دانشکده مهندسی برق

تکلیف کامپیوتری درس بازشناسی آماری الگو، سری سوم

استاد

دکتر ابریشمی مقدم

نیمسال اول ۱۴۰۵-۱۴۰۴

مهلت تحویل: ۱۴۰۴/۹/۱۷

با سلام و آرزوی شادی، موفقیت و سلامتی؛

لطفا در تحویل پاسخ‌های خود موارد زیر را مدنظر داشته باشید:

- فقط برنامه‌هایی که به زبان ترجیحا پایتون و یا متلب باشند قابل قبول خواهند بود.
- تحویل همزمان گزارش و کدها الزامی است.
- گزارش باید شامل خروجی‌های کدهای نوشته شده باشد، که موارد خواسته شده در سوالات هستند و سایر توضیحات خواسته شده دیگر در متن سوالات نیز پاسخ داده شود (از آوردن کد در گزارش خودداری کنید).
- لطفا کدهای برنامه به صورت ماجولار و همراه با توضیحات کافی باشند؛ طوری که بخش‌های مختلف برنامه کاملا قابل تفکیک بوده و اجرا و ارزیابی هر بخش توسط کاربر به آسانی و بدون نیاز به ورود به جزییات برنامه میسر باشد.
- فایل تحویلی پاسخ شما باید تنها یک فایل زیپ، تحت عنوان "SPR_CHW3_Student ID" محتوی دو پوشه باشد. گزارش خود را در پوشه اول با عنوان "Report" و کدهای خود را ترجیحا به فرمت Jupyter Notebook در پوشه "Codes" قرار دهید.
- با این که همکاری، مشورت، و استفاده از ابزارهای کمکی در حل سوالات پیشنهاد می‌شود، حتما به صورت مستقل به نوشتن کدها و گزارش بپردازید.
- ممکن است از دانشجویی خواسته شود در زمانی که تعیین خواهد شد جزئیات کدش را در جلسه‌ای مجازی توضیح دهد، نتایج را تحلیل کند و حتی تغییراتی در پارامترهای کد اعمال کند. در صورتی که دانشجویی تمارین را تحویل داده باشد اما نتواند کد خود را توضیح دهد و یا تغییراتی روی آن اعمال کند، و یا اینکه کد یا گزارش تحویلی به پاسخ دیگران شباهت غیرمنطقی داشته باشد، نمره تمرین صفر لحاظ شده و نمره‌ای منفی هم لحاظ خواهد شد.
- در صورت وجود هرگونه سوال یا ابهام، مشکل مربوطه را با آی دی تلگرام زیر در میان گذارید:

@omid_Emaa

Gaussian Mixture Models and the Annulus Problem – Parameter Estimation with EM

در این تمرین یک سوال متفاوت مطرح میشود: اگر مجموعه دادهای را در اختیار ما قرار دهند، چگونه میتوان توزیع آنها (یا به طور دقیقتر پارامترهای توزیع یک کلاس به خصوص) را پیدا کرد. آیا میتوان با داشتن چند نمونه تصادفی، توزیعی را پیدا کرد و اقرار کرد که این نقاط از همین توزیع تولید شدهاند؟ این سوال، سوال سختی است! برای هر کلاس میتوان توزیعهای متفاوتی را در نظر گرفت و به این ترتیب، برازش بهترین و کارآمدترین مدل، کاری دشوار است. یک راهبرد آن است که یک توزیع پارامتری را در نظر گرفته و به دنبال انتخاب پارامتر θ باشیم. پس هدف ما از یافتن توزیع، به یافتن پارامترهای θ به نحوی که توزیع متناظر با آن بیشترین همخوانی و تناسب را با دادهها داشته باشند، تغییر میکند. برای رسیدن به این هدف، روشهای زیادی وجود دارد برای مثال اگر فرض کنیم دادههای هر کلاس از یک توزیع پارامتری ساده (مثلاً نرمال) تولید شدهاند، می‌توانیم با روش بیشینه شباهت (Maximum Likelihood Estimation – MLE)پارامترهای آن توزیع را تخمین بزنیم. اما در بسیاری از مسائل واقعی، دادهها توسط یک توزیع ساده تولید نشدهاند و بهتر است از مدل‌های ترکیبی استفاده کنیم. یک مثال مهم، مدل ترکیبی گوسی (Gaussian Mixture Model – GMM) است. اگر برچسب این‌که هر نقطه از کدام مؤلفه تولید شده را نمی‌دانیم(Missing data)، استفادهی مستقیم از MLE ساده نیست. در این حالت از الگوریتم بیشینه امید (Expectation Maximization – EM) برای تخمین پارامترهای مدل استفاده می‌کنیم. الگوریتم EM به صورت iterative شامل دو گام است : گام Expectation و گام Maximization

