به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

دانشگاه تهران

تمرین سری دوم یادگیری ماشین

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

۱. در سوالات پیاده‌سازی حجم گزارش به هیچ عنوان معیار نمره دهی 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. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

۴. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.

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

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

۷. در هیچ یک از سوالات شبیه‌سازی امکان استفاده از پکیج‌های آماده (مثل sklearn و موارد مشابه) **وجود ندارد**، مگر اینکه در صورت سوال گفته شده باشد که مجاز به استفاده هستید.

۸. در صورت داشتن سوال، از طریق ایمیل های زیر سوال خود را مطرح کنید:

سوالات ۱، ۲، ۶، ۹ : m.dadkhah99@gmail.com

سوالات ۳، ۴، ۵، ۷، ۸ : mstfmasoudii@gmail.com

سوال ۱: (۱۰ نمره)

نشان دهید که کران بالای واریانس تابع pdf تخمین زده شده توسط فرمول ۱، مطابق فرمول ۲ بدست می آید.

$$\hat{p}(x) = \frac{1}{V_N}\left(\frac{1}{N}\sum_{i=1}^{N}\phi\left(\frac{x_i - x}{V}\right)\right)$$

فرمول ۱

$$\sigma_N^2(x) \leq \frac{\sup(\phi)\mathrm{E}[\hat{p}(x)]}{NV_N}$$

فرمول ۲

در فرمول ۲ $\sup(.)$ کران بالای تابع مورد نظر است.

سوال ۲: (۱۰ نمره)

توزیع یکنواخت p(x) و پنجره پارزن φ(x) به صورت زیر تعریف شده است.

$$p(x) \sim U(0, a)$$

$$\varphi(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

۱. نشان دهید میانگین چنین تخمینی از پنجره پارزن به صورت زیر می‌شود.

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}(1 - e^{-\frac{x}{h_n}}) & 0 \leq x \leq a \\ \frac{1}{a}(e^{\frac{a}{h_n}} - 1)e^{-\frac{x}{h_n}} & a \leq x \end{cases}$$

۲. h_n چه قدر باشد تا در بازه $0 < x < a$ مقدار بایاس کمتر از ۱ درصد باشد.

سوال ۳: (۱۰ نمره)

۱. یک رگرسیون خطی با L2 Regularization به صورت زیر در نظر بگیرید:

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda||\beta||_2^2$$

در رابطه بالا $\lambda \geq 0$ پارامتر L2 Regularization و $X_i = \left[X_i^{(1)} \dots X_i^{(d)}\right]$ است. با فرض $Y = [Y_1, \dots, Y_n]$, $A = [X_1, \dots, X_n]$، ثابت کنید جواب فرم بسته $\hat{\beta}$ برابر عبارت زیر است:

$$\hat{\beta} = (A^T A + \lambda I)^{-1} A^T Y$$

۲. L1 Regularization و L2 Regularization را تعریف کرده و تفاوت های آنها را بیان کنید. (حداقل ۳ مورد)

سوال ۴: (۱۰ نمره)

فرض کنید مجموعه‌ی زیر از بردارهای دوبعدی وجود دارد:

W1		W2		W3	
X1	X2	X1	X2	X1	X2
10	0	5	10	2	8
0	-10	0	5	-5	2
5	-2	5	5	10	-4

۱. مرز تصمیمی که نتیجه قانون نزدیک‌ترین همسایه(nearest-neighbor rule) فقط برای دسته‌بندی w_1 و w_2 را تولید می‌کند را رسم کنید. میانگین نمونه m_1 و m_2 را پیدا کرده و در همان نمودار مربوطه مرز تصمیمی را که متعلق به طبقه‌بندی x با اختصاص آن به دسته‌ی میانگین نمونه نزدیکتر است، رسم کنید. (نیاز نیست برای رسم شکل‌ها حساسیت به خرج دهید، با کمی تخمین هم قابل پذیرش است، مهم‌تر روش صحیح حل است)

۲. بخش ۱ را برای دسته‌بندی تنها w_1 و w_3 تکرار کنید.

۳. بخش ۱ را برای دسته‌بندی تنها w_2 و w_3 تکرار کنید.

۴. بخش ۱ را برای یک طبقه‌بند سه دسته‌ای تکرار کنید، که w_1، w_2 و w_3 را دسته‌بندی می‌کند.

