تکلیف کامپیوتری سوم - درس شناسایی الگو

وحید ملکی
شماره دانشجویی: ۴۰۳۱۳۰۰۴

۱۶ آذر ۱۴۰۴

چکیده

در این تکلیف، هدف پیاده‌سازی و بررسی عملکرد الگوریتم EM (Expectation-Maximization) بر روی مدل ترکیبی گوسی (GMM) است. مسئله‌ی مورد مطالعه، Problem Annulus (مسئله‌ی حلقه) است که در آن نقاط داده‌ای به‌صورت یکنواخت در ناحیه‌ی بین دو دایره قرار دارند. با استفاده از GMM با $C = 30$ مؤلفه‌ی گوسی و ماتریس‌های کوواریانس قطری، پارامترهای مدل تخمین زده شده و همگرایی مدل از طریق رسم کانتورهای بیضی در تکرارهای مختلف و بررسی نمودار Log-Likelihood تحلیل می‌شود.

۱ مقدمه و شرح مسئله

مسئله‌ی اصلی در این تمرین، یافتن توزیع احتمال (و پارامترهای آن) برای مجموعه‌ای از داده‌های دوبعدی است که از یک توزیع ساده‌ی نرمال تولید نشده‌اند، بلکه ساختاری پیچیده (شکل حلقوی) دارند. مدل ترکیبی گوسی (Model Mixture Gaussian یا GMM) یک راهکار قوی برای مدل‌سازی این‌گونه توزیع‌های پیچیده است.

۱.۱ تولید مجموعه‌ی داده (Data Annulus)

مجموعه‌ی داده شامل $N = 900$ نقطه دوبعدی است که به‌صورت یکنواخت در ناحیه‌ی بین دو دایره (حلقه) با شعاع داخلی $r_{\text{inner}} = 1$ و شعاع خارجی $r_{\text{outer}} = 2$ قرار دارند. برای تضمین یکنواختی نقاط در مساحت حلقه (و نه صرفاً در شعاع)، از روش تولید شعاع تصادفی با توزیع مناسب استفاده شده است. این روش شامل تولید r از توزیع یکنواخت برای مربع شعاع، یعنی $r^2 \sim \text{Unfiorm}(r_{\text{inner}}^2, r_{\text{outer}}^2)$ است. زاویه‌ها نیز به‌صورت یکنواخت در بازه‌ی $(0, 2\pi)$ تولید شده‌اند.

۲.۱ تعریف مدل ترکیبی گوسی

مدل GMM با $C = 30$ مؤلفه‌ی گوسی دوبعدی تعریف شده است. پارامترهای مدل شامل موارد زیر هستند:

- وزن‌های ترکیبی: π_k، به‌طوری که $\sum_{k=1}^{C} \pi_k = 1$.
- میانگین‌ها: $\mu_k \in \mathbb{R}^2$.
- ماتریس‌های کوواریانس: Σ_k که ماتریس‌های قطری 2×2 هستند.

۲ پیاده‌سازی الگوریتم EM و منطق حل

از آنجایی که برچسب هر نقطه (اینکه از کدام مؤلفه تولید شده) مجهول است (data Missing)، از الگوریتم تکراری EM (Expectation-Maximization) برای تخمین پارامترهای $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{C}$ استفاده می‌شود.

۱.۲ مقداردهی اولیه (Initialization)

مقداردهی اولیه مطابق دستورالعمل تمرین انجام شده است:

- **میانگین‌ها (μ_k):** 30 نقطه‌ی داده به‌صورت تصادفی از مجموعه‌ی X انتخاب شدند.
- **وزن‌ها (π_k):** همگی به‌صورت مساوی $\pi_k = 1/C = 1/30$ تنظیم شدند.
- **کوواریانس‌ها (Σ_k):** ماتریس‌های قطری با واریانس نسبتاً بزرگ (براساس واریانس کل داده‌ها) مقداردهی شدند تا از بروز پدیده‌ی singularity جلوگیری شود.

۲.۲ مراحل تکراری الگوریتم

الگوریتم EM در هر تکرار دو گام اصلی را اجرا می‌کند:

۱. **گام امید (E-Step):** در این گام، مسئولیت‌ها (Responsibilities) یا γ_{ik} محاسبه می‌شود. این مسئولیت نشان‌دهنده‌ی احتمال تعلق نقطه‌ی x_i به مؤلفه‌ی گوسی k-اُم، با توجه به پارامترهای فعلی (θ^{old}) است.

$$\gamma_{ik} = P(z_i = k|x_i, \theta^{\text{old}}) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{C} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}$$

۲. **گام بیشینه‌سازی (M-Step):** با استفاده از مسئولیت‌های γ_{ik} محاسبه شده در گام E، پارامترهای جدید (θ^{new}) به‌گونه‌ای محاسبه می‌شوند که امید ریاضی لگاریتم شباهت (که در E-Step محاسبه شده) بیشینه شود.

