Subject:
Date

مهدی جمال خواه
810100111

تمرین سری اول

یادگیری ماشین

سوال ①:

آ)

$$P(\omega_1|x) = P(\omega_2|x) \longrightarrow \frac{P(x|\omega_1)P(\omega_1)}{\cancel{P(x)}} = \frac{P(x|\omega_2)P(\omega_2)}{\cancel{P(x)}}$$

$$\underline{P(\omega_1) = P(\omega_2)}, \quad P(x|\omega_1) = P(x|\omega_2) \longrightarrow \frac{\cancel{1}}{\cancel{\pi} b}\,\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} = \frac{\cancel{1}}{\cancel{\pi} b}\,\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}$$

$$\longrightarrow \left(\frac{x-a_1}{b}\right)^2 = \left(\frac{x-a_2}{b}\right)^2 \longrightarrow \cancel{x^2} - 2a_1x + a_1^2 = \cancel{x^2} - 2a_2x + a_2^2$$

$$\longrightarrow 2\cancel{(a_2 - a_1)}x = \underbrace{\cancel{a_2^2 - a_1^2}}_{a_2+a_1} \longrightarrow \boxed{x = \frac{a_2 + a_1}{2}}$$

⟸ نمودار $P(\omega_1|x)$ و $P(\omega_2|x)$ در شکل (1.1) نشان داده شده است.

ب)

$$P(error) = P(\omega_1) - \int_{R_1} p(x)\left[P(\omega_1|x) - P(\omega_2|x)\right]dx$$

$$= P(\omega_1) - \int_{R_1}\left[P(\omega_1)\,p(x|\omega_1) - P(\omega_2)P(x|\omega_2)\right]dx$$

$$= P(\omega_1) - \int_{R_1}\left[P(\omega_1)\frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} - P(\omega_2)\frac{1}{\pi b}\,\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}\right]dx$$

$$= P(\omega_1) - \frac{1}{\pi}\left(\tan^{-1}\left(\frac{x-a_1}{b}\right)P(\omega_1) - \tan^{-1}\left(\frac{x-a_2}{b}\right)P(\omega_2)\right)\Big|_{R_1}$$

Subject:
Date

* بدست آوردن احتمال خطا به $P(\omega_1)$ و $P(\omega_2)$ وابسته است مثلاً اگر $P(\omega_1)=1$ باشد

احتمال خطا صفر خواهد شد اما معمولاً متقارن می باشد و در اینجا $P(\omega_1) = P(\omega_2)$ می کنیم

نیز برای رسیدن به جواب مدنظر باید فرض کنیم که $P(\omega_1)=P(\omega_2) = \frac{1}{2}$ می باشد

$$R_1: \quad P(\omega_1|x) > P(\omega_2|x) \longrightarrow P(x|\omega_1) > P(x|\omega_2)$$

$$\longrightarrow \frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} > \frac{1}{1+\left(\frac{x-a_2}{b}\right)^2} \xrightarrow[\text{T}]{\text{ساده}} \boxed{x < \frac{a_1+a_2}{2}}$$

$$P(\text{error}) = \frac{1}{2} - \frac{1}{2\pi}\left(\tan^{-1}\left(\frac{x-a_1}{b}\right) - \tan^{-1}\left(\frac{x-a_2}{b}\right)\right)\Bigg|_{-\infty}^{\frac{a_1+a_2}{2}}$$

$$= \frac{1}{2} - \frac{1}{2\pi}\left(\tan^{-1}\left(\frac{a_2-a_1}{2b}\right) - \tan^{-1}\left(\frac{a_1-a_2}{2b}\right)\right)$$

$$= \frac{1}{2} - \frac{1}{2\pi}\left(2\tan^{-1}\left|\frac{a_2-a_1}{2b}\right|\right)$$

$$= \boxed{\frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2-a_1}{2b}\right|} \blacksquare$$

پ) بیش‌ترین خطا زمانی رخ می دهد که classifier بهینه را برعکس کنیم یعنی اگر

classifier بهینه ~~[illegible]~~ کلاس 1 را پیش‌بینی کرد ما 2 پیش‌بینی می کنیم و برعکس) زیرا

اگر classifier دیگری داشته باشیم که خطای آن از classifier گفته شده بیش

PAPCO

Subject:
Date

تر باشد، کافیست که آن را برعکس کنیم تا یک classifier بهتری از بهینه بدست آوریم که این ممکن نیست. به بیان ریاضی:

Prove : $\boxed{P_{max}(error) = 1 - P^*(error)}$

فرض خلف $\Rightarrow$ $P'_{max}(error) > P_{max}(error)$

$\rightarrow P'_{max}(error) > 1 - P^*(error)$

حال آن classifier ادعا شده را برعکس می‌کنیم ($1 \to 2$ و $2 \to 1$)، خطای آن را با $P''(error)$ نشان می‌دهیم:

$$P''(error) = 1 - P'_{max}(error)$$

$$< 1 - (1 - P^*(error))$$

$$\Rightarrow P''(error) < P^*(error)$$

بنابراین بیش‌ترین مقدار $P(error)$ برابر است با: (با فرض $P(\omega_1) = P(\omega_2)$).

$$\boxed{P_{max}(error) = 1 - P^*(error)} = 1 - \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left|\frac{a_2 - a_1}{2b}\right|$$

$$\boxed{P_{max}(error) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left|\frac{a_2 - a_1}{2b}\right|} \blacksquare$$

$$P(\omega_1|x) \overset{1}{\underset{2}{\gtrless}} P(\omega_2|x) \rightarrow \frac{P(x|\omega_1)P(\omega_1)}{P(x)} \overset{1}{\underset{2}{\gtrless}} \frac{P(x|\omega_2)P(\omega_2)}{P(x)}$$ (ب

$$\rightarrow P(x|\omega_1) \overset{1}{\underset{2}{\gtrless}} P(x|\omega_2) \rightarrow \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} \overset{1}{\underset{2}{\gtrless}} \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}$$

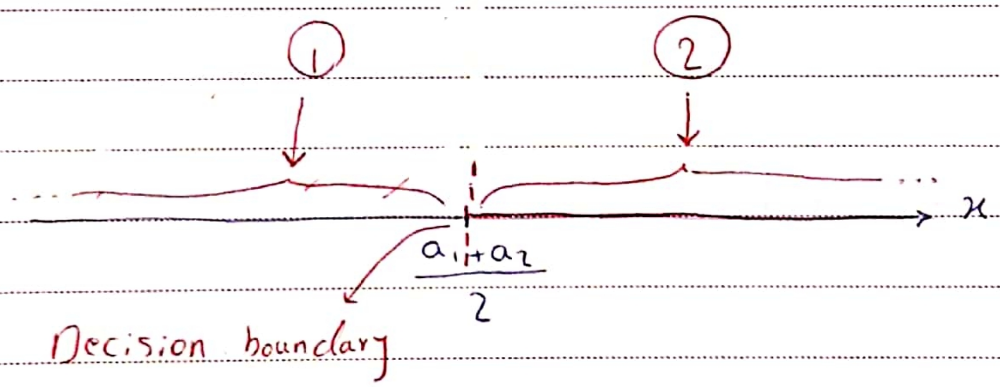
$$\rightarrow \left(\frac{x-a_2}{b}\right)^2 \overset{1}{\underset{2}{\gtrless}} \left(\frac{x-a_1}{b}\right)^2 \rightarrow -2a_2x+a_2^2 \overset{1}{\underset{2}{\gtrless}} -2a_1x+a_1^2$$

$$\rightarrow a_2^2-a_1^2 \overset{1}{\underset{2}{\gtrless}} 2x(a_2-a_1) \xrightarrow{a_2>a_1} 2x \overset{1}{\underset{2}{\lessgtr}} a_2+a_1$$

$$\rightarrow \boxed{x \overset{2}{\underset{1}{\gtrless}} \frac{a_1+a_2}{2}}$$

میزان احتمال خطا طبق محاسبات قسمت (ب) برابر است با :

$$\boxed{P(\text{error}) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2-a_1}{2b}\right|}$$



