سلام بر تمام دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
۲. نکته ی مهم در گزارش نویسی روشن بودن پاسخ ها می باشد، اگر فرضی برای حل سوال استفاده می کنید حتما آن را ذکر کنید ،اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
۳. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده از کنید. شکل ها، به طور واضح و در فرمت درست گزارش شود.
۴. انجام هر 3 سوال کامپیوتری اجباری است و از بخش تئوری می توانید سوالات را به دلخواه حل کنید.
۵. **هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر میشود.**
۶. در صورت داشتن سوال، از طریق ایمیل prs.ftd@gmail.com ، سوال خود را مطرح کنید.

1– (15 نمره) فرض کنید تعدادی نمونه ی تصادفی داریم که از توزیع زیر پیروی می کنند:

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & otherwise \end{cases}$$

به صورت پیش فرض می دانیم θ پارامتری محدوده در بازه ی $0 \le \theta \le 10$ است.

الف) با استفاده از مجموعه داده ی $D = \{4,7,2,8\}$ که به صورت تصادفی از توزیع یاد شده استخراج شده اند و روش های بازگشتی بیز (Recursive Bayes Methods) ، توزیع $p(\theta|D)$ و $\hat{\theta}$ را با حداکثر دقت ممکن تخمین بزنید.

$* \; p(\theta|D^0) = p(\theta) = U(0,10)$

ب) با نتایج بخش قبل برای $p(x|D)$ توزیع مناسبی پیشنهاد دهید.

2–(20 نمره) فرض کنید x یک بردار باینری (0 و 1) به طول d با توزیع چند متغیره ی برنولی است.

$$p(x|\theta) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

همچنین $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^t$ یک بردار با d متغیر مجهول است که هر θ_i احتمال 1 بودن x_i متناظرش را نشان می دهد. تخمین احتمال بیشینه برای θ (تخمین maximum likelihood) را به دست آورید.

3-(15 نمره) . مجموعه داده های {(1,1), (3,3), (2,*)} از یک توزیع دو بعدی جدایی پذیر با توزیع $p(x_1, x_2) = p(x_1)p(x_2)$ به دست آمده اند. $p(x_2)$ و $p(x_1)$ به شکل زیر هستند:

$$p(x1) = \begin{cases} \frac{1}{\theta 1} e^{-\theta 1 x1} if\ x1 > 0 \\ 0\ otherwise \end{cases}$$

$$p(x2) = U\left(0, \theta 2\right) = \begin{cases} \frac{1}{\theta 2} & if\ 0 \le x2 \le\ \theta \\ 0 & otherwise \end{cases}$$

* نمایانگر یک مقدار ویژگی نامعلوم است .

الف) با یک گام اولیه $\theta_0 = \binom{2}{4}$ آغاز کنید و به صورت تحلیلی $Q(\theta, \theta_0)$ (مرحله E الگوریتم EM) را محاسبه کنید .
دقت کنید که نرمالیزیشن توزیع را لحاظ کنید.

ب) پارامتر های θ را طوری بیابید که $Q(\theta, \theta_0)$ را به دست آورید. (مرحله ی M الگوریتم EM)

پ) داده ها را روی یک نمودار دو بعدی نشان دهید و تخمین های جدید از پارامتر ها را نمایش دهید.

4- (15 نمره) مساله ی یادگیری میانگین توزیع تک متغیره را در نظر بگیرید. $n_0 = \sigma^2/\sigma_0^2$ را معادل dogmatism در نظر بگیرید و تصور کنید μ_0 از میانگین گیری از n_0 نمونه x_k برای k ها به صورت $k = -n_0 + 1, -n_0 + 2, \ldots$ به دست آمده اند.

