لطفا نکات زیر را رعایت کنید:

۱) فایل گزارش به همراه تمامی کدها را در یک فایل فشرده و با عنوان HW#1_StudentNumber در سایت بارگذاری کنید.

۲) بخش‌های پیاده سازی مربوط به هر سوال را در فایل مربوطه با شماره‌ی آن سوال و در پوشه‌ای برای آن سوال قرار دهید. برای مثال، تمامی بخش‌های پیاده‌سازی سوال اول را در پوشه‌ی Q1 و کد مربوط به قسمت ب سوال اول را با نام Q1_b. py و در پوشه‌ی Q1 قرار دهید.

۳) از زبان برنامه‌نویسی پایتون در یکی از محیط‌های Jupyter notebook و یا Google colab برای کد نویسی و تست کدهای خود استفاده کرده و فایل کدها را با فرمت IPYNB ارسال کنید.

۴) گزارش نهایی باید شامل توضیح پیاده‌سازی و نتایج و تحلیل‌های خواسته‌شده در متن تمرین باشد. توجه کنید که در گزارش نهایی خود به تمامی سوال‌های پرسیده شده در متن تمرین (به‌خصوص در بخش عملی تمرین) پاسخ دهید.

۵) در صورت داشتن هرگونه سوال و ابهام با تدریس‌یاران درس در ارتباط باشید.

بخش نظری

۱- توزیع احتمال گاوسی

برای اینکه یک تابع بتواند نشان دهنده چگالی احتمال باشد می‌دانیم که بایستی در شرایط زیر صدق کند:

$$1.\ \forall x: p(x) \geq 0$$

$$2.\ \int_{-\infty}^{+\infty} p(x)\ dx = 1$$

شرط دوم به شرط نرمال بودن مشهور است. در این سؤال به اثبات شرط نرمال بودن برای تابع توزیع نرمال (گاوسی) تک متغیره می‌پردازیم. انتگرال زیر را درنظر بگیرید:

$$I = \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right)\ dx$$

محاسبه این انتگرال به کمک محاسبه انتگرال زیر به نوعی امکان‌پذیر می‌شود.

$$I^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)\ dx\,dy$$

حال از مختصات دکارتی (x, y) به مختصات قطبی (r, θ) تغییر متغیر انجام دهید و سپس در نظر بگیرید $u = r^2$. نشان دهید که با محاسبه انتگرال‌ها روی θ و u و سپس محاسبه جذر دو طرف، داریم:

$$I = (2\pi\sigma^2)^{\frac{1}{2}}$$

در نهایت از نتیجه بالا استفاده کرده و نشان دهید که توزیع $N(x|\mu, \sigma^2)$ شرط (۲) را ارضا می‌کند.

۲- تخمین با دیدگاه Maximum Likelihood

فرض کنید $X_1, X_2, \dots, X_N$ نمونه‌های تصادفی مستقل از یک توزیع با چگالی زیر هستند.

$$f(x|\theta) = \frac{1}{2}e^{-|x-\theta|}$$

با استفاده از روش MLE پارامتر θ را محاسبه کنید.

بخش عملی

۳. در این قسمت به پیاده سازی مسأله رگرسیون برای یک دیتاست ساده می‌پردازیم. برای این قسمت تنها استفاده از کتابخانه numpy برای انجام محاسبات برداری مجاز است.

در این تمرین هدف پیاده سازی فرم ساده معادله رگرسیون تک متغیره به فرم رابطه (۱) است:

$$y_n = f(x_{n1}) = w_0 + w_1 x_{n1} \qquad (1)$$

در این معادله x_{n1} ورودی مدل نشانگر وزن، y_n قد شخص و نشانگر قد شخص است. از تابع خطای میانگین مجموع مربعات (MSE) با رابطه (۲) برای یافتن مقادیر بهینه w_0 و w_1 استفاده می‌کنیم.

$$\mathcal{L}(w_0, w_1) = \frac{1}{2N}\sum_{n=1}^{N}\big(y_n - f(x_{n1})\big)^2 = \frac{1}{2N}\sum_{n=1}^{N}(y_n - w_0 - w_1 x_{n1})^2 \qquad (۲)$$

هدف ما پیدا کردن w_0^* و w_1^* به نحوی است که که تابع هزینه بالا حداقل شود. شیوه ذخیره سازی دادگان مورد استفاده در پایتون به صورت زیر است

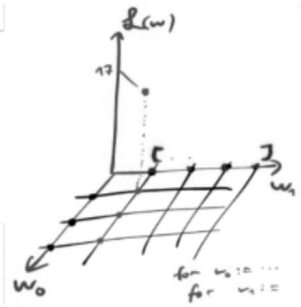
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & x_{31} \\ 1 & x_{N1} \end{bmatrix}$$

۳-الف) محاسبه تابع هزینه

با در نظر گرفتن ماتریس پارامترهای مدل به‌صورت $\boldsymbol{w} = [w_0\ w_1]^T$ و تعریف $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}$، تابع هزینه MSE را به صورت برداری بنویسید، سپس پیاده‌سازی تابع compute_loss(y, tx, w) را در فایل ex03-1402-1.IPYNB که در اختیارتان قرار گرفته است کامل کنید.