Subject:
Date

$$R(1|x) \underset{1}{\overset{2}{\gtrless}} R(2|x)$$ (ث

$$P(\omega_1|x)\lambda_{11} + P(\omega_2|x)\lambda_{12} \underset{1}{\overset{2}{\gtrless}} P(\omega_1|x)\lambda_{21} + P(\omega_2|x)\lambda_{22}$$

$$\rightarrow \underbrace{\frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{2}{\overset{1}{\gtrless}} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}}$$

$$\Rightarrow \frac{\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}}{\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}} \underset{2}{\overset{1}{\gtrless}} \frac{1}{2} \longrightarrow \frac{1+\left(\frac{x-a_2}{b}\right)^2}{1+\left(\frac{x-a_1}{b}\right)^2} \underset{2}{\overset{1}{\gtrless}} \frac{1}{2}$$

$$\rightarrow 2 + 2\left(\frac{x-a_2}{b}\right)^2 \overset{1}{>} 1 + \left(\frac{x-a_1}{b}\right)^2$$

$$\rightarrow b^2 + 2x^2 - 4a_2x + 2a_2^2 \overset{1}{>} x^2 - 2a_1x + a_1^2$$

$$\rightarrow x^2 + (2a_1 - 4a_2)x + 2a_2^2 - a_1^2 + b^2 \overset{1}{>} 0$$

$$\Delta = (2a_1 - 4a_2)^2 - 4(2a_2^2 - a_1^2 + b^2)$$

$$x_{1,2} = \frac{-b' \pm \sqrt{\Delta}}{2a'} = \frac{4a_2 - 2a_1 \pm \sqrt{\Delta}}{2}$$

$$= 2a_2 - a_1 \pm \sqrt{(a_1-2a_2)^2 - 2a_2^2 + a_1^2 - b^2}$$

$$= 2a_2 - a_1 \pm \sqrt{2(a_1-a_2)^2 - b^2}$$

$$R_2: \underbrace{2a_2 - a_1 - \sqrt{2(a_1-a_2)^2-b^2}}_{x_1} < x < \underbrace{2a_2 - a_1 + \sqrt{2(a_1-a_2)^2-b^2}}_{x_2}$$

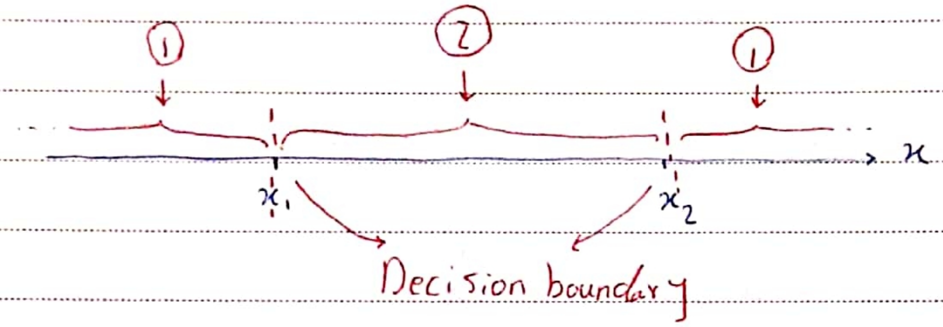
$$P(\text{error}) = \frac{1}{2} - \frac{1}{2\pi}\left(\tan^{-1}\left(\frac{x-a_2}{b}\right) - \tan^{-1}\left(\frac{x-a_1}{b}\right)\right)\Bigg|_{R_2}$$

$$= \frac{1}{2} - \frac{1}{2\pi}\left(\tan^{-1}\left(\frac{x_2-a_2}{b}\right) - \tan^{-1}\left(\frac{x_2-a_1}{b}\right)\right.$$

$$\left. - \tan^{-1}\left(\frac{x_1-a_2}{b}\right) + \tan^{-1}\left(\frac{x_1-a_1}{b}\right)\right)$$

$$;\ x_1 = 2a_2 - a_1 - \sqrt{2(a_1-a_2)^2 - b^2}$$

$$x_2 = 2a_2 - a_1 + \sqrt{2(a_1-a_2)^2 - b^2}$$



مقایسه بخش ت و ث : با توجه به ماتریس هزینه می توانیم بگوییم که هزینه اشتباه predict کردن یک داده متعلق به کلاس 1 بیش از کلاس 2 می باشد درحالی که این هزینه در بخش پ برابر بود در نتیجه ما برای کاهش ریسک در بخش ث سعی می کنیم که ناحیه ای که در آن کلاس 1 پیش بینی می شود (R_1) را بزرگ تر کنیم یعنی سخت تر می گوییم که داده ای متعلق به کلاس 2 است

این تفسیر در Decision boundary نشان داده شده در دو بخش ت و ث نیز قابل مشاهده است که ناحیه R_2 در بخش ث کوچک‌تر شده است.

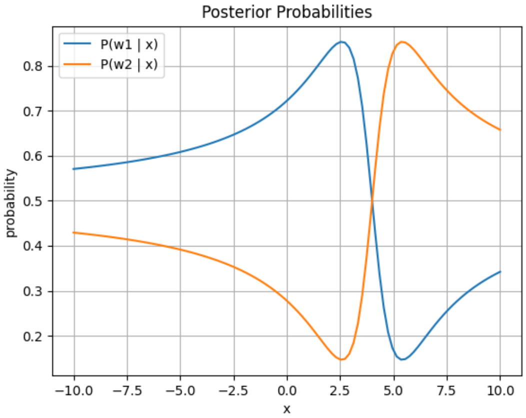


Figure 1.1: Posterior probabilities

Subject:
Date

سوال ②:

آ)

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix}$$

$$\begin{cases} R(1|x) = P(\omega_1|x)\lambda_{11} + P(\omega_2|x)\lambda_{12} \\ R(2|x) = P(\omega_1|x)\lambda_{21} + P(\omega_2|x)\lambda_{22} \end{cases}$$

$$R(1|x) \underset{1}{<} R(2|x)$$

$$\rightarrow P(\omega_1|x)\lambda_{11} + P(\omega_2|x)\lambda_{12} < P(\omega_1|x)\lambda_{21} + P(\omega_2|x)\lambda_{22}$$

$$\rightarrow P(x|\omega_1)P(\omega_1)\lambda_{11} + P(x|\omega_2)\lambda_{12}P(\omega_2) <$$

$$P(x|\omega_1)P(\omega_1)\lambda_{21} + P(x|\omega_2)P(\omega_2)\lambda_{22}$$

$$\rightarrow P(x|\omega_1)P(\omega_1)(\lambda_{11} - \lambda_{21}) < P(x|\omega_2)P(\omega_2)(\lambda_{22} - \lambda_{12})$$

$$\Rightarrow \frac{P(x|\omega_1)}{P(x|\omega_2)}(\lambda_{11} - \lambda_{21}) < (\lambda_{22} - \lambda_{12})\frac{P(\omega_2)}{P(\omega_1)}$$

$$\xrightarrow[\text{فرض}]{\lambda_{11} < \lambda_{21}} \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}} \frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow \boxed{\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} \quad : R_1}$$

Subject :
Date

ب)

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow \boxed{\frac{P(x|\omega_2)}{P(x|\omega_1)} < \frac{\lambda_{21}-\lambda_{11}}{\lambda_{12}-\lambda_{22}} \frac{P(\omega_1)}{P(\omega_2)}}$$

نتیجه: در ناحیه تصمیم گیری مربوط به کلاس یک $P(x|\omega_2)$ تا یک حدی می تواند از $P(x|\omega_1)$ بیش تر باشد که آن حد وابسته به هزینه به احتمال پیشین و ماتریس هزینه است.

