به نام خدا

یادگیری آماری

بهار ۱۴۰۴

مدرس: دکتر محمدزاده

دانشگاه صنعتی شریف
دانشکده مهندسی برق

تمرین اول

گردآورندگان: رامتین ملک‌پور، محمدحسین قریب (تئوری) / زینب شریفی ، محمدجواد محمدی (عملی)

نکات مهم پیش از انجام تمرین:

۱. تمرین‌ها پیش از تقدیم شدن به دانشجویان گرامی، پیاده‌سازی و بررسی شده‌اند. با این حال در صورت وجود هرگونه ابهام در تمرین، آن را در گروه تلگرام مربوط به درس مطرح فرمایید.
۲. هرگونه مشابهت بین تمرین‌های دانشجویان غیرمجاز است؛ لذا از کپی کردن صرف از منابع اینترنتی یا سایر دانشجویان خودداری فرمایید. در صورت مشاهده، نمره آن سوال به هیچکدام از عزیزان تعلق نخواهد گرفت.
۳. دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطا هنگام اجرا، حتی به دلیل اشتباه تایپی، نمره صفر به آن بخش تعلق خواهد گرفت.
۴. لطفا تصویری واضح از پاسخ سوالات تئوری خود آپلود کنید. در غیر اینصورت تمرین شما تصحیح نخواهد شد.
۵. کدها، نتایج و گزارشکار بخش عملی را میتوانید به فرمت ipynb. تحویل دهید.
۶. فایل های مربوط به تمرین را به فرمت HW#_StudentNumber_StudentName.zip آپلود کنید.

سوالات تئوری — **۱۰ + ۱۰۰ نمره**

تمرین اول: (۲۵ نمره)

داده‌های زیر را در نظر بگیرید.

x	1	2	3	4	5
y	2	3	5	7	8

اگر بخواهیم با مدل رگرسیون خطی زیر رابطه بین مقادیر x و y را مدل کنیم:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

الف) ابتدا عبارت مقادیر بهینه ضرایب رگرسیون $\widehat{\beta_0}$ و $\widehat{\beta_1}$ را بدست آورید. سپس با جایگذاری، مقادیر عددی آن ها را حساب کنید. (۱۰ نمره)

ب) مقدار R^2 را محاسبه کنید. (۷ نمره)

ج) آزمون فرضیه $H_0: \beta_1 = 1$ با $\alpha = 0.05$ را انجام دهید. (۸ نمره)

تمرین دوم: (۲۵ نمره)

در این تمرین میخواهیم نشان دهیم که در مدل رگرسیون خطی زیر، R^2 statistic برابر با مربع ضریب همبستگی(r_{xy}) میان X و Y است.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

به همین منظور برقراری تساوی های زیر را نشان دهید:

الف) (۷ نمره)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

ب) (۳ نمره)

$$\beta_1^2 = \frac{\sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

ج) (۵ نمره)

$$\beta_1 = r_{xy}\frac{\sigma_y}{\sigma_x}$$

د) (۱۰ نمره)

$$R^2 = r_{xy}^2$$

تمرین سوم: (۱۰ نمره)

در رگرسیون خطی ثابت کنید:

الف) بردار خطا ($e = Y - \hat{Y}$) متعامد (orthogonal) به فضای ستون های ماتریس X است. (۳ نمره)

ب) $\hat{Y}^T e = 0$ (۳ نمره)

ج) تحلیل هندسی نتایج قسمت الف و ب را توضیح دهید. (۴ نمره)

تمرین چهارم: (۱۵ نمره)

با توجه به مدل رگرسیون خطی، به سوالات زیر پاسخ دهید (جهت حل کردن این سوال، روابط را به فرم ماتریسی استفاده نمایید.):

مدل رگرسیون خطی: $Y = \beta_0+\beta_1X_1+\cdots+\beta_pX_p+\epsilon$

الف) ثابت کنید که مجموع مجذورات خطا (RSS) برای یک مدل رگرسیون خطی، یک تابع محدب از پارامترهای رگرسیون است. (۵ نمره)

