بسمه تعالی

دانشکده مهندسی برق و کامپیوتر

تمرین سری سوم درس یادگیری ماشین

دانشگاه تهران

پردیس دانشکده‌های فنی

سلام بر تمام دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید 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. از بین سوالات تحلیلی حتما به سه مورد پاسخ داده شود. از بین سوالات کامپیوتری پاسخ به سوالات ۷ و ۹ و ۱۰ لازم و بقیه اختیاری است. حداکثر تا نمره ۱۲۰ (۲۰ نمره امتیازی) لحاظ خواهد شد.
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 mo.pourrahimi@gmail.com ، سوال خود را مطرح کنید.

---

۱. مجموعه داده های $\{(1,1), (3,3), (2,*)\}$ از یک توزیع دو بعدی جدایی پذیر با توزیع $p(x_1, x2) = p(x_1)p(x_2)$ به دست آمده اند. $P(x_1)$ و $p(x_2)$ به شکل زیر هستند. (۲۰ نمره)

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

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

* 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) پارامترهای θ را طوری بیابید که $Q(\theta,\theta^0)$ را به دست آورید (مرحله M الگوریتم EM).

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

---

۲. X یک توزیع یکنواخت به شکل زیر دارد. (۲۰ نمره)

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

الف) $D = \{x1, x2, ..., xn\}$ را به عنوان مجموعه n تا نمونه در نظر بگیرید که به صورت مستقل از توزیع $p(x|\theta)$ به دست آمده اند. نشان دهید که تخمین Maximum Likelihood برای θ برابر $\max[D]$ است.
ب) فرض کنید $n = 5$ تا نمونه از این توزیع برداشته ایم که بیشینه آن ها برابر 0.6 است. نمودار likelihood ($p(D|\theta)$) را در بازه $0 <= \theta <= 1$ رسم کنید. توضیح دهید که چرا نیازی به دانستن بقیه نمونه ها ندارید.

---

۳. X یک بردار d بعدی باینری (عناصر آن ۰ یا ۱ هستند.) با توزیع برنولی چندمتغیره به شکل زیر است. (۲۰ نمره)

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

که در آن $\theta = (\theta_1, ..., \theta_d)^t$ بردار پارامتر های نامعلوم است. θ_i احتمال یک بودن x_i است.تخمین ML را برای θ به دست آورید.

---

۴. مساله یادگیری میانگین توزیع نرمال 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 $k = -n_0 + 1, -n_0 + 2, ..., 0$ به دست آمده اند. (۲۰ نمره)

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 برداشتی از یک توزیع پیشین (prior) به صورت $p \sim N(\mu_0, \sigma_0^2)$ استفاده کنید.

---

۵. متغیر تصادفی x 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. (۲۰ نمره)

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

---

۶. روش EM را برای توزیع پوآسن به دست آورید. (۱۰ نمره)

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

---

۷. (کامپیوتری) 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 (تیم فوتبال منچستر و چلسی) را بررسی می کنیم که بتوانیم از دو فیچر R, B (از RGB) به عنوان فیچر ها بهره بگیریم. (۲۰ نمره)

الف) با در نظر گرفتن k = 2 به عنوان تعداد مولفه های (component)، الگوریتم EM را برای تخمین پارامتر های توزیع های GMM مربوط به هر یک از دو کلاس پیاده سازی کنید. پارامتر های به دست امده برای GMM مربوط به هر کلاس را در گزارش 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) الف را برای چند مقدار مختلف k تکرار کنید. علاوه بر خواسته های الف، نمودار AIC و BIC بر حسب تعداد مولفه ها را رسم کرده و تعداد بهینه k را تعیین کنید.

---

۸. (کامپیوتری) در این سوال با دیتاست penguin کار می کنیم. در این دیتاست داده مربوط به ۶ ویژگی مختلف سه گونه پنگوئن فراهم شده است. (۲۰ نمره، اختیاری)

سه گونه پنگوئن Chinstrap, Adélie, or Gentoo 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.