Subject:
Date

$$P(\omega_1|x) \underset{2}{\overset{1}{\gtrless}} P(\omega_2|x) \longrightarrow \frac{P(x|\omega_1)P(\omega_1)}{P(x)} \underset{2}{\overset{1}{\gtrless}} \frac{P(x|\omega_2)P(\omega_2)}{P(x)}$$ سوال (3)

$$\Rightarrow P(x|\omega_1) \overset{1}{>} P(x|\omega_2) \Rightarrow \frac{x}{\sigma_1^2}\exp\left(-\frac{x^2}{2\sigma_1^2}\right) > \frac{x}{\sigma_2^2}\exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$

$$\xrightarrow{x>0} \sigma_2^2 \exp\left(-\frac{x^2}{2\sigma_1^2}\right) > \sigma_1^2 \exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$

$$\xrightarrow{\ln} 2\ln\sigma_2 - \frac{x^2}{2\sigma_1^2} > 2\ln\sigma_1 - \frac{x^2}{2\sigma_2^2}$$

$$\longrightarrow x^2\left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right) > 2\ln\sigma_1 - 2\ln\sigma_1$$

$$\xrightarrow[\text{فرض}]{\sigma_1>\sigma_2} x^2 > \frac{2\ln\sigma_1 - 2\ln\sigma_1}{\sigma_1^2-\sigma_2^2}\, 2\sigma_1^2\sigma_2^2$$

$$\Rightarrow x > 2\sigma_1\sigma_2\sqrt{\frac{\ln\sigma_1 - \ln\sigma_2}{\sigma_1^2-\sigma_2^2}}$$

if $\sigma_1 > \sigma_2 \Rightarrow x \underset{2}{\overset{1}{\gtrless}} 2\sigma_1\sigma_2\sqrt{\frac{\ln\sigma_1 - \ln\sigma_2}{\sigma_1^2-\sigma_2^2}}$

if $\sigma_2 > \sigma_1 \Rightarrow x \underset{1}{\overset{2}{\gtrless}} 2\sigma_1\sigma_2\sqrt{\frac{\ln\sigma_1 - \ln\sigma_2}{\sigma_1^2-\sigma_2^2}}$

مرز تصمیم $= 2\sigma_1\sigma_2\sqrt{\frac{\ln\sigma_1 - \ln\sigma_2}{\sigma_1^2-\sigma_2^2}}$ ■

Subject:
Date

سوال (4) :

$R_1: \quad P(\omega_1|x) > P(\omega_2|x)$

$$\rightarrow \frac{P(x|\omega_1)P(\omega_1)}{P(x)} > \frac{P(x|\omega_2)P(\omega_2)}{P(x)} \rightarrow \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

$$\rightarrow \frac{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)} > \frac{P(\omega_2)}{P(\omega_1)}$$

$$\xrightarrow{\ln} -\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_2)^2}{2\sigma^2} > \ln P(\omega_2) - \ln P(\omega_1)$$

$$\rightarrow (x-\mu_2)^2 - (x-\mu_1)^2 > 2\sigma^2(\ln P(\omega_2) - \ln P(\omega_1))$$

$$\rightarrow x(-2\mu_2 + 2\mu_1) + \mu_2^2 - \mu_1^2 > 2\sigma^2(\ln P(\omega_2) - \ln P(\omega_1))$$

if $\mu_2 < \mu_1$

$$\rightarrow x \gtrless \underbrace{\frac{2\sigma^2(\ln P(\omega_2) - \ln P(\omega_1)) + \mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)}}_{\text{مرز تصمیم (DB)}} \quad : R_1$$

if $\mu_1 < \mu_2$

$$DB = \frac{\sigma^2(\ln P(\omega_2) - \ln P(\omega_1))}{\mu_1 - \mu_2} + \frac{\mu_1 + \mu_2}{2}$$

* فرض می کنیم $\mu_1 < \mu_2$ می باشد. در واقع آن کلاس که میانگین آن بزرگ تراست را چپ می نامیم این به اثبات هیچ اشنایی وارد نمی کند و برای حالت $(\mu_1 > \mu_2)$ کافیست که جای ۱ و ۲ را در رابطه نهایی جا به جا کنیم.

- $DB \leq \mu_1 \longrightarrow \dfrac{\sigma^2(\ln P(\omega_2) - \ln P(\omega_1))}{\mu_1 - \mu_2} + \dfrac{\mu_2 - \mu_1}{2} \leq 0$

$$\longrightarrow \sigma^2(\ln P(\omega_2) - \ln P(\omega_1)) \geq \frac{(\mu_1 - \mu_2)^2}{2}$$

$$\longrightarrow \underline{\frac{P(\omega_2)}{P(\omega_1)} \geq \exp\left(\frac{(\mu_1 - \mu_2)^2}{2\sigma^2}\right)}$$

- $DB \geq \mu_2 \longrightarrow \dfrac{\sigma^2(\ln P(\omega_2) - \ln P(\omega_1))}{\mu_1 - \mu_2} + \dfrac{\mu_1 - \mu_2}{2} \geq 0$

$$\longrightarrow \sigma^2(\ln P(\omega_2) - \ln P(\omega_1)) \leq \frac{-(\mu_1 - \mu_2)^2}{2}$$

$$\longrightarrow \frac{P(\omega_2)}{P(\omega_1)} \leq \exp\left(\frac{-(\mu_1 - \mu_2)^2}{2\sigma^2}\right)$$

$$\boxed{\frac{P(\omega_2)}{P(\omega_1)} \geq \exp\left(\frac{(\mu_1 - \mu_2)^2}{2\sigma^2}\right) \ \underline{\text{OR}} \ \frac{P(\omega_2)}{P(\omega_1)} \leq \exp\left(\frac{-(\mu_1 - \mu_2)^2}{2\sigma^2}\right) \ ; \ \mu_1 \leq \mu_2}$$ ■

Subject:
Date

سوال ⑤ :
آ)

$$P(x|\omega_i) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$\mathcal{L}(\mu,\Sigma) = \sum_{i=1}^{n}\left[-\ln 2\pi|\Sigma|^{1/2} - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]$$

$$= -n \ln 2\pi|\Sigma|^{1/2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$\frac{d\mathcal{L}}{d\mu} = \sum_{i=1}^{n}\frac{2}{2}\Sigma^{-1}(x_i-\mu) = 0 \Rightarrow \Sigma^{-1}\left(\sum_{i=1}^{n}x_i - n\mu\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}x_i - n\mu = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i}$$

$$\frac{d\mathcal{L}}{d\Sigma} = \frac{n}{2}\Sigma^T - \frac{1}{2}\frac{d}{d\Sigma}\left(\sum_{i=1}^{n}tr\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right)$$

$$= \frac{n}{2}\Sigma^T - \frac{1}{2}\frac{d}{d\Sigma}\left(tr\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)\right)$$

$$= \frac{n}{2}\Sigma^T - \frac{1}{2}\left(\sum_{i=1}^{n}(x_i-\mu)^T(x_i-\mu)^T\right)^T = 0$$

$$\Rightarrow \boxed{\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T \quad ; \mu = \hat{\mu}}$$