الف) نشان دهید :

$$\mu_n = \frac{1}{n + n_0} \sum_{k=-n_0+1}^{n} x_k$$

و

$$\sigma_n^2 = \frac{\sigma^2}{n + n_0}$$

ب) از این نتیجه برای ارائه ی برداشتی از یک توزیع پیشین (prior) به صورت $p \sim N(\mu_0, \sigma_0^2)$ استفاده کنید.

5- (15 نمره) متغیر تصادفی با توزیع نرمال $N(\mu, \sigma^2)$ را در نظر بگیرید. قصد تخمین MAP برای پارامتر میانگین را داریم. توزیع پیشین میانگین را به صورت زیر در نظر بگیرید و مقدار تخمین MAP را به دست آورید.

$$f(\mu) = \frac{1}{\sigma^2{}_\mu}\mu \exp(-\frac{\mu^2}{2\sigma^2{}_\mu})$$

---

6 - (10 نمره) روش EM را برای توزیع پوآسون به دست آورید.

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

---

7 - (15 نمره) ماتریس های احتمال شرطی زیر تاثیر منطقه ی صید و آب و هوای فصل را بر نوع ماهی صید شده نشان می دهند. ماتریس های احتمال شرطی سطر بعد، مربوط به ویژگی های روشنایی (کم، متوسط، زیاد) و اندازه (لاغر یا پهن) بودن ماهی هستند . با توجه به آن ها به سوالات زیر پاسخ دهید.

$$P(x_i|a_j): \begin{array}{c} \\ \text{winter} \\ \text{spring} \\ \text{summer} \\ \text{autumn} \end{array} \begin{array}{c} \begin{array}{cc} \text{salmon} & \text{sea bass} \end{array} \\ \begin{pmatrix} .9 & .1 \\ .3 & .7 \\ .4 & .6 \\ .8 & .2 \end{pmatrix} \end{array}, \quad P(x_i|b_j): \begin{array}{c} \\ \text{north} \\ \text{south} \end{array} \begin{array}{c} \begin{array}{cc} \text{salmon} & \text{sea bass} \end{array} \\ \begin{pmatrix} .65 & .35 \\ .25 & .75 \end{pmatrix} \end{array}$$

$$P(c_i|x_j): \begin{array}{c} \\ \text{salmon} \\ \text{sea bass} \end{array} \begin{array}{c} \begin{array}{ccc} \text{light} & \text{medium} & \text{dark} \end{array} \\ \begin{pmatrix} .33 & .33 & .34 \\ .8 & .1 & .1 \end{pmatrix} \end{array}, \quad P(d_i|x_j): \begin{array}{c} \\ \text{salmon} \\ \text{sea bass} \end{array} \begin{array}{c} \begin{array}{cc} \text{wide} & \text{thin} \end{array} \\ \begin{pmatrix} .4 & .6 \\ .95 & .05 \end{pmatrix} \end{array}$$

الف) فرض کنید 20 دسامبر (اواخر پاییز و ابتدای زمستان) است، بنابراین در نظر بگیرید $p(a_1) = p(a_2) = 0.5$ ، ضمنا می دانیم که ماهی در آتلانتیک شمالی صید شده است. فرض کنید که روشنایی اندازه گیری نشده است، اما می دانیم که ماهی لاغر است. ماهی را به عنوان Salmon یا Sea bass طبقه بندی کنید. نرخ خطای مورد انتظار چقدر است.

ب) فرض کنید تمام چیزی که می دانیم این است که ماهی لاغر است و روشنایی اش متوسط است. به احتمال بیشتر الان کدام فصل از سال است؟ احتمال درست بودن این حدس را مشخص کنید.

پ) فرض کنید می دانیم ماهی لاغر است و روشنایی اش متوسط است و در آتلانتیک شمالی صید شده است، الان چه فصلی از سال است؟ احتمال درست بودن این حدس چقدر است؟

تمرینات کامپیوتری

8 - (20 نمره) در این سوال قصد داریم با استفاده از پیاده سازی الگوریتم EM و تخمین مدل GMM به طبقه بندی تصاویر بپردازیم. برای سادگی داده های دو کلاس (نیم فوتبال منچستر و چلسی) را بررسی می کنیم که بتوانیم از دو ویژگی (feature) R و B (از RGB) به عنوان فیچر ها بهره بگیریم.

