الگوریتم بر آن استوار است و ارایه دلایلی که چرا این الگوریتم ما را به برآوردهای درستنمایی ماکزیمم می‌رساند به عمل نمی‌آید. در این مقاله سعی ما بر این است که علاوه بر توضیح موارد فوق، مثالهای ساده‌ای که به درک عمیقتر این الگوریتم کمک می‌کند ارایه نماییم.

۱ مقدمه

فرض کنید مشاهدات ما بردار Y با تابع چگالی احتمال $g(y|\underline{\theta})$ که $\underline{\theta}$ پارامتر مورد نظر است باشد. در این الگوریتم، فرض بر این است که بردار تصادفی X با تابع چگالی احتمال $f(x|\underline{\theta})$ به گونه‌ای است که Y تابعی از X مانند $h(X)$ باشد در این صورت به X داده‌های کامل و به Y داده‌های ناقص گویند. معمولاً Y را داده‌های قابل مشاهده و X را داده‌های غیر قابل مشاهده نامند.

هر تکرار الگوریتم EM، شامل دو مرحله است: مرحله E و مرحله M. در مرحلهٔ E امید ریاضی شرطی، $Q(\underline{\theta}|\underline{\theta}^{(n)}) = E[lnf(X|\underline{\theta})|y;\underline{\theta}^{(n)}]$ تشکیل و در مرحلهٔ M این امید شرطی نسبت به $\underline{\theta}$ ماکزیمم می‌شود. $\underline{\theta}^{(n)}$ در عبارت امید ریاضی شرطی نشان‌دهندهٔ مقدار برآورد بدست آمده از مرحلهٔ تکرار (n)ام و $\underline{\theta}^{(n+1)}$ مقداری از $\underline{\theta}$ است که امید ریاضی شرطی فوق را با فرض ثابت بودن $\underline{\theta}^{(n)}$ ماکزیمم می‌کند و در نتیجه

[۱] سید محمد ابراهیم حسینی نسب، دانشجوی کارشناسی ارشد آمار دانشگاه فردوسی مشهد
[۲] دکتر ناصر رضا ارقامی عضو هیأت علمی گروه آمار دانشگاه فردوسی مشهد

$\underline{\theta}^{(n+1)}$، مقدار برآورد به دست آمده از مرحلهٔ $(n+1)$ام است.

توضیح این که چرا الگوریتم EM منجر به بیشینه شدن تابع درستنمایی می‌شود، با تعریف تابع $H(\underline{\theta}|\underline{\theta}^{(n)})$ با ضابطهٔ

$$H(\underline{\theta}|\underline{\theta}^{(n)}) = E_x[\ln f(X|\underline{\theta})|y;\underline{\theta}^{(n)}] - \ln g(y|\underline{\theta})$$

تسهیل می‌یابد و نقش اساسی این الگوریتم که همان افزایش درستنمایی از مرحله‌ای به مرحلهٔ بعد است بهتر روشن می‌گردد. در این تابع مقادیر y و $\underline{\theta}^{(n)}$ معلوم فرض می‌شود، X متغیر تصادفی و $\underline{\theta}$ متغیر غیر تصادفی است. ابتدا ثابت می‌کنیم این تابع ماکزیمم مقدارش را در $\underline{\theta} = \underline{\theta}^{(n)}$ اختیار می‌کند، سپس نشان خواهیم داد که مقدار تابع درستنمایی $g(y|\underline{\theta})$ به ازای $\underline{\theta} = \underline{\theta}^{(n+1)}$ بیشتر از مقدار $g(y|\underline{\theta})$ به ازای $\underline{\theta} = \underline{\theta}^{(n)}$ است و در نتیجه در هر مرحله، مقدار تابع درستنمایی افزایش می‌یابد.

