به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری دوم یادگیری ماشین

دانشگاه تهران

پردیس دانشکده های فنی

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.

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

٥. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.

٦. از بین سوالات **شبیه سازی** حتما به هر دو مورد 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 h.talebzadeh95@gmail.com سوال خود را مطرح کنید.

سوال ۱: (۱۵ نمره) جدول زیر نشان‌دهنده‌ی نمرات دانشجویان بر حسب میزان مطالعه هفتگی می باشد. رگرسیون خطی انجام داده و به سوالات پاسخ دهید:

میزان مطالعه	۵	۱۱	۱۸	۱۵	۲۱	۶	۱۷	۱۰	۲۴	۱۹
نمره	۳۴	۴۷	۶۶	۵۲	۸۰	۳۵	۶۶	۴۸	۸۷	۸۱

الف) در رگرسیون خطی ساده شیب خط را در نظر نگیرید و β_0 را بیابید.

$$y = \beta_0 + \varepsilon$$

ب) در رگرسیون خطی ساده عرض از مبدا را در نظر نگیرید و β_0 را بیابید.

$$y = \beta_1 x + \varepsilon$$

ج) تفاوت دو معادله‌ی زیر را توضیح دهید.

$$\hat{Y} = b_0 + b_1 X$$
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

د) بهترین تخمین از واریانس را محاسبه کنید.

ه) توضیح دهید اگر ورودی جدیدی از میزان مطالعه دانشجویی داشته باشیم، آیا می‌توانیم نمره‌ی او را با توجه به رابطه‌ی به دست آمده دقیقاً پیشبینی کنیم؟

سوال ۲: (۱۰ نمره) الف) L1 Regularization و L2 Regularization را تعریف کرده و با مقایسه‌ی روابط ریاضی، تفاوت‌های آن‌ها را توضیح دهید.

ب) یک رگرسیون خطی با L2 Regularization به صورت زیر را در نظر بگیرید و جواب فرم بسته برای $\hat{\beta}$ به دست آورید.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \| \beta \|_2^2$$

سوال ۳: (۱۵ نمره) یکی از راه‌های گسترش logistic regression به مجموعه‌های چندکلاسه بطور مثال برای کلاس K، این است که مجموعه‌های (K-1) از بردار وزن در نظر بگیریم و تعریف کنیم:

$$P(Y = y_k \mid X) \sim \exp(w_{k_0} + \sum_{i=1}^{d} w_{k_i} X_i) \quad for\ k = 1, \ldots, K-1$$

الف) این تعریف چه مدلی را برای $P(Y = y_k \mid X)$ نشان می‌دهد؟

ب) قانون طبقه بندی چه خواهد بود؟

سوال ۴: (۱۵ نمره) توزیع نرمال $P(x) \sim N(\mu, \sigma^2)$ و تابع پنجره‌ی پارزن $\varphi(x) \sim N(0,1)$ را در نظر بگیرید.

رابطه‌ی زیر تخمین پنجره پارزن را نشان می‌دهد:

$$\tilde{P}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \varphi(\frac{x - x_i}{h})$$

نشان دهید این تخمین برای h_n های کوچک دارای ویژگی‌های زیر است:

$$\tilde{P}_n \sim N(\mu, h_n^2 + \sigma^2)$$

$$P_n(x) - \tilde{P}_n(x) \simeq \frac{1}{2}(\frac{h_n}{\sigma^2})[1 - (\frac{x-\mu}{\sigma})^2]P(x)$$

سوال ۵: (۱۰ نمره) متریک فاصله اقلیدسی در d بعد را در نظر بگیرید:

$$D(a,b)=\sqrt{\sum_{k=1}^{d}(a_k-b_k)^2}$$

فرض کنید عناصر هر بعد را در یک مقدار حقیقی غیرصفر ضرب می‌کنیم:

$$x'_k = a_k x_k \quad for\ k=1,2,...,d$$

نشان دهید پس از ضرب نیز این متریک همچنان یک فاصله‌ی استاندارد است، یعنی ویژگی‌های یک فاصله‌ی استاندارد را داراست. در مورد تاثیر این ضرب بر طبقه بند KNN بحث کنید.

