pdf FarsiGauss 92/07/29 جلسه 7

در جلسه امتحان یک صفحه A4 را میتوانم هرچی میخواهم بنویسم و بیاورم. امتحان تکسری از فصل های خاصی از کتاب را شامل می شود، حداکثر ۲ فصل است، بقیه فصل ها باید در یه درجه خودش را نشان دهد. به ایمیل statdatamining@gmail.com ایمیل ها را بنویسند

روش امروزها دیگر سلسله مراتبی نیست ولی یک روش Based clustering است.

چند تا ایراد این مقاله دارد اما استاد pdf صحیح را قرار خواهد داد. ۱ Model Based clustering (GMM) را می خواهد یاد دهد.

هدفی که در این مقاله دنبال میشود این است که دو تا هدف دارد ۲- برای آموزش مدل بیست می توید یهایی Model Based را میتوانم به ۴ روش انجام دهم.

چهار روش براورد پارامترهاي يك مدل آميخته گاوسی

دنیا رحمانی عادل محمدپور

دانشکده‌ی ریاضی و علوم کامپیوتر دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)

چکیده

مدل آمیخته‌ی گاوسی پرکاربردترین مدل آمیخته‌ی متناهی است. خاصیت مهم این مدل انعطاف‌پذیری آن نسبت به توزیع‌های پیوسته با اشکال گوناگون می‌باشد. از آن‌جا که مهم‌ترین بخش یک مدل، برآورد پارامترهای آن می‌باشد، در این جا برآنیم تا پارامترهای مدل آمیخته‌ی گاوسی دو مولفه‌ای را از طریق چهار روش برآورد کنیم. در ابتدا مدل آمیخته‌ی متناهی و مدل آمیخته‌ی گاوسی را در حالت دو مولفه‌ای بیان می‌کنیم و سپس پارامترهای مدل را از دو روش گشتاوری و ماکسیمم درستنمایی تحت عنوان حل تحلیلی به دست آورده و در بخش بعد برای برآورد پارامترها از روش عددی استفاده می‌کنیم. در ادامه برآورد پارامترها را با استفاده از الگوریتم EM به دست آورده و در انتها نیز از الگوریتم نمونه‌گیر گیبز برای یافتن برآوردها استفاده کرده‌ایم. در بخش نتیجه‌گیری، نتایج به دست آمده از روش‌ها را با یکدیگر مقایسه می‌کنیم. به طور کلی سعی ما براین است که یک مسئله‌ی برآورد را با چهار روش معمول، به طور منسجم حل کرده و برتری‌ها و محدودیت‌های هر یک را برای کاربران مشخص کنیم.

کلمات کلیدی: حل تحلیلی، عددی، تقریبی، الگوریتم EM ، الگوریتم گیبز، مدل آمیخته گاوسی.

۱ مقدمه

مدل آمیخته‌ی متناهی اولین بار در قرن نوزدهم وارد ادبیات آمار شد، [۹]. ازجمله مدل‌های آمیخته متناهی می‌توان به مدل آمیخته‌ی پواسن برای گروه‌بندی اسناد در امر بازیابی اطلاعات و مدل آمیخته‌ی فیشر برای تحلیل متون و آزمایش‌های ژنی اشاره کرد. مشهورترین مدل آمیخته، مدل آمیخته‌ی گاوسی می‌باشد، [۴] و [۱۳]. مدل‌های آمیخته‌ی متناهی در خوشه‌بندی، برآورد تابع چگالی، تحلیل مولفه‌ای، تحلیل تصاویر و دیگر موارد کاربرد بسزایی دارد. یک مدل آمیخته‌ی متناهی به صورت زیر تعریف می‌شود:
فرض کنید $x_1, \ldots, x_n$ مشاهداتی از نمونه‌ی تصادفی مستقل و هم‌توزیع $X_1, \ldots, X_n$ با بردار پارامتر $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ باشند، آن‌گاه تابع چگالی آمیخته مشاهدات به صورت زیر خواهد بود:

$$f(x_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f_k(x_i; \theta_k), \qquad i = 1, \ldots, n. \qquad (1)$$

اگر بخواهیم GMM را به صورت ساده بیان کنیم، مثلاً اگر بخواهیم طول قد دانشجویان را در دانشگاه مطالعه کنیم، خیلی وقتها گفتیم که توزیع این قدها نرمال است، توزیع نرمال به این معنی است که آنهایی که طول قد کم دارند، کم هستند و همچنین آنهایی هم که بلند قامت هستند کم هستند بیشتر حول و حوش میانه، میانگین مد، به اندازه 3σ برابر اینها قرار دارند، یک شکل متقارن دارند، منحنی گوسی یا منحنی نرمال که توسط گوس ابداع شده، در واقع گوس یک قضیه ای را ابداع کرده به نام قضیه حد مرکزی، در قضیه حد مرکزی نشان داد اگر ما میانگین نمونه را هی ما بگیریم، مشاهده کنیم، ... $\bar{x}$ میانگین نمونه ما است) اگر اندازه نمونه به سمت بی نهایت برود، این

به سمت $e^{-\frac{1}{2}x^2}$ همگرا میشود که به این میگوییم تابع چگالی متغیر تصادفی نرمال استاندارد یا متغیر تصادفی گوسی که
این یک شکل زنگی دارد، میانگین - میانه و مدش برهم منطبق است و خواص خیلی خوبی دارد، جز مهم‌ترین توزیع هایی
است که در آمار بحثش رجوع میشود. خیلی از مسائل را میشه با این روش مدل بندی کرد. پرکاربردترین توزیع است.
ولی اگر بیاییم به طول قد دانشجویان دقیق‌تر نگاه کنیم متوجه میشویم که واقعا این داده‌ها توزیع نرمال نیستند چون درست است که
وقتی شکل‌اش را رسم میکنیم، شبیه توزیع نرمال است اگر آزمون تستی انجام دهیم (good of fitness) شاید ما را به نتیجه‌ای
برساند که توزیع اینها نرمال است، به عبارتی فرض نرمال بودن را نتوانیم رد کرد.

$f(x_i;\boldsymbol{\theta})$ را تابع چگالی آمیخته‌ی K مولفه‌ای نیز گویند. منظور از مولفه همان زیرجامعه‌ی تشکیل دهنده‌ی جامعه می‌باشد که تعدادشان را با K نشان می‌دهند. θ_k پارامتر مربوط به زیر جامعه‌ی k می‌باشد و α_k ضریب وزنی یا ضریب آمیخته‌ی مولفه‌ی kاُم است که در شرایط زیر صدق می‌کند:

$$0 \leq \alpha_k \leq 1, \qquad \sum_{k=1}^{K} \alpha_k = 1.$$

$f_k(x_i;\theta_k)$ چگالی مولفه‌ی kاُم با پارامتر θ_k می‌باشد، [8].

ولی اگر بیاییم دقیق کنیم و بگیم فکر کنید که طول قد دانشجویان چگونه حساب میشود، که دانشجویان ما دخترها و پسرها هستند، به طور تقریبا ۴۵٪ و ۵۵٪ هستند.
دخترها پسرها

۲ مدل آمیخته‌ی گاوسی

مدل آمیخته‌ی گاوسی برای مشاهدات مستقل و هم‌توزیع $x_1,\ldots,x_n$، مجموع وزن‌دار K مولفه، با تابع چگالی گاوسی است، که با معادله‌ی زیر نشان داده می‌شود:

$$f(x_i;\boldsymbol{\mu},\boldsymbol{\sigma}^2) = \sum_{k=1}^{K} \alpha_k \phi(x_i;\mu_k,\sigma_k^2)$$

در این‌جا $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_K)$ و $\boldsymbol{\sigma}^2 = (\sigma_1^2,\ldots,\sigma_K^2)$ بردار پارامترهای مدل آمیخته‌ی (۱) می‌باشند. $\phi(x_i;\mu_k,\sigma_k^2)$ تابع چگالی گاوسی با پارامترهای μ_k,σ_k^2 مربوط به مولفه‌ی kاُم می‌باشد که از رابطه‌ی زیر به دست می‌آید:

$$\text{تابع چگالی نرمال} \quad \phi(x_i;\mu_k,\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp[-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}]$$

در زیر مدل آمیخته‌ی گاوسی دو مولفه‌ای $K=2$، برای مشاهدات، بیان می‌شود:

α مثلاً درصد دخترهاست

$1-\alpha$ مثلاً درصد پسرهاست

مثال۱: برای مشاهدات $x_1,\ldots,x_n$، که ترکیبی از دو زیر جامعه گاوسی به ترتیب با میانگین‌های μ_1 و μ_2 و واریانس‌های σ_1^2 ، σ_2^2 باشند، آنگاه تابع چگالی آمیخته آن‌ها با ضرایب وزنی $\alpha_1 = \alpha$ و $\alpha_2 = 1-\alpha$ به صورت زیر خواهد بود:

$$\sum_{k=1}^{2} \alpha_k\phi(x_i;\mu_k,\sigma_k^2) = \overbrace{\frac{\alpha}{\sqrt{2\pi\sigma_1^2}}\exp[-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}]}^{\text{نرمال اولی}} + \underbrace{\frac{1-\alpha}{\sqrt{2\pi\sigma_2^2}}\exp[-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}]}_{\text{نرمال دومی}}$$

برای ساده‌تر کردن مسئله می‌توان مشاهداتی از مدل آمیخته‌ی گاوسی دو مولفه‌ای را بر اساس دستور ارایه شده در بخش ۱.۴ پیوست در نرم‌افزار R تولید می‌کنیم. مقادیر در نظر گرفته شده برای پارامترها در دستور آورده شده در پیوست، بسته به نظر کاربر انتخاب می‌شود.

همانطور که می‌دانیم مهم‌ترین مسئله در ارتباط با یک مدل، برآورد پارامترهای آن می‌باشد. از این رو در بخش ۳ این مقاله پارامترهای مدل آمیخته‌ی گاوسی دو مولفه‌ای را از دو روش گشتاوری و ماکسیمم درستنمایی به عنوان روشی تحلیلی بیان کرده، و سپس از روش عددی استفاده می‌کنیم. در ادامه نیز از الگوریتم EM و الگوریتم نمونه‌گیر گیبز برای یافتن برآوردها استفاده کرده‌ایم.

در همه جای دنیا غیر استثنایی طول قد دخترها از پسرها کمتر است، پس ظاهراً این هیستوگرامی که ما کشیدیم و به نظرمان آمده که دارای توزیع نرمال است این تشکیل شده از دوتا هیستوگرام، این دو تا هیستوگرام با هم دیگر مخلوط شده‌اند، ظاهر یک هیستوگرام درست کردند که شبیه توزیع نرمال بوده.

