به نام خدا

یادگیری ماشین — نیمسال اول ۹۸-۹۷

تمرین سری اول

موعد تحویل:بخش تئوری: ۱۳ مهر
بخش عملی: ۱۸ مهر — نمره: ۱۱۵ + ۳۰ امتیازی

سوال ۱: Linear Regression (20 Points)

فرض کنید n داده ی آموزش به صورت $D = (x^1, y^1), \cdots, (x^n, y^n)$ داریم که هر کدام از xها، d بعدی است. میخواهیم از رگرسیون خطی با تابع هزینه ی SSE استفاده کنیم. این تابع هزینه بدین صورت است:

$$J(w) = \sum_{i=1}^{n} (y^{(i)} - w^T x^{(i)})^2$$

۱,۱. (۵ نمره) ثابت کنید w ای که این رابطه را بهینه می کند برابرست با: $\widehat{w} = (X^T X)^{-1} X^T y$

۱,۲. (۲ نمره) استفاده ی مستقیم از این رابطه مشکلاتی دارد. دو مشکل بالقوه ی استفاده ی مستقیم از این رابطه را ذکر کنید و راه حلی برای هر یک ارائه دهید.

۱,۳. (۲ نمره) گر یکی از ابعاد داده ها ترکیب خطی ای از سایر ابعاد داده‌ها باشد، با ذکر دلیل توضیح دهید که چرا نمی توان از رابطه بالا استفاده کرد. راه حل شما برای این مشکل چیست؟

۱,۴. (۶ نمره) اگر مسئله را به صورت احتمالاتی بنویسیم، خواهیم داشت:

$$\widehat{w} = argmin_w E_{x,y}[(y - w^T x)^2]$$

مقدار بهینه ی w را بر حسب ماتریس خودهمبستگی ($R = E_x[xx^T]$) و بردار همبستگی $c = E_{xy}[xy]$ محاسبه کنید. سپس نشان دهید که خطای مورد انتظار را میتوان به صورت جمع خطای ساختاری structural error و خطای تقریب estimation error نوشت ، تعبیر هر کدام از جملات را بیان نمایید.

۱,۵. (۲ نمره) اگر به تابع خطای SSE یک جملهی منظم ساز L۲ (به صورت $||w||^2$) بیفزاییم، فرم بسته ی جواب w را بدست آورید.

۱,۶. (۳ نمره) رگرسیون خطی وزن دار ($Weighted\ Linear\ Regression$) یک تعمیم روی رگرسیون خطی است که در آن، هر کدام از نقاط داده‌ها یک ضریب وزن نیز میگیرد:

$$J(w) = \sum_{i=1}^{n} F_i (y^{(i)} - w^T x^{(i)})^2$$

فرم بسته ی w بهینه را برای این تابع هزینه بدست آورید.

سوال ۲: Gradient Descent (10 Points)

در این سوال قصد داریم تا رابطه‌ای برای روش گرادیان کاهشی برای رگرسیون خطی و غیرخطی بیابیم.

۲,۱. رابطه زیر را برای یک رگرسیون غیرخطی دو متغیره در نظر بگیرید:

$$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \epsilon \qquad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

(آ) (۱ نمره) رابطه‌ای برای $P(y \mid x_1, x_2)$ به دست آورید.

(ب) (۳ نمره) فرض کنید که مجموعه $(x_1^{(i)}, x_2^{(i)}, y^{(i)})\ for\ i = 1, \ldots, n$ ، داده‌های آموزشی می‌باشند. رابطه‌ی log likelihood را برای مجموعه داده‌های آموزش بنویسید.

(ج) (۱ نمره) با توجه به جواب به دست آمده در قسمت قبل، یک تابع به صورت $f(\omega_0, \omega_1, \omega_2, \omega_3)$ بنویسید که بتوان با مینیمم کردن آن، پارامترهای موجود را به دست آورد.

(د) (۵ نمره) گرادیان $f(\omega)$ نسبت به بردار $\omega = [\omega_0, \omega_1, \omega_2, \omega_3]$ محاسبه نمایید.

۲,۲. حال می‌توان با داشتن $\nabla_\omega f(\omega)$ رابطه گرادیان کاهشی را برای به روزرسانی پارامترهای موجود در توابع هزینه فوق، به دست آورد.

