به نام خدا

تمرین سری اول یادگیری آماری نیمسال اول ۰۳- ۰۲

۱- بردار خطا e در رگرسیون خطی، به صورت زیر تعریف می‌شود:

$$\vec{e} = \vec{y} - \boldsymbol{X}\hat{\vec{\beta}}$$

الف: نشان دهید $\boldsymbol{X}^T\vec{e} = 0$.

ب: نشان دهید $\vec{\hat{y}}^T\vec{e} = 0$.

۲- در صورتیکه داده های آموزش را در ماتریس مربعی و معکوس پذیر $\boldsymbol{Q}$ ضرب کنیم $(\boldsymbol{D} = \boldsymbol{XQ})$ ، نشان دهید بردار خطا در رگرسیون خطی در دوحالتی که از $\boldsymbol{D}$ و $\boldsymbol{X}$ برای تخمین $\vec{y}$ استفاده کنیم یکسان است. یعنی:

$$\vec{e} = \vec{y} - \boldsymbol{X}\hat{\vec{\beta}} = \vec{y} - \boldsymbol{D}\overrightarrow{\hat{\beta}'}$$

۳- فرض کنید مقدار بهینه $\vec{\beta}$ با استفاده از یک مجموعه داده محاسبه شده است. حال میخواهیم یک نمونه جدید $(\vec{x}_a, y_a)$ را به داده های آموزش اضافه کنیم. نشان دهید مقدار بهینه جدید $\vec{\beta}_{new}$ از رابطه زیر بدست می‌آید.

$$\vec{\beta}_{new} = \vec{\beta} + \frac{1}{1 + \vec{x}_a{}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\vec{x}_a}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\vec{x}_a(y_a - \vec{x}_a{}^T\vec{\beta})$$

توجه: در صورتی که $\boldsymbol{A}$ و $\boldsymbol{B}$ ماتریس مربعی و معکوس پذیر باشند و $Rank(\boldsymbol{B}) = 1$ باشد، رابطه زیر برقرار است:

$$(\boldsymbol{A} + \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1 + trace(\boldsymbol{B}\boldsymbol{A}^{-1})}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}.$$

راهنمایی ۱: می‌توانید ماتریس داده های ورودی و بردار خروجی جدید را به صورت زیر در نظر بگیرید:

$$\vec{y}_{new} = \begin{bmatrix} \vec{y} \\ y_a \end{bmatrix} \boldsymbol{X}_{new} = \begin{bmatrix} \boldsymbol{X} \\ \vec{x}_a \end{bmatrix}$$

راهنمایی ۲: برای سه ماتریس $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$،خاصیت $tr(\boldsymbol{ABC}) = tr(\boldsymbol{BCA})$ برقرار است و $tr()$ یک اسکالر با خود آن برابر خواهد بود.

۴- با در نظر گرفتن ورودی و خروجی به صورت زیر و به کمک رگرسیون خطی یک متغیره، $\hat{\vec{\beta}}$ و تخمین واریانس نویز ($\hat{\sigma}^2$) را محاسبه کنید.

X	Y
2	19.73
3	26.94
4	35.71

۵- فرض کنید داده آموزشی زیر را در اختیار داریم:

X	Y
1	3
2	1
3	0.5

و بخواهیم از مدل $\hat{y} = \hat{a}x^{\hat{b}}$ برای تخمین رابطه ی ورودی و خروجی استفاده کنیم. با استفاده رگرسیون خطی، ضرایب $\hat{a}$ و $\hat{b}$ را محاسبه کنید.

(**راهنمایی**: از طرفین معادله $\hat{y} = \hat{a}x^{\hat{b}}$، لگاریتم بگیرید.)

۶- هدف کلی این سوال آشنایی با عمل رگرسیون به عنوان یکی از مقدماتی ترین الگوریتم های ML می باشد. در این سوال ما از این dataset استفاده میکنیم تا مصرف سوخت ماشین ها را پیش بینی کنیم.
Dataset را به دقت مورد برسی قرار دهید و به نکات زیر توجه کنید.

- در این dataset ما ۹ ستون داریم که اسم آنها به ترتیب از چپ به راست به صورت زیر می باشد:

```
column_names = ['MPG', 'Cylinders', 'Displacement', 'Horsepower',
'Weight','Acceleration', 'Model Year', 'Origin', 'Car Name']
```

- ستون اول در واقع همان y ما می باشد و ستون های دیگر ویژگی های دیگر ما هستند (ستون های ماتریس $\boldsymbol{X}$می باشند)
- توجه کنید که car name به عنوان مثال به صورت bmw m5 در dataset آورده شده است اما ما در این سوال فرض میکنیم مدل ماشین در مصرف بنزین اهمیتی ندارد و تنها به برند ماشین دقت میکنیم یعنی bmw i20 تفاوتی با bmw m5 ندارد.
- نکته ی دیگری که در این سوال وجود دارد بعضی مقادیر در این dataset با ؟ نشان داده شده است که باید این مشکل را رفع کنید.
- توجه کنید که جنس Car Name از نوع string است و به طور مستقیم نمی توانید آنرا در مدل رگرسیون خود وارد کنید پس باید این مورد را هم در data processing خود مورد توجه قرار دهید.

بخش های سوال:

۱. Dataset را وارد کرده و تغییرات مورد نیاز را روی آن اعمال کنید. (بر اساس مراحل پردازش توضیح داده شده در بالا)
۲. Data را طوری تقسیم کنید که به صورت رندوم 80 درصد آن train و ۲۰ درصد آن داده هایtest باشد.(random_state=42)
۳. سه ویژگی را به دلخواه انتخاب کنید و نمودار مصرف بنزین را بر اساس هر یک از این ویژگی ها بکشید. (سه نمودار در یک figure رسم شود. همچنین در صورت لزوم توضیحات ارئه دهید)
۴. مدل خود را بر تمامی ویژگی ها آموزش دهید و مصرف سوخت نمونه های تست را پیش بینی کنید و MSE را بدست آورید و همچنین بردار ضرایب را گزارش کرده و $y_{t_{pre}}$ را بر اساس y_t رسم کرده و با نیم ساز ربع اول مقایسه کنید.
۵. در این بخش به آنالیز وابستگی میان ویژگی ها میپردازیم، ماتریس وابستگی را تشکیل داده سپس به کمک کتابخانه ی seaborn و دستور heatmap این ماتریس را نشان دهید.(نظرتان را درباره ی این

تصویر شرح دهید)، سپس ویژگی هایی که وابستگی زیادی دارند را حذف کنید و بار دیگر مرحله ی ۴ را تکرار کنید و نظر خود را درباره ی تفاوت بخش ۴ و ۵ شرح دهید.

در شکل های خواسته شده label بندی و... را به خوبی انجام دهید و در گزارش، موارد خواسته شده را به خوبی توضیح دهید.