سلام بر دانشجویان عزیز، چند نکته مهم:

۱. 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. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW3_StudentNumber داشته باشد.

۶. از بین سوالات **شبیه سازی** حتما به تمام موارد پاسخ داده شود.

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

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

۹. در صورت داشتن سوال، از طریق ایمیل سوال خود را مطرح کنید.

سوالات ۱، ۲، ۵ و ۶ arhosseini77@ut.ac.ir

سوالات ۳، ۴، ۷ و ۸ ftaherinezhad@ut.ac.ir

سوال ۱ : (۸ نمره)

با توجه به جدول زیر به سوالات ۱-۱ و ۱-۲ پاسخ دهید.

i	میزان مطالعه (x_i)	نمره (y_i)
1	16	46
2	27	80
3	11	36
4	20	52
5	30	98
6	25	75
7	5	10
8	24	70
9	21	64
10	10	30

۱-۱. در رابطه رگرسیون خطی شیب را در نظر نگیرید و در مدل زیر β_0 را بیابید.

$$y = \beta_0 + \varepsilon_i$$

۱-۲. در رابطه رگرسیون عبارت عرض از مبدا را در نظر نگیرید و در مدل زیر β_1 را بیابید.

$$y = \beta_1 x_i + \varepsilon_i$$

۱-۳. حال فرض کنید یک مدل رگرسیون به صورت $\hat{y} = 25 - 0.5x$ باشد. اگر یک نمره اضافی برای مشاهده جدید در $x = 6$ به دست آمده باشد ، آیا نمره آزمون برای مشاهده جدید لزوما ۲۲ خواهد بود ؟ دلیل خود را توضیح دهید.

۱-۴. اگر مجموع مربعات خطا برای این مدل ۷ باشد و به اندازه ۱۶ مشاهده وجود داشته باشد. بهترین تخمین برای σ^2 را ارائه دهید.

سوال ۲ : (۱۲ نمره)

۲-۱. تابع سیگموید $(\frac{1}{1+e^{-wX}})$ را بر حسب $X \in R$ ، برای $w \in \{1,5,100\}$ رسم کنید و استدلال کنید چرا جواب با وزن های بزرگ میتواند logistic regression را دچار overfit کند؟

۲-۲. با توجه به بخش ۲-۱ برای جلوگیری از overfit باید وزن ها کوچک باشند برای اینکار به جای حداکثر کردن احتمال شرطی تخمین MLE برای رگرسیون لجستیک،

$$\max_{w_0,..,w_d} \prod_{i=1}^{n} p(Y_i|X_i, w_0, \dots, w_d),$$

می توان حداکثر کردن احتمال شرطی تخمین MAP پسین را به صورت زیر در نظر گرفت.

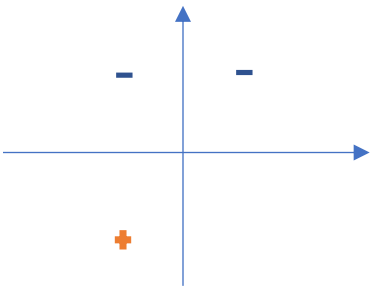
$$\max_{w_0,..,w_d} \prod_{i=1}^{n} p(Y_i|X_i, w_0, \dots, w_d) p(w_0, \dots, w_d)$$

$$p(w_0, \dots, w_d) \rightarrow prior\ on\ the\ weights$$

با فرض یک prior گوسی استاندارد $N(0, I)$ برای بردار وزن، قوانین به‌روزرسانی گرادیان را برای وزن‌ها استخراج کنید.

سوال ۳: (۱۵ نمره)

با استفاده از روش حل dual، معادله خط جدا کننده نمونه های زیر را بیابید.



$$x^1 = \begin{pmatrix}1\\2\end{pmatrix}, \quad y_1 = -1$$
$$x^2 = \begin{pmatrix}-1\\2\end{pmatrix}, \quad y_1 = -1$$
$$x^3 = \begin{pmatrix}-1\\-2\end{pmatrix}, \quad y_1 = 1$$

سوال ۴: (۱۰ نمره)

۴-۱. هدف استفاده از kernel ها در SVM چه بوده و مزیت مهم آنها در رابطه با فضای ویژگی ها را توضیح دهید. (۵ نمره)

۴-۲. معادله زیر را اثبات کنید. (۷ نمره)

$$k(x, x') = k_1(x, x')k_2(x, x')$$

که در آن $k_1(x, x')$ و $k_2(x, x')$ دو kernel معتبر هستند.