سوال ۶: **(شبیه سازی، ۱۵ نمره)** هدف از انجام این سوال بررسی overfitting و underfitting است. ابتدا با استفاده از کدهای زیر داده‌های مربوطه را تولید کنید:

```
X = np.arange(-10,10,0.2)
Y = 2*cos(x)/-pi + (2*x)/(2*pi)+2*cos(3*x)/(-3*pi)
```

حال Y این داده‌ها را در حالت اول نویز سفید گوسی جمع کنید و در حالت دوم با نویز پواسن با $\lambda = 2$ جمع کنید. نویزها را با ضریب تاثیر ۰.۱ به داده‌ها اضافه کنید. سعی کنید تابع درجه ۱ تا ۱۵ را به داده‌ها برازش کنید.

الف) بهترین و بدترین درجه را مشخص کنید.

ب) برای بهترین و بدترین درجه و درجات ۱، ۳، ۸ و ۱۵ نمودار برازش را رسم کرده و مقادیر MSE را گزارش کنید.

ج) با ذکر مقادیر بایاس و واریانس نتایج مشاهدات خود را شرح دهید.

سوال ۷: **(شبیه سازی، ۲۰ نمره)** در این سوال برای دو حالت زیر داده تولید کنید. در هر دو حالت دو دسته نقطه با مختصات (X,Y) داریم.

حالت اول:

دسته اول شامل ۲۰۰ نقطه درون دایره‌ای به مرکز $(1.5,0)$ محدود به شعاع‌های ۴ و ۹

دسته اول شامل ۲۰۰ نقطه درون دایره‌ای به مرکز $(1.5,0)$ محدود به شعاع‌های ۰ و ۶

حالت دوم:

دسته اول شامل ۱۰۰ نقطه با میانگین $(1,0)$ برای (X,Y)، نقاط تصادفی بوده و انحراف معیار ۱ دارند.

دسته اول شامل ۲۰۰ نقطه درون دایره‌ای به مرکز $(1.5,0)$ محدود به شعاع‌های ۲ و ۶

الف) نمودار داده‌ها در هر دو حالت را رسم کنید.

ب) با استفاده از الگوریتم Logistic Regression و استفاده از L2 Regularization دو کلاس این مجموعه داده را جدا کنید. واضح است که این داده‌ها به صورت خطی جداپذیر نیستند. بنابراین ابتدا فضای ویژگی‌ها را به ابعاد بالاتر ببرید. در زیر مثالی از افزایش ابعاد از ۲ به ۳۵ نشان داده شده است:

$$X = [x_1, x_2]^T$$

$$f(X) = [x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, \ldots, x_1x_2^6, x_2^7]^T, \quad f: R^2 \to R^{35}$$

دقت طبقه‌بند خود را برای داده‌های هر دو حالت گزارش کنید و مرز تصمیم به دست آمده را رسم کنید. در هر دو حالت بهترین درجه‌ای که با آن مرز تصمیم رسم شده است، گزارش کنید.

ج) نتایج به دست آمده را تحلیل کنید.

در صورت استفاده از پکج‌های آماده‌ی یادگیری ماشین، نصف نمره‌ی این سوال را خواهید گرفت.

سوال ۸: **(شبیه سازی، ۲۰ نمره)** در این سوال می‌خواهیم به پیاده‌سازی روش تخمین غیرپارامتری پارزن بپردازیم.

الگوریتم خواسته شده در این سوال را ابتدا بدون استفاده از کتابخانه‌های آماده‌ی موجود پیاده‌سازی کنید. ابتدا دیتاست ted talks را پیاده‌سازی کنید.

الف) ستون duration این دیتاست را استخراج کرده و توزیع دیتای این ستون را با استفاده از روش پنجره‌ی پارزن با کرنل گوسی به دست آورده و نتیجه را نمایش دهید. اندازه پنجره را برابر با ۱۰ در نظر بگیرید.

ب) تاثیر اندازه‌ی پنجره را با سه مقدار مختلف ۲۰، ۵۰ و ۱۰۰ بررسی کنید.

ج) با استفاده از کتابخانه‌های آماده توزیع ستوم duration را رسم کنید. با افزایش مقدار n روند تغییذ و همگرا شدن به توزیع اصلی را روی یک نمودار نشان دهید. مقدار n را در بازه‌ب ۲۵۰ نمونه تا کل دیتا با step ۲۵۰ بررسی کرده و همگرایی برای n های مختلف را روی یک نمودار نشان دهسد.

د) نتیجه قسمت الف را با نتیجه‌ی توابع کتابخانه‌های آماده مقایسه کنید.