لطفا قبل از شروع تمرین به نکات زیر توجه فرمایید:

۱. حجم گزارش به هیچ عنوان معیار نمره‌دهی 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 کنید.
۴. شکل ها، به طور واضح و در فرمت درست گزارش شود.
۵. مجموع نمرات تمرین ۱۵۰ نمره است که ۱۱۰ نمره (۱۰ نمره امتیازی) لحاظ خواهد شد. از بین تمارین شبیه‌سازی، دو مورد را حتما انجام دهید.
۶. **هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر میشود.**
۷. در صورت داشتن سوال، از طریق ایمیل ati.noorzad@gmail .com و یا farhoodetaati@gmail.com ، سوال خود را مطرح کنید.

۱ فرض کنید $a, b > 0$:

أ. نشان دهید در حالی که $a \leq b$ باشد، شرط $a \leq (ab)^{\frac{1}{2}}$ برقرار است. (۵ نمره)

ب. نشان دهید که اگر ناحیه‌ی تصمیم‌گیری یک مسئله‌ی طبقه‌بندی دو کلاسه را در نظر داشته باشیم، به صورتی که این نواحی به گونه‌ای انتخاب شده باشند که خطای طبقه‌بندی حداقل شود، شرط زیر برقرار خواهد بود. (۵ نمره)

$$p(mistake) \leq \int \{p(x, C_1)p(x, C_2)\}^{\frac{1}{2}} dx$$

۲ به سوالات زیر پاسخ دهید.

أ. ابتدا هر یک از معیارهای Accuracy، Precision، Recall را توضیح دهید. (۵ نمره)

ب. برای تشخیص بیماری کویید ۱۹ یک دستگاه تشخیص ساخته شده است که هر فرد را به یکی از دو کلاس بیمار (۱) و سالم (۰) تقسیم می‌کند. معیاری را انتخاب کنید که با بهینه‌سازی عملکرد دستگاه برای آن معیار، بتوان به کادر درمان در تشخیص بیماری بیشترین اطمینان را داد. چرا؟ (توجه کنید که ریسک تشخیص ندادن بیماری فرد بیمار، با ریسک تشخیص دادن بیماری برای یک فرد سالم یکی نیست.) (۵ نمره)

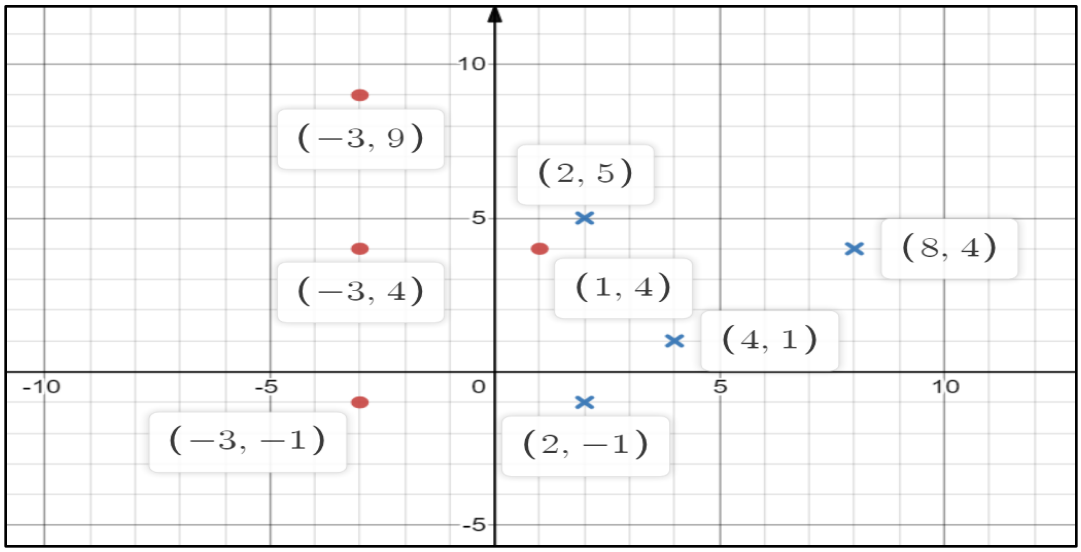
ج. مفهوم Discriminability را توضیح دهید و بیان کنید برای مقایسه‌ی دو متغیر تصادفی نرمال Discriminability را به چه فرمتی تعریف می‌کنیم. سپس حساسیت این مفهوم را به تغییر ممان‌های متغیرهای تصادفی تشریح کنید. (۵ نمره)

د. هدف از طبقه‌بندی را بیان کنید و تفاوت رویکردهای Generative و Discriminative برای حل مسائل مربوط به آن را شرح دهید. (۵ نمره)

ه. ویژگی‌های منحنی ROC را برای طبقه‌بند ایده‌آل بیان کرده و علت صعودی بودن این منحنی را به صورت شهودی بیان کنید. (۵ نمره)