ب) نشان دهید که چگونه می توان از الگوریتم نزول گرادیان (gradient descent) برای به حداقل رساندن RSS برای مدل رگرسیون خطی استفاده کرد. رابطه به‌روزرسانی ضرایب را برای این منظور به دست آورید. سپس توضیح دهید که چرا این الگوریتم به مقدار حداقل سراسری (global minimum) برای تابع RSS همگرا می شود؟ (۵ نمره)

ج) ثابت کنید که واریانس $\widehat{\beta_j}$ با رابطه زیر محاسبه می شود که در آن σ^2 واریانس ϵ است. (۵ نمره)

$$Var(\widehat{\beta_j}) = \sigma^2(X^TX)^{-1}{}_{jj}$$

تمرین پنجم: (۱۵ نمره)

مدل رگرسیون خطی ساده زیر را در نظر بگیرید.

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0,\sigma^2)$$

الف) ثابت کنید که مجموع مربعات خطا (RSS) از توزیع زیر پیروی میکند. (۱۰ نمره)

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$

ب) برآوردگر بی‌طرف واریانس نویز σ^2 به صورت $s^2 = \frac{\text{RSS}}{n-2}$ تعریف می‌شود. واریانس s^2 را محاسبه کنید. (۳ نمره)

ج) رفتار $\text{Var}(s^2)$ را وقتی $n \to \infty$ نشان دهید. (۲ نمره)

راهنمایی: اگر A یک ماتریس خودتوان باشد ($A^2 = A$) آنگاه $\text{rank}(A)=\text{trace}(A)$ و مقادیر ویژه آن ۰ یا ۱ است.

تمرین ششم: (۱۰ نمره)

دو مدل زیر برای محاسبه وزن افراد را در نظر بگیرید:

$$Y = \beta_0 + \beta_1X_1 + \epsilon, \quad \epsilon \sim N(0,\sigma^2) \qquad (1)$$

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \epsilon, \quad \epsilon \sim N(0,\sigma^2) \qquad (2)$$

که در آن X_1 اندازه قد فرد و X_2 متغیری به صورت زیر است:

$$X_2 = \begin{cases} 1 & female \\ 0 & male \end{cases}$$

الف) عبارت میانگین وزن زنان با استفاده از هرکدام از مدل های ۱ و ۲ را بدست آورید. (۲ نمره)

ب) اگر در مدل ۱ مقادیر $\beta_0 = 5, \beta_1 = 10, \sigma^2 = 4$ را داشته باشیم، $Var(Y|X_1)$ و $Cov(Y, X_1)$ را محاسبه کنید. (۴ نمره)

ج) اگر در مدل ۲ مقادیر $\beta_0 = 0, \beta_1 = 50,\ \beta_2 = -5, \sigma^2 = 4$ را داشته باشیم، احتمال این که مردی با قد ۱.۷ متر وزنی بیش از ۸۹ کیلوگرم داشته باشد چقدر است؟ (۴ نمره)

تمرین هفتم: (امتیازی) (۱۰ نمره)

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

$$\overrightarrow{\beta_{new}} = \vec{\beta} + \frac{1}{1 + \overrightarrow{x_a}^T (X^T X)^{-1} \overrightarrow{x_a}} (X^T X)^{-1} \overrightarrow{x_a} (y_a - \overrightarrow{x_a}^T \vec{\beta})$$

راهنمایی: در صورتی که A و B ماتریس مربعی و معکوس پذیر باشند و Rank(B) = 1 باشد، رابطه زیر برقرار است:

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

سوالات عملی **۳۰ + ۱۵۰ نمره**

تمرین اول: (۸۰ نمره + ۱۰ نمره)

در این تمرین قصد داریم بر روی دیتاست real_estate.csv مفاهیم ساده تا پیچیده مبحث رگرسیون را مورد بررسی قرار دهیم. برای این منظور به نوتبوک real_estate.ipynb مراجعه نمائید و **مطابق توضیحات** فایل را تکمیل کنید.