اینجا بوده که بحث پیش آمده که ما به جای اینکه با یک توزیع نرمال کار کنیم، بیاییم به طول قد دانشجویان، دوتا توزیع نرمال برازش
دهیم یک توزیع نرمال برای دخترها و یک توزیع نرمال برای پسرها $\boxed{2}$ ولی اینکه چقدر ارزش داشته باشه که پارامترهای اضافی وارد مدل
- مان کنیم، یک بحثی است که همیشه باید بین اینها بالانس برقرار کنیم، یعنی ما میتوانیم همیشه یک مدل خیلی دقیق ارائه دهیم ولی
تعداد پارامترها که زیاد شود، خیلی خوشایند نیست. در این مسئله هم می‌بینیم که وقتی ما یک توزیع نرمال داریم، با پارامترهایش،
راحتی قابل برآورد شده هستند، مثلاً میانگین اش را با $\bar{x}$ برآورد میکنیم و واریانس اش را هم با s^2 ولی اگر دو تا توزیع

زمان داشته باشیم، که معمولا این دیتا را به صورت آمیخته در نظر میگیریم، آمیخته به این معناست که تابع چگالی شان
را به صورت وزنی با هم جمع کنیم، ما می‌دانیم مساحت زیر تابع چگالی ها، با هم جمعشان می‌شود ۱ احتمال یک است، مساحت
زیر تابع چگالی ها هم یک است، پس اگر قرار باشد یک تابع چگالی جوری برای این متغیر تصادفی آمیخته تعریف کنیم، باید یک
convence convination این از دیتا تابع چگالی باشد که انتگرالش هم یک شود. پس باید یک سری از این مواجه میشویم به جای اینکه یک

توزیع نرمال داشته باشیم دو تا توزیع نرمال یا چند تا توزیع نرمال داریم، حالا اینجا توزیع نرمال هایی که ما در نظر گرفتیم تک متغیره است، در مسائل خوشه‌بندی معمولا چند متغیره است،

۳ برآورد پارامترهای مدل آمیخته‌ی گاوسی به روش تحلیلی

در این بخش ابتدا روش گشتاوری و سپس روش ماکسیمم درستنمایی، را برای برآورد پارامترهای مدل آمیخته‌ی گاوسی به کار می‌بریم.

۱.۳ روش گشتاوری

پی‌یرسن (۱۸۹۴) برای برازش مدل به داده‌هایی که نسبت طول پیشانی به طول بدن ۱۰۰۰ خرچنگ را بیان می‌داشتند و توسط ولدن در سال ۱۸۹۲ جمع‌آوری شده بودند، از مدل آمیخته‌ی گاوسی تک متغیره استفاده کرد، [۹]. او داده‌ها را به ۲۹ بازه با فراوانی y_i برای $i = 1, \ldots, 29$؛ تقسیم‌بندی کرد وبرای مدل‌بندی آن‌ها دو تابع چگالی نرمال با میانگین‌های μ_1 و μ_2 ، واریانس‌های σ_1^2 و σ_2^2 را با ضرایب وزنی α و $1 - \alpha$ با یکدیگر آمیخت. او برای برآورد پارامترهای مدل از روش گشتاوری استفاده کرد. پی‌یرسن، در ابتدا گشتاورهای مرکزی مرتبه‌ی اول تا پنجم مشاهدات را از روابط زیر به دست آورد:

فعلا این مسئله را خیلی ساده بیان می‌کنیم چون مسائل مسئله ای نیست از نظر ریاضی. با دست تقریبا غیرممکن است این مسئله
را بتوانیم حل کنیم. هدف از GMM، clustering است.
هدف بحث امروز این است که با بیاییم GMM را معرفی کنیم و
بعد پارامترهایش را برآورد کنیم.
برای حل این مسائل که پارامترها را برآورد

$$
\begin{aligned}
\tilde{\mu}_1 &= 0 \\
\tilde{\mu}_2 &= \mu_2' - {\mu_1'}^2 \\
\tilde{\mu}_3 &= \mu_3' - 3\mu_1'\mu_2' + 2{\mu_1'}^3 \\
\tilde{\mu}_4 &= \mu_4' - 4\mu_1'\mu_3' + 6{\mu_1'}^2\mu_2' - 3{\mu_1'}^4 \\
\tilde{\mu}_5 &= \mu_5' + 10{\mu_1'}^2\mu_3' - 5\mu_1'\mu_4' - 10{\mu_1'}^3\mu_2' + 4{\mu_1'}^5
\end{aligned}
$$

کنیم، به ۴ روش میتوانیم این مسئله را حل کنیم. پس در واقع بخش عمده ای از مقاله معطوف شده به حل یک مسئله با روش های مختلف

به طوری‌که μ_n'، n اُمین گشتاور مشاهدات حول مبدا می‌باشد و به طریق زیر محاسبه می‌شود :

$$\mu_n' = \frac{1^n \times y_1 + 2^n \times y_2 + 3^n \times y_3 + \cdots + 29^n \times y_{29}}{1000}$$

سپس با استفاده از روش گشتاوری به ۵ رابطه برای ۵ پارامتر مجهول رسید:

$$
\begin{gathered}
\mu_1\alpha + \mu_2(1-\alpha) = 0 \\
{\mu_1}^2\alpha(1+{\sigma_1}^2) + {\mu_2}^2(1-\alpha)(1+{\sigma_2}^2) = \tilde{\mu}_2 \\
{\mu_1}^3\alpha(1+3{\sigma_1}^2) + {\mu_2}^3(1-\alpha)(1+3{\sigma_2}^2) = \tilde{\mu}_3 \\
{\mu_1}^4\alpha(1+6{\sigma_1}^2+3{\sigma_1}^4) + {\mu_2}^4(1-\alpha)(1+6{\sigma_2}^2+3{\sigma_2}^4) = \tilde{\mu}_4 \\
{\mu_1}^5\alpha(1+10{\sigma_1}^2+15{\sigma_1}^4) + {\mu_2}^5(1-\alpha)(1+10{\sigma_2}^2+15{\sigma_2}^4) = \tilde{\mu}_5
\end{gathered}
$$

مدل های آمیخته یک سابقه ی خیلی طولانی در دنیای آمار دارند که از قرن ۱۹ این مسئله به وجود آمده و دو نفر تنها مشهورترین
کسی بود که به این مسئله پرداخته، مدل های آمیخته کاربردهای خیلی زیادی دارند، یکی از کاربردهایش خوشه‌بندی است.
اینجا n تا مشاهده داشته باشیم، که یک متغیره در نظر میگیریم، که این مشاهدات ما از یک سری متغیرهای تصادفی آمده اند x_1 تا x_n.
یعنی این مشاهدات بعد می‌کنیم مسئله خوشه‌بندی سلسله مراتبی، آن مثالی که در مورد جرم و جنایت در ایران بود که x_1 مثلا شهر اول
x_n مثلا شهر آخر، اول x_1 مولفه هایش مثلا دزدی بوده، قتل، قاچاق بوده، حاشیه نشینی دزدی بوده و... اینها مولفه های

x_1 هستند و x_n هم واسه شهر آخر است. پس x_1 تا x_n متغیرهای ما هستند مستقل و هم توزیع.

مستقل و هم توزیع یعنی این که در رخ داد جرم و جنایت از یک شهر به شهر دیگر ربط نباشد، فرض خیلی سنگینی است.

در آمار

و خیلی وقت‌ها هم غیرواقعی است ولی مجبوریم این فرض را اعمال کنیم تا کار را راحت تر فرمول بندی اش کنیم. هم توزیع بودن یعنی

متغیرها که از x_1 تا x_n یا توزیع نرمال دارند یا توزیع آمیخته نرمال دارند (که توزیع آمیخته کشی ندارد) یعنی توزیع شان را باید تابع چگال بیان کنیم، تابع چگالی شان را

با حل دستگاه معادله‌ی بالا، به معادله‌ی درجه‌ی ۹ زیر دست یافت:

$$24p_2^9 - 28\lambda_4 p_2^7 + 36\tilde{\mu}_3^2 p_2^6 - (24\tilde{\mu}_3\lambda_5 - 10\lambda_4^2)p_2^5 - (148\tilde{\mu}_3^2\lambda_4 + 2\lambda_5^2)p_2^4 +$$
$$(288\tilde{\mu}_3^4 - 12\lambda_4\lambda_5\tilde{\mu}_3 - \lambda_4^3)p_2^3 + (24\tilde{\mu}_3^3\lambda_5 - 7\tilde{\mu}_3^2\lambda_4^2)p_2^2 + 32\tilde{\mu}_3^4\lambda_4 p_2 - 24\tilde{\mu}_3^6 = 0$$

میتوانیم بنویسیم f. هم توزیع بودن فرض سنگینی نیست به نسبت اولی و θ هم یا پارامترهای ما هستند (صفحه ۱، فرمول ۱)

به طوری‌که

$$p_1 = \mu_1 + \mu_2, \qquad p_2 = \mu_1\mu_2. \qquad (2)$$

$$\lambda_4 = 9\tilde{\mu}_2^2 - 3\tilde{\mu}_4،$$
$$\lambda_5 = 30\tilde{\mu}_2\tilde{\mu}_3 - 3\tilde{\mu}_5.$$

آن موقع تابع چگالی آمیخته مشاهدات را به صورت یک ترکیب محدب از تابع چگالی ها در نظر می‌گیریم (فرمول ۱) یعنی مثلا در مسائل طول قد دانشجویان

پی‌یرسن با حل معادله‌ی درجه ۹ بالا مقدار p_2 را به دست آورد. بعد از به دست‌آوردن مقدار p_2 مقدار متناظر با آن، یعنی p_1، از رابطه‌ی زیر حاصل شد:

$$p_1 = \frac{2\tilde{\mu}_3^3 - 2\tilde{\mu}_3\lambda_4 p_2 - \lambda_5 p_2^2 - 8\tilde{\mu}_3 p_2^3}{p_2(4\tilde{\mu}_3^2 - \lambda_4 p_2 + 2p_2^3)}$$

ما دو تا f داریم f_1 و f_2 و ضرایبشان هم می‌شود α و $1-\alpha$.

سپس با توجه به رابطه‌ای (۲) μ_1 و μ_2 را معادل ریشه‌های معادله‌ی (۳) در نظر گرفت:

$$\mu^2 - p_1\mu + p_2 = 0 \qquad (3)$$

پارامترهایی که داریم؟

و همچنین α و $\alpha - 1$ را معادل ریشه‌های معادله‌ی (۴) قرار داد:

$$\alpha^2 - \alpha - \frac{p_2}{{p_1}^2 - 4p_2} = 0 \qquad (4)$$

میانگین و انحراف معیار

در انتها نیز، ${\sigma_1}^2$ و ${\sigma_2}^2$ از رابطه‌ی زیر به دست آمدند:

$$(\mu_1\sigma_1)^2 = \frac{\tilde{\mu}}{\mu_1} - \frac{1}{3}\frac{\tilde{\mu}_3}{\mu_1\mu_2} - \frac{1}{3}(\mu_1 + \mu_2) + \mu_2$$
$$(\mu_2\sigma_2)^2 = \frac{\tilde{\mu}}{\mu_2} - \frac{1}{3}\frac{\tilde{\mu}_3}{\mu_1\mu_2} - \frac{1}{3}(\mu_1 + \mu_2) + \mu_1$$

دانشجویان پسر و دانشجویان دختر است. پس 5 پارامتر داریم این ساده ترین مدل آمیخته است.

پی‌یرسن بعد از حل معادله‌ی درجه‌ی ۹، دو مدل برای داده‌ها به دست آورد، سپس با رسم نمودارهای دو مدل، مشاهده کرد که هر دو مدل برای داده‌ها مناسب است. به عنوان معیاری برای مقایسه‌ی دو مدل، گشتاور مرتبه‌ی ششم دو مدل را به دست آورد و نتیجه گرفت که مدل ۱ نسبت به مدل ۲ بهتر است، زیرا گشتاور مرتبه‌ی ششم کمتری دارد. همانطور که مشاهده می‌شود روشی که پی‌یرسن برای برآورد گشتاوری پارامترهای مدل در نظر گرفته، برای داده‌های چندمتغیره، نیاز به محاسبات زیادی دارد که در عمل کاربرد چندانی نخواهد داشت. بدین منظور از روش ماکسیمم درستنمایی برای برآورد پارامترهای مدل استفاده می‌شود. برآوردگر ماکسیمم درستنمایی تحت برقراری شرایط نظم، کاراتر از برآوردگر گشتاوری می‌باشد. قابل ذکر است که معادلات فوق را می‌توان با استفاده از نرم‌افزارهای محاسبات جبری نظیر متمتیکا (Mathematica) حل کرد. در این جا معادلات به دست‌آمده از روش گشتاوری را در نرم‌افزار متمتیکا

k تعداد مؤلفه‌ها است.