P4PCO

Subject:
Date

کافیست که با جایگذاری sample ها در دو فرمول صفحه پیش میانگین و کوواریانس هر دو کلاس را محاسبه کنیم. بعد از محاسبه با python (که آن بارگذاری شده) مقادیر زیر بدست می آیند:

$$\hat{\mu}_2 = \begin{bmatrix} 1.333 \\ 1.6111 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 0.555 & 0.185 \\ 0.185 & 0.987 \end{bmatrix}$$

$+ \Rightarrow 2$

$\bullet \Rightarrow 1$

$$\hat{\mu}_1 = \begin{bmatrix} -0.3 \\ -0.15 \end{bmatrix} \qquad \hat{\Sigma}_1 = \begin{bmatrix} 2.06 & -0.02 \\ -0.02 & 0.5025 \end{bmatrix}$$

ب)

$$P(\omega_1|x) \overset{1}{>} P(\omega_2|x) \rightarrow P(x|\omega_1) \overset{1}{>} P(x|\omega_2)$$

$$\rightarrow \frac{\frac{1}{|\hat{\Sigma}_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x-\hat{\mu}_1)\right)}{\frac{1}{|\hat{\Sigma}_2|^{1/2}} \exp\left(-\frac{1}{2}(x-\hat{\mu}_2)^T \hat{\Sigma}_2^{-1}(x-\hat{\mu}_2)\right)} \overset{1}{>} 1 : R_1$$

$$\rightarrow -\frac{1}{2}\ln|\hat{\Sigma}_1| - \frac{1}{2}(x-\hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x-\hat{\mu}_1) + \frac{1}{2}\ln|\hat{\Sigma}_2| + \frac{1}{2}(x-\hat{\mu}_2)^T \hat{\Sigma}_2^{-1}(x-\hat{\mu}_2) > 0$$

مرز تقسیم

$\Leftarrow$ مرز تقسیم بالا توسط python محاسبه شده و در شکل (5.1) رسم شده است

PAPCO

Subject:
Date

طبق شکل (5.1) تعداد نقاط اشتباه پیش‌بینی شده برابر است با

Class 2 $\Rightarrow$ 2 تا

Class 1 $\Rightarrow$ 1 عدد

$$\text{accuracy} = \frac{19-3}{19} \times 100 = \boxed{84\%}$$

$$\text{درصد خطا} = \frac{3}{19} = \boxed{16\%}$$

پ)

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} = \frac{2a}{a} \frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow \frac{P(x|\omega_1)}{P(x|\omega_2)} > 2\frac{P(\omega_2)}{P(\omega_1)}$$

$$\Rightarrow -\frac{1}{2}\ln|\hat{\Sigma}_1| - \frac{1}{2}(x-\hat{\mu}_1)^T \hat{\Sigma}_1^{-1}(x-\hat{\mu}_1)$$

$$+ 2\frac{P(\omega_2)}{P(\omega_1)}\left[\frac{1}{2}\ln|\hat{\Sigma}_2| + \frac{1}{2}(x-\hat{\mu}_2)^T \hat{\Sigma}_2^{-1}(x-\hat{\mu}_2)\right] > 0$$

$\Leftarrow$ مرز تصمیم بالا با فرض $P(\omega_2)=P(\omega_1)$ در شکل (5.2) رسم شده است

تحلیل: با قرار دادن این ریسک ما هزینه اشتباه پیش‌بینی کردن کلاس 2 را بیشتر می‌کنیم و در طبقه‌بند جدید تنها یک نمونه از کلاس 2 اشتباه Predict شده است یعنی یکی نسبت به قبل کاهش پیدا کرده است.

Subject:
Date

ت) مرز تصمیم آن با مرز تصمیم مورد قبل یکسان است و فقط باید به جای

$P(\omega_2)$ و $P(\omega_1)$ به ترتیب $\frac{2}{3}$ و $\frac{1}{3}$ قرار دهیم

⟸ مرز تصمیم آن در شکل (5.3) رسم شده است.

تحلیل: چون احتمال کلاس 2 بیش تر است پس فضای R_2 نسبت به مورد

قبل بزرگ تر می شود. و تمام نمونه های کلاس 2 درست Predict شده اند

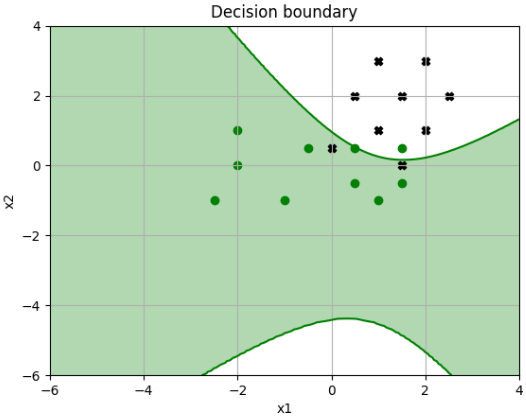


Figure 5.1: Decision Boundary (part b)

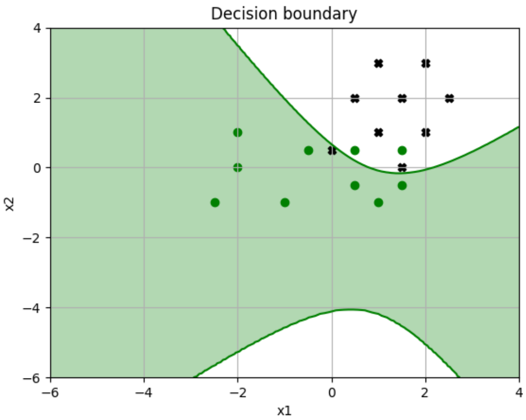


Figure 5.2: Decision Boundary (part c)

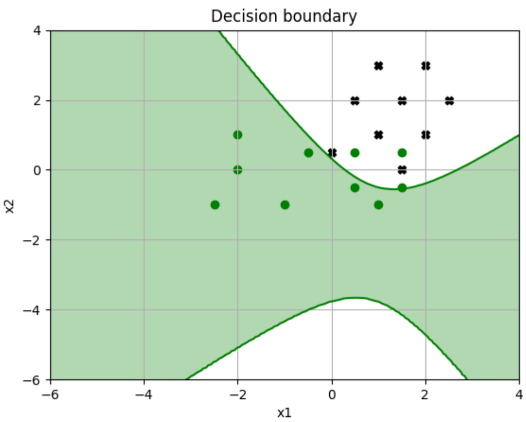


Figure 5.3: Decision Boundary (part d)

Subject:
Date

سوال ⑥:

الف)

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log P(x_i|\theta) \Rightarrow \sum_{i=1}^{n} \log \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \mathcal{L}(\lambda)$$

$$\mathcal{L}(\lambda) = \sum_{i=1}^{n} (x_i \log\lambda - \lambda - \log x_i!) = \boxed{\log\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log x_i!}$$

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} \mathcal{L}(\lambda)$$

$$\frac{d\mathcal{L}}{d\lambda} = 0 \Rightarrow \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0 \longrightarrow \boxed{\hat{\lambda}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$ ■

ب)

$$P(\lambda|D) \propto P(D|\lambda)P(\lambda)$$

$$= \prod_{i=1}^{n} P(x_i|\lambda)P(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} C \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} C \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$= \underbrace{\frac{C}{\prod_{i=1}^{n} x_i!}}_{C'} \lambda^{\overbrace{\sum_{i=1}^{n} x_i + \alpha}^{\alpha'} - 1} e^{-\overbrace{(\beta+n)}^{\beta'}\lambda}$$

$$= \boxed{C' \lambda^{\alpha'-1} e^{-\beta'\lambda} \quad ; \quad \alpha' = \alpha + \sum_{i=1}^{n} x_i, \quad \beta' = \beta + n}$$ ■

Subject:
Date

پ) بله. چون توزیع احتمال پیشین و پسین آن یک توزیع یکسان شد و با دیدن نمونه ها فقط پارامتر های توزیع آپدیت می شوند.