A) ابتدا فایل دیتاست را بخوانید و چند سطر اول آن را نمایش دهید. سپس: (۲۰ نمره)

- اطلاعات اولیه دیتاست مانند ابعاد دیتاست، نام ستون ها، تعداد مقادیر نامعلوم هر ستون و ... را نمایش دهید. (۲ نمره)
- پیش پردازش هایی مانند مدیریت مقادیر غیرعددی، ستون‌های فاقد اهمیت، مقادیر از دست رفته و Scale کردن مقادیر عددی را انجام دهید. (۳ نمره)
- برای آشنایی بیشتر با دیتاست هیستوگرام هر ستون عددی دیتاست را رسم کنید. سپس به کمک تابع pairplot از کتابخانه seaborn هیستوگرام دو به دو ستون ها را هم رسم کنید. (۵ نمره)

- ماتریس correlation را مطابق دستورالعمل نوتبوک بدون استفاده از تابع آماده پیاده‌سازی کنید و نمایش دهید. کدام متغیرها با یکدیگر correlation بیشتری دارند؟ (۵ نمره)
- در نهایت، معیار VIF (Variance Inflation Factor) [1] را به کمک کتابخانه statsmodels محاسبه کنید. نتایج را با ماتریس correlation مقایسه کنید و مقادیر را تحلیل کنید. (۵ نمره)

B) با استفاده از الگوریتم Gram–Schmidt [2] و توضیحات نوتبوک، ماتریس X را تجزیه Q.R نمائید. در نهایت، بیشینه correlation بین داده‌ها را در این دو حالت مقایسه نمائید. (۱۰ نمره)

C) حال دیتاست را به بخش‌های آموزش، صحت سنجی و تست به نسبت ۰.۶ – ۰.۲ – ۰.۲ تقسیم نمائید. سه مدل چند جمله ای (با درجه‌های ۱، ۲ و ۳) را آموزش دهید. در ادامه، مدل‌های آموزش داده شده را با کمک معیارهای MSE، R^2 و F-Statistics ارزیابی نمائید. (۲۵ نمره)

D) در ادامه با پیاده‌سازی برخی معیارهای پیچیده‌تر، مدل‌های بخش قبل را دقیق تر بررسی کنید. این معیارها عبارتند از: Cook's distance [3]، Shapiro–Wilk & Q–Q Plot [4] و Partial F-Tests [5]. (توجه شود بجز Shapiro–Wilk & Q–Q Plot باید بقیه الگوریتم‌ها را خودتان – بدون کمک توابع آماده – پیاده‌سازی کنید.) (۲۰ نمره)

E) با توجه به نتایجی که تا به اینجا کسب نموده‌اید سوالات پاسخ کوتاهی که مطرح شده است را پاسخ دهید. (۵ نمره)

F) (امتیازی) در این بخش بایستی تبدیل (های) ریاضی را پیدا کنید که اعمال آن به یک یا چند ویژگی ورودی در شرایط یکسان از نظر درجه چند جمله‌ای، باعث افزایش دقت مدل شود. سعی کنید علت را به صورت خلاصه توضیح بدهید. واضح است هر چه تاثیر این عملیات (های) ریاضی بیشتر و تحلیل آن بهتر باشد، نمره بالاتری دریافت میکنید. (کد این بخش بایستی در ادامه فایل نوتبوک اضافه شود.) (۱۰ نمره)

تمرین دوم: (۷۰ نمره + ۲۰ نمره)