حال اگر جامعه‌ی ما توزیع نرمال داشته باشد به این می‌گوییم mixture of gossian model

یعنی با Mixture معمولی

هیچ فرقی ندارد، تنها تفاوتش این است که تابع چگالی ها به صورت یک ترکیب محدب از توابع چگالی نرمال است که اینجا μ ها میانگین های ما است، σ^2 هم واریانس های ما است (صفحه ۲) ما اگر بخواهیم از جامعه آمیخته یک سری عدد تصادفی تولید کنیم یا شبیه‌سازی کنیم P

یعنی به ما بگویند مثلا متغیرهای x_1 تا x_n را شبیه‌سازی کنید!؟ آیا ما اصلا میتوانیم از یک جامعه شبیه‌سازی کنیم توی متلب مثلا دستور

rand (n) را استفاده میکردیم، که یک ماتریس به ما می‌دهد. حالا یکی از زیرمجموعه‌هایش را استفاده کنیم (در متلب کار کنیم) به عنوان یک بردار می

دهد یعنی اگر $n=20$ ... randn(1,20) ...

حالا سوال این است؟ ما اگر بخواهیم ۲۰ تا عدد از طول قد دانشجویان تولید کنیم، یک راه اش این است که یک کیت دم درب قرار دهیم و طول قد دانشجویان را به تصادف وقتی که وارد میشوند اندازه گیری کنیم، که هزینه بر است و استفاده نمی کنیم ولی از بررسی های قبلی ما میانگین و انحراف معیار طول قد دانشجویان را میتوانیم حدس بزنیم.

پس به جای اینکه برسیم داده ها را گردآوری کنیم برای کارهای تحقیقاتی آینده مان دقت کنید باید داده ها را شبیه سازی کنیم.

حل کرده و نتیجه به دست‌آمده را، در جدول ۱ با روش پی‌پرسن مورد مقایسه قرار داده‌ایم. همانطور که می‌بینیم، نتایج تا حدودی به یکدیگر نزدیکند.

حتی اگر ما بخواهیم یک روشی k-means که یک روش پارتیشنینگ است، یا یک روش سلسله مراتبی را به کار ببریم روی یک سری عدد یا داده هیچ راهی نداریم به جز اینکه بیاییم داده ها را با روش های آماری شبیه سازی کنیم.

جدول ۱: برآورد پارامترهای داده‌های پی‌پرسن

	برآورد پی‌پرسن		Mathematica	
پارامترها	مدل ۱	مدل ۲	مدل ۱	مدل ۲
α_1	۰/۴۱۴۵	۰/۴۶۷۲	۰/۴۲۴۰۶۱	۰/۴۶۷۰۲۸
α_2	۰/۵۸۵۵	۰/۵۳۲۸	۰/۵۷۵۹۳۹	۰/۵۳۲۹۷۲
μ_1	-۳/۵۱۷	۲/۱۷۶۹	-۳/۴۰۰۹۷	۲/۷۶۹۷۴
μ_2	۲/۴۹۰	-۲/۴۲۸	۲/۵۰۴۱۱	-۲/۴۲۷۰۴
σ_1	۴/۴۶۸۵	۲/۸۷۸	۴/۵۱۱۶۵	۲/۸۷۷۲۷
σ_2	۳/۴۱۱۵۴	۴/۱۷۷۰۲	۳/۱۰۹۴۱	۴/۷۷۰۲۵

پس این روش بین دو تا روش دیگر قرار میگیرد، یک روشی که از لحاظ فنی در آمار بررسی شود و یک روشی که کاملاً ریاضی است ولی در آن دو روش هیچ روشی برای (25) تولید اعداد تصادفی از چند تا کلاستر ما نداریم، ما هر وقت می‌خواهیم بیاییم عدد تصادفی تولید کنیم

۲.۳ روش ماکسیمم درستنمایی

روشی متداول در آمار برای برآورد پارامتر، روش ماکسیمم درستنمایی است. در این روش پارامترها به گونه‌ای برآورد می‌شوند که تابع درستنمایی مدل آمیخته‌ی گاوسی را ماکسیمم نمایند. تابع درستنمایی برای مشاهدات $x_1, \dots, x_n$ با تابع چگالی (یا تابع جرم احتمال) $f(x_i;\boldsymbol{\theta})$ با بردار پارامتر $\boldsymbol{\theta}$ می‌باشند، به صورت زیر تعریف می‌شود:

$$L(\boldsymbol{\theta}|x_1,\dots,x_n)=\prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

در این‌جا $\boldsymbol{\theta}$ مقداری مجهول، اما ثابت می‌باشد.

در بیشتر موارد برای سهولت کار $\ln L(\boldsymbol{\theta}|x_1,\dots,x_n)$ را ماکسیمم می‌کنند. تابع درستنمایی توزیع آمیخته‌ی گاوسی تک متغیره (مثال ۱) تحت تبدیل لگاریتمی به صورت زیر است:

به چه منظور؟ به این مقاله ... با آن یک کار خیلی ساده انجام بدهیم یا بیاییم دو روش یا دو مدل خوشه بندی را با هم دیگر مقایسه کنیم. ما مجبوریم داده ها را شبیه سازی کنیم، از هم باید شبیه سازی کنیم در کلاسترینگ؟ از تصادفی، تفکیک؟ باید بگوییم هیچ راهی نداریم و مجبوریم که

GMM یعنی توزیع های با اول معانی که آماری هستند. پس اینکه ما از این روش استفاده میکنیم برای تولید اعداد

$$\begin{aligned} l(\boldsymbol{\theta}) &= \ln L(\alpha,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2|x_1,\dots,x_n) \qquad (5)\\ &= \sum_{i=1}^{n}\ln\Big(\frac{\alpha}{\sqrt{2\pi\sigma_1{}^2}}\exp[-\frac{(x_i-\mu_1)^2}{2\sigma_1{}^2}]+\frac{(1-\alpha)}{\sqrt{2\pi\sigma_2{}^2}}\exp[-\frac{(x_i-\mu_2)^2}{2\sigma_2{}^2}]\Big)\end{aligned}$$

برای به دست آوردن برآوردگر ماکسیمم درستنمایی از معادله‌ی (۵)، نسبت به $\boldsymbol{\theta}=(\alpha,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)$ مشتق گرفته و ۵ معادله‌ی زیر حاصل می‌شود:

مهم مان برای تولید کلاسترهایی که میخواهیم بعداً روش هایمان را بر رویشان اعمال کنیم، باید حتماً از روش های آماری استفاده کنیم.

چرا این روش خوب است؟ به خاطر اینکه ما می آییم یک label به داده هایمان می دهیم، ببینید ما یک سری عدد از توزیع نرمال با یک پارامتر تولید می کنیم، یک سری عدد با پارامتر دیگر، اون اعدادی که با پارامتر اولی تولید کردیم، خودمان می دانیم که کدام اعداد است، در واقع label خوشه بندی ها را داریم، یعنی کاری که ما در خوشه بندی انجام می دهیم، می خواهیم کنار این x هایی که مشاهدات ما هستند یک عدد ۱ یا ۲ بگذاریم، که به آن متغیر پنهان می گویند hidden variable

یعنی متغیر پنهان در واقع اتفاقی که در خوشه بندی می‌افتد این است که ما یک بردار مشاهدات داریم که به ما label اش را نداده (به ما طول قد دانشجو را داده نگفته که دختر است یا پسر) در خوشه بندی هدف ما باید این باشد که با استفاده از داده‌ها بگوییم که این طول قدی که به ما داده طول قد مربوط به دختر بوده یا پسر، اگر ما این را مشخص کنیم، عملاً خوشه بندی انجام داده‌ایم. پس مسئله خوشه بندی این است که به ما یک سری عدد داده‌اند، ولی label هایشان را نداده‌اند و hidden و پنهان است، ما می‌خواهیم این hidden ها را برآورد کنیم.

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \alpha}=\sum_{i=۱}^{n}\Big(\frac{\frac{۱}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]-\frac{۱}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}\Big)=۰$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mu_۱}=\sum_{i=۱}^{n}\Big(\frac{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}][\frac{(x_i-\mu_۱)}{\sigma_۱^۲}]}{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}\Big)=۰$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \mu_۲}=\sum_{i=۱}^{n}\Big(\frac{\frac{\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}][\frac{(x_i-\mu_۲)}{\sigma_۲^۲}]}{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}\Big)=۰$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_۱^۲}=\sum_{i=۱}^{n}\Big(\frac{\frac{-\alpha}{۲\sqrt{۲\pi}}(\sigma_۱^۲)^{-\frac{۳}{۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}][\frac{(x_i-\mu_۱)^۲}{۲(\sigma_۱^۲)^۲}]}{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}\Big)=۰$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_۲^۲}=\sum_{i=۱}^{n}\Big(\frac{\frac{-(۱-\alpha)}{۲\sqrt{۲\pi}}(\sigma_۲^۲)^{-\frac{۳}{۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}][\frac{(x_i-\mu_۱)^۲}{۲(\sigma_۱^۲)^۲}]}{\frac{\alpha}{\sqrt{۲\pi\sigma_۱^۲}}\exp[-\frac{(x_i-\mu_۱)^۲}{۲\sigma_۱^۲}]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^۲}}\exp[-\frac{(x_i-\mu_۲)^۲}{۲\sigma_۲^۲}]}\Big)=۰$$

اصلاً بحث ما برآورد پارامتر μ و Σ و α نیست، این‌ها پارامتر در واقع رابطی هستند برای اینکه ما بتوانیم آن label ها را مشخص کنیم. هیچ وقت در خوشه بندی از ما نمی‌خواهند که بگوییم مثلاً μ و Σ و α چقدر است!! حالا α را گاهی اوقات ممکن است که بخواهند. به ما می‌گویند که داده‌ها را خوشه بندی کنیم که یعنی label هایشان را برآورد کنیم.

معادلات به دست‌آمده توابعی غیر خطی از پارامترها می‌باشند و ماکسیمم کردن آن‌ها به روش مستقیم امکان‌پذیر نمی‌باشد، همچنین برای برآورد پارامترها فرم تحلیلی وجود ندارد. از این رو این برآوردگرها نیز همیشه رفتار خوبی ندارند، [۳].

۱.۲.۳ برآورد پارامترهای مدل آمیخته‌ی گاوسی به روش عددی

از آن‌جا که برای دستیابی به برآورد ماکسیمم درستنمایی محاسبات زیادی مورد احتیاج است، می‌توان از روش‌های عددی، مقداری تقریبی برای برآورد پارامترهای مدل به دست آورد. دستور optim در نرم‌افزار R، با مینیمم کردن قرینه‌ی تابع درستنمایی، به روش عددی Nelder–Mead پارامترهای مدل را برآورد می‌کند. می‌توان برای برآورد پارامترهای داده‌ها به روش عددی از دستور بخش ۲.۴ استفاده کرد.
با توجه به این‌که روش‌های عددی نسبت به مقدار اولیه حساس هستند، از این رو نمی‌توان جواب‌های قابل قبولی از آن‌ها انتظار داشت. برای رفع این حساسیت، روش‌های استوارتری نسبت به روش‌های عددی ارایه شده که در بخش‌های بعد به معرفی آن‌ها می‌پردازیم.