الف) با در نظر گرفتن k=2 به عنوان تعداد مولفه ها (components)، الگوریتم EM را برای تخمین پارامتر های توزیع GMM مربوط به هر یک از دو کلاس پیاده سازی کنید. پارامتر های به دست آمده برای GMM هر کلاس را در گزارش کار خود ذکر کنید. نمودار های داده های هر دو کلاس و کانتور های مدل های GMM فیت شده به آن ها را رسم کنید.

ب) بخش الف را برای چند مقدار مختلف k تکرار کنید. علاوه بر خواسته های بخش الف، نمودار AIC و BIC بر حسب تعداد مولفه ها را رسم کرده و تعداد بهینه ی k را پیدا کنید.

---

9 - (20 نمره) در این سوال با دیتاست penguin کار می کنیم. در این دیتاست داده های مربوط به 6 ویژگی مختلف سه گونه پنگوئن فراهم شده است. سه گونه پنگوئن Gentoo، Adelie و Chinstrap با ویژگی های:

- **culmen_length_mm**: culmen length (mm)
- **culmen_depth_mm**: culmen depth (mm)
- **flipper_length_mm**: flipper length (mm)
- **body_mass_g**: body mass (g)
- **island**: island name (Dream, Torgersen, or Biscoe) in the Palmer Archipelago (Antarctica)
- **sex**: penguin sex

در این سوال می خواهیم با بررسی دو به دو تعدادی فیچر ها به بهترین زوج فیچر برای جداسازی گونه های مختلف پنگوئن برسیم. چهار حالت مختلف به شرح زیر را بررسی می کنیم:

1- Culmen_length_mm, culmen_depth_mm

2- Flipper_length_mm, Culmen_length_mm
3- Body_mass_g, Flipper_length_mm
4- Flipper_length_mm, culmen_depth_mm

الف) scatter plot مربوط به هر جفت فیچر را رسم کنید. توجه کنید که محورها لیبل مناسب داشته باشند و هر نمودار عنوان مشخص داشته باشد. برای داده های سه گونه ی مختلف، سه رنگ مختلف در نظر بگیرید و آنها را با تابع legend مشخص کنید. با توجه به نمودارها تحلیل کنید در کدام یک GMM، discriminability بهتری ایجاد می کند.

ب) برای هر یک از کلاس ها در هر حالت یک مدل GMM فیت کنید و پارامترهای آن را در گزارش خود ذکر کنید. هم چنین کانتورهای آن را روی نمودار های scatter plot رسم کنید.

پ) برای مدل های گوسی فیت شده، خطاها را با هم مقایسه کنید و بهترین حالت را با ذکر دلیل تعین نمایید.

ت) برای بهترین حالت تعداد مولفه های گوسی را بالا ببرید (2 و 3 و 4 و5) و نمودار AIC و BIC بر حسب تعداد مولفه ها را رسم کنید. عملکرد ها را مقایسه کنید و بهترین تعداد مولفه های گوسی را تعیین کنید.