در این تمرین قصد داریم رابطه بین flexibility مدل با پدیده های Overfitting و Underfitting را مشاهده کنیم و در کنار آن با bias و variance نیز بیشتر آشنا شویم. برای این هدف مسئله رگرسیون تک بعدی $y = \sin(4\pi x) + \epsilon$ را در نظر بگیرید که در آن ϵ یک توزیع گاوسی با میانگین صفر و واریانس ۰.۰۹ است. مدل رگرسیون مورد استفاده، یک چند جمله ای درجه P به صورت زیر است:

$$f_p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

[1] https://en.wikipedia.org/wiki/Variance_inflation_factor
[2] https://en.wikipedia.org/wiki/Gram%E2%80%93Schmidt_process
[3] https://en.wikipedia.org/wiki/Cook%27s_distance
[4] https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
[5] https://en.wikipedia.org/wiki/F-test

تابع خطا نیز از رابطه زیر پیروی میکند: (همان رابطه MSE)

$$Err_{train} = \frac{1}{N}\sum_{i=1}^{N}(f_p(x_i) - y_i)^2$$

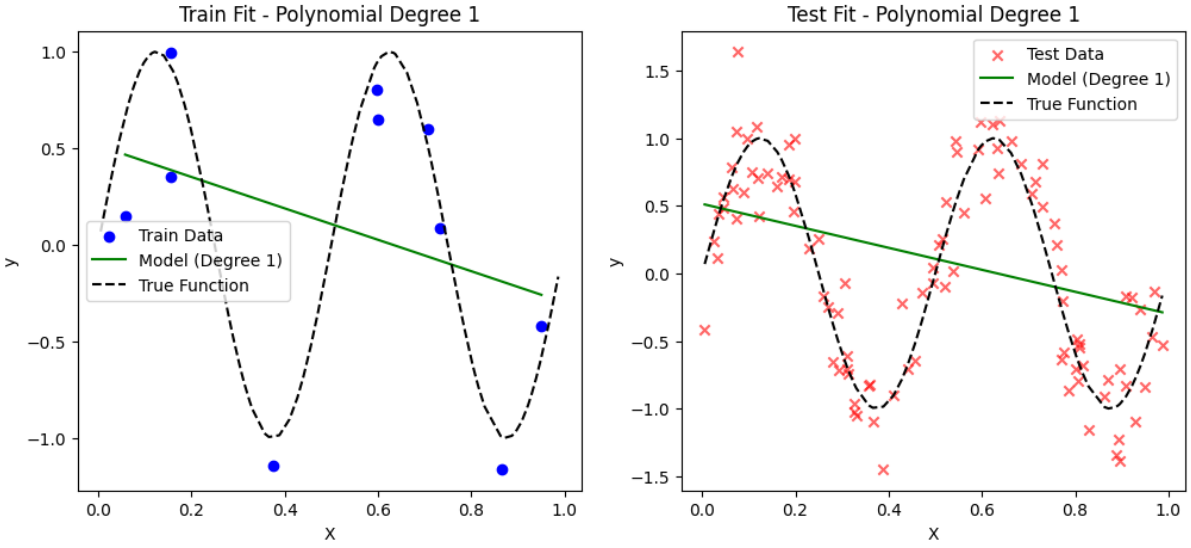
تعداد **داده های آموزش** برابر ۱۰ و تعداد **داده های تست** برابر **۱۰۰** است. قصد داریم درجه توان چند جمله ای را از **۱ تا ۹** تغییر دهیم و اثر افزایش flexibility را بررسی کنیم.

توجه: تمرین دارای دو بخش است. در بخش اول (A) برای انجام عملیات های ریاضی تنها مجاز به استفاده از کتابخانه Numpy هستید اما در بخش دوم (B) استفاده از توابع آماده کتابخانه scikit-learn مجاز است.