کنار داده یک عدد بگذاریم که این شماره‌ی خوشه‌اش باشد. این عدد هم قراردادی است، ما می‌توانیم به جای عدد حرف بگذاریم، هر عددی دلمان می‌خواهد بگذاریم.

ما می‌خواهیم یک سری عدد تولید کنیم (می‌دهیم به) که بعداً بیاییم روش‌هایمان را روی‌شان آزمایش کنیم و ببینیم واقعاً این اعدادی که ما داریم کدام روش بهتر می‌تواند label ها را پیدا کند (هر کس یک روشی پیشنهاد می‌دهد که این label ها را مشخص کند، یک روش خوشه بندی داده‌هاست) معیار تست کردن این روش‌ها تولید داده‌ها از اعداد شناخته شده است البته Benchmark هم داریم، یعنی یک سری

داده های تست هم هستند که هم روشمان این کارها را انجام داده‌اند، اما به خاطر اینکه ارزش مان را نچسبه کنیم معمولا آن

نمیتوانیم از آن داده ها استفاده کنیم. اگر ارزش ما در این آزمایش، سرافراز خارج شد آن وقت ما میتوانیم روی داده های

واقعی هم کار کنیم. میخواهیم بیاییم از یک جامعه نرمال آمیخته، عدد تصادفی تولید کنیم، یعنی ما این آمیخته این کار را انجام میدهیم

میگوییم مثلا ۴۵٪ دخترها هستند، پس ما بیاییم ما ۴۵٪ داده هایمان را از توزیع نرمال با میانگین مثلا ۶۵ و ۱۶ بگیریم

و ۵۵٪ باقی را از میانگین

۲.۲.۳ برآورد پارامترهای مدل آمیخته‌ی گاوسی با الگوریتم EM

۶۷ و ۱۶ بگیریم این کار از نقطه نظر تئوری غلط است اما توی عمل بد نیست و جواب میدهد ولی از لحاظ ریاضی وقتی این را می‌نویسیم، این ارزش غلط است دقت کنید چرا؟

دمپستر و همکارانش (۱۹۹۷) الگوریتم EM را ارایه دادند. این الگوریتم روشی برای محاسبه‌ی برآوردگر ماکسیمم درستنمایی است، هنگامی که داده‌ی گمشده وجود داشته باشد یا روش‌های ساده‌ی بهینه‌سازی با شکست مواجه شوند، [۳]. از مهمترین کاربردهای الگوریتم EM ، یافتن برآورد پارامترهای مدل آمیخته متناهی می‌باشد.

برای یافتن پارامترهای مدل آمیخته گاوسی در این روش علاوه بر مجموعه مشاهدات، از متغیر تصادفی چندجمله‌ای z_i با احتمال

$$\begin{aligned} \alpha &= p(z_i = ۱) \\ ۱ - \alpha &= p(z_i = ۰) \end{aligned}$$

استفاده می‌شود. به z_i متغیر پنهان یا برچسب گفته می‌شود. به عبارت ساده‌تر با متناظر کردن یک برچسب به مشاهده‌ی x_i، می‌توان نشان داد که این مشاهده به کدام زیرجامعه تعلق دارد.

hidden variable

ما در ارزش یک متغیر تصادفی است، متغیر تصادفی برنالی است، یعنی ما x_1 را اگر می‌گیریم، یک آزمایش تصادفی باید انجام دهیم، شیر یا خط کنیم ولی احتمال شیر آمدن باید ۰٫۴۵ باشد نه ۵۰٪، اگر شیر آمد می‌گوییم این طول قد دخترها است، اگر خط

الگوریتم EM با در نظر گرفتن متغیرهای پنهان از چرخه‌ی مکرر برای برآورد پارامترها استفاده می‌کند. این الگوریتم با در نظرگرفتن مقدار اولیه برای پارامترهای مدل شروع می‌شود، که به این مرحله، مرحله‌ی آغازین گویند، درگام بعد که مرحله‌ی تکرار نامیده می‌شود، این پارامترها به روز می‌شود و چرخه تا جایی تکرار می‌شود که الگوریتم همگرا شود. مرحله‌ی تکرار از دو گام تشکیل می‌شود: محاسبه‌ی امید ریاضی و ماکسیمم‌سازی. درگام اول به جای محاسبه‌ی مستقیم تابع درستنمایی، امید آن بر حسب بردار متغیرهای پنهان $z = (z_1, \ldots, z_n)$ به صورت زیر محاسبه می‌گردد:

$$E_{f(z|x_1,\ldots,x_n,\theta^{(t)})}[\ln f(x_1, \ldots, x_n, z|\theta)]$$

درگام بعد پارامترهایی انتخاب می‌شوند که بر اساس آن‌ها امید به دست آمده از مرحله‌ی قبل ماکسیمم مقدار شود یا به عبارتی:

$$\theta^{(t+1)} = \sup E_{f(z|x_1,\ldots,x_n,\theta^{(t)})}[\ln f(x_1, \ldots, x_n, z|\theta)]$$

در اینجا منظور از $\theta^{(t)}$ برآورد θ در تکرار tاُم می‌باشد. از آن‌جا که مقدار تابع درستنمایی در هر تکرار افزایش می‌یابد، از این رو این الگوریتم، همگراست. بنابراین برآوردهای به دست آمده از این روش به مقدار ماکسیمم درستنمایی آن‌ها میل می‌کند.

در مثال ۱ فرض کنید که $\mu_1^{(t)}, \mu_2^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}, \alpha^{(t)}$ برآورد به دست آمده از مرحله‌ی tام الگوریتم باشند. سپس امید تابع درستنمایی مثال ۱ را که با $Q = Q(\alpha^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$ نشان

آمد می‌گوییم طول قد پسرها است. به عبارتی اگر ما، متغیر تصادفی تعریف کنیم طول قد دخترها (x_1)، متغیر تصادفی x_2 را

طول قد پسرها تعریف کنیم: $x_1|z$ دارای توزیع نرمال μ_1 و Σ_1 است، $x_2|z$ دارای توزیع نرمال μ_2 و Σ_2 است،

z همان آزمایش تصادفی است، یعنی اگر ما میخواهیم بیاییم، n تا عدد تصادفی تولید کنیم که شبیه سازی کنیم، Generate

کنیم از توزیع GMM با ضریب میکس ۰٫۴۵ و ۰٫۶، اول ما باید بیاییم مثلا اگر میخواهیم نمونه ۲۰ تایی بگیریم که ۲۰ تا برنولی تولید

کنیم، این ۲۰ تا برنولی را که تولید کردیم که (برنولی یا صفر می‌گیرد یا یک، یک موفق است و صفر شکست) اگر یک گرفت ما میدونیم یک

... طول قد دخترها که آن ... یک توزیع نرمال تولید میکنیم با ... طول قد ... و به این

روش ما میتوانیم عدد تصادفی از توزیع GMM تولید کنیم. نتیجه این روش با روشی که به صورت deterministic گفتیم، وقتی n زیاد شود، اصلاً محسوس نیست ولی از لحاظ ریاضی آن روش غلط است، این روش درست است، یعنی اگر ما یک کار دم دستی داریم، میخواهیم سریع کارمان را انجام بدهیم، میتوانیم α را بگیریم ۰٫۴۵ و اگر بخواهیم ۱۰۰ تا عدد تولید کنیم، ۴۵ تا با تولید قد دخترها تولید کنیم، ۱۰۰ تا با طول قد پسرها. ولی گزارش ندهید با این روش حل کردید چون از لحاظ نظری ایراد دارد.

می‌دهیم، می‌توان به صورت زیر نوشت:

$$
\begin{aligned}
Q &= E[\ln\prod_{i=1}^{n}(\frac{\alpha^{(t)}}{\sqrt{2\pi\sigma_1^{2(t)}}}\exp[-\frac{(x_i-\mu_1^{(t)})^2}{2\sigma_1^{2(t)}}])^{z_i}(\frac{1-\alpha^{(t)}}{\sqrt{2\pi\sigma_2^{2(t)}}}\exp[-\frac{(x_i-\mu_2^{(t)})^2}{2\sigma_2^{2(t)}}])^{1-z_i}\\
&= \sum_{i=1}^{n}E(z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)})(\ln\alpha^{(t)}-\frac{1}{2}\ln(2\pi\sigma_1^{2(t)})-\frac{(x_i-\mu_1^{(t)})^2}{2\sigma_1^{2(t)}})\\
&+ (1-E(z_i|x_i,\mu_2^{(t)},\sigma_2^{2(t)})(\ln(1-\alpha^{(t)})-\frac{1}{2}\ln(2\pi\sigma_2^{2(t)})-\frac{(x_i-\mu_2^{(t)})^2}{2\sigma_2^{2(t)}}) \qquad (6)
\end{aligned}
$$

ولی اگر بخواهیم کارمان را اصولی انجام دهیم، ما هم اول ۱۰۰ تا عدد تصادفی از توزیع برنولی تولید میکنیم، که احتمال موفقیت اش ۰٫۴۵ است، اگر یک مشاهده کردیم، متغیر آن می‌رود یک دونم نرمال تولید می‌کند و صفر مشاهده کردیم، متغیر آن می‌رود یک دونم نرمال تولید می‌کنیم. این کار را ما انجام دادیم و برنامه‌اش را خواهید دید که خط بسته نیست.

سپس $E(z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)})$ محاسبه می‌شود:

$$
\begin{aligned}
E(z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)}) &= f(z_i=1|x_i,\mu_1^{(t)},\sigma_1^{2(t)})\\
&= \frac{\alpha^{(t)}\phi(x_i;\mu_1^{(t)},\sigma_1^{2(t)})}{\alpha^{(t)}\phi(x_i;\mu_1^{(t)},\sigma_1^{2(t)})+(1-\alpha^{(t)})\phi(x_i;\mu_2^{(t)},\sigma_2^{2(t)})}
\end{aligned}
$$

و در نهایت از رابطه‌ی (۶) نسبت به پارامترها مشتق می‌گیریم.

$$
\begin{aligned}
\frac{\partial Q}{\partial\alpha^{(t)}} &= \frac{\sum_{i=1}^{n}E(z_i=1|x_i,\mu_1^{(t)},\sigma_1^{2(t)})-n\alpha^{(t)}}{\alpha^{(t)}(1-\alpha^{(t)})}\\
\frac{\partial Q}{\partial\mu_1^{(t)}} &= \frac{\sum_{i=1}^{n}E(z_i=1|x_i,\mu_1^{(t)},\sigma_1^{2(t)})(x_i-\mu_1^{(t)})}{2\sigma_1^{2(t)}}\\
\frac{\partial Q}{\partial\mu_2^{(t)}} &= \frac{\sum_{i=1}^{n}[1-E(z_i=1|x_i,\mu_2^{(t)},\sigma_2^{2(t)})](x_i-\mu_2^{(t)})}{2\sigma_2^{2(t)}}\\
\frac{\partial Q}{\partial\sigma_1^{2(t)}} &= \frac{\sum_{i=1}^{n}E(z_i=1|x_i,\mu_1^{(t)},\sigma_1^{2(t)})((x_i-\mu_1^{(t)})^2-\sigma_1^{2(t)})}{2\sigma_1^{4(t)}}\\
\frac{\partial Q}{\partial\sigma_2^{2(t)}} &= \frac{\sum_{i=1}^{n}[1-E(z_i=1|x_i,\mu_2^{(t)},\sigma_2^{2(t)})]((x_i-\mu_2^{(t)})^2-\sigma_2^{2(t)})}{2\sigma_2^{4(t)}}
\end{aligned}
$$

به گونه‌ی ساده‌تر می‌توان الگوریتم EM را برای مثال ۱، به صورت زیر بیان کرد:

الگوریتم EM برای مدل آمیخته‌ی گاوسی دو مولفه‌ای:

گام اول: انتخاب مقادیر اولیه برای پارامترهای مدل $\mu_1^{(t)},\mu_2^{(t)},\sigma_1^{2(t)},\sigma_2^{2(t)},\alpha^{(t)}$، به ازای $t=0$.