ابتدا به بیان یک لم و اثبات یک قضیه می‌پردازیم:

لم ۱ (نامساوی آنتروپی): فرض کنید f و g دو تابع چگالی احتمالی، نسبت به اندازهٔ μ و همچنین هر دو تقریباً همه جا (a.e) نسبت به μ مثبت باشند. اگر E_f امید ریاضی نسبت به اندازهٔ احتمال $fd\mu$ باشد، آنگاه $E_f(\ln f) \geq E_f(\ln g)$ و تساوی برقرار است اگر و تنها اگر $f = g$ a.eμ (اثبات در مرجع [۴])

قضیه ۱ (مرجع [۳]): $H(\underline{\theta}|\underline{\theta}^{(n)})$ در نقطه $\underline{\theta} = \underline{\theta}^{(n)}$ ماکزیمم می‌شود.

اثبات :

$$\begin{aligned} E_x[\ln g(Y|\underline{\theta})|y;\underline{\theta}^{(n)}] &= \int_{\{x;h(x)=y\}} \ln g(y|\underline{\theta}) f(x|y;\underline{\theta}^{(n)})dx \\ &= \ln g(y|\underline{\theta}) \int_{\{x;h(x)=y\}} f(x|y;\underline{\theta}^{(n)})dx \\ &= \ln g(y|\underline{\theta}) \end{aligned}$$

بنابراین

$$H(\underline{\theta}|\underline{\theta}^{(n)}) = E_x[\ln f(X|\underline{\theta})|y;\underline{\theta}^{(n)}] - E_x[\ln g(Y|\underline{\theta})|y;\underline{\theta}^{(n)}]$$

$$\begin{aligned} &= E_x\left\{\ln\left[\frac{f(X|\underline{\theta})}{g(Y|\underline{\theta})}\right]|y;\underline{\theta}^{(n)}\right\} \\ &= \int_{\{x;h(x)=y\}} \ln\left[\frac{f(x|\underline{\theta})}{g(y|\underline{\theta})}\right] f(x|y;\underline{\theta}^{(n)})dx \\ &= \int_{\{x;h(x)=y\}} \ln\left[\frac{f(x|\underline{\theta})}{g(y|\underline{\theta})}\right]\left[\frac{f(x|\underline{\theta}^{(n)})}{g(y|\underline{\theta}^{(n)})}\right] dx \end{aligned}$$

بنابه لم ۱ می‌توان نوشت:

$$\begin{aligned} &\int_{\{x;h(x)=y\}} \ln\left[\frac{f(x|\underline{\theta})}{g(y|\underline{\theta})}\right]\left[\frac{f(x|\underline{\theta}^{(n)})}{g(y|\underline{\theta}^{(n)})}\right] dx \\ \leq &\int_{\{x;h(x)=y\}} \ln\left[\frac{f(x|\underline{\theta}^{(n)})}{g(y|\underline{\theta}^{(n)})}\right]\left[\frac{f(x|\underline{\theta}^{(n)})}{g(y|\underline{\theta}^{(n)})}\right] dx \\ = &\int_{\{x;h(x)=y\}} [\ln f(x|\underline{\theta}^{(n)}) - \ln g(y|\underline{\theta}^{(n)})] f(x|y;\underline{\theta}^{(n)})dx \\ = &\; E_x[\ln g(Y|\underline{\theta}^{(n)})|y;\underline{\theta}^{(n)}] - \ln g(y|\underline{\theta}^{(n)}) \end{aligned}$$

یعنی این که :

$$H(\underline{\theta}|\underline{\theta}^{(n)}) \leq H(\underline{\theta}^{(n)}|\underline{\theta}^{(n)})$$

با توجه به قضیهٔ ۱ می‌توان نوشت:

$$\begin{aligned} &H(\underline{\theta}^{(n)}|\underline{\theta}^{(n)}) \geq H(\underline{\theta}^{(n+1)}|\underline{\theta}^{(n)}) \Rightarrow \\ &E_x[\ln f(X|\underline{\theta}^{(n)})|y;\underline{\theta}^{(n)}] - \ln g(y|\underline{\theta}^{(n)}) \geq \\ &E_x[\ln f(X|\underline{\theta}^{(n+1)})|y;\underline{\theta}^{(n)}] - \ln g(y|\underline{\theta}^{(n+1)}) \Rightarrow \\ &E_x[\ln f(X|\underline{\theta}^{(n)})|y;\underline{\theta}^{(n)}] - E_x[\ln f(X|\underline{\theta}^{(n+1)})|y;\underline{\theta}^{(n)}] \\ &\geq \ln g(y|\underline{\theta}^{(n)}) - \ln g(y|\underline{\theta}^{(n+1)}) \qquad (1.1) \end{aligned}$$