ت)

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} P(\lambda \mid D) = \arg\max_{\lambda} \log P(\lambda \mid D)$$

$$= \arg\max_{\lambda} [\underbrace{\log C' + (\alpha'-1)\log\lambda - \beta'\lambda}_{\mathcal{L}}]$$

$$\frac{d\mathcal{L}}{d\lambda} = \frac{\alpha'-1}{\lambda} - \beta' = 0 \longrightarrow \boxed{\hat{\lambda}_{MAP} = \frac{\alpha'-1}{\beta'} = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{\beta + n}}$$

ث)

$$\lim_{n\to\infty} \hat{\lambda}_{MAP} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{\beta + n} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\lambda}_{ML}$$

بنابراین وقتی که تعداد داده ها به بی نهایت می رود تخمین گر MAP به MLE میل می کند و این اتفاق در اکثر موارد می افتد. یعنی وقتی تعداد داده ها به اندازه کافی زیاد باشد، فرقی نمی کند که از کدام روش استفاده کنیم، هر دو جواب یکسانی خواهند داشت. در واقع MAP سعی می کند به همان جواب MLE برسد اما از آنجا که هر بار توزیع پیشین را هم در نظر می گیرد، کم کم به آن میل می کند (با دیدن داده ها)

Subject:
Date

ج) دیدگاه Beysain، قطعیت کمتری دارد. اما ML به صورت قطعی یک مقدار به عنوان پارامتر به ما می دهد. بنابراین در شرایطی که نمونه کم داریم بهتر است از دیدگاه Beysain برویم چون اگر ML استفاده کنیم احتمال آنکه آن عدد بدست آمده به عنوان پارامتر با پارامتر واقعی تفاوت زیادی داشته باشد بالاست ولی وقتی نمونه به اندازه کافی داریم بهتر است که از روش ML استفاده کنیم چون وقتی تعداد نمونه ها زیاد شود در اکثر مواقع $\hat{\theta}_{MAP}$ و θ_{ML} یکسان می شود پس بهتر است که از اول همان ML را انتخاب کنیم چون روش ساده تری است و نیاز به محاسبات و اطلاعات کمتری دارد (دیگر لازم نیست که از توزیع پارامترها نیز آگاه باشیم.)

7 Naive Bayes classifier

7.1 Naive Bayes Classifier Algorithm

Naive Bayes classifier is based on Bayes' Theorem. It assumes independence between features and calculates the probability of a given input belonging to a particular class. The naive Bayes classifier is a simplified version of the Bayes classifier, in which we assume that the features are conditionally independent given the class instead of modeling their full conditional distribution given the class.

Denote the class C and the vector of features $(F_1, F_2, \ldots, F_k)$ the predicted class is:

$$\begin{aligned}\hat{C} &= argmax_c \ P(C = cF_1 = f_1, \ldots, F_k = f_k) \\ &= argmax_c \ P(F_1 = f_1, \ldots, F_k = f_k | C = c)P(C = c)\end{aligned}$$

By Naive Bayes assumption we have:

$$C = argmax_c P(C = c) \prod_{i=1}^{k} P(F_i = f_i | C = c)$$

So we need to estimate $P(F_i = f_i | C = c)$ and the Naive Bayes uses the maximum likelihood estimation.

7.2 Naive Bayes Classifier VS. Bayes Classifier

The main difference between the two lies in the assumption about the relationship between the features. The Naive Bayes classifier assumes that the features are conditionally independent of each other given the class label. This assumption simplifies the computation. However, it may not hold in all cases, and if the features are highly correlated, the Naive Bayes classifier's performance may suffer.

On the other hand, the general Bayes classifier does not make this assumption. This allows for more accurate modeling but comes with a higher computational cost, as it requires estimating and handling the full joint distribution of the features. Another issue is that usually we need more knowledge about the samples to be able to estimate the posterior probability.

In practice, the choice between the Naive Bayes classifier and the general Bayes classifier depends on the specific characteristics of the dataset. If the independence assumption is reasonably satisfied, Naive Bayes can be a good choice. However, if the features are known to be correlated or the independence assumption is violated, Bayesian classifier may be more appropriate, even though it comes with a higher computational cost.

Using Naive Bayes can significantly reduce computation costs; however, there is no free lunch, and the accuracy of the classifier may be compromised compared to Bayes classifiers.

7.3 Preprocessing

One important step in preparing our data is dealing with missing values. In Naive Bayes, we rely on the count of each feature, so it's crucial to handle missing values properly. We have two options: either we remove the rows with missing values or we fill in the missing values. In this case, Since the amount of missing data is not significant, it's better to fill in the missing values rather than removing the entire rows, we will choose the second option and fill the missing values with the most common value for that feature(mode).

7.4 Implementing from Scratch

We have implemented the Naive Bayes algorithm as explained in the previous section, and here are the classification results obtained on the lung_cancer dataset:

The majority of "YES" labels were predicted correctly, resulting in high precision and recall values for "YES" class. However, it's important to note that the number of "NO" labels in the dataset is smaller compared to the "YES" labels. As a result, even a small number of mispredictions can have a significant impact on the precision and recall metrics for the "NO" class. Although the

Table 7.1: Missing Data of lung_cancer Dataset

Feature	count	Percentage
GENDER	309	100.00
AGE	308	99.68
SMOKING	308	99.68
YELLOW$_FINGERS$	309	100.00
ANXIETY	309	100.00
PEER$_PRESSURE$	308	99.68
CHRONIC DISEASE	307	99.35
FATIGUE	308	99.68
ALLERGY	308	99.68
WHEEZING	308	99.68
ALCOHOL CONSUMING	307	99.35
COUGHING	308	99.68
SHORTNESS OF BREATH	308	99.68
SWALLOWING DIFFICULTY	307	99.35
CHEST PAIN	309	100.00
LUNG$_CANCER$	309	100.00

Table 7.2: Classification Report (lung_cancer from scratch)

	precision	recall	f1-score	support
NO	0.80	0.73	0.76	11
YES	0.97	0.98	0.98	113
accuracy			0.96	124
macro avg	0.89	0.85	0.87	124
weighted avg	0.96	0.96	0.96	124

overall model accuracy is good, the precision and recall for the "NO" class are not as strong as the accuracy.