گام دوم: مرحله‌ی محاسبه‌ی امید ریاضی: محاسبه‌ی احتمال متعلق بودن مشاهده iام به مولفه اول (که

ما چهار روش می‌خواهیم بیاییم این مسئله را حل کنیم یعنی پارامترهای GMM وقتی که ۲ تا پارامترهم بسته ندارد (۲ تا مولفه بسته ندارد) بیاییم این پارامترها را برآورد کنیم، اول روش دقیق است؛ یعنی ما بیاییم exact (دقیق) این مسئله را حل کنیم، بعد می‌آییم این مسئله را به صورت تقریبی حل اش می‌کنیم و بعد می‌آییم با روش EM الگوریتم که روش نسبتاً آماری است حل اش میکنیم و بعد با استفاده از شبیه‌سازی که اصطلاحاً بهش میگویند Gip sampling با آن روش حل اش می‌کنیم. هر روش برای حل این مسئله واقعاً [illegible]

است، دو روش آخر است یعنی EM و GibsSampling است. که Gibs [redacted] یک نوع خاص MCMC

است. پس کاری که ما در واقع داریم انجام می‌دهیم در این صفحه، بررسی این دو روش است. اگر برگردیم به روش اول یا روش تحلیلی است

ببینید این مسئله را اصلاً نمی‌توانیم به روش تحلیلی حل کنیم (روش گشتاوری) پیرسون در سال ۱۸۹۴ برای اولین بار این کار

را انجام داد، یک سری داده داشت که نسبت طول پیشانی به طول بدن ۱۰۰۰ خرچنگ بود، منطقاً بیاید این خرچنگ‌ها را براساس

اینها به دو دسته تقسیم کند

جدول ۲: توزیع پیشین و توزیع پسین برای پارامترهای مدل آمیخته‌ی گاوسی $(k = 1,2)$

پارامترها	توزیع پیشین	توزیع شرطی کامل
α	$D(\delta_1, \delta_2)$	$D(\delta_1 + n_1, \delta_2 + n)$
μ_k	$N(\mu_0, \tau^2)$	$N(\frac{\tau^2 \sum_{i=1;z_i=k}^{n} x_i + \mu_0 \sigma_k^2}{n_k \tau^2 + \sigma_k^2}, \frac{1}{n_k \tau^2 + \sigma_k^2})$
σ_k^2	$IG(\omega_0, \beta_0)$	$IG(\omega_0 + \frac{1}{2} n_k, \frac{1}{2} \sum_{i=1;z_i=k}^{n} (x_i - \mu_k)^2 + \beta_0)$

البته این داده ها در سال ۱۸۹۲ توسط ولدن جمع آوری شده بود (صفحه ۳) او نه کاری که انجام داد، آمد داده ها را دسته بندی کرد و بعد برای مدل بندی آن ها دو تابع چگالی نرمال

آن را $\gamma_i^{(t)}$ نامیده‌ایم.)

$$\gamma_i^{(t)} = \frac{\alpha^{(t)} \phi_{\theta_1^{(t)}}(x_i)}{(1 - \alpha^{(t)}) \phi_{\theta_1^{(t)}}(x_i) + \alpha^{(t)} \phi_{\theta_2^{(t)}}(x_i)}$$

گام سوم: مرحله‌ی ماکسیمم‌سازی: محاسبه‌ی میانگین و واریانس بر حسب $\gamma_i^{(t)}$ به دست آمده از گام دوم.

$$\mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_i^{(t)} x_i}{\sum_{i=1}^{n} \gamma_i^{(t)}}, \qquad \mu_2^{(t+1)} = \frac{\sum_{i=1}^{n} (1 - \gamma_i^{(t)}) x_i}{\sum_{i=1}^{n} (1 - \gamma_i^{(t)})},$$

$$\sigma_1^{2(t+1)} = \frac{\sum_{i=1}^{n} \gamma_i^{(t)} (x_i - \mu_1^{(t)})^2}{\sum_{i=1}^{n} \gamma_i^{(t)}}, \qquad \sigma_2^{2(t+1)} = \frac{\sum_{i=1}^{n} (1 - \gamma_i^{(t)}) (x_i - \mu_2^{(t)})^2}{\sum_{i=1}^{n} (1 - \gamma_i^{(t)})},$$

$$\alpha^{(t+1)} = \sum_{i=1}^{n} \frac{\gamma_i^{(t)}}{n},$$

با میانگین های μ_1, μ_2 و واریانس های σ_1^2 و σ_2^2 را با ضرایب وزنی α و $1-\alpha$ با یکدیگر ادغام کرد. او از روش گشتاوری که ساده‌ترین روش آماری است برای برآورد پارامترهایش استفاده کرد. (ادامه در صفحه ۳ زیرش خط کشیدم)

گام چهارم: تکرار گام ۲ و ۳ تا رسیدن به همگرایی.

برای برآورد پارامترهای داده‌ها بر اساس الگوریتم EM کدهایی در بخش ۴.۳ در نرم‌افزار R نوشته شده است.

ما تا اینجا این که مطمئن شدیم پیرسون کارش را درست انجام داده، دستی که نمی‌توانستیم حلش کنیم چون خیلی سخت بود، آمدیم با Matematica این را انجام دادیم، دیدیم واقعاً جواب هایش درست بوده یعنی ۱۸۹۴ با دست

۳.۲.۳ برآورد پارامترهای مدل آمیخته‌ی گاوسی با روش بیزی

روش نمونه‌گیر گیبز از جمله روش‌های مونت کارلوی زنجیر مارکوفی می‌باشد که بر اساس توزیع شرطی مشاهدات، زنجیر مارکوفی از آن‌ها تولید می‌کند. این روش اولین بار در سال ۱۹۸۴ در مقاله‌ای توسط برادران گمن برای مدل‌های پردازش تصویر بیان شد، اما الگوریتمی که امروزه به عنوان الگوریتم گیبز در مسایل آماری از آن استفاده می‌کنیم، در سال ۱۹۹۰ توسط گلفند و اسمیت ارایه شد. نمونه‌گیری گیبز روشی برای تولید متغیرهای تصادفی بر اساس توزیع شرطی آن‌ها است، [۱۲].

در آمار بیز، توزیع پیشین مزدوج برای پارامترها در نظر گرفته می‌شود، [۱]. به پارامترهای توزیع

همچنین ۹ معادله پیچیده ای را آمده انجام داده و جوابی هم که گرفته درست بوده یعنی نتیجه ای که می‌توان گرفت این است که ما معمولاً وقتی GMM استفاده می‌کنیم ۴ مؤلفه در نظر می‌گیریم، اینجا ۲ تا در نظر گرفتیم، تو دو تا مؤلفه به معادله درجه ۹ رسیدیم، تو چهار تا مؤلفه به معادله درجه چند می‌رسیم؟!!! با تناسب نمیشه این را حل کرد اما میشه فهمید که اگر ما دو تا مؤلفه بیشتر داشتیم روش تحلیلی، اصلاً جواب نمیده و خیلی پیچیده می‌شود.

پیشین ابرپارامتر گفته می‌شود، بنابراین اولین گام برای اجرای الگوریتم گیبز تعیین مقادیر اولیه برای ابرپارامترها می‌باشد. در جدول ۲ توزیع پیشین برای مثال ۱ آورده شده است، [۵]. همان‌طور که می‌بینیم $\boldsymbol{\delta} = (\delta_1, \delta_2), \omega_0, \beta_0, \nu_0, \tau^2$ ابرپارامترها برای توزیع‌های پیشین در نظر گرفته شده‌اند. بعد از تولید متغیرهای تصادفی بر اساس توزیع پیشین، وارد مرحله‌ی تکرار می‌شویم. در اولین گام از تکرار tاُم، احتمال متعلق بودن هر مشاهده به دو زیرجامعه، بر حسب پارامترهای قبل‌تر یعنی مرحله‌ی $(t-1)$اُم محاسبه می‌گردد. سپس بر اساس این اطلاعات به دست آمده، پارامترها بر اساس توزیع شرطی کامل‌شان از جدول ۲ تولید می‌گردند. حال به کمک روش‌های بیزی و جدول ۲ الگوریتم نمونه‌گیری گیبز را برای برآورد پارامترهای مدل آمیخته‌ی گاوسی (مثال ۱)، می‌توان به صورت ساده‌تر بیان کرد:

الگوریتم نمونه‌گیر گیبز برای مدل آمیخته‌ی گاوسی دو مولفه‌ای:

- **مرحله‌ی آغازین:**

  در این مرحله مقادیر اولیه برای ابرپارامترهای توزیع پیشین $\boldsymbol{\delta} = (\delta_1, \delta_2), \omega_0, \beta_0, \mu_0, \tau^2$ انتخاب می‌شوند، سپس براساس این ابرپارامترها از توزیع پیشین داده تولید می‌کنیم. توجه کنید که در این‌جا مقدار اولیه برای ابرپارامترهای هر دو زیرجامعه یکسان در نظر گرفته شده است.

- **مرحله‌ی تکرار:**

  این مرحله برای هر t، $t = 1, 2, \ldots, T$؛ که T تعداد تکرار الگوریتم و بسته به نظر کاربر تعریف می‌شود، در سه گام انجام می‌شود:

گام اول: ابتدا z_i را از توزیع چند جمله‌ای با احتمال زیر تولید می‌شود.

$$p(z_i = 1|\mu_1^{(t-1)}, \sigma_1^{2(t-1)}, \alpha^{(t-1)}) = \frac{\alpha^{(t-1)}\phi_{\theta_1^{(t-1)}}(x_i)}{\alpha^{(t-1)}\phi_{\theta_1^{(t-1)}}(x_i) + (1-\alpha^{(t-1)})\phi_{\theta_2^{(t-1)}}(x_i)}$$

در این جا ذکر این نکته لازم است که $\mu_1^{(t-1)}, \sigma_1^{2(t-1)}, \alpha^{(t-1)}$ در تکرار اول یعنی $t = 1$ همان متغیرهای تصادفی تولید شده از توزیع پیشین در مرحله‌ی آغازین می‌باشند.

گام دوم:

$\alpha^{(t)}$، $\mu_1^{(t)}$، $\mu_2^{(t)}$، $\sigma_1^{2(t)}$ و $\sigma_2^{2(t)}$ بر اساس توزیع شرطی کامل‌شان در جدول ۲ تولید می‌شوند.

برای برآوردپارامترهای داده‌ها با استفاده از الگوریتم گیبز، کدهای بخش ۴.۴ در نرم‌افزار R نوشته شده است. نقطه‌ی داغیدن در این الگوریتم ۵۰۰ فرض شده است. در انتها، در بخش نتیجه‌گیری، نتایج به دست آمده از روش‌ها را با یکدیگر مورد ارزیابی قرار می‌دهیم و برتری‌ها و محدودیت‌های هر یک را برای کاربران مشخص می‌کنیم.