اما چون در تکرار $(n+1)$ام، $\underline{\theta}^{(n+1)}$ مقداری از $\underline{\theta}$ است که $E[\ln f(X|\underline{\theta})|y;\underline{\theta}^{(n)}]$ را ماکزیمم می‌کند لذا:

$$E_x[\ln f(X|\underline{\theta}^{(n)})|y;\underline{\theta}^{(n)}] \leq E_x[\ln f(X|\underline{\theta}^{(n+1)})|y;\underline{\theta}^{(n)}]$$

لذا از نامساوی (۱.۱) داریم:

$$ln\, g(y|\underline{\theta}^{(n)}) \leq ln\, g(y|\underline{\theta}^{(n+1)})$$

به عبارت دیگر، الگوریتم EM باعث افزایش لگاریتم درستنمایی از مرحله‌ای به مرحله بعد از آن می‌شود و این روند تا همگرا شدن الگوریتم به جواب، که ثابت می‌شود تحت شرایطی همان برآورد درستنمایی ماکزیمم است ادامه خواهد داشت. لازم به ذکر است که همگرایی، به انتخاب نخستین $\underline{\theta}$ (یعنی $\underline{\theta}^{(0)}$) برای شروع بستگی ندارد.

تذکر ۱: هنر استفاده از الگوریتم EM در انتخاب و تشخیص داده‌های کامل X است. اگر چه روشهای زیادی برای محاط کردن Y در یک فضای نمونهٔ بزرگتر وجود دارد اما اغلب طبیعت مسئله یا ملاحظات دیگر به یک تعریف روشن و صریح از X منجر می‌شود.

مثالهای زیر ما را در جهت استفادهٔ صحیح از این الگوریتم با در نظر گرفتن نکات و ریزه‌کاری‌های آن یاری می‌دهد.

مثال ۱: فرض کنید هر یک از والدین دارای سه نوع ژن O, B, A با احتمالات P_A و P_B و P_O که $P_A + P_B + P_O = 1$ باشند. بر اساس اصول ژنتیک، گونه‌های ژنی بوجود آمده توسط آنها عبارتند از: $O|O, B|B, B|O, A|O, A|B, A|A$ است. به عنوان مثال گونه ژنی $A|B$ نشان دهندهٔ آنست که ژن دریافتی از مادر A و ژن دریافتی از پدر B می‌باشد یا بالعکس.

بر اساس نمونه‌های خون گرفته شده از افراد، ما قادر به مشاهده فراوانی گونه‌های ژنی بالا نمی‌باشیم اما با تحت تأثیر قرار دادن نمونه‌های خونی بوسیلهٔ آنتی‌ژنهای A و B می‌توان اطلاعاتی دربارهٔ گونه‌های ژنی بدست آورد. به‌این ترتیب که در هر نمونه اگر آنتی‌ژن A به تنهایی نمایان شود یکی از گونه‌های $A|A$ یا $A|O$، اگر آنتی‌ژن B به تنهایی ظاهر شود یکی از گونه‌های $B|B$ یا $B|O$، اگر آنتی‌ژنهای A و B هر دو با هم نمایان شوند گونهٔ $A|B$ و اگر هیچ‌کدام از آنتی‌ژنهای A یا B نمایان نگردد گونهٔ $O|O$ مشخص خواهد شد. جدول زیر مطالب گفته شدهٔ بالا را بطور خلاصه نشان می‌دهد.

گونه‌های ژنی	آنتی ژن
$A\|A$ و $A\|O$	A
$B\|B$ و $B\|O$	B
$A\|B$	AB
$O\|O$	O

در این مسأله، تعداد افراد مشاهده شده از دسته‌های سمت چپ جدول، داده‌های Y را تشکیل می‌دهند در صورتی که تعداد افرادی که هریک از ۶ گونهٔ ژنی سمت راست را دارند و غیر مشخص و نامعلوم هستند داده‌های X را مشخص می‌کنند.

