سلام بر 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 می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. کدهای ارسال شده بدون گزارش فاقد نمره می باشند.

۴. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW1_StudentNumber داشته باشد.

۶. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل تمرین برای طرفین **صفر** خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل mahdi.taherkhani@ut.ac.ir، سوال خود را مطرح کنید.

سوال ۱: (۱۵ نمره)

در یک مسئله تک بعدی دو کلاسه، توزیع های نرمال برای دو کلاس به ترتیب برابر $N(0,\sigma^2)$ و $N(1,\sigma^2)$ می باشد. نشان دهید که آستانه x_0 برای به حداقل رساندن میانگین ریسک ، با فرض اینکه $\lambda_{11} = 0$ و $\lambda_{22} = 0$ است ، برابر است با:

$$x_0 = 1/2 - \sigma^2 \ln \frac{\lambda_{21} P(\omega_2)}{\lambda_{12} P(\omega_1)}$$

سوال ۲: (۲۰ نمره)

در مسئله طبقه بندی چند کلاسه Multiclass classification :

الف : نشان دهید که قانون تصمیم گیری Bayes احتمال خطا را کمینه می کند.

ب: نشان دهید که در حالت M کلاسه حد بالای احتمال خطا بصورت زیر می باشد:

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

ج : راه حلی برای رسم نمودار ROC در حالت چند کلاسه ارائه دهید.

د : فرض کنید که فضای ویژگی x دارای ستون های کاملا مستقل از هم باشد. با توجه به فرض استقلال ویژگی ها در طبقه بند Bayes naïve آیا این طبقه بند عملکرد بهینه بر روی این داده خواهد داشت یا روش های دیگر می توانند به جواب بهتری دست پیدا کنند.؟ علت جواب خود را توضیح دهید.

سوال ۳: (۱۵ نمره)

متغیر تصادفی x دارای توزیع نرمال $N(\mu,\sigma^2)$ می باشد ، که μ یک پارامتر می باشد که pdf آن به شکل زیر توصیف می شود:

$$p(\mu) = \frac{\mu \exp(-\mu^2/2\sigma_\mu^2)}{\sigma_\mu^2}$$

نشان دهید که تخمین MAP ، μ برابر است با :

$$\hat{\mu}_{MAP} = \frac{Z}{2R}\left(1 + \sqrt{1 + \frac{4R}{Z^2}}\right)$$

با فرض این که:

$$Z = \frac{1}{\sigma^2}\sum_{k=1}^{N} x_k, \quad R = \frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}$$

سوال ۴: (۲۰ نمره)

فرض کنید که در مساله دو کلاسه ، هر کامپوننت x_i از x میتواند مقادیر باینری (۰ یا ۱) بگیرد و $P(\omega_1) = P(\omega_2) = 0.5$ می باشد. فرض کنید که احتمال به دست آوردن ۱ در هر کامپوننت برابر :

$$p_{i1} = p$$

$$p_{i2} = 1 - p$$

برای قطعیت $p > ½$ را فرض کنید. با نزدیک شدن ابعاد d به بی نهایت، احتمال خطا به صفر نزدیک می شود.در این مساله قصد داریم افزایش تعداد ویژگی ها را در یک نمونه بررسی نماییم.

الف : فرض کنید از کلاس ω_1 یک نمونه $x = (x_1, \ldots, x_d)^T$ گرفته شده است.نشان دهید که maximum likelihood estimate برای p برابر است با :

$$\hat{p} = \frac{1}{d}\sum_{i=1}^{d} x_i$$

ب : رفتار $\hat{p}$ را با نزدیک شدن d به بی نهایت توصیف کنید. نشان دهید که چرا چنین رفتاری به این معنی است که اگر تعداد ویژگی ها ، بدون محدودیت افزایش یابد ، می‌توانیم یک طبقه بند بدون خطا به دست آوریم، حتی اگر از هر کلاس فقط یک نمونه داشته باشیم.

ج : فرض کنید $T = \frac{1}{d}\sum_{j=1}^{d} x_j$ نمایانگر نسبت 1 ها در یک نمونه می باشد.نمودار $P(T|\omega_i)$ برحسب T را برای حالتی که P = 0.6، برای d کوچک و برای d بزرگ (به عنوان مثال،به ترتیب d = 11 و d = 111) رسم کنید و پاسخ خود را توضیح دهید.

سوال ۵: (شبیه سازی، ۲۰ نمره)

برای این قسمت از دیتاست data_cancer_Breast استفاده می نماییم ، که این دیتاست شامل ۲ کلاس و ۵ ویژگی می باشد .

الف : طبقه بندی naïve bayes و optimal bayes را به صورت مختصر شرح دهید و با هم مقایسه نمایید.
سپس با فرض گاوسی بودن داده ها الگوریتم naïve bayes را **بدون استفاده از کتابخانه** پیاده سازی نمایید.

ب : الگوریتم پیاده سازی شده را بر روی دادگان ، تست نمایید و نتایج (ماتریس آشفتگی و دقت و precision و recall) را بررسی و تحلیل نمایید.

ج : حال در این قسمت دو الگوریتم ذکر شده را با کمک کتابخانه بر روی دادگان ، تست نمایید و نتایج را با قسمت قبل مقایسه و بررسی نمایید.

سوال ۶: (شبیه سازی، ۲۰ نمره)

در این قسمت از دیتاست image استفاده نمایید و اقدام به طراحی طبقه بند ۲ کلاسه برای دو تیم **منچستریونایتد** و **چلسی** نمایید. الگوریتم پیاده سازی شده را بر روی دادگان ، تست نمایید و نتایج (ماتریس آشفتگی و دقت و precision و recall) را بررسی و تحلیل نمایید.

راهنمایی: برای طراحی طبقه بند می توانید ، میانگین رنگ در هر عکس را محاسبه نمایید، و با رنگ آبی و قرمز مقایسه نمایید.