



پردیس دانشکده های فنی

به نام خدا  
دانشکده‌ی مهندسی برق و کامپیوتر  
تمرین سری سوم یادگیری ماشین



دانشگاه تهران

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 حتماً 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** حتماً به تمام موارد پاسخ داده شود.
۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ ( **۱۰ نمره امتیازی** ) می‌توانید 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](mailto:arhosseini77@ut.ac.ir)

سوالات ۳، ۴، ۷ و ۸ [ftaherinezhad@ut.ac.ir](mailto: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 بگیرید و در مدل زیر  $\beta_0$  را بیابید.

$$y = \beta_0 + \varepsilon_i$$

۱-۲. در رابطه رگرسیون عبارت عرض از مبدا را در نظر بگیرید و در مدل زیر  $\beta_1$  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 لزوماً 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  $\sigma^2$  را ارائه دهید.

سوال ۲: (۱۲ نمره)

۲-۱. تابع سیگموئید  $\left(\frac{1}{1+e^{-wX}}\right)$  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 ۱-۲ برای جلوگیری از overfit باید وزن ها کوچک باشند برای اینکار به جای حداکثر

کردن احتمال شرطی تخمین MLE برای رگرسیون لجستیک،

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

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

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

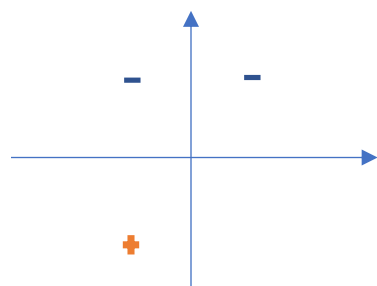
$$p(w_0, \dots, w_d) \rightarrow \text{prior on the weights}$$

با فرض یک prior گوسی استاندارد  $N(0, I)$  برای بردار وزن، قوانین به روزرسانی گرادیان را برای وزن ها استخراج

کنید.

سوال ۳: (۱۵ نمره)

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



$$\begin{aligned} x^1 &= \begin{pmatrix} 1 \\ 2 \end{pmatrix}, & y_1 &= -1 \\ x^2 &= \begin{pmatrix} -1 \\ 2 \end{pmatrix}, & y_2 &= -1 \\ x^3 &= \begin{pmatrix} -1 \\ -2 \end{pmatrix}, & y_3 &= 1 \end{aligned}$$

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 آنلاین 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 اختیار تان 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 اپلیکیشن 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 پیش پردازش های 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 داده های تست پیش بینی 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 پیش بینی به دست آورید.

۵-۱۱. یک نمودار توزیع آماری برای تفاضلات مقادیر واقعی و پیش بینی 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 کوئفشیونت های هر فیچر را به دست آورید و توضیح دهید چرا نرمالیزه کردن داده ها در پیش پردازش 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: (شبيه سازى، ۲۰ نمره)

در اين سوال هدف بررسى ديتاست يك شركت تبليغاتى و پيش بينى اينكه كاربر بر روى يك تبليغ كليك مىكند يا نه ميباشد.

۱-۶. ديتاست را به صورت ديتا فريم خوانده و اطلاعات كلى و آمارى آن را نمايش دهيد.

۲-۶. نمودار توزيع آمارى بر حسب سن را رسم كنيد.

۳-۶. جوينت پلات درآمد بر حسب سن را رسم كنيد.

۴-۶. KDE پلات ميزان زمان گذاشته شده روى سايت تبليغ بر حسب سن را رسم كنيد.

۵-۶. در نهايت يك Pair Plot كلى براى ديتاست، با مشخص كردن تفاوت كسانى كه روى تبليغ كليك كرده اند و نكرده اند ( 'hue='Clicked on Ad' ) را رسم كنيد.

۶-۶. از تمام پلات هاى كه ترسيم كرديد چه نتيجه هاى ميگيريد ؟

۷-۶. حال پيش پردازش هاى لازم را با ذكر توضيح بر روى داده ها انجام دهيد و سپس آنها را به ۲ بخش فيچر و ليل تقسيم كرده و در نهايت آنها را به ۲ بخش آموزش و تست تقسيم كنيد.

۸-۶. بدون استفاده از كتابخانه هاى آماده يك مدل logistic regression بسازيد و داده ها را به كمك آن آموزش دهيد. ( در صورت استفاده از كتابخانه هاى آماده ۰.۷ نمره اين بخش را خواهيد گرفت )

۹-۶. به كمك مدلى كه آموزش داده ايد، براى داده هاى تست پيش بينى كنيد كه کدام يك روى تبليغ ها كليك مى كنند.

۱۰-۶. معيار هاى ارزيابى همچون Score و Confusion Matrix و Classification Report را براى مدل آموزش داده شده گزارش دهيد.

\* 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 پیامک های 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 با استفاده از یک مدل مبنی بر SVM آموزش دهید (پایه سازی مدل SVM نیاز نیست). در این سوال لازم است تا دو کرنل `linear` و `rbf` مورد بررسی قرار گرفته و همچنین ۳ یا ۴ مقدار مختلف برای پارامترهای `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` به طور کلی توضیح داده و مقایسه کنید.

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

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 پیش بینی کند. ابتدا مراحل پیش پردازش (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 `regression` مبنی بر SVM را بر این دادگان آموزش دهید. می توانید از کتابخانه `scikit learn` استفاده کنید. سپس دقت و خطای مدل را بر دادگان آموزش، اعتبار سنجی و تست بدست آورده و مقایسه نمایید.

موفق باشید.