A-1) ابتدا با توجه به بسط عبارت $\sin(4\pi x)$ بیان کنید که چرا رگرسیون چندجمله ای گزینه مناسبی است؟ با توجه به تعداد داده های آموزش پیش بینی میکنید مدل در کدام درجه توان کاملا به داده های آموزش fit شود؟ چرا؟ (۵ نمره)

A-2) حال داده های آموزش و تست را ایجاد کنید. به همین منظور ۱۰ نمونه رندم از توزیع یکنواخت در بازه [0,1] برای داده های آموزش ایجاد کنید و مقادیر y_i ها را از رابطه $y = \sin(4\pi x) + \epsilon$ بدست آورید. برای داده های تست نیز به صورت مشابه عمل کنید. داده های آموزش و تست و نمودار تابع $\sin(4\pi x)$ را در یک شکل رسم کنید. (۵ نمره)

A-3) حال برای درجه توان های ۱ تا ۹ و به کمک روابطی که در اسلایدهای درس با آن ها آشنا شده اید، مدل را به داده های آموزش fit کنید و سپس نتایج را بر روی داده های تست آزمایش کنید. نمودار خروجی مدل بر روی داده های آموزش و تست را برای هر درجه توان رسم کنید. خروجی مورد انتظار برای هر درجه توان مطابق شکل زیر است:



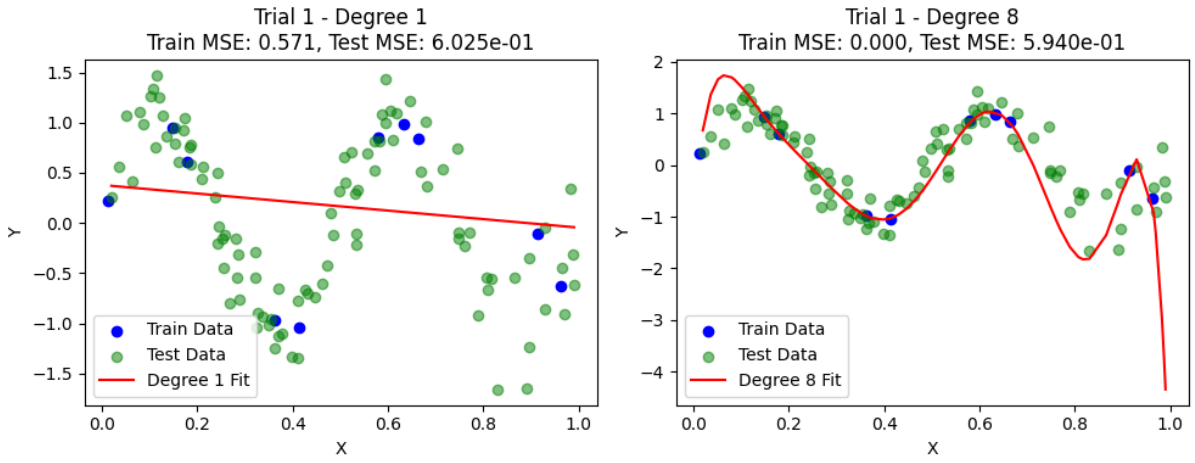
به نظر شما در کدام درجه مدل بهتر توانسته است نوسانات داده های تست را مدل کند؟ (۱۵ نمره)

A-4) حال با محاسبه خطای مدل بر روی داده های آموزش و تست به ازای هر درجه توان، نمودار *MSE* آموزش و تست بر حسب درجه توان را در یک شکل نمایش دهید. نقطه مینیمم خطای تست را مشخص کنید. آیا با پیش بینی قسمت قبل شما منطبق بود؟ نمودار را تحلیل کنید و درباره *Overfitting* و *Underfitting* مدل ها توضیح دهید. (۱۰ نمره)

A-5) (امتیازی) به نظر شما چرا در این سوال برخلاف روال معمول، داده های آموزش بسیار کمتر از داده های تست انتخاب شده است؟ (۵ نمره)