۴ نتیجه‌گیری

نتایج به دست آمده در جدول‌های ۳، ۴ و ۵ مربوط به داده‌های شبیه‌سازی شده از مدل آمیخته‌ی گاوسی دو مولفه‌ای به حجم ۱۰۰۰ در نرم‌افزار R می‌باشد. همان‌طور که می‌بینیم، برآوردها را در کنار مقدار واقعی پارامترها، انحراف معیار و میانگین مربع خطای نمونه‌ای در ر ستون نشان داده شده‌اند. میانگین مربع خطای نمونه‌ای به عنوان معیاری برای مقایسه‌ی دقت روش‌های شبیه‌سازی در نظر گرفته شده است، که از رابطه‌ی زیر به دست می آید:

$$\frac{۱}{۱۰۰۰}\sum_{i=۱}^{۱۰۰۰}(\hat{\theta}_i-\theta)^{۲}$$

که θ مقدار واقعی و $\hat{\theta}_i$ مقدارهای برآورد شده در هر الگوریتم می‌باشد. تعداد تکرار در هر روش ۱۰۰۰ در نظر گرفته شده و متوسط اختلاف مقدار واقعی و مقدار برآوردشده مبنای مقایسه روش‌هاست. همان‌گونه که انتظار داشتیم روش MCMC از EM و EM از عددی پاسخ جزئی بهتری به ما می‌دهد. اما پیاده‌سازی روش EM از عددی و گیبز از EM دشوارتر و شرایط همگرایی آن‌ها سنگین‌تر است. در حالت کلی برآورد بیز پاسخ قابل قبول تری در مقایسه با دو روش دیگر ارایه داده است. (که در این جا در مورد آن‌ها بحث نمی‌شود.)

در انتها نیز برای قابل لمس بودن موارد استفاده از مدل‌های آمیخته کاربردهایی از آن، ارایه شده است.

جدول ۳: برآورد ماکسیمم درستنمایی با روش عددی

پارامترها	مقدار واقعی	مقدار اولیه	مقدار برآوردشده	انحراف معیار	میانگین مربع خطای نمونه‌ای
$\alpha_۱$	۰/۷	۰/۵	۰/۵۷۳۳۱۸۹	۰/۲۴۵۸۲۱۲	۰/۱۴۷۷۶۹۸
$\alpha_۲$	۰/۳	۰/۵	۰/۴۲۶۶۸۱۱	۰/۲۴۵۸۲۱۲	۰/۱۴۷۷۶۹۸
$\mu_۱$	۳	۴	۲/۹۳۰۶۹	۰/۳۳۵۹۰۱۴	۰/۱۰۱۸۳۳۴
$\mu_۲$	۱	۱/۶	۱/۵۱۵۷۶۸	۰/۹۰۸۲۸۲۵	۰/۹۱۵۳۴۲۲
$\sigma_۱$	۱	۰/۳	۰/۸۷۵۷۶۲۹	۰/۲۲۶۰۸۹۵	۰/۱۶۱۴۳۹۶۸
$\sigma_۲$	۱.۲	۰/۸	۱/۲۵۵۱۶۸۸	۰/۳۵۳۷۲۸۷	۰/۸۷۷۷۲۲۶۸

کاربردها:

در سال‌های اخیر برای شناسایی گوینده (تشخیص صدا)، در متون مستقل، از مدل‌های آمیخته‌ی گاوسی استفاده می‌کنند. شناسایی گوینده، هنگامی که هیچ پیش فرضی از آن‌چه گوینده به زبان می‌آورد وجود ندارد، اصطلاحا شناسایی گوینده با متون مستقل گفته می‌شود. برای هر گوینده مدل آمیخته‌ی گاوسی به گونه‌ای در نظر گرفته می‌شود، که تابع احتمال پسین‌اش ماکسیمم مقدار شود. رینولد و همکارانش در سال ۱۹۹۵ در مقاله‌ی [۱۱] نشان دادند که مدل آمیخته‌ی گاوسی برای شناسایی گوینده در متون مستقل، مدلی

جدول ۴: برآورد ماکسیمم درستنمایی بر اساس الگوریتم EM

میانگین مربع خطای نمونه‌ای	انحراف معیار	مقدار برآوردشده	مقدار اولیه	مقدار صحیح	پارامترها
۰/۰۷۰۴۳۳۳۵	۰/۱۹۴۳۷۷	۰/۵۲۶۳۹۶	۰/۵	۰/۷	α_1
۰/۰۷۰۴۳۳۳۵	۰/۱۹۴۳۷۷	۰/۴۷۳۶۰۴	۰/۵	۰/۳	α_2
۰/۰۶۵۳۵۰۰۳	۰/۲۴۳۸۲۳۱	۳/۱۰۸۸۳۶	۴	۳	μ_1
۰/۷۷۰۷۰۷۵۵	۰/۳۱۹۹۵۳۱	۱/۸۲۳۷۵۶	۱/۶	۱	μ_2
۰/۱۲۸۴۹۵۱	۰/۳۳۴۴۰۵۷	۰/۷۲۶۲۶۳۲	۰/۳	۱	σ_1
۰/۷۳۱۸۶۴۶	۰/۲۱۳۹۵۶۳	۱/۹۸۶۶۵۲۱	۰/۸	۱/۲	σ_2

جدول ۵: برآورد بیز بر اساس الگوریتم گیبز

میانگین مربع خطای نمونه‌ای	انحراف معیار	مقدار برآوردشده	مقدار صحیح	پارامترها
۰/۰۴۲۳۴۱۱۵	۰/۱۱۶۴۴۱۷	۰/۶۳۷۲۹۱۵	۰/۷	α_1
۰/۰۴۲۳۴۱۱۵	۰/۱۱۶۴۴۱۷	۰/۳۶۲۷۰۸۵	۰/۳	α_2
۰/۰۴۵۰۳۸۲	۰/۱۴۶۳۴۳۴	۲/۹۴۸۳۰۵	۳	μ_1
۰/۳۴۶۹۳۰۳	۰/۲۱۶۹۶۳۴	۱/۲۷۹۸۰۰	۱	μ_2
۰/۰۴۳۲۵۲۸۸	۰/۱۰۹۲۲۳۱	۱/۰۸۴۸۲۸	۱	σ_1
۰/۶۸۲۷۹۸۳۱	۰/۲۹۵۵۳۸۳	۱/۳۷۷۲۹۶	۱/۲	σ_2

استوار می باشد.

علاوه بر تشخیص گوینده، از مدل آمیخته‌ی گاوسی برای شناسایی چهره‌ی افراد نیز استفاده می‌شود، [۱۰]. مهم‌ترین مشکل در تشخیص چهره افراد و بازیابی آن، سایه روشن‌ها، تغییرات نور و پس‌زمینه‌های همرنگ است . تشخیص چهره در مسایل امنیتی، تشخیص تغییرات در افراد و فهرست‌گذاری در تصاویر ویدئویی، کاربرد دارد. برای مدل‌بندی رنگ چهره‌ی (پوست) افراد نیز، می‌توان از مدل آمیخته‌ی گاوسی استفاده کرد.

در روش خوشه‌بندی مبتنی بر مدل، که برای مشاهدات، مدلی احتمالاتی در نظر گرفته می‌شود، از مدل آمیخته‌ی گاوسی استفاده می‌شود. بدین صورت که در این روش، هر خوشه به وسیله یک توزیع پارامتری نشان داده می‌شود. آن‌گاه مدلی که برای کل داده‌ها ارایه می‌شود ترکیب آمیخته‌ی متناهی از این توزیع‌ها، می‌باشد. با استفاده از مدل آمیخته‌ی گاوسی، اطلاعات کامل‌تری درباره‌ی خوشه‌ها به دست می‌آوریم.

به دلیل انعطاف‌پذیری مدل آمیخته‌ی گاوسی برای انواع مختلفی از توزیع‌ها، در یافتن الگوهایی برای امور مالی تجربی نیز، از مدل آمیخته‌ی گاوسی استفاده می‌شود. در مدل‌سازی مالی و کاربردهای آن، توزیع نرخ سود (بازده) در دارایی‌های مالی نقش مهمی دارد. متداول‌ترین فرض این است که نرخ سود دارایی‌ها، توزیع گاوسی دارد و از آنجا که دیگر توزیع‌ها نیز می‌توانند به خوبی با یک مدل آمیخته‌ی گاوسی متناهی تقریب زده شوند، این مدل در امور مالی مورد توجه بسیار قرار گرفته است.

همچنین مدل آمیخته‌ی گاوسی در نجوم، زیست‌شناسی، پزشکی و مهندسی نیز کاربرد بسیاری دارد که برای جزییات بیشتر می‌توان به [۴]، [۱۳]، [۷] و [۶] مراجعه نمود.

مراجع

[۱] Casella، G.، GeorgeSource E.I. ،(۱۹۹۲) Explaining the Gibbs Sampler، *The American Statistician*، ،۴۶ ۱۶۷-۱۷۴ .

[۲] Dempster، A.P.، Laird N.M. and Rubin D.B. ،(۱۹۷۷) Maximum likelihood from incomplete data via the EM algorithm، *journal of the Royal Statistical Society*، ،۳۹ ۱-۳۸ .

[۳] Everitt، B. S. and Hand D. J. ،(۱۹۸۱) *Finite Mixture Distributions*، Chapman and Hall.

[۴] Greenspan، H.، Goldberger، J. and Eshe، I. ،(۲۰۰۱) Mixture model for face-color modeling and segmentation، *Pattern Recognition Letters archive*، ۲۲ ، .۱۵۲۵-۱۵۳۶

[۵] Gonzalez، D.S.، Kuruoglu، E. and Ruiz، D.P. ،(۲۰۱۰) with mixture of symmetric stable distributions using Gibbs sampling، *Modelling Signal Processing*، ،۹۰ .۷۷۴-۷۸۳

[۶] Kon، S. ،(۱۹۸۴) Models of Stock Returns a Comparison، The *Journal of Finance*، ،۳۹ .۱۴۷-۱۶۵

[۷] McKenna، S.، Gong، S. and Raja، Y. ،(۱۹۹۸) Modelling facial color and identity with Gaussian mixtures، *Pattern Recognition*، ،۳۱ .۱۸۸۳-۱۸۹۲

[۸] Melnykov، V.، Maitra، R. ،(۲۰۱۰) Finite mixture models and model-based clustering، *Statistics Surveys*، ۴ ، .۱۱۶-۸۰

[۹] Pearson، K.، ،(۱۸۹۴) Contributions to the Mathematical Theory of Evolution Source ، *Philosophical Transactions of the Royal Society of London. A*، ،۱۸۵ .۷۱-۱۱۰

فرض کنید ما میخواهیم GMM تولید کنیم، ما میخواهیم دو جمله تولید کنیم، اینجا $\alpha=0.7$ ، $\mu_1=3$ ، $\mu_2=1$ ،
$\Sigma_1=1$ و $\Sigma_2=1.2$ ، میخواهیم بیاییم به اندازه‌ی n که اندازه‌ی نمونه ما است، بیاییم عدد تصادفی تولید کنیم (n مثلاً ۲۰.
۱۰۰) ما اول باید بیاییم n تا برنولی تولید کنیم، در این دو برنولی ندارد اما binomal دارد، binomal با پارامتر ① یعنی
برنولی، پس ما می‌نویسیم که n تا برنولی می‌خواهیم و α پارامتر موفقیت مان است (ادامه روی کدها نوشته شده)

[۱۰] Reynolds، D.A.، Quatieri، T.F. and Dunn R.B. ،(۲۰۰۰) Speaker Verification Using Adapted Gaussian Mixture Models، *Digital Signal Processing*، .۴۱–۱۹

[۱۱] Reynolds، D.A. and Rose، R.C. ،(۱۹۹۵) Robust text-independent speaker identification using Gaussian mixture speaker models، *IEEE Trans. Speech Audio Process* ، .۸۳–۷۲

[۱۲] Robert، C.P and Casella، G. ،(۲۰۰۴) *Monte Carlo Statistical Methods*، New York، Springer.