7.5 Implementing with Libraries

We utilized the `GaussianNB` classifier from the Sklearn library to classify the same dataset, and the resulting classification metrics are as follows:

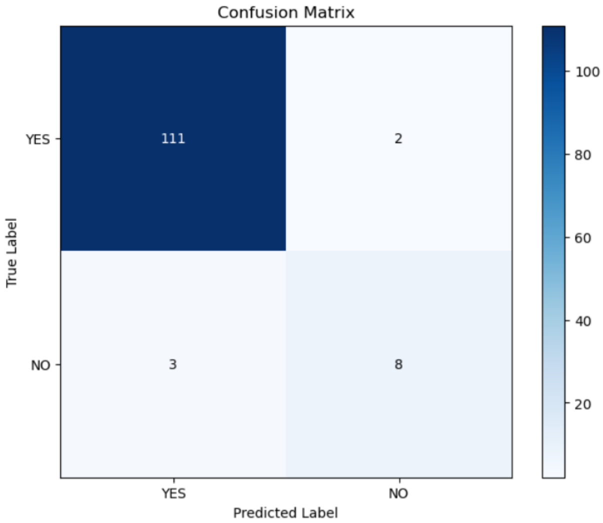


Figure 7.1: Classification of lung_cancer dataset with from scratch algorithm

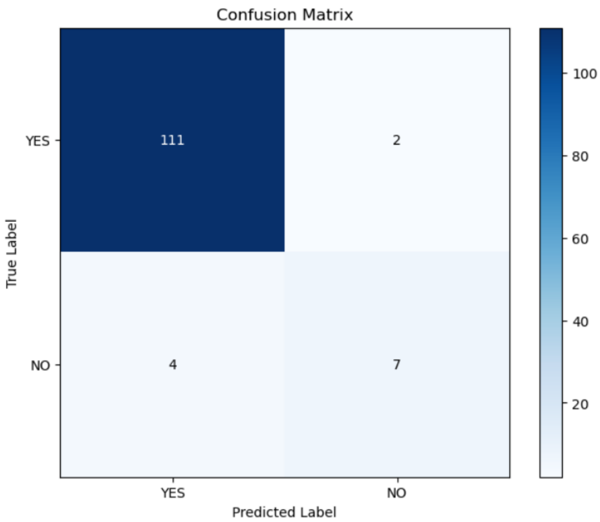


Figure 7.2: Classification of lung_cancer dataset with library

Table 7.3: Classification Report (lung_cancer with library)

	precision	recall	f1-score	support
NO	0.78	0.64	0.70	11
YES	0.97	0.98	0.97	113
accuracy			0.95	124
macro avg	0.87	0.81	0.84	124
weighted avg	0.95	0.95	0.95	124

The results of this section are nearly identical to the previous section, with a small difference due to the utilization of the `GaussianNB` algorithm from the library, compared to our implementation of the Bernoulli Naive Bayes from scratch. The `GaussianNB` algorithm assumes a Gaussian distribution for each feature.

Overall, the classifier's performance is not particularly strong. This may be attributed to the presence of correlations between certain features, which is evident from the correlation matrix shown in Figure 7.3. For instance, features "ANXIETY" and "YELLOW_FINGERS" exhibit a high correlation. As we previously discussed, Naive Bayes disregards these correlations and assumes independence among the features.

Considering the limitations of Naive Bayes in handling correlated features, it is expected that the classifier's performance may be affected.

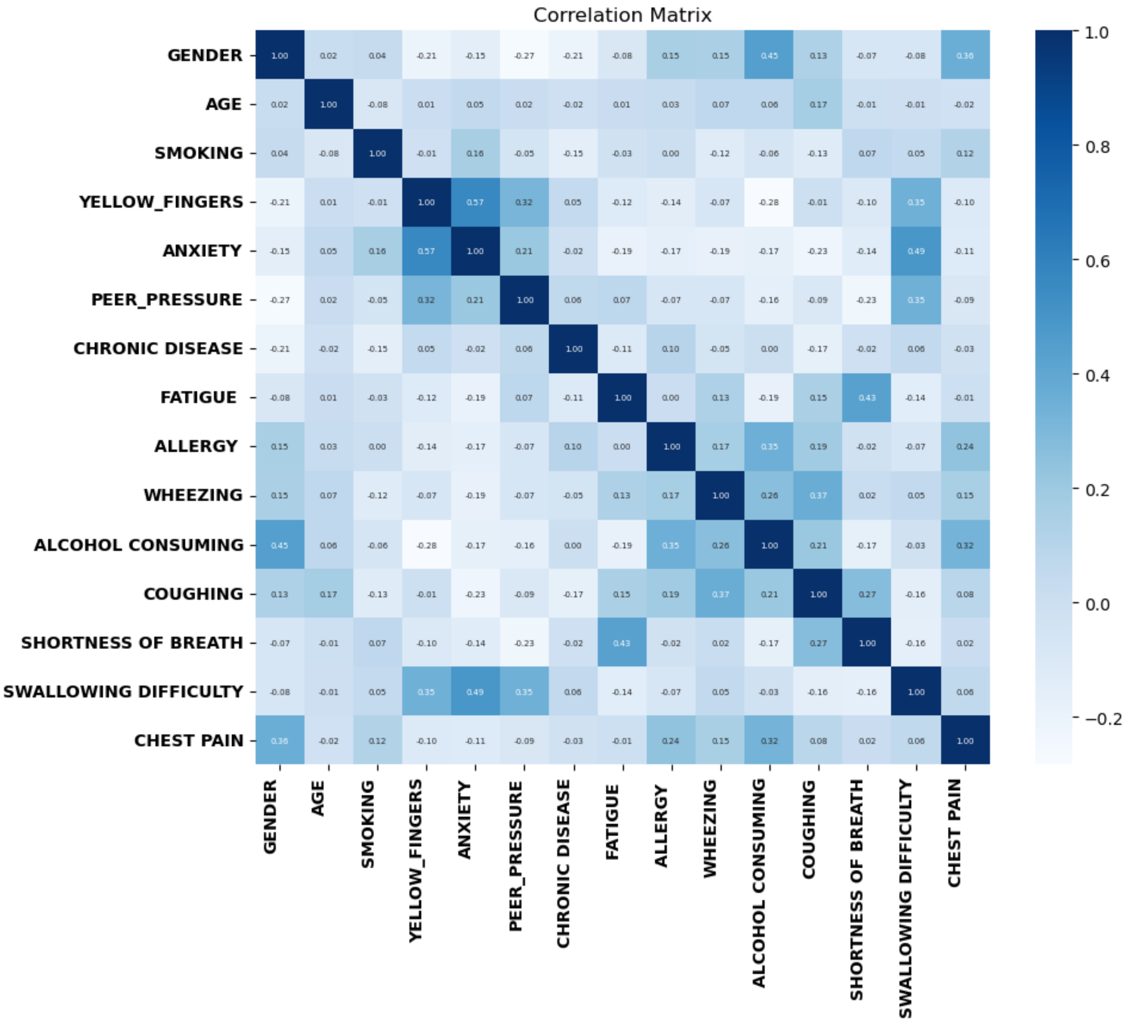


Figure 7.3: Correlation Matrix of lung_cancer dataset

7.6 Web-Page-Phishing Dataset

In this section, we have performed classification on another dataset called "web-page-phishing." The results of the classification are as follows:

