۱- در یک مسئله طبقه‌بندی دو کلاسی با یک ویژگی، با فرض توزیع گوسی برای این ویژگی در دو کلاس با متوسط‌های صفر و یک و واریانس‌های 0.5 و 0.25، با روش مینیمم کردن ریسک، مرز تصمیم‌گیری را در هریک از حالات زیر بدست آورید:

الف) $P(\omega_1) = P(\omega_2), \quad \lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

ب) $P(\omega_1) = P(\omega_2), \quad \lambda = \begin{pmatrix} 0 & 0.5 \\ 1 & 0 \end{pmatrix}$

پ) $P(\omega_1) = 2P(\omega_2), \quad \lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

ت) $P(\omega_1) = 0.5P(\omega_2), \quad \lambda = \begin{pmatrix} 0 & 0.5 \\ 1 & 0 \end{pmatrix}$

۲- یک مسئله طبقه‌بندی دو کلاسی دو بعدی با احتمال وقوع یکسان دو کلاس با بردارهای ویژگی با توزیع گوسی با متوسط‌های $\underline{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underline{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$ را داریم. در هریک از حالت‌های زیر معادله مرز تصمیم‌گیری بیز را بدست آورید و رسم کنید.

$$a)\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad b)\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \quad c)\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix} \quad d)\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

۳-یک مسئله طبقه‌بندی سه کلاس هم‌احتمال دو بعدی با بردارهای ویژگی با توزیع گوسی با متوسط‌ها و کواریانس زیر را داریم:

$$\underline{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underline{\mu}_2 = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \underline{\mu}_3 = \begin{pmatrix} 0 \\ 4 \end{pmatrix}, \Sigma = \begin{pmatrix} \frac{2}{3} & b \\ b & \frac{2}{3} \end{pmatrix}$$

الف) فرض کنید $b = \frac{1}{3}$. معادله مرزهای تصمیم‌گیری بیز را بدست آورید و در صفحه مختصات به‌طور دقیق رسم کرده و برچسب هر ناحیه را تعیین کنید.

ب) فرض کنید $b = 0$. بدون هیچ‌گونه محاسبه‌ای و با استدلال معادله مرزهای تصمیم‌گیری بیز را نوشته و در صفحه مختصات به‌طور دقیق رسم کرده و برچسب هر ناحیه را تعیین کنید.

۴- یک مسئله طبقه‌بندی سه کلاسی دو بعدی، بردارهای ویژگی با توزیع گوسی با متوسط‌ها و کواریانس زیر را داریم:

$$\underline{\mu}_1 = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}, \underline{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underline{\mu}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$$

الف) اگر $P_1 = 0.3,\ P_2 = 0.3,\ P_3 = 0.4$، برای داده $\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ احتمال پ سین سه کلاس را بد ست آورده و سپس با طبقه‌بند بیز به این داده برچسب بزنید.

ب) اگر $P_1 = 0,\ P_2 = 0.6,\ P_3 = 0.4$ مرز تصمیم‌گیری بیز بین کلاس ۲ و ۳ را بدست آورید و رسم کنید.

۵- نشان دهید که در یک مسئله طبقه‌بندی با M کلاس، احتمال خطای طبقه‌بند بیز (که به صورت $P_e = 1 - \arg\underset{\omega_i}{Max}\, P(\omega_i|\underline{x})$ تعریف می‌شود) در رابطه زیر صدق می‌کند.

$$P_e \leq \frac{M-1}{M}$$

۶- یکی از کاربردهای طبقه‌بند Naïve Bayes برای طبقه‌بندی متن است. در این مسئله از شما خواسته شده تا عبارات زیر را در دو کلاس «سیاسی» و «غیرسیاسی» طبقه بندی کنید.

- یک مسابقه بسیار نزدیک
- یک انتخابات بسیار نزدیک

مجموعه آموزش نیز به این صورت است.

| *عبارت* | *کلاس* |
|---|---|
| *بسکتبال یک بازی فوق‌العاده برای بازی‌کردن است* | غیرسیاسی |
| *مناظره بسیار روشن* | سیاسی |
| *یک مسابقه نزدیک اما فراموش‌شدنی* | غیرسیاسی |
| *انتخابات یک مسابقه است* | سیاسی |

با استفاده از این مجموعه آموزشی، هدف طبقه‌بند Naïve Bayes محاسبه:
$P$(«یک مسابقه بسیار نزدیک»|سیاسی) و $P$(«یک مسابقه بسیار نزدیک»|غیرسیاسی) و طبقه‌بندی این عبارت در یکی از دو کلاس سیاسی یا غیرسیاسی است.
یکی از روش‌های انجام این طبقه بندی، ایجاد ویژگی‌هایی بر اساس کلمات است، در واقع با فرض مستقل بودن رویداد هر کلمه، می‌توان نوشت:

$$P\left(\text{یک مسابقه بسیار نزدیک}\right) = P\left(\text{یک}\right)P\left(\text{مسابقه}\right)P\left(\text{بسیار}\right)P(\text{نزدیک})$$

الف) با استفاده از توضیحات بالا، احتمال‌های شرطی زیر را مشابه قضیه بیز (با فرض استقلال کلمات) بنویسید (بدون محاسبه عددی).