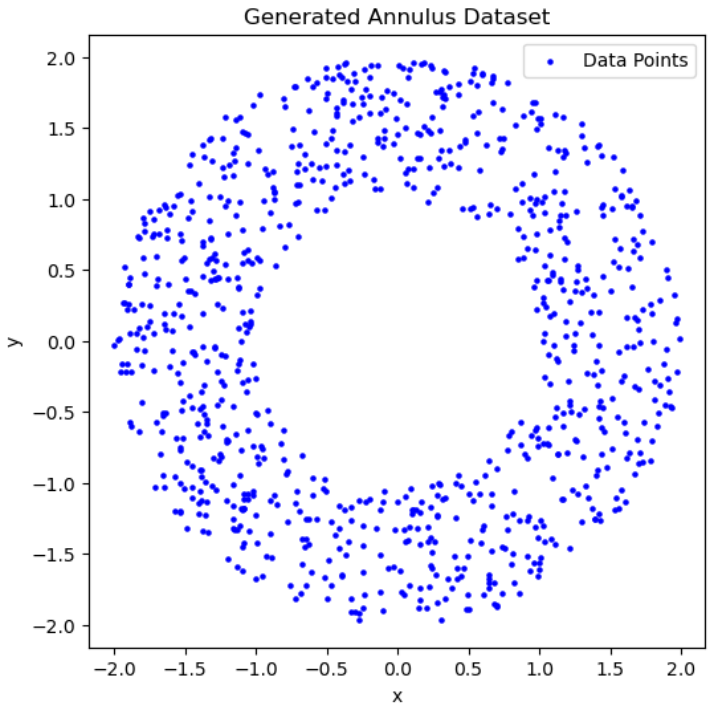
- تعداد مؤثر نقاط اختصاص داده شده به مؤلفه‌ی k: $N_k = \sum_{i=1}^{N} \gamma_{ik}$
- به‌روزرسانی میانگین: $\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i$
- به‌روزرسانی کوواریانس: $\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k^{\text{new}})(x_i - \mu_k^{\text{new}})^T$
- به‌روزرسانی وزن: $\pi_k^{\text{new}} = \frac{N_k}{N}$

در پیاده‌سازی، برای Σ_k^{new} تنها عناصر روی قطر اصلی نگهداری شدند تا شرط ماتریس کوواریانس **قطری** حفظ شود.

۳ تحلیل خروجی‌ها

خروجی‌های کد پایتون شامل سه بخش اصلی زیر است:

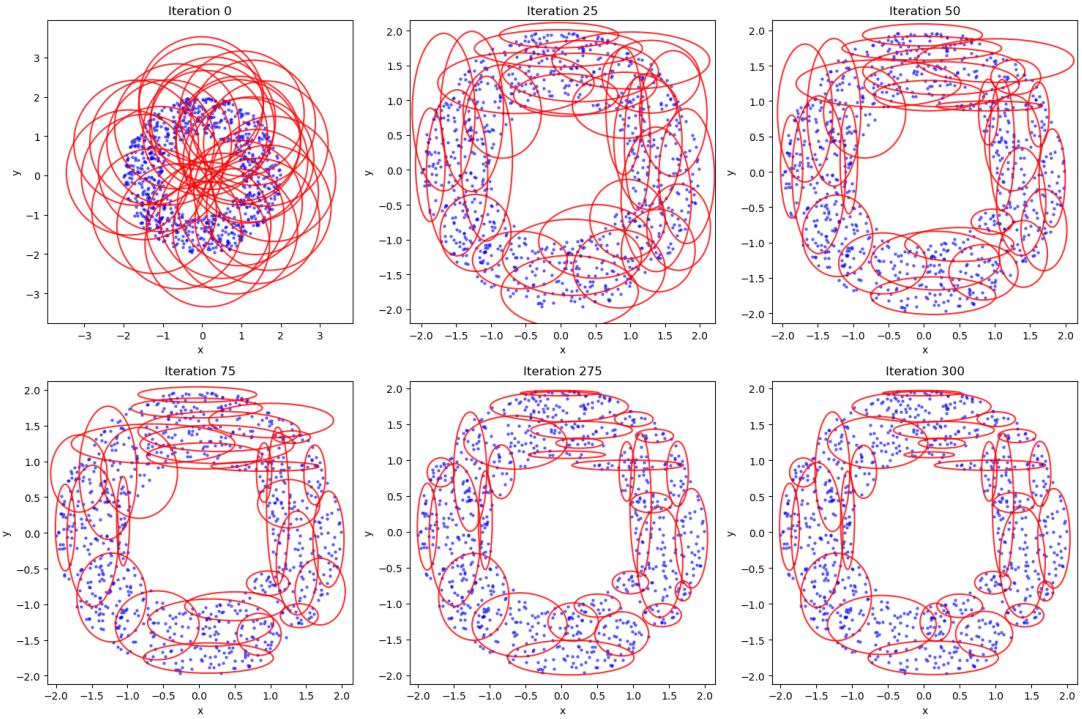
۱.۳ نمودار داده‌های تولید شده (Dataset Annulus Generated)



شکل ۱: نقاط داده‌ای Annulus تولید شده. یکنواختی سطحی حفظ شده و شعاع داخلی 1 و خارجی 2 به‌خوبی قابل مشاهده است.