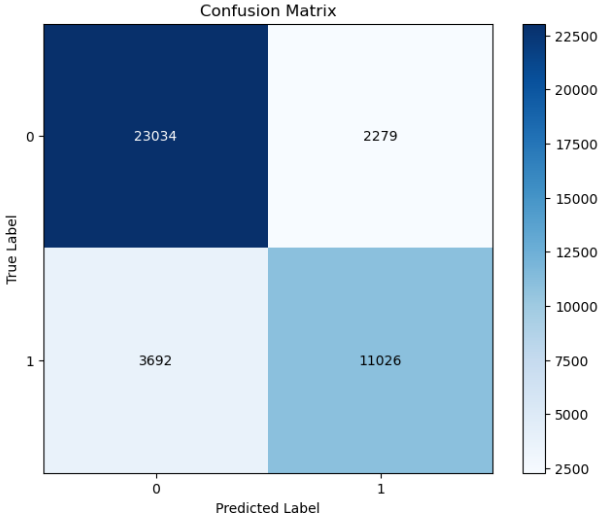


Figure 7.4: Classification of web-page-phishing dataset with from scratch algorithm

Table 7.4: Classification Report (web-page-phishing from scratch)

	precision	recall	f1-score	support
0	0.86	0.91	0.89	25313
1	0.83	0.75	0.79	14718
accuracy			0.85	40031
macro avg	0.85	0.83	0.84	40031
weighted avg	0.85	0.85	0.85	40031

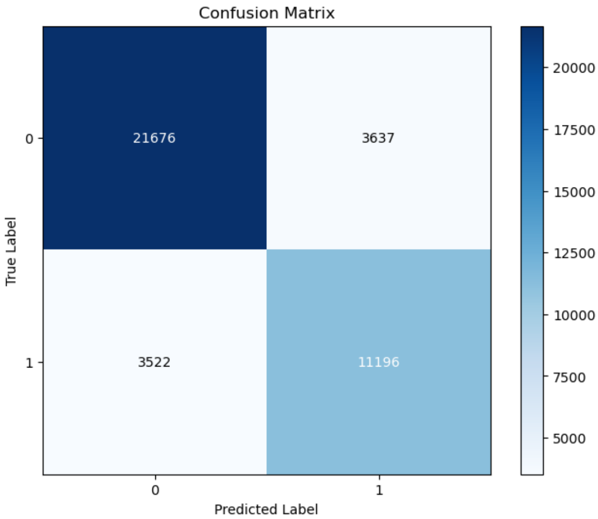


Figure 7.5: Classification of web-page-phishing dataset with library

The performance of the classifier on the previous dataset is better compared to the current one. This difference in performance can be attributed to the higher degree of correlation among the features in the web-page-phishing dataset, as depicted in Figure 7.6. As we mentioned earlier, the Naive Bayes algorithm ignores the correlation between features and assumes them to be independent. Consequently, the presence of strong correlations among the features in the web-page-phishing dataset might negatively impact the performance of the Naive Bayes classifier.

Table 7.5: Classification Report (web-page-phishing with library)

	precision	recall	f1-score	support
0	0.86	0.86	0.86	25313
1	0.75	0.76	0.76	14718
accuracy			0.82	40031
macro avg	0.81	0.81	0.81	40031
weighted avg	0.82	0.82	0.82	40031

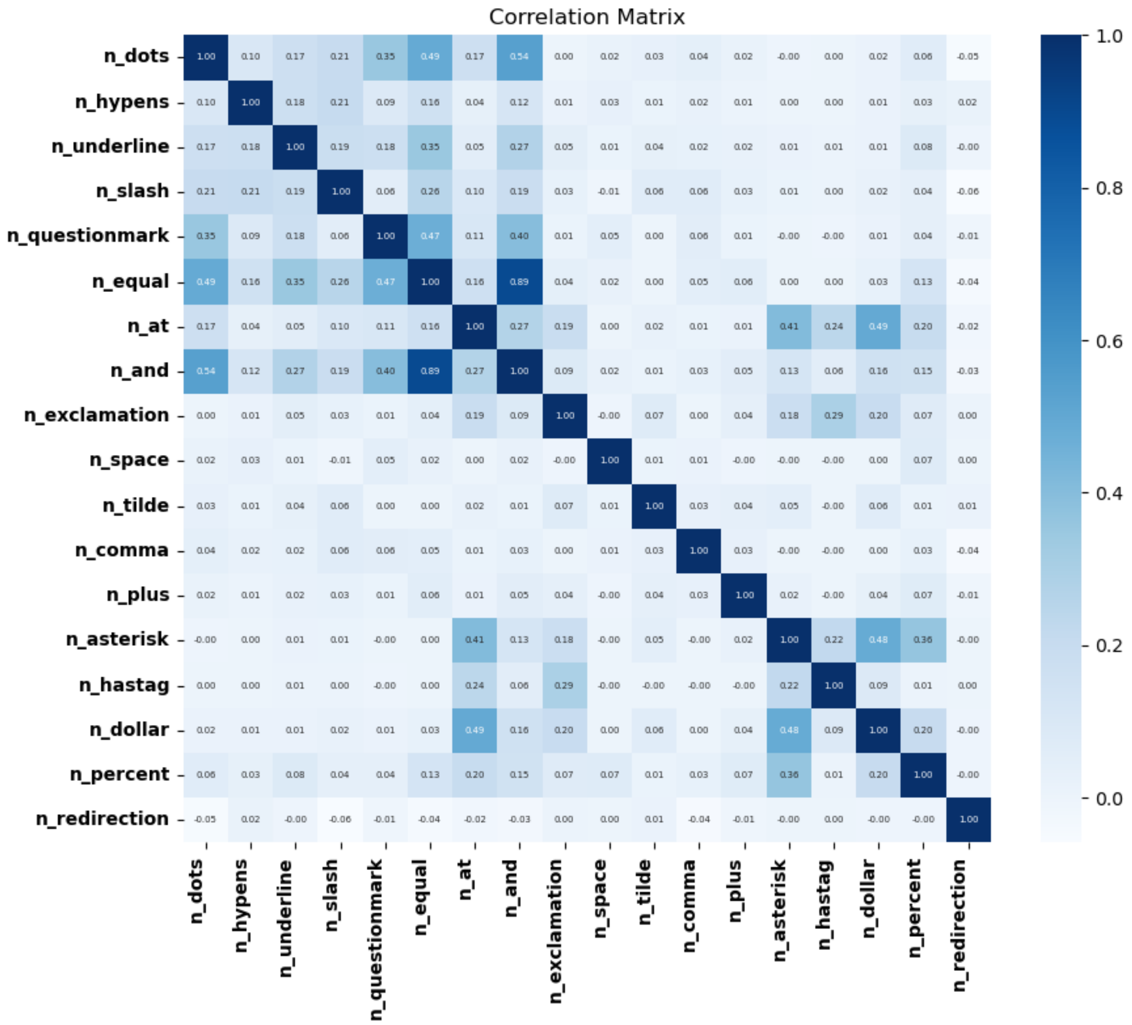


Figure 7.6: Correlation Matrix of web-page-phishing dataset

7.7 Conclusion

In this section, we applied the Naive Bayes algorithm using both the from-scratch implementation and libraries to classify two different datasets. While there wasn't a significant difference between the two implementations of the algorithm, there was a noticeable disparity in the classification results between the two datasets.

The dataset with a higher degree of correlation among its features demonstrated lower accuracy compared to the other dataset. This observation aligns with our expectation that Naive Bayes, does not consider any correlations between features.

Overall, neither classification performed well, which can be attributed to the inherent simplicity and naivety(!) of the Naive Bayes classifier. The naive Bayes algorithm is known for its simplicity, and it assumes feature independence, making it less effective in handling complex relationships within the data.

