دستورالعمل

قبل از حل سوالات این بخش، لطفا به جدول زیر توجه کنید:

الزامات	توضیحات	ملاحظات
استفاده از ابزارهای هوش مصنوعی	استفاده از این ابزارها در این تمرین بلامانع است ولی لازم است تا مفاهیم پشت الگوریتم‌ها و مفاهیم را کاملا متوجه شده باشید.	اگر چه فضا برای همفکری باز است ولی لطفا فقط پاسخ‌های خودتان را ارسال کرده و از کپی کاری پرهیز کنید.
آدرس ایمیل	**Mohammadjavad.mousavi۹۷@gmail.com**	**Feel free to say hello!**

[سوال] رگرسیون خطی

در یک مسئله رگرسیون می‌خواهیم رابطه بین ورودی و مقدار خروجی را به صورت زیر مدل کنیم:

$$y = e^{wx}$$

که در رابطه بالا $y \in \mathbb{R}$ و $\mathbf{x} \in \mathbb{R}$ و $\mathbf{w} \in \mathbb{R}$ پارامتر مدل است. فرض کنید مجموعه داده‌ی آموزشی زیر را در اختیار داریم:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

أ. تابع هزینه‌ی مجموع مجذور خطا (**MSE**) را برای مجموعه داده‌ی **D** تشکیل دهید.

ب. اگر بخواهیم با استفاده از روش گرادیان کاهشی مقدار بهینه **w** را بدست آوریم، رابطه بروزرسانی **w** چه خواهد بود.

ت. با انجام محاسبات نشان دهید که برای کمینه کردن تابع هزینه، مقدار بهینه پارامتر **w** در رابطه زیر باید صدق کند.

$$\sum_{i=1}^{n} x_i e^{2wx_i} = \sum_{i=1}^{n} x_i y_i e^{wx_i}$$

(راهنمایی: از رابطه $\frac{\partial L}{\partial w} = 0$ استفاده کنید.)

[سوال] گرادیان نزولی

در این بخش به سوالات مربوط به مبحث بهینه سازی خطا و گرادیان نزولی پاسخ دهید.

أ. چرا در روش گرادیان نزولی **(Gradient Descent)** باید نرخ یادگیری **(Learning Rate)** را به درستی انتخاب کنیم؟ نرخ یادگیری بیش از حد کوچک یا بزرگ چه مشکلاتی ایجاد می‌کند؟

ب. ابتدا با رسم یک شکل ساده نشان دهید که روش گرادیان نزولی ممکن است در کمینه‌سازی محلی دچار مشکل شود. سپس، یک راه حل برای این مشکل به طور خلاصه پیشنهاد دهید.

[سوال] رگرسیون خطی(شبیه سازی)

تصور کنید که تحصیلات خود را به پایان رسانده‌اید و از گذشته علاقه به تحصیل در یک دانشگاه خاص داشته‌اید. حال برای دریافت پذیرش این دانشگاه اطلاعات مربوط به دادگان ضمیمه شده با نام **Admission** شامل ویژگی-های مختلف برای محاسبه احتمال قبولی در این دانشگاه خاص در اختیار شماست. در این تمرین به دنبال آن هستیم که به کمک این ویژگی‌ها رگرسیون خطی انجام بدهیم. برای این منظور مراحل زیر را دنبال کنید:

أ. دادگان را خوانده و مقادیر آن را مشاهده کنید. در یک نمودار، توزیع چند ویژگی دلخواه را رسم کنید. سپس به کمک یک **pairplot** بررسی کنید که بین ویژگی‌ها کدام دو ویژگی بیشترین همبستگی را با هم دارند و نتایج بررسی را گزارش کنید.

ب. داده‌ها را به دو قسمت آموزش و آزمون تقسیم کنید. ستون **GRE Score** را به عنوان هدف (متغیر وابسته) و بقیه ستونها را به عنوان ویژگی‌ها (متغیرهای مستقل) در نظر بگیرید. رابطه فرم بسته رگرسیون خطی را نوشته و با استفاده از آن مدل رگرسیون خطی را بسازید.

ت. مقدار خطای مربعات (**MSE**) مدل را روی دادگان اموزش و آزمون گزارش کنید. (انجام پیش پردازش های لازم قبل از این مرحله ضروری است).

ث. در این قسمت می‌خواهیم تاثیر توابع خطای مختلف را بر مقدار ضرایب مشاهده کنیم. یک بار با **lasso** و یکبار با **ridge** این مساله را حل کنید. مقادیر ضرایب را در دو حالت بررسی کنید و بگویید آیا این تغییرات مطابق انتظار شما بوده است یا خیر.

[سوال] رگرسیون خطی(شبیه سازی)

هدف از این سوال، استفاده از یک مدل **Linear Regression** جهت پیش بینی تعداد خریدهای کالای مشتریان یک فروشگاه است. داده‌هایی که در این سوال از آنها استفاده می کنیم، شامل اطلاعات مربوط به خریدهای مشتریان یک فروشگاه و نیز ویژگی‌های شخصیتی آنها می‌باشد. (دیتاست این سوال در ضمیمه فایل تمرین با نام **marketing_campaign.csv** آورده شده است).

یکی از مسائلی که در تحلیل داده‌ها وجود دارد، پاک‌سازی آن‌ها است بطوری که مدل‌سازی بر اساس آن‌ها با صحت همراه باشد. در این بین تشخیص صحت مقادیر متغیرهای موجود، احتیاج به تحلیل کاوشگرانه داده **(Explanatory Data Analysis)** یا به اختصار **EDA** دارد. به بیان دیگر یک متخصص تحلیل داده **(Data Scientist)** باید به چیزی که در پس مقادیر و اعداد قرار دارد پی ببرد و ریشه‌های اصلی تناقضات و مشکلات مجموعه داده **(Data Set)** را پیدا کند.

أ. به عنوان یک تحقیق کوچک چند مورد (حداقل ۴) از مراحلی که در **EDA** باید مورد بررسی قرار گیرد را به اختصار توضیح دهید.

ب. با توجه به تحقیق بخش قبل، برای دید بهتر خودتان و سهولت استفاده از دیتاست در مراحل بعدی، نمودارهای لازم برای دیتاست این سوال را ترسیم کرده و آنها را تحلیل کنید. مواردی که انتظار می رود حتما به آنها اشاره شود به صورت زیر است:

۱) نسبت داده‌های از دست رفته برای هر ویژگی

۲) نمودار **scatter plot** و **histigram** برای ویژگی‌ها

۳) بررسی وابستگی میان ویژگی‌ها و نیز وابستگی هر ویژگی با ستون هدف

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

۱) **Handling missing values**

۲) **Train/Test Split**

(راهنمایی: شما در این مرحله برای ستون‌هایی از ویژگی که مقادیر از دست رفته دارند، از **Random Sampling** استفاده کنید و برای نرمالیزه کردن داده‌ها از روش **min/max** استفاده کنید.)

ث. در این قسمت هدف پیش بینی کردن تعداد خرید های یک مشتری از فروشگاه می باشد که در ستون **NumPurchases** مقدار واقعی آن، آمده است:

در این بخش 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 $b + ax = y$ 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 **R^2 Score** مقادیر پیش بینی شده را ارزیابی کنید)