۳-ب) الگوریتم جست‌وجوی شبکه برای یافتن پارامترهای بهینه

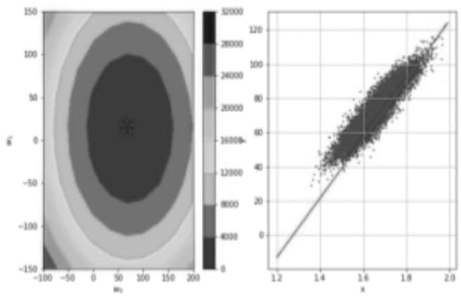
یکی از روش‌هایی که در کلاس درس برای حل مسأله بهینه سازی از آن صحبت نشده است روش جست و جوی شبکه و یا grid search برای پارامترهای بهینه است. در این روش با گسسته سازی فضای مربوط به هر پارامتر مدل، تک تک ترکیب‌های ممکن برای پارامترها را امتحان می‌کنیم و در نهایت ترکیبی که بهترین پاسخ (در این مسأله کمترین هزینه) را به ما می‌دهد به عنوان پاسخ نهایی مسأله درنظر می‌گیریم. به شکل۱ توجه کنید، درواقع می‌خواهیم تابع هزینه را به ازای مقادیر مختلف $\boldsymbol{w}$ در فضای جست و جو محاسبه کنیم و سپس بردار $\boldsymbol{w}$ ای را انتخاب کنیم که به ازای آن مدل کمترین میزان هزینه را دارد. در این قسمت از تمرین هدف پیاده‌سازی این الگوریتم است.



شکل ۱: الگوریتم grid search برای یافتن بردار $\boldsymbol{w}$ بهینه: گسسته سازی فضای دو بعدی w_0 و w_1 و امتحان کردن تمامی حالت های ممکن برای این دو پارامتر

برای پیاده سازی الگوریتم grid search تابع grid_search(y, tx, w0, w1) را کامل کنید. شما بایستی یک حلقه for برای هر کدام از ابعاد بردار w بنویسید و مقدار تابع هزینه را برای هر جفت w_0 و w_1 محاسبه کنید. همه مقادیر حاصل از محاسبه تابع هزینه را در آرایه loss ذخیره کنید. کد نوشته شده تقریبی از نقطه بهینه برای مسأله را پیدا می‌کند.

کد شما بایستی مقدار مینیمم بدست آمده برای تابع هزینه به همراه مقادیر بدست آمده برای w_0^* و w_1^* را در خروجی چاپ کند. همچنین در نهایت بایستی کانتوری از مقادیر تابع هزینه به همراه خطی بدست آمده حاصل از حل مسأله بهینه سازی مطابق شکل ۲ نمایش داده شود.



شکل ۲ نمایشی از نتیجه انجام الگوریتم grid search

در گزارش نهایی خود در مورد مطالب زیر بحث کنید.

- آیا پاسخ بدست آمده تقریب مناسبی از پاسخ بهینه برای مسأله است؟ تمرین بالا را با تغییر سایز شبکه از ۱۰ به ۵۰ مجددا تکرار کنید. پاسخ جدید بدست آمده را با پاسخ قبلی مقایسه کنید.
- برای بدست آوردن یک خط مناسب گذرنده از داده‌ها نیاز به یک شبکه درشت داریم یا یک شبکه ریز؟
- افزایش تعداد نقاط شبکه چه اثری در هزینه محاسباتی مسأله دارد؟ سرعت اجرای کد چگونه تغییر میکند؟ اگر تعداد نقاط شبکه را از ۱۰ به ۵۰ افزایش دهید تعداد عملیات‌های لازم برای یافتن بردار پارامتر بهینه چند برابر می‌شود؟ (پیشنهاد می‌شود برای آشنایی بیشتر با پیچیدگی محاسباتی الگوریتم‌ها در مورد نماد O بزرگ – Big O – جست و جو کنید.)

۳-ج) الگوریتم گرادیان نزولی

با توجه به مطالب کلاس درس می‌توانیم گرادیان تابع خطای MSE را برای مسأله رگرسیون خطی تک متغیره به صورت روابط (۳) و (۴) بنویسیم.

$$\frac{\partial \mathcal{L}}{\partial w_0} = -\frac{1}{N}\sum_{i=1}^{N}(y_0 - w_0 - w_1 x_{n1}) = -\frac{1}{N}\sum_{n=1}^{N} e_n \qquad (۳)$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = -\frac{1}{N}\sum_{i=1}^{N}(y_0 - w_0 - w_1 x_{n1})x_{n1} = -\frac{1}{N}\sum_{n=1}^{N} e_n x_{n1} \qquad (۴)$$

در نتیجه بردار گرادیان تابع هزینه MSE را برای این سوال از رابطه (۵) بدست می‌آید.

$$\nabla \mathcal{L}(\boldsymbol{w}) = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_0} & \frac{\partial \mathcal{L}}{\partial w_1} \end{bmatrix} = -\frac{1}{N} \begin{bmatrix} \sum_{n=1}^{N} e_n \\ \sum_{n=1}^{N} e_n x_{n1} \end{bmatrix} = -\frac{1}{N} \boldsymbol{X}^T \boldsymbol{e} \qquad (۵)$$

با استفاده از رابطه تابع compute_gradient(y, tx, w) را کامل کنید. در پیاده سازی این تابع حتماً توجه کنید که تابع شما خروجی درستی باز می‌گرداند. این امر را به صورت دستی چک کنید. به عبارتی ابتدا مقادیر دلخواه برای $\boldsymbol{y}$ و $\boldsymbol{X}$ و $\boldsymbol{w}$ در نظر بگیرید و مقدار مشتق بالا را محاسبه کنید، سپس مقدار بدست آمده را با خروجی تابع خود مقایسه کنید.