سوال ۵: (۱۰ نمره)

طبقه‌بندهای مبتنی بر نمونه که از توزیع‌های زیر پیروی می‌کنند را در نظر بگیرید.

$$p(x|w_1) = \begin{cases} \frac{3}{2}, & 0 \le x \le \frac{2}{3} \\ 0, & otherwise, \end{cases}$$

$$and$$

$$p(x|w_2) = \begin{cases} \frac{3}{2}, & \frac{1}{3} \le x \le 1 \\ 0, & otherwise. \end{cases}$$

۱. قانون تصمیم بیز (Bayes Decision Rule) و خطای طبقه‌بند بیز چیست؟

۲. فرض کنید یک نقطه را به صورت تصادفی از w_1 و یک نقطه را از w_2 انتخاب کرده و یک طبقه‌بند بر اساس نزدیک‌ترین همسایه (Nearest-Neighbor Classifier) ایجاد کنیم. همچنین فرض کنید یک نقطه تست از یکی از دسته‌ها انتخاب می‌کنیم (برای مثال w_1). نرخ خطای مورد انتظار $P_1(e)$ برای این طبقه بند را بدست آورید.

سوال ۶: (**شبیه سازی**، ۱۰ نمره)

برای متغیر تصادفی X، تابع pdf در

فرمول ۳آمده است:

$$p(x) = \begin{cases} \frac{1}{2} & for\ 0 < x < 2 \\ 0 & otherwise \end{cases}$$

فرمول ۳

با استفاده از پنجره پارزن با کرنل نرمال استاندارد، تابع p(X) را تخمین بزنید. برای اینکار پارامتر V_N را برابر مقادیر 0.05 و 0.2 قرار داده و برای هر کدام تخمین خود را به ازای تعداد نقاط ۳۲، ۲۵۶ و ۵۰۰۰ نمایش دهید.

سوال ۷: (**شبیه سازی**، ۱۰ نمره)

در این سوال قصد داریم با استفاده از Parzen-window estimates and classifiers به حل یک مسئله classification با ۳ کلاس بپردازیم. ابتدا شما باید با استفاده از کتابخانه **Numpy**، ۵۰ داده برای هریک از سه کلاس بصورت رندوم و با توزیع نرمال با میانگین به ترتیب [۰,۰]،[۲-,۲-] و[۲,۲] و واریانس [۱,۱] تولید کنید. دقت کنید که داده ها ۲ بعدی هستند. سپس مراحل زیر را انجام دهید:

۱. داده های تولید شده برای هر کلاس را در **یک** نمودار نمایش دهید.

۲. یک Parzen-window classifier با تابع کرنل زیر را با استفاده از نقاط تولید شده آموزش دهید.

۳. ۴ نقطه تست دلخواه در محدوده کلاس ها در نظر بگیرید.

۴. مقدار V_n را برابر ۱ و ۰.۱ قرار داده و برای هرکدام، ۴ نقطه تست دلخواه را طبقه بندی کنید.

۵. این‌بار برای هر کلاس ۵۰۰ داده تولید کرده و مراحل قبل را تکرار کنید. نتایج بدست آمده را با نتایج قبلی مقایسه کرده و تحلیل خود را ارائه دهید.

سوال ۸: (شبیه سازی، ۲۰ نمره)

در این سوال می خواهیم با استفاده از دیتاست Wheat Seeds[1] به حل مسئله classification بپردازیم. در این دیتاست ویژگی های مربوط به ۳ نوع دانه گندم آورده شده است.

الف) EDA

در این قسمت برای درک بهتر دادگان، نمودار scatter plot هر دوتایی از ویژگی ها را ترسیم کنید. توجه داشته باشد که نمودار ترسیم شده باید با درج تمامی اطلاعات مورد نیاز(برچسب مناسب برای محورها، عنوان مناسب برای هر نمودار و ...) همراه باشد. حال از میان نمودار های رسم شده مشخص کنید که کدام یک از دو ویژگی ها می تواند با دقت بیشتری کلاس ها را از هم جدا کند. همچنین برای هر ویژگی، توزیع کلاس های آن را با استفاده از نمودار histogram ترسیم نمایید. توجه کنید که تحلیل نمودار ها در این سوال اهمیت بالایی دارد.

ب) Preprocessing & Normalization

در این مرحله هرگونه پیش پردازش و یا نرمال سازی که لازم است را بر روی دادگان انجام داده و در گزارش علت استفاده از هر روش را ذکر کنید.

