دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳
استاد: علی شریفی زارچی

مسئول تمرین:

تمرین اول | فصل اول | مهلت ارسال نهایی: ۲۵ مهر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روزهای مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر مجاز داشته باشید و در مجموع حداکثر ۱۵ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت جداگانه محاسبه می‌شود و مجموع تأخیر هر دو نباید بیشتر از ۱۵ روز شود. پس از اتمام زمان مجاز، دو روز اضافی برای آپلود غیرمجاز در نظر گرفته شده است که در این بازه به ازای هر ساعت تأخیر، ۲ درصد از نمره تمرین کسر خواهد شد.
- اگر بخش عملی یا تئوری تمرین را قبل از مهلت ارسال امتیازی آپلود کنید، ۲۰ درصد نمره اضافی به آن بخش تعلق خواهد گرفت و پس از آن، ویدئویی تحت عنوان راهنمایی برای حل تمرین منتشر خواهد شد.
- حتماً تمرین‌ها را بر اساس موارد ذکرشده در صورت سوالات حل کنید. در صورت وجود هرگونه ابهام، آن را در صفحه تمرین در سایت کوئرا مطرح کنید و به پاسخ‌هایی که از سوی دستیار آموزشی مربوطه ارائه می‌شود، توجه کنید.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- فایل پاسخ‌های سوالات نظری را در قالب یک فایل pdf به فرمت $HW1\_T\_[STD\_ID].pdf$ آماده کنید و برای سوالات عملی، هریک را در یک فایل zip جداگانه قرار دهید و فایل zip اول را به فرمت $HW1\_P1\_[STD\_ID].zip$ و فایل zip دوم را به فرمت $HW1\_P2\_[STD\_ID].zip$ نامگذاری کرده و هرکدام را به صورت جداگانه آپلود کنید.
- گردآورندگان تمرین: مبینا سلیمی‌پناه، محمد مولوی، امیرعلی لقمانی، فاطمه السادات موسوی، عرشیا قارونی

سوالات نظری (۱۰۰ نمره)

۱. (۲۰ نمره) به سوالات زیر پاسخ کوتاه دهید.

الف) چرا از تابع softmax اغلب برای مسائل دسته‌بندی استفاده می‌شود؟

ب) بالا بودن واریانس در مدل چه معنایی دارد؟ یک روش ممکن برای کاهش واریانس در مدل خود بیان کنید.

پ) چرا در حالتی که تمام ویژگی ها تا حد خوبی با خروجی مرتبط هستند، رگرسیون Ridge به رگرسیون Lasso ترجیح داده می‌شود؟

ت) چگونه رگولاریزیشن L_2 در $classifier$های خطی بر روی تعادل بایاس-واریانس تأثیر می‌گذارد؟

حل.

الف) خروجی softmax محدودیت‌های یک توزیع احتمالی را برآورده می‌کند (یا پاسخ‌های دیگری که اشاره می‌کنند خروجی softmax مقادیر بین ۰ تا ۱ تولید می‌کند که جمع آن‌ها برابر ۱ است).

ب) این بدان معناست که مدل به داده‌های آموزش بیش از حد تطبیق یافته است (overfitting) و قابلیت تعمیم ندارد. برای بخش دوم هر پاسخی که به قابلیت تعمیم کمک کند، قابل قبول است مانند اضافه کردن داده‌های بیشتر، dropout ، ساختن مدل کوچکتر و غیره.

پ) زیرا در این حالت رگولاریزیشن Lasso از بین متغیرهایی که ارتباط زیادی باهم دارند یکی را نگه می‌دارد و بقیه را حذف می‌کند در حالیکه ممکن است تمام این متغیرها در خروجی اثر بزرگی داشته باشند. در رگولاریزیشن Ridge حتی اگر متغیرها همبستگی زیادی داشته باشند تمام متغیرها حفظ می‌شوند. بنابراین امکان حذف متغیر پراهمیت وجود ندارد.

ت) رگولاریزیشن $L_۲$ با اعمال جریمه‌ای به مقادیر بزرگ وزن‌ها، پیچیدگی مدل را کنترل می‌کند. این امر از بیش‌برازش $(overfitting)$ جلوگیری کرده و واریانس مدل را کاهش می‌دهد بدون اینکه بایاس به‌طور قابل‌توجهی افزایش یابد. در نتیجه، $Generalization$ مدل بهبود می‌یابد.