تحلیل: همان‌طور که در شکل مشاهده می‌شود، $N = 900$ نقطه به‌صورت یکنواخت در فضای حلقوی بین $r = 1$ و $r = 2$ توزیع شده‌اند. این نشان می‌دهد که منطق تولید داده با استفاده از تبدیل تصادفی $r \sim \sqrt{U}$ به درستی اعمال شده است.

۲.۳ کانتورهای گوسی در تکرارهای مختلف

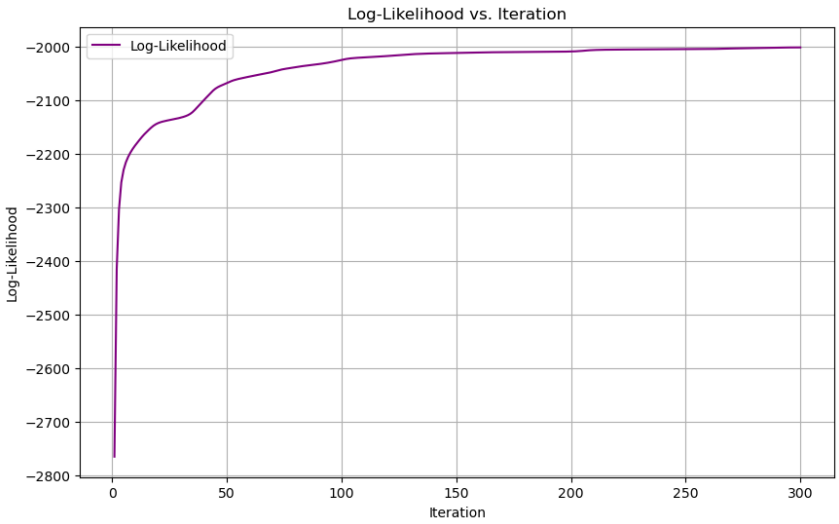


شکل ۲: بیضی‌های GMM در تکرارهای 0255075275 و 300. نقاط آبی داده‌ها و بیضی‌های قرمز، کانتور 2σ هر مؤلفه گوسی را نشان می‌دهند.

تحلیل همگرایی:

- **تکرار 0 (Initialization):** بیضی‌ها به‌صورت اولیه بزرگ و متمرکز در نزدیکی مرکز هستند و تقریباً همه‌ی ناحیه‌ی داخلی را پوشش می‌دهند، که نتیجه‌ی مقداردهی اولیه با واریانس بزرگ است.
- **تکرار 25 و 50 (Convergence Early):** مؤلفه‌ها به‌سرعت از مرکز فاصله گرفته و شروع به قرارگیری در ناحیه‌ی متراکم داده‌ها (حلقه) می‌کنند. بیضی‌ها در این مرحله، داده‌ها را به‌صورت نودولار (گره‌ای) پوشش می‌دهند.
- **تکرار 75 تا 300 (Convergence Final):** با پیشرفت الگوریتم، 30 مؤلفه‌ی گوسی به‌صورت منظم در امتداد حلقه پخش می‌شوند و هر مؤلفه مسئول مدل‌سازی بخشی کوچک از انحنای حلقه است. در تکرارهای 275 و 300، تفاوت بسیار اندک است که نشان‌دهنده‌ی رسیدن مدل به یک بیشینه‌ی محلی از شباهت (Likelihood Local Maximum) است. این رفتار مطلوب و مطابق با انتظار مسئله است.

۳.۳ نمودار Log-Likelihood در مقابل تکرار



شکل ۳: نمودار Log-Likelihood در برابر تکرارها.

تحلیل:

- **صعود مونوتون:** همان‌طور که در تئوری الگوریتم EM مورد انتظار است، مقدار Log-Likelihood در هر تکرار افزایش می‌یابد (صعودی یکنواخت یا Increase Monotonic).
- **همگرایی:** شیب افزایش در تکرارهای اولیه (تا حدود تکرار 100) بسیار تند است و پس از آن کاهش می‌یابد. نمودار در حوالی تکرار 200 تا 250 عملاً **اشباع (Saturation)** شده و مقدار آن به نزدیکی مقدار نهایی (≈ -2005) می‌رسد. این همگرایی قوی نشان‌دهنده‌ی موفقیت الگوریتم در تخمین پارامترها و رسیدن به بیشینه‌ی محلی شباهت است.