۱) Culmen_length_mm, culmen_depth_mm
۲) Flipper_length_mm, culmen_length_mm
۳) Body_mass_g, flipper_length_mm
۴) Flipper_length_mm, culmen_depth_mm

الف) scatter plot مربوط به هر جفت فیچر را رسم کنید. توجه کنید که محور ها لیبل مناسب داشته و هر نمودار عنوان مشخص داشته باشد. برای داده های سه گونه مختلف، سه رنگ مختلف در نظر بگیرید و با legend آن ها را مشخص کنید. با توجه به نمودار ها 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 پارامتر های آن را در گزارش خود بیاورید. هم چنین کانتور های آن را روی نمودار های scatter plot رسم نمائید.

پ) برای مدل های گوسی فیت شده، خطا ها را با هم مقایسه کنید و بهترین حالت را با ذکر دلیل تعیین نمائید.

ت) برای بهتری حالت، تعداد مولفه های گوسی را بالا ببرید (۲و۳ و۴ و۵) و نمودار AIC, BIC بر حسب تعداد مولفه ها را رسم کنید. عملکرد ها را مقایسه کرده و بهترین تعداد مولفه های گوسی را تعیین نمائید.

---

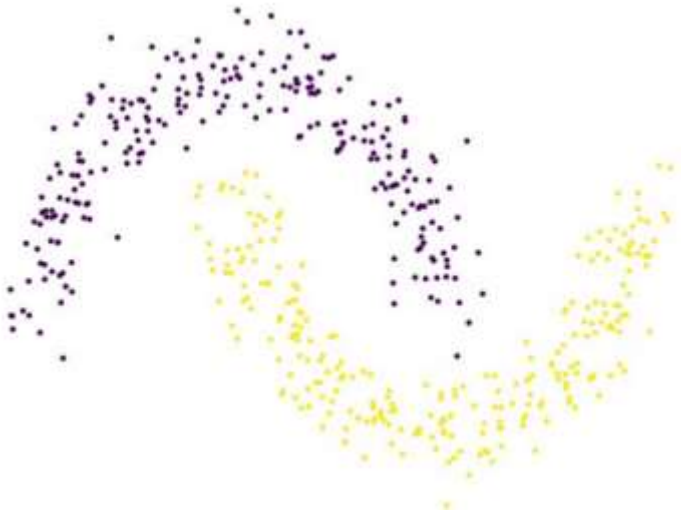
۹. (کامپیوتری) ابتدا دیتاست شکل زیر را با استفاده از قطعه کد زیر ایجاد کنید. (۱۰ نمره)

```
from sklearn import cluster, datasets, mixture

noisy_moons = datasets.make_moons(n_samples=500, noise=.11)
```

فایل moons.csv برای استفاده در متلب آپلود شده است. ستون سوم شامل لیبل نقاط است. داده های حاصل مطابق شکل ۱ هستند.

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



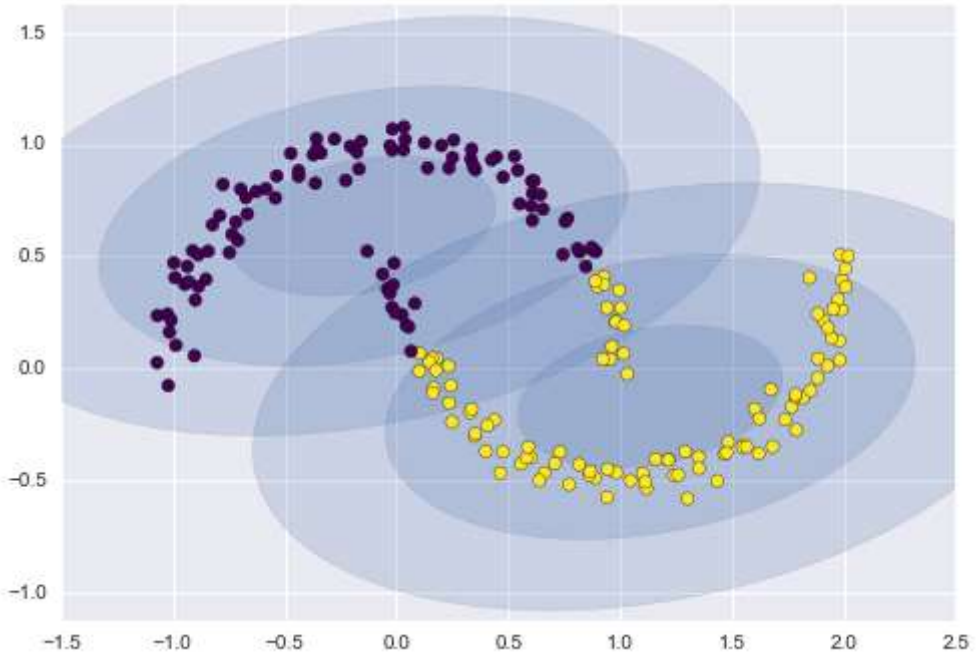
شکل ۱

حال در روش تخمین بیزی:

الف) یک بار هر کلاس را با یک توزیع نرمال تخمین بزنید و پارامتر های آن را به دست آورده و کانتور های مربوطه را رسم کنید. در شکل زیر یک نمونه برای تعداد مولفه برابر ۲ آورده شده است.

ب) این بار دیگر از روش GMM استفاده کنید. در روش GMM با تعداد مولفه های مختلف ۱-۱۶ تست کنید و شکل داده ها و کانتور ها را برای تعداد مولفه های برابر۳و ۸ و ۱۶ بیاورید. شکل زیر نمودار AIC, BIC بر حسب تعداد مولفه ها را رسم کرده و تعداد بهینه مولفه ها را تعیین کنید.

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



شکل ۲

---

۱۰. **(کامپیوتری) پردازش زبان طبیعی (NLP)**. مدل HMM به دلیل توانایی در مدل سازی دنبال داده ها در مدل کردن زبان طبیعی به کار گرفته می شود. در این سوال قصد داریم تا با استفاده از HMM به پیش بینی صعود یا سقوط ارزش سهام بر اساس اخبار روز بپردازیم. با بررسی اخبار reddit و تاثیر آن ها بر تغییر ارزش شاخص DJIA (Dow Jones Industrial Average یک شاخص سهام که میزان عملکرد سهام ۳۰ شرکت بزرگ در لیست بازار بورس آمریکا می سنجد.) به پیش بینی تغییرات این شاخص بر اساس اخبار جدید می پردازیم. (۱۰ نمره)

برای این کار از combinedNewsDJIA.csv که در کنار فایل های تمرین قرار دارد استفاده می کنیم. در این دیتاست ۲۵ خبر برتر (طبق رای کاربران) در تاریخ های مشخص ثبت شده است. ستون اول داده تاریخ، ستون دوم لیبل و ستون های سوم تا ۲۷ ام داده ها ۲۵ تیتر خبری آن تاریخ هستند. لیبل نشان دهنده افزایش و ثابت ماندن (۱) یا کاهش ارزش DJIA است. داده های تاریخ 2008-08-08 تا 2014-12-31 را به عنوان داده آموزش در نظر بگیرید، و داده های دو سال باقی مانده (2015-01-02 تا 2016-07-01) را به عنوان داده های تست در نظر بگیرید.

ابتدا باید داده های متنی را پیش پردازش کنید. برای این کار لازم است تا مراحل زیر را طی کنید:

- تبدیل تمام حروف تیتر های خبری به lower case .
- تقسیم کردن جمله به لیستی از کلمات
- حذف علائم نگارشی و کلمات بی معنی

- تبدیل لیست کلمات به جدولی شامل کلمات یکتا و تعداد تکرار آن ها در جمله

برای انجام این مراحل نیازی به پیاده سازی ندارید و می توانید از CountVectorizer در scikit-learn استفاده کنید. سپس با مدل سازی HMM داده ها را طبقه بندی کنید و صعود / ثابت ماندن یا کاهش را برای داده های تست پیش بینی کنید. برای حل این سوال مجاز به استفاده از کتابخانه های scikit-learn و مشابه آن هستید.