B-1) در بخش دوم میخواهیم اثر افزایش *flexibility* بر *bias* و *variance* مدل ها را بررسی کنیم. این کار از طریق تکرار چندین باره آزمایش بخش قبل میسر است. اما در گام اول با توجه به اسلایدهای مقدماتی درس، درباره بایاس و واریانس توضیح دهید. سپس با رسم شکل مفهوم آن ها را برای یک مدل با درجه آزادی کم و یک مدل با درجه آزادی زیاد در حین تکرار آزمایش ها بیان کنید. (۱۰ نمره)

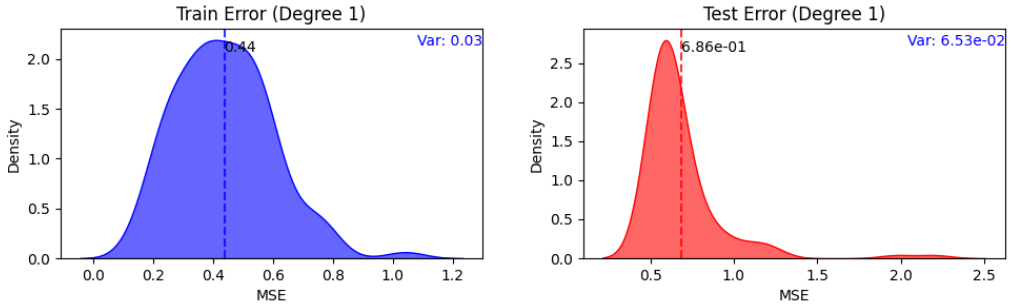
B-2) برای هر کدام از درجه توان های ۱ و ۸، آزمایش بخش قبل را ۳ بار تکرار کنید. به همین منظور یک مجموعه ۱۰۰ تایی واحد برای داده های تست در نظر بگیرید و در هر تکرار، ۱۰ داده متفاوت برای آموزش ایجاد کنید؛ سپس یک مدل با درجه توان ۱ و یک مدل با درجه توان ۸ *fit* کنید و در یک نمودار داده های آموزش و تست و خروجی مدل برای داده های تست را نمایش دهید. مقادیر *MSE* برای داده های آموزش و تست را هم نمایش دهید و این آزمایش را ۳ بار تکرار کنید. خروجی مورد انتظار مطابق زیر است:



سپس با توجه به توضیحات خود در B-1، نتایج را توجیه کنید. (۱۵ نمره)

B-3) حال ۳ نمونه رندم از داده های تست را جدا کنید و برای تمامی درجه های توان ۱ تا ۹، بخش B-2 را ۱۰۰ بار تکرار کنید، میزان خطای *MSE* را در هر تکرار برای این ۳ نمونه ذخیره کنید. در نهایت میانگین و واریانس خطای مدل در هر نمونه را برای هر درجه توان محاسبه کنید و نتایج را تحلیل کنید. (۱۰ نمره)

B-4) (امتیازی) بخش B-4 را این بار برای تمامی داده های تست و آموزش انجام دهید. خطای *MSE* داده های تست و آموزش را در هر تکرار ذخیره کنید و سپس به کمک تابع kdeplot از کتابخانه *seaborn* توزیع خطاهای آموزش و تست به دست آمده برای هر درجه توان را نمایش دهید. میانگین و واریانس داده ها را نیز نمایش دهید. نمونه خروجی مورد انتظار برای هر درجه توان مطابق زیر است:



تغییرات میان نمودارها را توجیه کنید. به نظر شما شیب تغییرات از کدام درجه توان بیشتر شده است؟ (نمودارهای آموزش را با یکدیگر و نمودارهای تست را هم با یکدیگر مقایسه کنید.) از آنجا که در شرایط آزمایش این تمرین توزیع داده ها محدود و واقعی نیست نمیتوان به مقایسه دو به دو میان توزیع خطای تست و آموزش پرداخت اما به نظر شما در یک آزمایش واقعی، با حرکت از درجه توان ۱ تا ۹، تغییرات این دو توزیع نسبت به یکدیگر چگونه خواهد بود؟ (۱۵ نمره)