[۱۳] Titterington، D.M. ،Smith، A.F.M.، and Makov، U.E.(۱۹۸۵)، *Statistical Analysis of Finite of Mixture Distributions*، John Willey Son Ltd.

پیوست

۱.۴ کد تولید اعداد تصادفی مدل آمیخته گاوسی

```
#Generating random sample from Gaussian mixture of 2 component
#n: sample size
#alpha1:  mixing proprotions of  first component:
#mean1:   mean of first components.
#mean2:   mean of second components.
#sigma1:  sd of first component, default is 1.
#sigma2:  sd of second component, default is 1.
alpha=.7
mean1=3
mean2=1
sigma1=1
sigma2=1.2
z=rbinom(n,1,alpha)
x=rnorm(n, ifelse(z==1,mean1,mean2), ifelse(z==1,sigma1,sigma2))
```

بعد توزیع نرمال تولید میکنیم، چجوری تولید میکنیم؟ میگوییم نرمال هایی که تولید میکنیم باید این خاصیت را داشته باشد: اگر z ها ۱ بود، میانگین ۱،

۲.۴ کد برآورد پارامترهای مدل آمیخته گاوسی با روش عددی

اگر z یک نبود میانگین ۲، دوباره اگر $z=1$ بود انحراف معیار ۱ و اگر $z\neq 1$ بود، انحراف معیار ۲، با این ترفند می‌آییم GMM تولید

```
#At first it is defined likelihood function for maximizing
#par: parameter of model:
#par=c(mean1,mean2,sigma1,sigma2,alpha)
#x: univariate Gaussian mixture of k component.
Likelihood=function(par,x){
f=par[5]*dnorm((x-par[1])/par[3])/par[3]+
(1-par[5])*dnorm((x-par[2])/par[4])/par[4]
if(any (f<=0)) Inf
else -sum(log(f))}
#initial value
intpar=c(4,1.6,.3,0.8,0.5)
optim(intpar,Likelihood,x=x)$par
```

میکنیم با دو مؤلفه، برای سه مؤلفه هم به همین نحو ما
می‌توانیم در ریم جلو، به جای اینکه از binom استفاده
کنیم یک دستور داریم بنام multinominal، یک توزیع احتمالی
هست به نام multinominal از این توزیع عدد تولید میکنیم.

این کد برای نرم افزار R است، اما نرم افزار متلب هم میشه کار کرد. یک جزوه‌ای در google هست که اگر search کنیم R- matlab
میتوانیم دستورهای معادل در R و matlab را در آن بیابیم. وقتی میخواهیم با بردار کار کنیم بهتره با R کار کنیم و اگر خواستیم با ماتریس
ها کار کنیم از متلب استفاده کنیم. روش گشتاوری یک روش آماری، ساده‌ترین روش هست که با ماشین پارامترها را برآورد کرد.

۳.۴ کد برآورد پارامترهای مدل آمیخته گاوسی با روش الگوریتم EM

```
#Considering initial values for parameters:
#mean1=2,mean2=-2,sigma1=1,sigma2=1,alpha=0.6
#EM Algorithm:
Em=function(par){
#stage 1: (E-step)
gamma1=NULL
gamma1=par[1]*dnorm(x,par[2],sqrt(par[4]))/((1-par[1])*
dnorm(x,par[3],sqrt(par[5]))+par[1]*dnorm(x,par[2],sqrt(par[4])))
#stage 2: (M-step)
  par[1] <- (mean(gamma1))
  par[2] <- sum(gamma1*x)/sum(gamma1)
  par[3] <- sum((1-gamma1)*x)/sum(1-gamma1)
  par[4] <- sum(gamma1*((x-par[2])^2))/sum(gamma1)
  par[5] <- sum((1-gamma1)*((x-par[3])^2))/sum(1-gamma1)
  c(par[1],par[2],par[3],par[4],par[5])}
## Initial values for EM algorithm:
par0<-c(.5,4,1.6,.3,.8)
## Running EM algorithm
dis <- 1
iter <- 1
while (dis > 0.01 || iter <= 200){
  iter <- iter+1
  param <- Em(par0)
  dis <- max(abs(par0-param))
par0 <- param
}par0
```

۴.۴ کد برآورد پارامترهای مدل آمیخته گاوسی با روش الگوریتم گیبز

```
#iteration: number of repeat for algorithm.
##define initial values for parameters##
library(rgenoud)
library(multinomRob)
k=2
iteration=1000
mix.new=matrix(0,iteration,k)
mu.new=matrix(0,iteration,k)
var.new=matrix(0,iteration,k)
z.new=matrix(0,length(x),k)
mix=NULL
var=NULL
mu=NULL
mix=NULL
##define hyperparameter for prior distibution###
delta=rep(1,k)
mu.0=rep(1,k)
tau2=rep(9,k)
omega.0=rep(1,k)
betta.0=rep(1,k)
##generate random variable from prior distribution##
     for(i in 1:k){
```

```
        mix[i]<-rgamma(n=1,shape=delta[i],rate=1)
        mu[i]=rnorm(1,mu.0[i],sqrt(tau2[i]))
        var[i]=1/rgamma(1,omega.0[i],betta.0[i])
       }
mix=mix/sum(mix)
numer=matrix(rep(0,0),nrow=length(x),ncol=k)
## Iteration step#####
for(it in 1:iteration)
{
 ##find the latent variable z#####
   for(i in 1:k) {
      numer[,i]=(mix[i]*dnorm(x,mean=mu[i],sd=sqrt(var[i])))
                  }
   prob=numer/matrix(rep(rowSums(numer),k),ncol=k,byrow=F)
   z=matrix(0,length(x),k)
     for(j in 1:length(x)){
    z[j,]=t(rmultinomial(1,prob[j,]))
                 }
    n.mix=apply(z,2,sum)
 ##generate parameters from Posterior distribution####
  for(i in 1:k) {
  mix[i]=rgamma(1,shape=delta[i]+n.mix[i],rate=1)
  mu[i]=rnorm(1,(tau2[i]*sum(x[z[,i]==1])+mu.0[i]*var[i])/
  (n.mix[i]*tau2[i]+var[i]),sqrt((var[i]*tau2[i])/
  (n.mix[i]*tau2[i]+var[i])))
  var[i]=1/rgamma(1,shape=omega.0[i]+
 .5*n.mix[i],rate=betta.0[i]+
 .5*sum(z[,i]*(x-mu[i])^2))
          }
 mix=mix/sum(mix)
##Save###
 z.new=z.new+z
 mix.new[it,]=mix
 mu.new[it,]=mu
 var.new[it,]=var
}
##END OF ITERATION STAGE###
##Compute mean between estimated parameter #####
apply(mix.new[(iteration/2):(iteration-10),],2,mean)
apply(mu.new[(iteration/2):(iteration-10),],2,mean)
apply(var.new[(iteration/2):(iteration-10),],2,mean)
```

92,07,29 Data mining ادامه جلسه 7 (هفتم 5)

روش ماکزیمم درست نمایی یک روش دیگری است که ما میتوانیم با استفاده از آن پارامترها را به صورت دقیق‌تر برآورد کنیم، اساس این روش این است که ما تابع چگالی را به چشم یک تابعی از پارامتر ببینیم، چون تابع چگالی معنی و مفهوم احتمال دارد، اگر ما بتوانیم پارامتری انتخاب کنیم که این تابع چگالی بیشترین مقدارش را بگیرد، این برآوردی که پیدا میکنیم اسمش را می‌گذاریم، ماکزیمم درست نمایی، پیدا کردن ماکزیمم کار پیچیده‌ای نیست، در حالتی که مسئله پیچیده نباشد، یعنی وقتی ما می‌خواهیم likelihood function را بنویسیم ما آن را likelihood function می‌نویسیم likelihood function تابعی از θ (تتا) است و x_1 تا x_n که حاصل ضرب تابع چگالی هاست، که θ مقدار مجهول اما ثابت است. میخواهیم $L(\theta | x_1, \dots, x_n)$ (L به شرط x) را ماکزیمم کنیم، دقت کنید میخواهیم بگوییم به جای θ برحسب x چه مقداری بگذاریم که این L ماکزیمم شود، چرا این حتما خوب است؟! چون f احتمال است، اگر ما بتوانیم برحسب x به جای θ یک عددی بگذاریم، یک تابعی برحسب x بگذاریم که بعدا عدد میشه که این L ماکزیمم شود، یعنی احتمال زیاد شده است، یعنی ما آمدیم θ را بهتر برآورد کردیم، این معنی شهودی اش

$$L(\theta | x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i ; \theta)$$ فرمول ماکزیمم درست نمایی

است. در عمل خیلی خوب جواب می‌دهد، یعنی بهترین برآوردگری که ما معمولاً استفاده میکنیم در کارهایمان، برآورد ماکزیمم درست نمایی است. ایده‌اش خیلی ساده است و جزء قدیمی‌ترین برآوردگرهاست یعنی برآوردگر جدیدی نیست. مشکل ما در حل این مسئله چیست؟!! این است که ماکزیمم کردن θها که تا پارامتر (در ساده‌ترین حالت) ما داریم 2 تا مؤلفه را برآورد میکنیم که کار راحتی نیست، مثال ساده بزنیم: فکر کن که توزیع طول قد دانشجویان نرمال است، نرمال معمولی، یعنی توزیع طول قد پسرها نرمال است، یک پارامتر μ و σ^2 را برآورد کنید، برآورد کردن یعنی چی؟! یعنی به جایش یک تابعی از مشاهدات قرار دهید، به این میگیم برآورد، یعنی اگر ما بتوانیم یک تابعی از مشاهدات را برآورد کنیم که به جای μ و σ^2 قرار دهیم میشه برآورد پارامترها. روش ماکزیمم درست نمایی میگه ما بیاییم، چند نفر را دارید طول قدشان را اندازه‌گیری میکنید؟! میگوییم: 20 نفر، میگوید بیایید تابع چگالی این 20 نفر را در هم دیگر ضرب کنید، چرا؟! چون این تابع چگالی توأم شان بدست می‌آید، تابع چگالی توأم شان را به چشم یک تابعی از پارامترها نگاه کنید، یعنی پارامتر μ و σ^2، وقتی به این چشم نگاه کردیم، ما باید به جای μ و σ^2 برحسب x_1 تا x_n جور ماکزیمم کنیم این تابع را، خب چطور ماکزیمم می‌کنیم یک تابع دو پارامتری را؟! نسبت به μ مشتق میگیریم، نسبت به σ^2 مشتق میگیریم، مشتقات اول، مساوی صفر می‌گذاریم، برابر صفر، ماکزیمم را می‌دهد، خوشبختانه در اکثر همه‌ی این کارها بر وفق مراد است، ماست کنیم نقطه ماکزیمم را می‌دهد ☺

حالا چکار کنیم؟ مشتق بگیریم، برابر صفر قرار دهیم، و دستگاه حل کنیم، جوابش خیلی جالب میشود. μ برآوردش میشه $\bar{x}$

(1)

و ۲ کلاه‌ش $\frac{1}{n}$ را اساسیش $(x_i-\bar{x})^2$، بعضی مثلی S_{rB}، S_2^{B} (Baez) همان S_2 معمولی است (sample variance)
براوردگر