---

۱۱. Arc Reversal

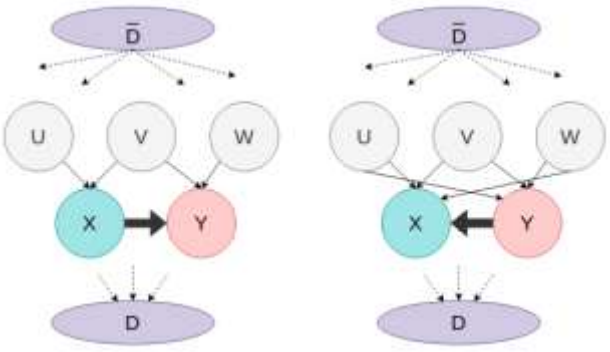
عملیات reversal arc در یک شبکه بیزی این امکان را ایجاد می کند که جهت یک یال که به صورت $X \rightarrow Y$ بوده است را به $Y \rightarrow X$ تغییر دهیم. برای این که چگالی مشترک شبک بیزی تغییری نکند لازم است که تمامی والد های X ، والدهای Y بشوند و تمامی والد های Y ، والدهای X بشوند. در نتیجه این عملیات ممکن است تعدادی یال به شبکه بیزی افزوده شود . فرض کنیم والد های X را به صورت $U \cup V$ و والد های Y را به صورت $W \cup V$ نشان دهیم به طوری که $W \cap U = \emptyset$. همچنین مجموعه نود هایی از شبکه که X یا Y را به عنوان اجداد خود دارند با D و آن هایی که X یا Y را به عنوان اجداد خود ندارند با $\bar{D}$ نشان می دهیم. نشان دهید که چگالی مشترک متغیر های شبکه بیزی با انجام عملیات reversal arc تغییری نمیکند($f'(D, D, X, Y, W, U, V) = f(D, D, X, Y, W, U, V)$). (۱۰ نمره)

**راهنمایی: ابتدا گزاره های زیر را اثبات کنید.

$$f'(Y|U, V, W) = \sum_x f(Y|V, W, x)f(x|U, V)$$

$$f'(X|U, V, W, Y) = f(Y|X, V, W)f(X|U, V)/f(Y|U, V, W)$$

$$f'(X, Y|U, V, W) = f(X, Y|U, V, W)$$



شکل ۳. نحوه تغییر شبکه بیزی توسط arc reversal

---

۱۲. ماتریس های احتمال شرطی زیر تاثیر منطقه 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. (۲۰ نمره)

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

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

الف) فرض کنید ۲۰ دسامبر (اواخر پاییز و ابتدای زمستان) است، بنابراین در نظر بگیرید $P(a_1) = P(a_2) = 0.5$. بعلاوه 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؟

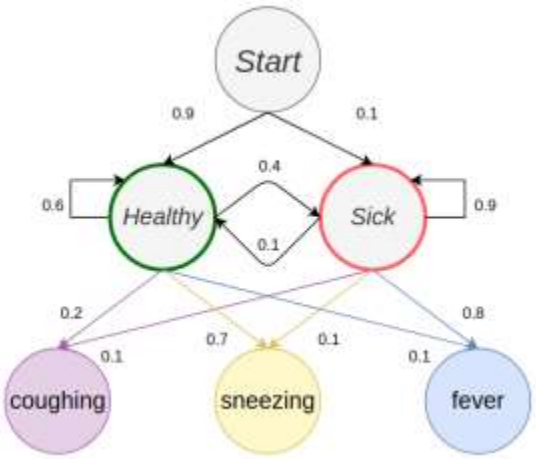
ب) فرض کنید تمام چیزی که می دانیم این است که ماهی لاغر است و روشنایی اش متوسط است. به احتمال بیشتر الان کدام فصل از سال است؟ احتمال درست بودن این حدس ؟

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 ؟

---

۱۳. (کامپیوتری) برای مدل کردن علامت های یک بیماری، مدل شکل۲ پیشنهاد شده است که در آن فرد به صورت احتمالاتی در یکی از شرایط سلامتی و یا بیماری قرار می گیرد و در هر حالت علامت هایی از خود نشان می دهد. احتمال دنباله FFSCFCSCSCFF را بااستفاده از الگوریتم forward بدست آورید. این بار احتمال های اولیه ورود به حالت بیماری و سلامتی را جا به جا کنید و احتمال مشاهده دنباله FFSCFCSCSCFF را مجددا محاسبه کنید. بنظر شما کدام یکی از مدل های HMM برای توصیف وضعیت فرد مناسب تر است؟ (۲۰ نمره، اختیاری)

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



شکل ۴ . مدل پیشنهادی برای علامت های بیماری