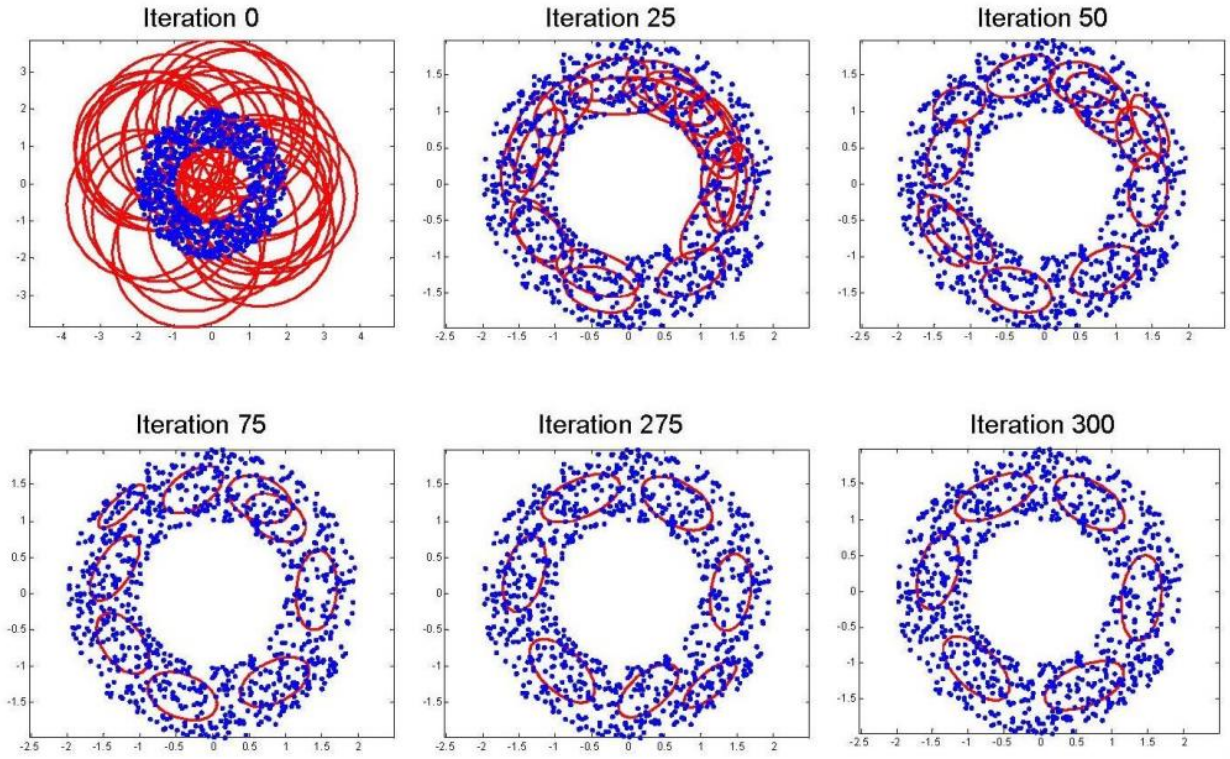
در این تمرین، هدف شما پیادهسازی و بررسی این الگوریتم روی یک مسئلهی مشهور به نام مسئلهی حلقه (annulus problem) است. در رابطه با الگوریتم EM در گزارش خود توضیح دهید و فرمول و نحوه کار ان را بیان کنید.

توضیح مسئله و دیتاست

در این تمرین یک **مجموعه‌داده‌ی مصنوعی دوبُعدی** تولید می‌کنیم که نقاط آن داخل یک حلقه (annulus) قرار دارند :

- شعاع داخلی حلقه r-inner = 1
- شعاع خارجی حلقه r-outer = 2
- تعداد نمونه‌ها N=900

نقاط روی این حلقه از یک توزیع تصادفی یکنواخت تولید می‌شوند (uniform pdf). سپس یک مدل ترکیبی گوسی با C=30 مؤلفه تعریف کرده و با الگوریتم EM پارامترهای آن را تخمین می‌زنیم. در طول آموزش، انتظار می‌رود مؤلفه‌های گوسی به‌تدریج روی حلقه قرار گرفته و شکلی نسبتا مشابه تصاویر زیر بسازند.



نمونه خروجی

تولید داده و آشنایی با آن

در پایتون یا متلب، N=900 نقطه دوبعدی تولید کنید که به‌طور یکنواخت در ناحیه‌ی بین دو دایره‌ی با شعاع‌های ۱ و ۲ قرار گیرند. راهنمایی : یک شعاع تصادفی r بین ۱ و ۲ (با توزیع مناسب برای یکنواختی سطحی) - یک زاویه تصادفیθ بین 0 و 2π سپس $x=r\cos\theta$ ، $y=r\sin\theta$.

نقاط تولید شده را در یک نمودار دوبعدی رسم کنید.(scatter plot) و محورهای x و y را هم‌مقیاس تنظیم کنید تا شکل حلقه به‌خوبی دیده شود.

تعریف مدل ترکیبی گوسی و مقداردهی اولیه

Use a GMM with **C = 30 Gaussian components** in 2D.

Parameterization:

- Mixture weights: $\pi_k, k = 1, \ldots, C$
- Means: $\mu_k \in \mathbb{R}^2$
- Covariances: Σ_k are **2×2 diagonal matrices**.

Initialization (following the slide description):

- Choose 30 random data points as the **initial means**.
- Initialize all mixture weights equally: $\pi_k = 1/C$.
- Initialize all covariance matrices as diagonal with relatively large variance on each dimension (e.g. based on the data variance), to avoid singularities.

در صورت ابهام به اسلایدهای درس مرتبط با مبحث زیر رجوع شود.

Mixture models and the annulus problem

پیاده‌سازی الگوریتم EM برای GMM

در این بخش، الگوریتم EM را برای تخمین پارامترهای مدل ترکیبی گوسی پیاده‌سازی می‌کنید. مراحل پیاده سازی خود را تشریح کرده و به مانند خروجی نمونه برای هر یک از تکرارهای- iteration (۰, ۲۵, ۵۰, ۷۵, ۲۷۵, ۳۰۰) نقاط داده (آبی) را رسم کنید. روی همان محورها، کانتورهای بیضی برای هر مؤلفه گاوسی (قرمز) رسم کنید، که نشان‌دهنده فاصله ماهالانوبیس ثابت از میانگین است (مثلاً ۲برابر انحراف معیار).

log-likelihood vs. iteration را رسم کنید و در مورد رفتار آن توضیح دهید.

"موفق باشید"