۳ مسئله‌ی دو کلاسه‌ای را در نظر بگیرید که $P(x|w_1)$ دارای توزیع نرمال با میانگین μ و انحراف معیار σ و $P(x|w_2)$ بین دو مقدار غیر صفر a و b توزیع یکنواخت دارد. حد بالای خطای تصمیم‌گیری بیز را برحسب نرمال استاندارد بدست آورید. (۱۰ نمره)

۴ کلاس مربوط به نقطه‌ی (۲و۲) براساس روش‌های طبقه‌بندی، معیارهای فاصله‌سنجی و داده‌های ترسیم‌شده را در جدول زیر مشخص کنید. (۱۸ نمره)



	Nearest Neighbor	Nearest Centroid
Chebyshev Distance		
Manhattan Distance		
Euclidean Distance		

۵ متغیر تصادفی N بعدی گوسی X با توزیع به فرم $N(x|\mu,\Sigma)$ درنظر بگیرید، به‌طوری‌که در آن μ و Σ به ترتیب بردار میانگین و ماتریس کواریانس این توزیع می‌باشند. ماتریس کواریانس این توزیع را می‌دانیم و می‌خواهیم با استفاده از یک فضای نمونه‌ی Kتایی، $X=\{x_1,x_2,\dots,x_k\}$ بردار میانگین را تخمین بزنیم. در این صورت اگر $p(\mu)=N(\mu|\mu_0,\Sigma_0)$، توزیع $p(\mu|X)$ را به دست آورید. (۱۱ نمره)

۶ **(شبیه‌سازی)** با دستور زیر یک مجموعه داده که با نویز جمع شده است درست کنید:

```
import numpy as np

x = np.linspace(-5, 5, num=20)
rng = np.random.default_rng(42)
y = - 0.5 *(x ** 3) + (2 * x**2) + x + 4
y_noisy = y + 5 * rng.normal(loc=0, scale=1, size=len(x))
```

$$y = -\frac{x^3}{2} + 2x^2 + x + 4$$

أ. با استفاده از پیاده‌سازی روش حل معادله‌ی نرمال و تخمین کمینه‌ی مربعات، داده‌ی y_{noisy} را با یک چندجمله‌ای درجه ۳ تخمین بزنید و مقادیر ثابت‌های تخمین‌زده شده‌ی چند جمله‌ای را مخابره کنید. هم‌چنین داده‌های اصلی، نویزی و خروجی تخمین‌گر خود را در یک scatter plot رسم کنید. (۱۰ نمره)

ب. با استفاده از پیاده‌سازی روش تخمین نزول گرادیان، مسئله‌ی قبل را حل نمایید. (۱۰ نمره)

۷ **(شبیه‌سازی)** در این سوال از دادگان random_dataset.csv استفاده کنید و در صورت استفاده از پکیج‌های آماده‌ی یادگیری ماشین نمره‌ای به شما تعلق نمی‌گیرد.

ابتدا مجموعه دادگان را نمایش دهید. سپس به کمک روش‌های One vs All و Logistic Regression کلاس‌های مختلف را جدا کنید. در انتها، خطوطی که کلاس‌ها را از هم جدا می‌کنند را توسط پارامترهای آموزش‌داده‌شده نمایش دهید و توضیح دهید مشکل روش One vs All را با استفاده از خطوط ترسیمی بیان کنید.

دقت کنید که ستون اول مربوط به ویژگی اول، ستون دوم مربوط به ویژگی دوم و ستون سوم برچسب داده‌هاست. (۱۵ نمره)

۸ **(شبیه‌سازی)** دادگان digits را با استفاده از دستور زیر در پایتون کنید:

```
from sklearn.datasets import load_digits
X, y = load_digits(return_X_y = True)
```

با استفاده از دستور train_test_split، فضای داده را با نسبت ۰/۸ و ۰/۲ به مجموعه‌ی آموزش و ارزیابی تقسیم کنید. در این تابع random_state را برابر ۴۲ قرار دهید تا توانایی بازتولید همین مجموعه را در چند بار اجرا داشته باشید.

أ. با استفاده از یک طبقه‌بند چندهمسایگی، سعی کنید تا دادگان آموزش را طبقه‌بندی کنید. برای دادگان تست ماتریس آشفتگی را در یک نمودار رسم کنید و با استفاده از دستور classification_report در ماژول sklearn.metrics عملکرد طبقه‌بند را بر روی دادگان ارزیابی گزارش کنید. (۱۵ نمره)

ب. با استفاده از طبقه‌بند GassianNB سعی کنید تا مرحله‌ی قبل را تکرار کنید. (۶ نمره)

۹ **(شبیه‌سازی)** با استفاده از multivariate_normal در کتابخانه scipy.stats دو توزیع نرمال با میانگین و ماتریس کواریانس دلخواه شبیه‌سازی کنید. این کار را در دو حالتی که دو توزیع درهمرفتگی زیاد و کم دارند تکرار کنید. برای هر دو حالت بیان شده Decision Boundary را پیدا کرده و به همراه دو توزیع ترسیم کنید. (ترسیم را در حالت سه‌بعدی یا به صورت کانتور انجام دهید.) (۲۰ نمره)