سوال ۳: Regression Shrinkage Methods(15 Points)

در این سوال می‌خواهیم منظم‌سازها را در رگرسیون خطی مورد تحلیل و بررسی قرار دهیم.

۱,۳. (۵ نمره) نشان دهید رگرسیون خطی با منظم ساز $L2$ را می‌توان با اضافه کردن تعدادی داده به فرم رگرسیون خطی کمینه مربعات نوشت. این اضافه کردن داده و تبدیل کردن مساله به رگرسیون منظم شده را توجیه کنید.

۲,۳. (۴ نمره) می‌دانیم در منظم‌سازی $L1$ در موقعیت‌های متعددی موفق رفتار می‌کند و جواب تنکی را تولید می‌کند. آیا می‌توانید سناریو ای را ذکر کنید که این روش با محدودیت روبرو شود و موفق عمل نکند؟

۳,۳. (۶ نمره) برای رفع محدودیت‌هایی که در بالا ذکر شد منظم ساز دیگری پیشنهاد شد که ترکیبی از هر دو روش قبلی است. تابع هزینه آن به شکل زیر است

$$L(w, \lambda_1, \lambda_2) = |y - Xw|^2 + \lambda_1 ||w||_2^2 + \lambda_2 ||w||_1$$

در این روش هم نشان دهید می‌توان با اضافه کردن تعدادی داده مساله را به شکل رگرسیون با منظم ساز $L1$ نوشت.

سوال ۴: Regression and Projection (20 Points)

در این سوال میخواهیم رگرسیون خطی (به دست آوردن بردار وزن بهینه و تخمین خروجی به صورت ترکیب خطی از ویژگی‌ها) را از دید هندسی بررسی کنیم.

۱,۴. (۱۳ نمره) فضای برداری $\mathbb{R}^n$ را در نظر بگیرید. فرض کنید مجموعه S مجموعه تمام بردارهایی است که میتوان به صورت ترکیب خطی بردارهای $\{X_1, ..., X_r\}$ نوشت. S یک زیرفضای برداری $\mathbb{R}^n$ است. برای مثال برای حالت $n = 3$ بردارهای X_1, X_2 می توانند بردارهای $[1, 0, 0]^T$ و $[0, 1, 0]^T$ باشند. در اینصورت S نشان دهنده صفحه افقی $x - y$ است.
هر بردار $u \in \mathbb{R}^n$ را میتوان روی زیرفضای برداری S تصویر کرد. در مثال قبل تصویر بردار $[x, y, z]^T$ روی صفحه S برابر است با بردار $[x, y, 0]^T$. تعریف دقیق تر عملگر تصویر به صورت زیر است. عملگر تصویر بر روی زیرفضای S، یک تابع به صورت زیر

$$f_S : \mathbb{R}^n \to \mathbb{R}^n$$

$$f_S(u) = \underset{x \in S}{\operatorname{argmin}} |u - x|^2$$

است. در واقع این تابع به هر بردار $u \in \mathbb{R}^n$ نزدیک بردار ممکن در S را نسبت میدهد (که مطابق شهود هندسی مان همان پای عمود تصویر بردار در یک صفحه است).

(آ) (۸ نمره) ثابت کنید عملگر تصویر یک تبدیل خطی است . همچنین نشان دهید بردار $f_S(u) - u$ به تمامی بردارهای موجود در زیرفضای برداری S عمود است.
میتوان نشان داد که هر تبدیل خطی $f : \mathbb{R}^n \to \mathbb{R}^n$ متناظر با یک ماتریس $P_{n \times n}$ است، به طوری که به ازای هر بردار $u \in \mathbb{R}^n$ داریم

$$f(u) = Pu$$

برای مثال تصویر بر روی صفحه $x - y$ ماتریس $P_{3 \times 3}$ متناظر با این تبدیل را ارائه دهید.

(ب) (۵ نمره) تابع $f : \mathbb{R}^n \to \mathbb{R}^n$ را در نظر بگیرید. عملکرد این تابع رو بردار $u = [u_1, ..., u_n]^T$ به صورت زیر است.

$$\begin{cases} m_u = \frac{1}{n} \sum_{i=1}^n u_i \\ f_S(u) = [u_1 - m_u, u_2 - m_u, ..., u_n - m_u] \end{cases}$$

ثابت کنید این تابع بردار u را بر یک زیرفضای برداری S تصویر میکند، همچنین زیرفضای برداری S را توصیف کنید و بُعد آن را به دست آورید. درایه‌های ماتریس $P_{n \times n}$ متناظر با این عملگر تصویر را به صورت دقیق بنویسید.