فرض کنید تعداد افرادی که دارای گونه‌های $O|O, A|B, B|O, B|B, A|O, A|A$ هستند به ترتیب با $x_1, x_2, \ldots, x_6$ نشان دهیم. n حجم نمونهٔ گرفته شدهٔ ما و y_1, y_2, y_3, y_4 به ترتیب تعداد افراد مشاهده شده برای دسته‌های O, AB, B, A هستند که $n = y_1 + y_2 + y_3 + y_4$. توجه کنید که در این جا $y_1 = x_1 + x_2$ ، $y_2 = x_3 + x_4$ ، $y_3 = x_5$ و $y_4 = x_6$ می‌باشد.

بنابراین داریم:

$$\begin{aligned} f(x|p) &= \binom{n}{x_1, x_2, \ldots, x_6} (P_A^2)^{x_1} (2P_AP_O)^{x_2} \\ &\times (P_B^2)^{x_3} (2P_BP_O)^{x_4} (2P_AP_B)^{x_5} (P_O^2)^{x_6} \end{aligned}$$

که $P_O^2, 2P_AP_B, 2P_BP_O, P_B^2, 2P_AP_O, P_A^2$ به ترتیب احتمالات متناظر با گونه‌های ژنی بالا است. در گام E از الگوریتم EM ما باید $E[lnf(X|P)|y; P^{(m)}]$ را تشکیل دهیم که $P^{(m)} = (P_A^{(m)}, P_B^{(m)}, P_O^{(m)})^T$ بردار جاری P می‌باشد.

$$\begin{aligned} Q(P|P^{(m)}) &= E(lnf(X|P)|y; P^{(m)}) \\ &= x_1^{(m)} ln(P_A^2) + x_2^{(m)} ln(2P_AP_O) \\ &+ x_3^{(m)} ln(P_B^2) + x_4^{(m)} ln(2P_BP_O) \\ &+ y_3 ln(2P_AP_B) + y_4 ln(P_O^2) \end{aligned}$$

که در آن $i = ۱,۲,۳,۴,۵,۶$ ، $x_i^{(m)} = E(X_i|y;P^{(m)})$ و $x_۶^{(m)} = y_۴, x_۵^{(m)} = y_۳$ است.

اما در گام M از الگوریتم EM تابع $Q(P|P^{(m)})$ ماکزیمم می‌شود، که در این جا بدلیل قید $P_A + P_B + P_O = ۱$ این کار با استفاده از روش لاگرانژ انجام می‌گردد.

$$H(P,\lambda) = Q(P|P^{(m)}) + \lambda(P_A + P_B + P_O - ۱)$$

اگر نسبت به P_O, P_B, P_A و λ مشتق گرفته و مساوی صفر قرار دهیم، پس از ساده کردن داریم:

$$P_A^{(m+۱)} = \frac{۲x_۱^{(m)} + x_۲^{(m)} + y_۳}{۲n}$$
$$P_B^{(m+۱)} = \frac{۲x_۳^{(m)} + x_۴^{(m)} + y_۳}{۲n}$$
$$P_O^{(m+۱)} = \frac{x_۲^{(m)} + x_۴^{(m)} + ۲y_۴}{۲n}$$

بر اساس داده‌های ۱۸۶ $= y_۱$، ۳۸ $= y_۲$، ۱۳ $= y_۳$ و ۲۸۴ $= y_۴$ و با بهره‌گیری از رایانه، برآوردهای زیر حاصل می‌شوند:

$$\hat{P}_A = ۰/۲۱۳۶, \quad \hat{P}_B = ۰/۰۵۰۱, \quad \hat{P}_O = ۰/۷۳۶۳$$

(این مثال در مرجع [۴] آمده است).