سوال ۵: (شبیه سازی، ۲۰ نمره)

طرز کار یک شرکت در حوزه فروش انلاین لباس به این شکل است که جلسات مشاوره ای حضوری برای مشاوره دادن به مشتریان برگزار می کند و پس از جلسه، مشتری از طریق اپلیکیشن یا وب سایت سفارش خود را ثبت می کند. یک دیتاست به نام Ecommerce Customers شامل موارد زیر در اختیارتان قرار گرفته است.

Email : ایمیل

Address : آدرس

Avatar : رنگ پروفایل

Avg. Session Length : میانگین طول جلسات مشاوره

Time on App : میانگین تایم گذرانده شده در اپلیکشین بعد از جلسه

Time on Website : میانگین تایم گذرانده شده در سایت بعد از جلسه

Length of Membership : تعداد سال های اشتراک مشتری

Yearly Amount Spent : میانگین مبلغ خرید در یک سال

۵-۱. ابتدا دیتاست را به صورت یک دیتافریم خوانده و به صورت کلی و آماری دیتافریم را بررسی کنید. (از متود های info و describe استفاده کنید.)

۵-۲. از استایل whitegrid و پالت GnBu_d یک joint plot برای میانگین خرید سالانه و بر حسب زمان گذرانده شده در وب‌سایت ترسیم نمایید. (هدف یادگیری و کار با seaborn میباشد)

۵-۳. همان ترسیم بخش ۲-۷ را برای زمان گذرانده شده در اپلیکیشن انجام دهید.

۵-۴. یک pairplot برای دیتاست رسم کرده و بر اساس آن بگویید کدام فیچر بیشترین تاثیر را روی میزان خرید سالیانه دارد؟

۵-۵. حال داده ها را به ۲ بخش فیچر و لیبل ('Yearly Amount Spent') تقسیم کنید و پیش‌پردازش‌های لازم را با ذکر توضیح انجام دهید.

۵-۶. با استفاده از متد train_test_split از کتابخانه و بخش sklearn.model_selection فیچر ها و لیبل ها را به دو بخش آموزش (۷۰ درصد) و بخش تست (۳۰ درصد) با رندوم استیت ۱۰۱(برای مشابه شدن نتایج) تقسیم کنید.

۵-۷. یک مدل رگرسیون خطی بسازید و داده ها را بر روی آن آموزش دهید. (استفاده از کتابخانه های آماده بلامانع است)

۵-۸. به کمک مدل آموزش داده شده مبلغ خرید نهایی را برای داده‌های تست پیش‌بینی کنید.

۵-۹. یک پلات اسکتر برای واقعی خرید و پیش بینی شده ترسیم کنید..

۵-۱۰. مقادیر خطاهای Mean Absolute Error، Mean Squared Error و the Root Mean Squared Error را برای این پیش‌بینی به دست آورید.

۵-۱۱. یک نمودار توزیع آماری برای تفاضلات مقادیر واقعی و پیش‌بینی شده ترسیم کنید و توضیح دهید چرا در صورتی که مشکل خاصی در آموزش وجود نداشته باشد و مدل به خوبی فیت شده باشد این ترسیم باید شکل توزیع نرمال داشته باشد.

۵-۱۲. به کمک .coef_جدول کوئفشینت های هر فیچر را به دست آورید و توضیح دهید چرا نرمالیزه کردن داده ها در پیش‌پردازش برای این بخش ضروری بود.