8 Image Classification

In this section, we will discuss the classification of a binary image dataset. Our goal is to extract a numeric feature from each picture and use it for training and testing. To obtain the center of each class, we will calculate the mean of all the features in the training data of that class. For testing, we will determine the class by comparing the feature of our sample with the center of each class and selecting the class whose center is closest to the feature.

Now, let's address the question of how to extract the feature. We will explore two approaches, which will be explained in the following sections.

8.1 Extract Feature: Mean

We can calculate the mean value of all pixels in an image and use it as the feature for that particular image.

Here are the results obtained from this classifier:

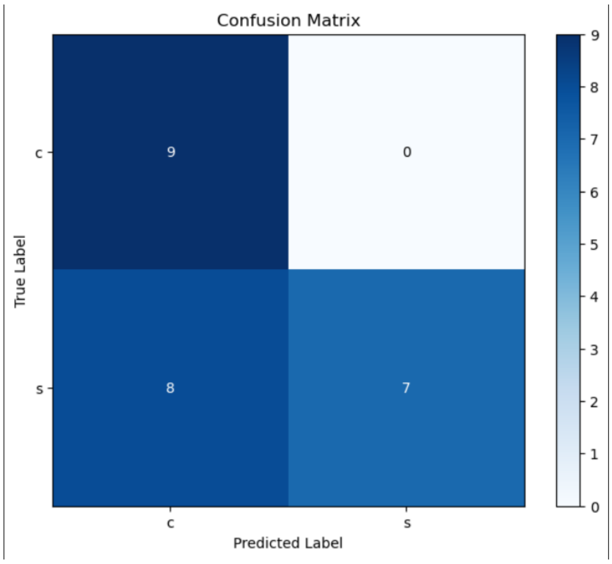


Figure 8.1: Confusion Matrix (feature: mean)

Table 8.1: Classification Report (feature: mean

	precision	recall	f1-score	support
c	0.53	1.00	0.69	9
s	1.00	0.47	0.64	15
accuracy			0.67	24
macro avg	0.76	0.73	0.66	24
weighted avg	0.82	0.67	0.66	24

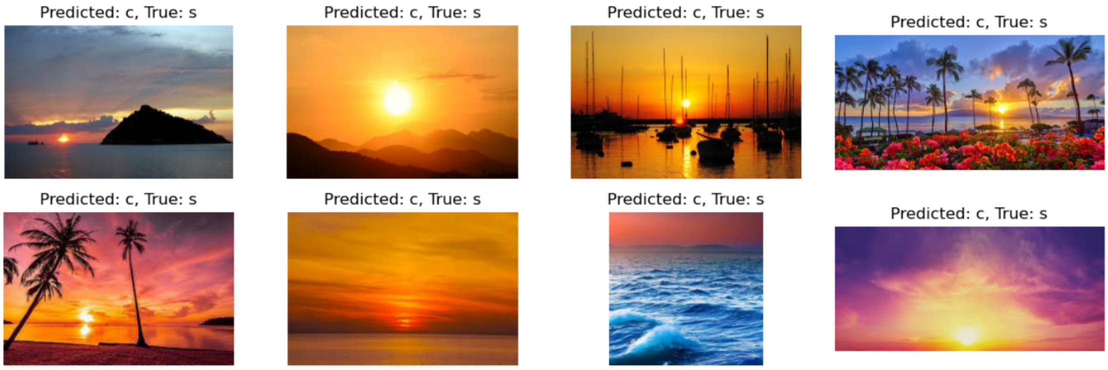


Figure 8.2: False Predicted (feature: mean)

The cloudy images tend to have a higher presence of white color and appear brighter, while the sunny images tend to exhibit more orange tones. Consequently, we expect the center value for the cloudy images to be larger. In practice, this expectation is met.

As we observe zero false predictions for the cloudy label. However, we encounter a significant number of misclassifications, specifically 8, for the sunny label. This can be attributed to the fact that we consider all pixels in the image, including those in sunny images where the orange area is relatively small and other colors dominate. As a result, these images are wrongly predicted.

It seems that the current approach of considering only the mean of all pixels as the feature may not be sufficient to accurately distinguish between cloudy and sunny images.

8.2 Extract Feature: Mode

Let's attempt to classify the images visually, as it is a straightforward process, isn't it? What features does our brain rely on? A more effective method for feature extraction involves examining specific regions of the images and identifying the predominant colors within those regions.

To achieve better results, we divide all pixels in the image by a constant and round down the resulting values. Next, we extract the mode, which represents the color that appears most frequently, from these adjusted pixel values. We experiment with different constants to determine the optimal value that yields the best classification outcome. The reason behind using a constant is to treat the close number as the same number.

Now, let me present the results of this classifier:

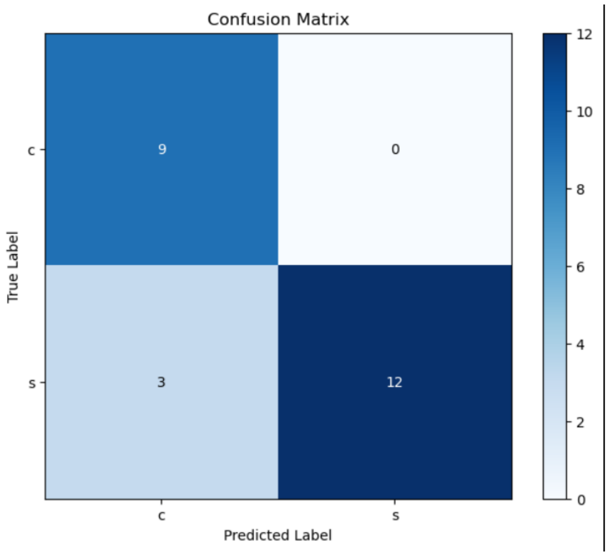


Figure 8.3: Confusion Matrix (feature: mode)

Table 8.2: Classification Report (feature: mode)

	precision	recall	f1-score	support
c	0.75	1.00	0.86	9
s	1.00	0.80	0.89	15
accuracy			0.88	24
macro avg	0.88	0.90	0.87	24
weighted avg	0.91	0.88	0.88	24

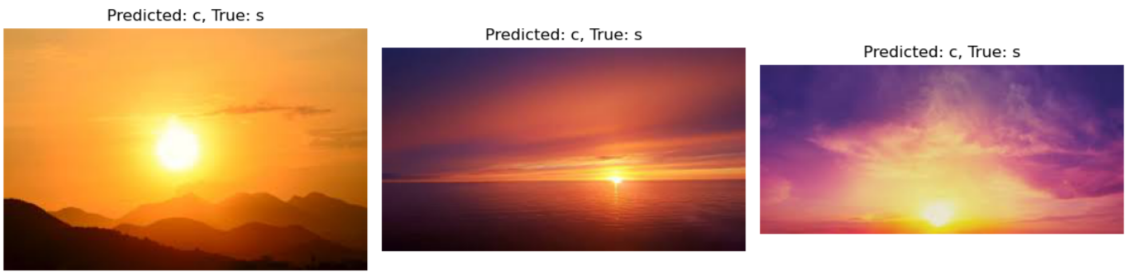


Figure 8.4: False Predicted (feature: mode)

Now, with the improved classification approach, we have achieved better results. All the cloudy images are correctly predicted, and only three sunny images are misclassified. This can be attributed to the fact that in these 3 images, the black parts of the images have more similarity and are effectively captured by the feature extraction process.