که به جای اینکه مخرج $n-1$ باشد مخرجش n است. تفاوت چندانی ندارند، اولی نااریب است، این اریب است. پس روش ماکزیمم likelihood روش خیلی پیچیده‌ای، به قول ما خیلی خفنی نیست. روش راحتی است، وقتی که یک مولفه داریم ولی ما وقتی چند مولفه داریم، مسئله ما سخت‌تر میشه. پیچیدگی کار در mixture of Gossian چیست؟! این است که بین تابع چگالی‌ها جمع است، اگر ضرب بود خیلی راحت این مسئله حل میشد ولی چون بین اینها جمع است، تابع چگالی‌ها را داریم در هم ضرب میکنیم، ضرب و جمع میشه، این سختی کار ما است به خاطر همین میگیم Σ قبل این در انتها می‌آید، یعنی اصلاً (همینطوری) در حالت تک متغیره نرمال مسئله خیلی ساده است. پس مشکل ما در دستگاه معادله likelihood این است که ما ۵ تا معادله داریم، ۵ تا مجهول، معادلات اینقدر پیچیده است که حتی با Matematica هم نمیشه حل اش کرد به صورت تحلیلی. ولی به صورت عددی میتوانیم این دستگاه را حل کنیم، ما در R یک دستوری داریم بنام optim شبیه این در متلب هم داریم، با دستور optim ما معادله را بدهیم، برایت می‌تواند ماکزیمم را بدست آورد بدون اینکه مشتق بگیرد، البته اگر بتوانیم گرادیان بهش بدهیم، دقیق‌تر این کار را انجام می‌دهد ولی بدون این مشتقات هم میتواند این کار را انجام دهد. حالا بروید کدش را ببینید، میخواهیم همان مسئله را با این روش حل کنیم که کدش را می‌بینید (کد 2.4 هفته ۱۴) وقتی ماکزیمم را بدست بیاوریم در واقع پارامترها را بدست آوردیم، و براساس پارامترها میتوانیم label ها را حساب کنیم و در نهایت خوشه‌بندی که هدف‌مان است را انجام دهیم.

داده پرت در توزیع نرمال وجود ندارد. وقتی ما فرض‌مان این است که GMM داریم، اینجا عملاً فرض بر این است که داده پرت نداریم. ولی روش‌هایی داریم که داده پرت را تشخیص دهیم ولی باید این‌ها را کنار بگذاریم.

کد 2.4: با استفاده از دستور optim داریم پارامترها را برآورد می‌کنیم، ما اسم تابع چگالی نرمال را می‌نویسیم بر حسب پارامترها علاوه تابع چگالی اولی یکی دیگر نرمال می‌کنیم بر حسب پارامترها منفی sum(log(f)) میکنیم، (دقت کنید ما وقتی میخواهیم یک تابعی را ماکزیمم کنیم، با اینکه بیاییم منفی آن را مینیمم کنیم، هیچ فرقی نمیکند. دستور optim مینیمم پیدا میکند، به خاطر همین ما گفتیم به جای اینکه خود تابع را ماکزیمم کنیم، می‌آییم لگاریتم اش را ماکزیمم می‌کنیم، چون لگاریتم یک تابع صعودی است. حالا اگر sum اش را بگیریم، میشه likelihood function مان. محدودیت روش عددی چیست؟! بزرگترین محدودیت روش‌های عددی، مقدار اولیه است. یعنی اگر ما مقدار اولیه در روش‌های عددی نداشتیم واقعاً ممکن بود هیچ وقت به جواب خوب نرسیم. پس ما باید یک مقدار اولیه بدهیم، معمولاً مقدار اولیه را براساس میانگین می‌گذاریم. α را نمی‌دانیم چقدر (62) α را وقتی نمی‌دانیم 0.5 قرار می‌دهیم، اگر می‌دانستیم دخترها از پسرها بیشتر بودند یا پسرها از دخترها بیشتر بودند، براساس آن مقدار α را عوض میکردیم. (۲)

مقدار α را اورپیب می‌کردیم به سمت اوتی که بهینه است. در مورد میانگین‌ها هم، حدس اولیه می‌زنیم، انحراف معیار هم همین‌طور
دستور optim صبر، چکار میکند؟ P صبر این ها تا پارامتر را برآورد می‌کند، به ما گزارش می‌دهد این روش ابزاری است
برای خوشه‌بندی، حالا روش‌های ترکیبی این روش Adak است، Adak آنها را میگذارم جدا، ما اصلاً در درس مان به روش
های Adak کاری نداریم. ما می‌آییم جدا الگوریتم‌های پارتیشنینگ را می‌گوییم، دیگه پایین ترتیبی بریم. وقتی که از optim استفاده
میکنیم این روش فوق‌العاده به مقدار اولیه حساس است یعنی اگر مقدار اولیه را عوض کنیم ممکن است به جواب درست نرسد اما در مطلب
آخر که گزارش داریم (همان مشتق) کار کنیم به مقدار اولیه کمتر حساس است روش. پس ما با روش عددی می‌توانیم پارامترها را تخمین
بزنیم، حالا روش عددی چه مشکلی دارد؟ مشکل عمده‌اش این بود که به مقدار اولیه شدیداً وابسته است، ما دلمان می‌خواهد روشی
بدهیم که به مقدار اولیه وابستگی‌اش خیلی کم باشد، EM الگوریتم، الگوریتمی هست که با استفاده از آن، ما می‌توانیم همان مسئله چیزیم
به اکرون را با ماکزیمم به اکران انجام دهیم بدون اینکه داده‌های اولیه با مقادیر اولیه خیلی وابسته باشند. این روش، روش مرسومی
است، خیلی هم کاربرد دارد و از روش استفاده میشود، ولی پیاده‌سازی اش به سادگی روش عددی نیست، یعنی همچنین یکم کدنویسی
دارد، ولی برتری اش این است که سرعتش بالاست و وابسته نبودن به مقدار اولیه. البته سرعتش از روش عددی کمتر است ولی باز
هم جزء روش‌های سریع است. حال روش EM الگوریتم را نمیخواهیم اینجا توضیح دهیم و نرم‌افزارهای ما معمولاً خودشان به صورت
اتوماتیک این کارها را انجام می‌دهند یعنی Model Based clustering را هم به صورت اتوماتیک انجام میدهند و نیازی نداریم که همیشه کد
بنویسیم ولی همین قدر که بدانیم چه اتفاق‌هایی داره پشت صحنه می‌افتد، این بد نیست.
روش آخر هم فقط اشاره میکنیم، روش گیبز سمپلینگ است (Gibs sampling) چکار میکند؟ در واقع ما تا الان مسئله را آمدیم
در چارچوب آمار کلاسیک حل کردیم، ولی از الان به بعد میخواهیم بیاییم مسئله را در چارچوب آمار بیز حل کنیم، مسئله آمار بیزی فرض
مطرح می‌شود که ما یک اطلاعات پیشینی هم در مورد پارامترها داشته باشیم، برای توزیع نرمال معمولاً این اطلاعات پیشین را طوری
در نظر میگیرند که اطلاعات پسین مان هم همان توزیع را داشته باشد یعنی توزیع پیشین و پسین مان یکی باشد، به این‌ها میگویند
توزیع‌های پیشین مزدوج، یک مسئله استاندارد است. حالا در این مسئله استاندارد ما می‌آییم توزیع پیشین را تعریف میکنیم، پسین
را بدست می‌آوریم ولی برآوردهای بیزی نمیتوانیم حساب کنیم، برآورد بیز را در حالت یک الگوریتم MCMC می‌رویم تخمین اش میزنیم که چون
روشی که استفاده میکنیم حالت خاصی از MCMC است به نام گیبز، بنابراین میگویند برآورد پارامترهای مدل آمیخته نرمال را
با روش گیبز سمپلینگ، این روش اصلاً به مقدار اولیه وابسته نیست، البته محدودیت این روش کندتر، کدنویسی اش است
ولی خوبیش دقت بالایی که دارد است. یعنی در این مقاله هم آخرش نشسته‌ایم که گرفته و با روش Gib sampling

از نظر دقت از بقیه خیلی روش بهتری است و مقادیر اولیه ای که داره حتی اگر شانسی به عنوان مقدار اولیه انتخاب کنیم، مقادیر
تصادفی هم
اولیه اش را شانسی دهیم، باز هم این روش همگرا میشود و جواب خوبی می دهد.
تصادفی
تمرین: تولید عدد از Mixture of Gassian ما در اینجا یک متغیره را گفتیم، شما چند متغیره را انجام دهید!!

(k-means-hour.pdf) Data mining 92/08/6 جلسه 8 (غایب)

یک کتاب داده کاوی دیگر که روش نوشتن به فارسی است میتوانید تهیه و استفاده کنید (که ترجمه Tan است) = کتاب داده کاوی کاربردی دکتر محمد صنیعی زاده دانشگاه تربیت مدرس.

این جلسه میخواهیم درمورد k-means صحبت کنیم، در واقع درمورد روش های clustering که صحبت کردیم، ما جلسه ی پیش یک اشاره ای کردیم به روش clustering، Model Based. در Model Based آمدیم گفتیم: این روش به خوبی روش سلسله مراتبی نیست ولی اگر که داده های ما فرض نرمال بودن را داشته باشند، جوابش با سلسله مراتبی تفاوت چندانی ندارد، و تقریباً به یک جواب می رسد اگر در سلسله مراتبی از norm اقلیدسی استفاده کنیم و آمدیم مسئله را با چهار روش سعی کردیم حل کنیم البته این بیشتر جنبه ی آموزشی داشت نه پژوهشی، که بدانیم یک مسئله را واقعاً ما میتوانیم با چهار روش بوک خیلی از جاها حل کنیم: به روش دقیق - با روش ریاضی - با الگوریتم EM و نهایتاً با استفاده از شبیه سازی. اینکه ما چطوری این کار را Model Based clustering انجام میدهیم از حوصله کلاس و درس ما خارج است. ولی ما باید بتوانیم با نرم افزار این کار را انجام دهیم، یعنی استاد از ما نمی خواهد بیاییم مثلاً یک برنامه بنویسیم که model Based clustering انجام بدهد، ممکن است که ما در درس clustering همچین کاری انجام بدهیم ولی در درس داده کاوی ما نمی خواهیم این قدر ریز بشویم و در clementine میتوانیم این کار (model Based clustering) را انجام دهیم. ولی سلسله مراتبی را استاد نمی خواهد که با دست انجام دهیم. امروز میخواهیم نوع سوم clustering بپردازیم البته انواع clustering سه روش نیست، سه روش عمده اش را ما به ترتیب میگوییم: اول = سلسله مراتبی، دوم = دقیق، سوم = Model Based، چهارم = آماری
پنجم = k-means که یکی نسل اول تعداز آماری است ولی این روش هم اگر که داده های ما از توزیع نرمال تبعیت کنند و norm ای که در نظر میگیریم، norm اقلیدسی باشد توی این k میانگین و یا اینکه فاصله ای که در این k میانگین در نظر میگیریم فاصله ی اقلیدسی باشد، آن موقع جواب این روش هم یکی میشود ولی خوبی k-میانگین این است که سرعتش از Model Based هم بیشتر است، بخاطر همین هم است که خیلی وقتها با الگوریتم در بحث شروع clustering همیشه ما با k میانگین شروع می کنیم، پس امروز ما میخواهیم روش سوم پارتیشنینگ را درمورد روش صحبت کنیم، روش سوم که پارتیشنینگ است.

4