ج) Classification

داده ها را به صورت تصادفی و با نسبت مشخص به داده های آموزش و تست تفکیک کنید و یک طبقه بند چند کلاسه با استفاده از Logistic Regression و تکنیک one against all پیاده سازی کنید. برای این قسمت موارد زیر را گزارش کنید:

Accuracy, Precision, Recall, F۱-score, Confusion Matrix

[1] https://www.kaggle.com/datasets/jmcaro/wheat-seedsuci

حتما توجه داشته باشید که تمامی مراحل این قسمت باید بصورت دستی پیاده سازی شود و امکان استفاده از پکیج های آماده را ندارید.

د) KNN Classifier

همانطور که می‌دانید می‌توان با استفاده از الگوریتم KNN و بررسی K تا نزدیک ترین نمونه‌های آموزشی به هر نمونه‌ی تست، طبقه‌بندی انجام داد. یک طبقه بند با همین الگوریتم پیاده سازی کرده و نمودار دقت داده‌های تست، برای مقادیر مختلف k (از ۱ تا ۱۰) را بدست آورده و رسم نمایید و بر اساس آن بهترین K را گزارش کنید.

سوال ۹: (**شبیه سازی**، ۲۰ نمره)

هدف از این سوال، استفاده از یک مدل Linear Regression جهت پیش بینی تعداد خرید های کالای مشتریان یک فروشگاه است. داده هایی که در این سوال از آنها استفاده می کنیم، شامل اطلاعات مربوط به خریدهای مشتریان یک فروشگاه و نیز ویژگی های شخصیتی آنها میباشد. (دیتاست این سوال در ضمیمه فایل تمرین آورده شده است).

الف)EDA

مشابه سوال قبل، در این قسمت نیز لازم است تا برای داشتن دید بهتر نسبت به داده ها، نمودار های لازم را ترسیم کرده و آنها را تحلیل کنید. مواردی که انتظار می رود حتما به آنها اشاره شود به صورت زیر است:

- نسبت داده های از دست رفته برای هر ویژگی
- نمودار scatter plot و histogram برای ویژگی ها
- بررسی وابستگی میان ویژگی ها و نیز وابستگی هر ویژگی با ستون هدف

بعلاوه می توانید هر بررسی دیگری که به شناخت داده ها کمک می کند را پیاده سازی و تحلیل کنید.

ب) Preprocessing & Normalization

در این مرحله لازم است تا هرگونه پیش پردازش و نرمال سازی که لازم است را بر روی دادگاه پیاده سازی و تحلیل کنید. مواردی که انتظار می رود حتما به آنها اشاره شود به صورت زیر است:

- Handling missing values
- Train/Test Split (0.8/0.2)

ج) مدل سازی

در این قسمت که خود از ۲ بخش تشکیل شده است، هدف پیش بینی کردن تعداد خرید های یک مشتری از فروشگاه می باشد که در ستون NumPurchases مقدار واقعی آن، آمده است.

در بخش اول به ساخت یک مدل linear regression مرتبه ۱ می‌پردازیم. شما لازم است برای این قسمت یک تابع بصورت **دستی** پیاده سازی کرده و یک ویژگی را به عنوان ورودی این تابع انتخاب نمایید. شما باید با استفاده از فرمول $y = ax + b$ و با در نظر گرفتن تابع خطای RMSE اقدام برای پیدا کردن مقادیر بهینه a, b کنید و با حل دستگاه دو معادله دو مجهول به یک فرمول برای مقادیر بهینه برسید و از این طریق جواب مناسب را پیدا کنید. با توجه به تحلیل هایی که در بخش EDA انجام دادید، ویژگی مورد نظر را انتخاب کرده و علت انتخاب خود را توضیح دهید. سپس مدل مورد نظر را رو داده های Train آموزش داده و خروجی آن را بر روی داده های تست، ارزیابی کنید. (با استفاده از روش RMSE و R2 Score مقادیر پیش بینی شده را ارزیابی کنید)

در مرحله قبل، توانستیم با استفاده از دو معادله و دو مجهول به مقادیر بهینه وزن‌ها برسیم. با افزایش تعداد ویژگی‌ها، حل این دستگاه بسیار دشوار میشود و نیاز به روشی هست که بتوان مرحله به مرحله به وزن‌های بهینه نزدیک شویم. شما در بخش ۲، رگرسیون را روی چندین ویژگی انجام می‌دهید. در این قسمت با استفاده از روش گرادیان کاهشی، یک مدل multiple regression ساخته و مدل را به ازای ۳ ویژگی اجرا کنید. انتخاب ویژگی ها دست شماست ولی باید برای انتخاب خود دلیل داشته باشید. دقت مدل جدید را با بخش قبل مقایسه کرده و عملکرد آن را توضیح دهید.