۲,۴. (۷ نمره) در رگرسیون خطی هدف به دست آوردن یک تخمین به صورت $\hat{y} = w_1 x_1 + ... + w_r x_r$ برای مقدار خروجی y است. فرض کنید n مشاهده از مقدار خروجی و مقادیر ویژگی‌های مربوط به هرکدام را داریم. فرض کنید بردارn تایی Y نشان دهنده مقادیر خروجی مطلوب و بردار n تایی X_i نشان دهنده مقادیر ویژگی i ام به ازای تمام نمونه‌های آموزش باشد. هدف تخمین زدن بردار Y با تابعی خطی از ویژگی‌ها $\hat{Y} = w_1 X_1 + ... + w_r X_r$ به صورتی است که مقدار $|\hat{Y} - Y|^2$ کمینه شود. میدانیم که بردار $\hat{Y}$ به صورت ترکیب خطی ویژگی‌ها است و بنابراین در زیرفضای برداری S قرار میگیرد. در قسمت قبل نشان دادیم که تابعی که $|\hat{Y} - Y|^2$ را کمینه میکند یک تبدیل خطی و به صورت دقیق تر یک عملگر تصویر است. میخواهیم در حالتی که زیرفضای برداری S مجموعه تمام بردارهای ساخته شده توسط ترکیب خطی بردارهای $\{X_1, ..., X_r\}$، ماتریس P_S را برحسب بردارهای $\{X_1, ..., X_r\}$ به دست آوریم.

(آ) (۵ نمره) میتوانیم $\hat{Y}$ را به صورت $\hat{Y} = XW$ بنویسیم که در آن $X = [X_1, ..., X_r]$ ماتریس داده‌ها و $W = [w_1, ..., w_r]^T$ بردار ضرایب است. با توجه به اینکه $\hat{Y}$ بهینه برابر با تصویر بردار Y بر روی زیرفضای S است و شهود هندسی تصویر کردن (که در بخش (آ) قسمت قبل اثبات کردید) شرطی بر روی ماتریس و بردارهای X, W, Y بنویسید به طوری که بردار $\hat{Y} = XW$ برابر با تصویر بردار Y بر روی زیرفضای برداری S باشد.

(ب) (۲ نمره) حال برای سادگی فرض کنید که مجموعه بردارهای $\{X_1, ..., X_r\}$ مستقل خطی هستند و با استفاده از رابطه ای که در قسمت قبل نوشته اید ماتریس P_S را به دست آورید. نشان دهید که این ماتریس تمام ویژگی‌هایی را که در بخش (ج) قسمت قبل ثابت کردید دارد.

سوال ۵: Linear Regression and LOOCV (20 Points)

همانطور که می‌دانید، می‌توان با استفاده از روش Cross Validation خطای واقعی یک روش یادگیری را تخمین زد. یک راه حل که تقریبا تخمین ناریبی از این مقدار خطای واقعی به ما می‌دهد روش Leave-One-Out Cross Validation (LOOCV) می‌باشد. مشکل این روش آن است که زمان نسبتا زیادی طول می‌کشد که خطای LOOCV محاسبه شود. همانطور که در سوال ۱ دیدید، برای n نمونه آموزشی $(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ رگرسیون خطی ناریب به صورت $Y = X\omega + \epsilon$ است که با توجه به تابع هزینه مربوطه مقدار بهینه برای ω به صورت زیر می‌باشد:

$$\hat{\omega} = (X^T X)^{-1} X^T Y$$

که در واقع تخمین ما از Y به صورت زیر است:

$$\hat{Y} = X\hat{\omega} = X(X^T X)^{-1} X^T Y = PY$$

که در واقع ما ماتریس P (که در ادامه از آن استفاده خواهیم کرد) را به صورت زیر در تعریف می‌کنیم:

$$P = X(X^T X)^{-1} X^T$$

همانطور که از قبل می‌دانیم خطای مجموع مربعات به صورت زیر می‌باشد:

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

حال خطای LOOCV به صورت زیر تعریف می‌شود:

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

که در تعریف فوق $\hat{y}_i^{(-i)}$ در واقع تخمین نمونه i ام به ازای تخمین‌گر $\hat{Y}^{(-i)}$ می‌باشد. تخمین‌گر $\hat{Y}^{(-i)}$ در واقع همان تخمین‌گر Y است که در رابطه مربوط به SSE آن مورد i ام حذف شده است یعنی در واقع مسئله بهینه سازی این تخمین‌گر تابع هدف زیر را مینیمم می‌کند:

$$\sum_{j \neq i} (y_j - \hat{y}_i^{(-i)})^2$$

به عبارت دیگر به ازای تخمین‌گر i ام، از نمونه i ام صرف‌نظر می‌شود.