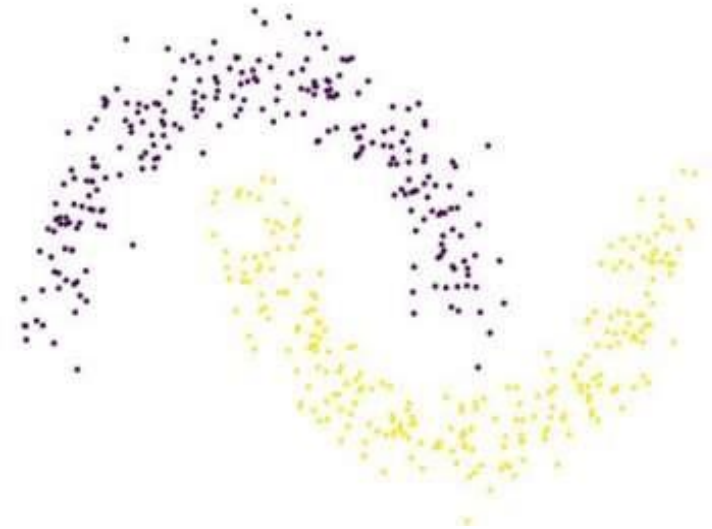
---

10 – (20 نمره) ابتدا دیتاست شکل زیر را با استفاده از قطعه کد زیر ایجاد کنید.

```
from sklearn import cluster, datasets, mixture
noisy_moons=datasets.make_moons(n_samples=500, noise=0.11)
```

فایل moons.csv برای استفاده در متلب آپلود شده است. ستون سوم شامل لیبل نقاط است. داده های حاصل مطابق شکل 1 هستند.

* در این سوال الگوریتم خواسته شده را باید خودتان پیاده سازی کنید و مجاز به استفاده از کتابخانه های آماده نیستید.

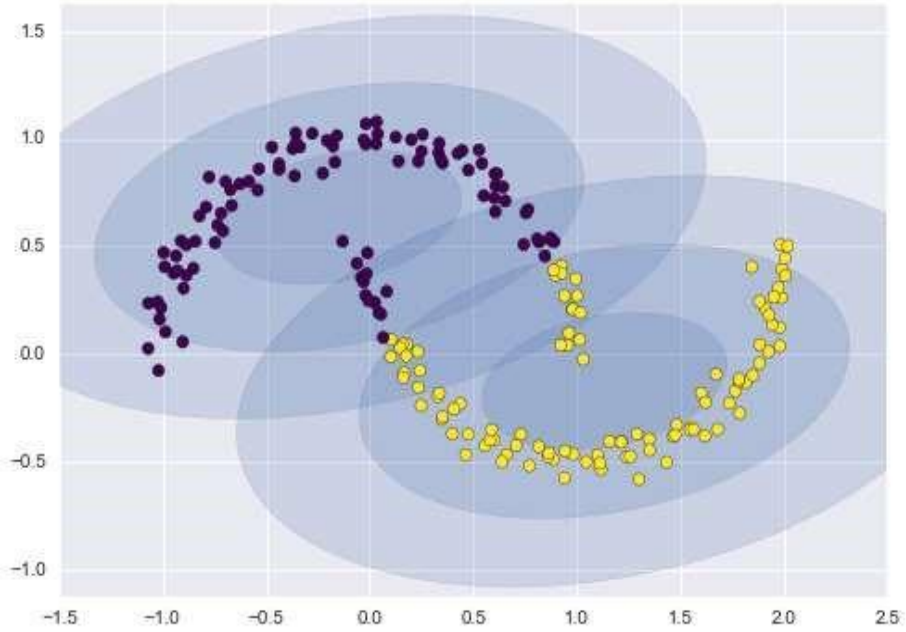


شکل 1

حال در روش تخمین بیزی :

الف) یک بار هر کلاس را با توزیع نرمال تقریب بزنید و پارامترهای آن را به دست آورده و کانتورهای مربوطه را رسم نمایید. در شکل زیر یک نمونه برای تعداد مولفه ی مساوی 2 آورده شده است.

ب) این بار از روش های GMM استفاده کنید. در روش GMM با تعداد مولفه های مختلف 1 تا 16 تست کنید و شکل داده ها و کانتورها را برای تعداد مولفه های برابر 3، 8 و 16 بیاورید. سپس مانند شکل زیر نمودارهای AIC و BIC را بر حسب تعداد مولفه ها رسم کرده و تعداد بهینه ی مولفه ها را بیابید.



شکل 2