پس از اطیمنان از اینکه تابع شما گرادیان تابع هزینه را به درستی محاسبه می‌کند، مقدار گرادیان را به ازای مقادیر زیر برای w_0 و w_1 محاسبه کنید. در کدام حالت مقدار گرادیان تابع بیشتر است و این چه معنایی می‌دهد؟

$w_0 = 100, w_1 = 20$

$w_0 = 50, w_1 = 10$

راهنمایی: یک تابع درجه دوم و مقدار گرادیان این تابع در نزدیک نقطه مینیمم و دور از آن را درنظر بگیرید.

درگام بعدی تابع gradient_descent(y, tx, initial_w, …) را کامل کنید. کد را اجرا کرده و تکرارهای الگوریتم را نمایش دهید. به مقادیر تابع هزینه و وزن‌های به‌روزرسانی شده در هر گام توجه کنید.

- آیا تابع هزینه مینیمم شده است؟
- آیا الگوریتم به همگرایی رسیده است؟ در مورد سرعت همگرایی الگوریتم چه می‌توانید بگویید؟
- مقادیر نهایی بدست آمده برای w_1 و w_0 مناسب هستند؟

۳-د) بررسی اثر مقدار دهی اولیه پارامترها

در این قسمت به بررسی اثر مقدار اولیه وزن‌ها و مقدار گام در همگرایی الگوریتم گرادیان نزولی می‌پردازیم. به‌صورت تئوری برای یک تابع محدب اگر مقدار گام به‌صورت مناسب انتخاب شود، الگوریتم گرادیان نزولی به نقطه مینیمم همگرا می‌شود.

به ازای مقادیر 0.001, 0.01, 0.5, 1, 2, 2.5 برای مقدار گام همگرایی الگوریتم را بررسی کنید. آیا همواره همگرایی حاصل می‌شود؟ به ازای مقادیر اولیه زیر برای پارامترهای w_0 و w_1 و اندازه گام $\eta = 0.1$ همگرایی الگوریتم را بررسی کنید. آیا همواره همگرایی حاصل می‌شود؟

$w_0 = 0, w_1 = 0$

$w_0 = 100, w_1 = 10$

$w_0 = -1000, w_1 = 1000$

۴. در این سوال می‌خواهیم یک مدل رگرسیون خطی توسعه دهیم و با آن تنش تسلیم بتن را به کمک ویژگی‌های ترکیبات ساخته شده آلیاژ آن پیش‌بینی کنیم. در این سوال می‌توانید از کتابخانه‌های زبان پایتون استفاده کنید. داده‌های این سوال در فایل Concrete_Data.xlsx در اختیار شما قرار دارند. پیشنهاد می‌شود که این فایل را با فرمت csv ذخیره کرده و سپس فایل csv را در پایتون بخوانید.
ستون‌های اول تا هشتم این فایل ویژگی‌های آلیاژ ساخته شده هستند. هفت ستون اول غلظت مواد سازنده در آلیاژ نهایی و ستون هشتم عمر قطعه بتنی است. ستون نهم نشان دهنده تنش تسلیم فشاری بتن‌ها است که می‌خواهیم مقدار آن را با استفاده از ویژگی‌های ورودی مدل تخمین بزنم.

الف) تنش تسلیم قطعه ساخته شده بیشترین وابستگی را به کدام ویژگی دارد؟ با استفاده از این ویژگی یک مدل خطی تک متغیره به داده‌ها برازش کنید، خطای RMSE مدل نهایی را برحسب MPa گزارش کنید.

راهنمایی: ضریب همبستگی میان ستون‌های مختلف فایلی که در اختیار دارید را محاسبه کنید و سپس برای رسم این مقادیر از heatmap استفاده کنید.

ب) با استفاده از heatmap بدست آمده در قسمت قبل و انتخاب ۳ ویژگی مهم که بیشترین همبستگی را با مقدار تنش تسلیم بتن دارند یک مدل رگرسیون خطی چند متغیره به داده‌ها برازش کنید، می‌توانید از ویژگی‌های غیرخطی مانند حاصلضرب دو ویژگی در یکدیگر را در مدل درنظر بگیرید. نمودار MSE را برای تمامی مراحل الگوریتم gradient descent گزارش کنید.

توجه: نرمالایز کردن داده‌ها می‌تواند به یادگیری کمک کرده و دقت مدل را بهبود ببخشد.