پس برای محاسبه خطای LOOCV لازم است n تا تخمین‌گر محاسبه شوند.

۱,۵. (۳ نمره) پیچیدگی محاسباتی برای محاسبه خطای LOOCV در حالت ساده محاسبه نمایید.(منظور از حالت ساده این است که در یک حلقه به تعداد نمونه‌ها هر مرتبه یک تخمین‌گر بر روی سایر نمونه‌ها ایجاد شود.) راهنمایی: محاسبه وارون یک ماتریس $k \times k$ پیچیدگی از مرتبه $O(k^3)$ دارد.

۲,۵. (۱ نمره) $\hat{y}_i$ را بر حسب P و Y محاسبه نمایید.

۳,۵. (۳ نمره) نشان دهید که تخمین‌گر $\hat{Y^{(-i)}}$ یک تخمین‌گر برای Z است که SSE آن را مینیمم می‌کند.

$$Z_j = \begin{cases} y_i, & j \neq i \\ \hat{y_i}^{(-i)}, & j = i \end{cases}$$

۴,۵. (۲ نمره) $\hat{y_i}^{(-i)}$ را بر حسب P و Z محاسبه نمایید. (جوابی مشابه به قسمت ۲ همین سوال مد نظر است.)

۵,۵. (۶ نمره) نشان دهید که:

$$\hat{y_i} - \hat{y_i}^{(-i)} = P_{ii}y_i - P_{ii}\hat{y_i}^{(-i)}$$

P_{ii} همان درایه قطری i ام در ماتریس P است.

۶,۵. (۵ نمره) نشان دهید که:

$$LOOCV = \sum_{i=۱}^{n} (\frac{y_i - \hat{y_i}^{(-i)}}{۱ - P_{ii}})^{۲}$$

پیچیدگی محاسباتی خطای LOOCV طبق رابطه فوق چه می‌باشد؟ رابطه به دست آمده را تحلیل کنید.

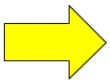
سوال ۶: Hands On Regression! (30 Points)

در این مساله می‌خواهیم از رگرسیون خطی برای پیش‌بینی هزینه‌ی پزشکی افراد بر اساس ویژگی‌های شخصی آنها استفاده کنیم. توجه کنید که برای این مسئله نمی‌توانید از کتابخانه‌های آماده یادگیری ماشین استفاده کنید و تنها مجاز به استفاده از numpy هستید. داده ها در دو فایل train.csv و test.csv ذخیره شده‌اند. هر داده دارای ۶ ویژگی ورودی (x) است و یک خروجی هزینه‌ی پزشکی (y) دارد. برای خواندن فایل‌های csv استفاده از کتابخانه pandas مجاز است.

همانطور که در داده‌های سوال دیده می‌شود، ویژگی‌های جنیست، منطقه و سیگاری بودن از نوع دسته‌بندی شده هستند (categorical) و عددی نیست؛ برای انکود کردن این ویژگی‌ها روش‌های متفاوتی وجود دارد. در این تمرین از دو روش استفاده می‌کنیم:

- integer encoding: به هر دسته، یک عدد یکتا (معمولا بین ۰ تا n-۱) اختصاص داده می‌شود.
- one hot encoding (OHE): برای n دسته در یک ویژگی، n متغیر از نوع binary (مقدار ۰ یا ۱ میگیرد) در نظر گرفته می‌شود و برای ورودی که مقدار امi را دارد، متغیر i ام ۱ می‌شود و دیگر متغیرها ۰ می‌شوند. همانند شکل زیر:

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

در داده‌های سوال، ویژگی‌ها را به صورت زیر تغییر دهید:

- جنسیت: مقادیر female و male را به ترتیب تبدیل به ۰ و ۱ میکنیم
- سیگاری‌بودن: مقادیر no و yes را به ترتیب تبدیل به ۰ و ۱ میکنیم
- منطقه :(region) در این ویژگی ۴ ستون وجود دارد. از روش OHE برای این ویژگی استفاده کنید. دلیلی که برای region از OHE استفاده می‌شود این است که در روش integer encoding ، ترتیب بین مقادیر عددی ممکن است خطا ایجاد کند (درحالی که بین مناطق هیچ ترتیبی وجود ندارد) اما در OHE این مشکل برطرف می‌شود.