- $P$(«یک مسابقه بسیار نزدیک»|سیاسی)
- $P$(«یک انتخابات بسیار نزدیک»|غیرسیاسی)

ب) اگر بخواهیم با استفاده از آموزشی که از مجموعه داده بالا می‌توانیم بدهیم یک طبقه‌بند بسازیم، یکی از راه‌های ساده استفاده از فرکانس وقوع کلمات به ازای هر کلاس است. به این صورت می‌توانیم تعریف کنیم:

$$tf(w,c) = c \text{ تعداد بار مشاهده کلمه } w \text{ در کلاس}$$

یک روش پیچیده‌تر نیز استفاده از متد زیر است، به نام *Term Frequency – Inverse Document frequency*:

$$tfidf(w,c) = tf(w,c)\ln\left(\frac{N}{n(w)}\right)$$

که در آن $N$ تعداد کل نمونه‌های آموزش است و $n(w)$ تعداد باری که نمونه‌های آموزش شامل کلمه $w$ است. با این ویژگی می‌توانیم احتمال وجود کلمه $w$ در صورت وقوع کلاس $c$ را به صورت زیر محاسبه کنیم:

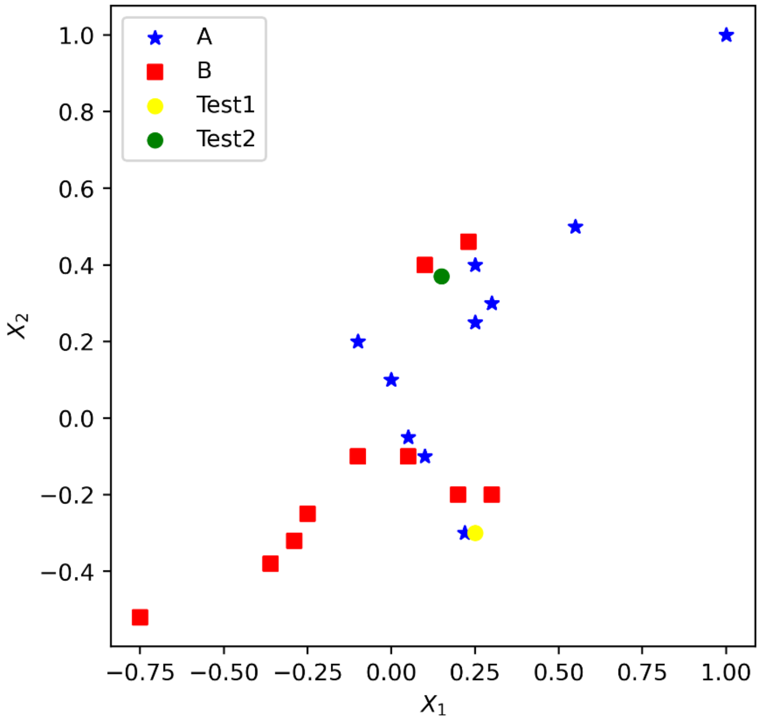
$$P(w|c) = \frac{tfidf(w,c)}{N(c)}$$

که مخرج کسر نشان دهنده تعداد کلمات در مجموعه داده آموزشی به ازای این کلاس است. برای مثال در این مسئله داریم:
$tf$(سیاسی و یک)=۱ و $tf$(غیرسیاسی و یک)=۲ و $N$(سیاسی)=۷ و $N$(غیرسیاسی)=۱۲
با استفاده از معیار دوم عبارت‌های زیر را طبقه بندی کنید.

- یک مسابقه بسیار نزدیک
- یک انتخابات بسیار نزدیک

(در واقع احتمالات مشابه $P$(«یک مسابقه بسیار نزدیک»|سیاسی) را مقایسه کنید.)

۷– در یک طبقه‌بندی دو کلاسی با دو ویژگی، داده‌های آموزشی دو کلاس با حروف A, B و دو داده تست با شکل دایره و رنگ‌های زرد و سبز مشخص شده‌اند. می‌خواهیم این داده‌های تست را با طبقه‌بند kNN و معیار فاصله اقلیدسی، طبقه‌بندی کنیم. برای k = 1,3,5 کلاس داده‌های تست را بدست آورده و آن‌هارا مقایسه کنید. آیا افزایش k موجب تغییر یا بهبود در نتیجه می‌شود؟




۸– الف) در م سئله ۲۹ از ف صل دوم کتاب تخمین متو سط و واریانس را به روش $ML$ بد ست آورید. فرض کنید N م شاهده م ستقل از این متغیر تصادفی را داریم.

ب) مسئله ۲۸ از فصل دوم کتاب را حل کنید. فرض کنید N مشاهده مستقل از این متغیر تصادفی را داریم که با پارامتر یکسان تولید شده‌اند.

۹– متغیر x دارای توزیع ارلانگ است، چنانچه: $p(x|\theta) = \theta^2 x e^{-\theta x} u(x)$ که تابع $u(x)$ تابع پله واحد است. نشان دهید تخمین ML برای متغیر $\theta$ با N مشاهده از رابطه زیر بدست خواهد آمد.

$$\widehat{\theta_{ML}} = \frac{N}{\sum_{k=1}^{N} x_k}$$

۱۰– فرض کنید X یک متغیر تصادفی گسسته با تابع احتمال زیر است و ۱۰ مشاهده مستقل از این توزیع گرفته شده است (۰و۲و۳و۱و۱و۰و۲و۲و۲و۳). تخمین ML پارامتر θ چقدر است؟ $(0 \leq \theta \leq 1)$

| X | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| P(X) | $\frac{(1-\theta)}{6}$ | $\frac{(1-\theta)}{3}$ | $\frac{2\theta}{3}$ | $\frac{(3-\theta)}{6}$ |