۵-۱۳. با توجه به جدول بخش ۱۲-۷ این شرکت باید روی کدام قسمت سرمایه گذاری بیشتری انجام دهد؟

سوال ۶: (شبیه سازی، ۲۰ نمره)

در این سوال هدف بررسی دیتاست یک شرکت تبلیغاتی و پیش‌بینی اینکه کاربر بر روی یک تبلیغ کلیک میکند یا نه میباشد.

۶-۱. دیتاست را به صورت دیتافریم خوانده و اطلاعات کلی و آماری آن را نمایش دهید.

۶-۲. نمودار توزیع آماری بر حسب سن را رسم کنید.

۶-۳. جوینت پلات درآمد بر حسب سن را رسم کنید.

۶-۴. KDE پلات میزان زمان گذاشته شده روی سایت تبلیغ بر حسب سن را رسم کنید.

۶-۵. در نهایت یک Pair Plot کلی برای دیتاست، با مشخص کردن تفاوت کسانی که روی تبلیغ کلیک کرده اند و نکرده اند ('hue='Clicked on Ad') را رسم کنید.

۶-۶. از تمام پلات هایی که ترسیم کردید چه نتیجه هایی میگیرید ؟

۶-۷. حال پیش پردازش های لازم را با ذکر توضیح بر روی داده ها انجام دهید و سپس آنها را به ۲ بخش فیچر و لیبل تقسیم کرده و در نهایت آنها را به ۲ بخش آموزش و تست تقسیم کنید.

۶-۸. بدون استفاده از کتابخانه های آماده یک مدل logistic regression بسازید و داده ها را به کمک آن آموزش دهید. (در صورت استفاده از کتابخانه های آماده ۰.۷ نمره این بخش را خواهیدگرفت)

۶-۹. به کمک مدلی که آموزش داده‌اید، برای داده های تست پیش بینی کنید که کدام یک روی تبلیغ ها کلیک می‌کنند.

۶-۱۰. معیار های ارزیابی همچون Score و Confusion Matrix و Classification Report را برای مدل آموزش داده شده گزارش دهید.

* در حل دو سوال شبیه سازی زیر، شما مجاز به استفاده از تمامی کتابخانه های مورد نیاز هستید.

سوال ۷: (**شبیه سازی**، ۱۵ نمره)

در این سوال قصد داریم تا مدلی جهت تشخیص پیامک‌های spam بسازیم. فایل spamSMS شامل دو ستون v1 و v2 است ستون v1 برچسب پیامک دریافت شده و v2 نیز متن پیامک است. پس از لود کردن این فایل، گزارشی از توزیع برچسب های آن ارائه دهید. سپس از تابع CountVectorizer استفاده نموده و از ستون پیامک، ویژگی استخراج کنید. دادگان را با نسبت ۳۰٪ به دادگان آموزش و تست تقسیم کرده و آموزش دهید. در این سوال لازم است تا دو کرنل linear و rbf مورد بررسی قرار گرفته و همچنین ۳ یا ۴ مقدار مختلف برای پارامتر های C و gamma تست شوند. این امر باید با استفاده از grid search و random search صورت گیرد. روش های grid search و random search به‌طور کلی توضیح داده و مقایسه کنید.

پس از آموزش مدل های خواسته شده، مقادیر پارامتر های مناسب را نمایش داده و سپس دقت را بر دادگان تست بدست آورده و گزارش کنید (با استفاده از ماتریس آشفتگی).

سوال ۸: (**شبیه سازی**، ۱۰ نمره)

فایل پیوست housePricing شامل ویژگی های یک خانه و قیمت آن است. در این سوال قصد داریم تا مدلی آموزش دهیم که با استفاده از این دادگان، قیمت خانه را پیش‌بینی کند. ابتدا مراحل پیش‌پردازش (حذف ستون هایی با missing value متعدد و جایگزینی سایر missing value ها با مقادیر مناسب) را بر دادگان انجام داده و سپس با استفاده از تابع SelectKBest از کتابخانه sklearn، ویژگی های مناسب برای آموزش دادگان را انتخاب نمایید. دادگان را به سه قسمت آموزش، اعتبار سنجی و تست تقسیم کرده و مدل را آموزش دهید.

موفق باشید.