۱,۶. (۱۰ نمره) رگرسیون بدون منظم سازی:

(آ) (۵ نمره) رگرسیون خطی تعمیم یافته را بدون جمله ی منظم ساز پیاده سازی کرده، نتایج را روی داده های تست گزارش کنید. تابع basis برای ویژگی سن به صورت زیر باشد

$$\phi_{age}(x_{age}) = x_{age}^{۲}$$

و برای بقیه ویژگی‌ها از تابع همانی

$$\phi_(x) = x$$

استفاده کنید. همچنین از فرمول بسته رگرسیون خطی تعمیم یافته برای محاسبه ی w استفاده کنید. مقدار نهایی w را نیز در گزارش یادداشت کنید.

(ب) (۵ نمره) در ابتدا تعداد دادههای تمرین را ۱۰ در نظر بگیرید و با گامهای ۱۰ تایی، تا ۱۰۰۰ افزایش دهید. تغییرات خطای تست و آموزش را با افزایش داده آموزش بررس کنید. برای مقایسه‌پذیر بودن خطا، از MSE برای تابع هزینه استفاده کنید. نمودار این تغییرات را در گزارش بیاورید.

۲,۶. (۱۰ نمره) روش‌های batch gradient descent و stochastic gradient descent را پیاده‌سازی کنید و مقدار w و خطا را در هر حالت به همراه خطا گزارش کنید.

۳,۶. (۱۰ نمره) رگرسیون با منظم سازی: در این قسمت، به تابع هزینه ی SSE ی جمله ی منظم ساز L۲ با ضریب λ بیفزایید. با استفاده از $۵ - fold\,Cross - validation$ ، بهترین مقدار پارامتر λ را از بین اعضای مجموعه ی $۱۰^{-۴}, ۱۰^{-۳}, \cdots, ۱, \cdots, ۱۰^{۴}$ در هر مورد بیابید و سپس نتیجه را روی داده های تست بدست آورید. نمودار خطا را برحسب لگاریتم λ رسم کنید. در این قسمت مجددا از فرمول بسته‌ی w استفاده کنید. نتیاج نهایی و مقدار خطا را برای داده‌های تست و آموزش در گزارش بنویسید.

سوال ۷: **Kaggle (30 Points Extra)**

در این سئوال قرار است در یکی از رقابت‌های عملی وب سایت Kaggle شرکت کنید (دقت کنید که این وبسایت با آی پی ایران قابل دسترس نیست و برای دسترسی به آن باید از وی پی ان استفاده کنید). این وب سایت نقطه شروع خوبی برای فراگیری کاربردهای عملی یادگیری ماشین در مسائل دنیای واقعی است. یک اکانت با نام کاربری "ML۹۷#" بسازید که در آن # شماره دانشجویی اتان است. برای این سئوال قرار است در این رقابت شرکت کنید. نمره دهی به این سئوال به صورت نسبی بر اساس نتایج تمام دانشجویان درس انجام خواهد شد. در صورتی که رتبه شما بیشتر از ۸۰۰ شود نمره ای به شما تعلق نخواهد گرفت. میتوانید از قسمت Kernels ایده‌های دیگر شرکت کنندگان و روش‌هایی که استفاده کرده اند را ببینید و از آن‌ها برای بهبود نتایج خود استفاده کنید. برای این مسئله میتوانید از کتابخانه‌های یادگیری ماشین در پایتون استفاده کنید. برای این سئوال کد خود و فایل csv خروجی برای داده‌های تست (همان فایلی که در سایت آپلود میکنید) را ضمیمه کنید. همچنین لازم است که تمام ایده‌ها و روش‌هایی را که استفاده کرده اید در قالب یک فایل گزارش به همراه تمرین ارسال کنید.