مثال ۲: داده‌های جدول زیر مربوط به سالهای ۱۹۱۰ تا ۱۹۱۲ شهر لندن است. ستون اول این جدول تعداد مرگ و میر در بین زنان ۸۰ و بیشتر از ۸۰ سال را نشان می‌دهد، ستون دوم y_iها یعنی فراوانی روزهایی را که در آن روزها i مرگ و میر داشته‌ایم مشخص می‌سازد. بدلیل الگوهای متفاوت مرگ و میر در زمستان و تابستان، یک توزیع پواسن به تنهایی نمی‌تواند برازش خوبی بر داده‌ها داشته باشد، لذا به‌نظر می‌رسد ترکیبی از دو توزیع پواسن، برازش بهتری را ایجاد کند. تحت مدل آمیختهٔ پواسن، تابع درستنمایی مشاهدات به صورت زیر است:

$$L(y;\underline{\theta}) = \prod_{i=۰}^{۹}[\alpha e^{-\mu_۱}\frac{\mu_۱^i}{i!} + (۱-\alpha)\frac{e^{-\mu_۲}\mu_۲^i}{i!}]^{y_i}$$

که α پارامتر ترکیب و $\mu_۱$ و $\mu_۲$ میانگین‌های دو توزیع پواسن و $\underline{\theta} = (\alpha, \mu_۱, \mu_۲)^T$ هستند.

فراوانی (y_i)	تعداد مرگ و میر (i)	فراوانی (y_i)	تعداد مرگ و میر (i)
۶۱	۵	۱۶۲	۰
۲۷	۶	۲۶۷	۱
۸	۷	۲۷۱	۲
۳	۸	۱۸۵	۳
۱	۹	۱۱۱	۴

اگر فرض کنیم $Z_i(\underline{\theta})$ احتمال پسین آن که یک روز با i مرگ و میر به جامعهٔ پواسن ۱ متعلق باشد، آنگاه $۱ - Z_i(\underline{\theta})$ احتمال پسین تعلق یک روز با i مرگ و میر به جامعهٔ پواسن ۲ خواهد بود و به صورت زیر بدست می‌آید:

$$Z_i(\underline{\theta}) = \frac{\alpha e^{-\mu_۱}\mu_۱^i}{\alpha e^{-\mu_۱}\mu_۱^i + (۱-\alpha)e^{-\mu_۲}\mu_۲^i} \quad i = ۰,۱,۲,\ldots,۹$$

با تعریف H به صورت:

$$H = \begin{cases} ۱ & \text{اگر روزی با هر تعداد مرگ و میر از جامعه پواسن ۱ باشد.} \\ ۰ & \text{در غیر اینصورت} \end{cases}$$

داریم $P(H = ۱) = \alpha$.

در این مثال $X = (U,H)$، که متغیر تصادفی U تعداد مرگ و میر است و مقادیر ۹ ، ...، ۲ ، ۱ ، ۰ را می‌گیرد.

$$\begin{aligned} f(x|\underline{\theta}) &= f_{\underline{\theta}}(u,h) \\ &= (\alpha g_۱(u))^h[(۱-\alpha)g_۲(u)]^{۱-h} \quad h = ۰,۱ \end{aligned}$$

که

$$g_j(u) = \frac{e^{-\mu_j}(\mu_j)^u}{u!} \qquad j = ۱,۲$$
$$A(u) = \frac{g_۱(u)}{g_۲(u)} \qquad u = ۰,۱,\ldots,۹$$

لذا:

$$\begin{aligned} f(x_۱,\ldots,x_n) &= \prod_{r=۱}^{n} f(x_r|\underline{\theta}) \\ &= (۱-\alpha)^{\sum_{i=۰}^{۹} y_i}\left(\frac{\alpha}{۱-\alpha}\right)^{\sum_{i=۰}^{۹} q_i} \\ &\times \prod_{i=۰}^{۹}[g_۲(i)]^{y_i}[A(i)]^{q_i} \end{aligned}$$

و در آن $S_i = \{r|u_r = i\}$ ، $i = ۰,۱,۲,\ldots,۹$ و $q_i = \sum_{r\in S_i} H_r$ می‌باشند، واضح است که $N(S_i) = y_i$.

$$\begin{aligned} E[\ln f(X|\underline{\theta})\,;\underline{\theta}^{(m)}] &= \sum_{i=۰}^{۹} y_i[\ln(۱-\alpha) + \ln g_۲(i)] \\ &+ \sum_{i=۰}^{۹}[\ln(\frac{a}{a-۱}) + \ln A(i)]E[q_i|y;\underline{\theta}^{(m)}] \end{aligned}$$

و اما :

$$\begin{aligned} E[q_i|y;\underline{\theta}^{(m)}] &= \sum_{r\in S_i} E[H_r|y;\underline{\theta}^{(m)}] \\ &= \sum_{r\in S_i} Z_i(\underline{\theta}^{(m)}) \\ &= y_i Z_i(\underline{\theta}^{(m)}) \end{aligned}$$

که $\underline{\theta}^{(m)} = (\alpha^{(m)}, \mu_۱^{(m)}, \mu_۲^{(m)})^T$ پارامترهای بدست آمده از تکرار mام الگوریتم (پارامتر جاری) است.

با مقدارگذاری به جای $A(i)$ و $g_۲(i)$ در امید شرطی بالا، ساده کردن و در نهایت مشتق‌گیری داریم:

$$\begin{aligned} \alpha^{(m+۱)} &= \frac{\sum_{i=۰}^{۹} y_i Z_i(\underline{\theta}^{(m)})}{\sum_{i=۰}^{۹} y_i} \\ \mu_۱^{(m+۱)} &= \frac{\sum_{i=۰}^{۹} y_i Z_i(\underline{\theta}^{(m)})i}{\sum_{i=۰}^{۹} y_i Z_i(\underline{\theta}^{(m)})} \\ \mu_۲^{(m+۱)} &= \frac{\sum_{i=۰}^{۹} y_i(۱-Z_i(\underline{\theta}^{(m)}))i}{\sum_{i=۰}^{۹} y_i(۱-Z_i(\underline{\theta}^{(m)}))} \end{aligned}$$

با استفاده از رایانه جوابهای عددی بدست آمده عبارتند از:

$$\hat{\alpha} = ۰/۳۵۹۹ \;,\; \hat{\mu}_۱ = ۱/۲۵۶۱ \;,\; \hat{\mu}_۲ = ۲/۶۶۳۴$$

تذکر ۲: در این دو مثال، بوضوح می‌توان دید که Y تابعی از X است. مجدداً تأکید می‌شود که مشاهدات X در دسترس نیستند و تنها مشاهدات Y است که در اختیار ما قرار دارد.

تذکر ۳: گاهی اوقات ممکن است بنا به دلایلی اجرای گام M از الگوریتم EM امکان‌پذیر نباشد، در این گونه مواقع می‌توان از الگوریتم GEM (الگوریتم EM تعمیم یافته) که آن نیز برای نخستین بار توسط Dempster و سایرین در سال ۱۹۷۷ (در [۱]) ارایه شده، استفاده نمود. روند این الگوریتم طوری است که:

$$Q(\underline{\theta}^{(n+۱)}|\theta^{(n)}) \geq Q(\underline{\theta}^{(n)}|\underline{\theta}^{(n)})$$

بنابراین در این جا نیز از مرحله‌ای به مرحلهٔ بعد افزایش مقدار تابع درستنمایی را خواهیم داشت و ملاحظه می‌شود که الگوریتم EM حالت خاصی از الگوریتم GEM می‌باشد.

تذکر ۴: الگوریتم EM و GEM موارد استفاده فراوانی دارند که به عنوان مثال می‌توان به استفاده از آنها در بازسازی تصویر در سی‌تی‌اسکن، مسایل ژنتیک و نیز مسایلی که در آن دادهٔ گمشده وجود دارد، اشاره نمود.

مراجع

[1] Dempster AP , Laird NM, Rubin DB. Maximum likelihood from incomplete data via the FM algorithm. *JR Stat Soc Series B* 1977; 39: 1-38.

[2] Wu CF. On the convergence properties of the EM algorithm. *Ann stat* 1983; 11: 95-103.

[3] K.Lange and R.Carson, EM Reconstruction Algorithms for Emission and Transmission Tomograohy. *J. of Computer Assisted Tomograohy*, Vol. 8, No. 2, 1984.

[4] K. Lange. Mathematical and Statistical Methods for Genetic Analysis. Springer, 1997, p 22-32.