۲. (۲۰ نمره) در یک مسئله رگرسیون خطی داریم:

$$y = \underline{w}^T \underline{x}, \quad \underline{x} \in \mathbb{R}^L, \quad y \in \mathbb{R}$$

$$X = [\underline{x}_۱, \underline{x}_۲, \cdots, \underline{x}_N], \quad \underline{y} = \begin{pmatrix} y_۱ \\ y_۲ \\ \vdots \\ y_N \end{pmatrix}, \quad \underline{w} = \begin{pmatrix} w_۱ \\ w_۲ \\ \vdots \\ w_L \end{pmatrix}$$

الف) اگر رگرسیون را فقط بر روی ویژگی j انجام دهیم، نشان دهید که:

$$w_j = \frac{X_j \underline{y}}{X_j X_j^T}$$

که X_j سطر j ماتریس داده‌ها است.

ب) فرض کنید ویژگی‌ها مستقل هستند (یعنی سطرهای ماتریس داده‌ها مستقل هستند). ثابت کنید که پارامترهای بهینه از آموزش رگرسیون بر روی همه ویژگی‌ها با پارامترهای بهینه حاصل از آموزش روی هر ویژگی به طور مستقل یکسان است.

پ) فرض کنید $y = \underline{w}^T \underline{x} + w.$ و رگرسیون را فقط بر روی ویژگی j انجام دهیم. $w.$ و w_j را بدست آورید.

حل.

الف) تخمین‌گر OLS برای بردار وزن به صورت معادله زیر بیان می‌شود:

$$w = (X^T X)^{-۱} X^T y$$

در سوال مورد نظر، رگرسیون خطی ساده با یک متغیر پیش‌بینی‌کننده بررسی شده است. در این حالت، ماتریس X شامل یک بردار ستونی X_j است و وزن w_j برای متغیر پیش‌بینی‌کننده X_j به صورت زیر قابل محاسبه است:

$$w_j = \frac{X_j y}{X_j X_j^T}$$

ب) در این حالت، ماتریس کوواریانس $X^T X$ یک ماتریس قطری است، چرا که عناصر غیر قطری که معرف کوواریانس بین ویژگی‌های مختلف هستند، به دلیل فرض استقلال صفر هستند. بنابراین، معکوس $X^T X$ برابر است با معکوس هر عنصر قطری.

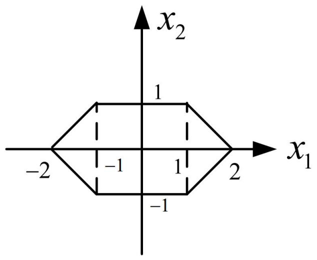
وزن برای هر ویژگی w_j می‌تواند به صورت مستقل با استفاده از فرمول بخش قبلی محاسبه شود. این به دلیل فرض استقلال است که نشان می‌دهد اطلاعات مشترکی بین ویژگی‌ها وجود ندارد که بر محاسبه وزن‌ها تاثیر بگذارد.

پ)

$$w_. = \bar{y} = E\{y\}, \quad w_j = \frac{X_j(y - w_.)}{X_j X_j^T}$$

در این بخش، $w_.$ برابر با میانگین y است و w_j به صورت فوق محاسبه می‌شود.

۳. (۲۰ نمره) یک شبکه عصبی با دو گره در لایه ورودی و دولایه مخفی و تابع فعالسازی پله داریم. وزن‌ها و بایاس‌های این شبکه را به گونه‌ای تعیین کنید که در ناحیه داخل ۶ ضلعی و روی اضلاع خروجی شبکه ۱ و در باقی نواحی ۰ باشد.

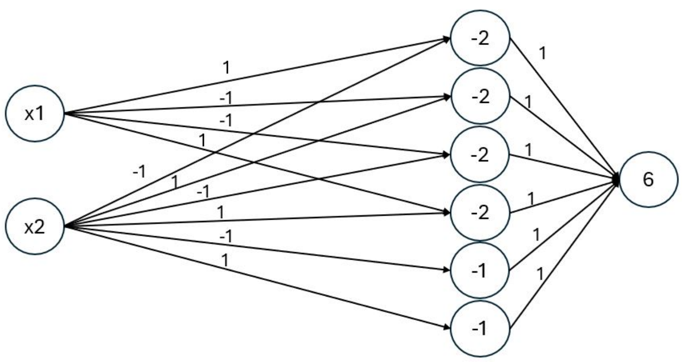


حل.

باتوجه به شکل داده شده می‌توانیم نامساوی‌های زیر را بنویسیم:

$$x_۱ - x_۲ \geq -۲, x_۲ - x_۱ \geq -۲, -x_۲ \geq -۱, x_۲ \geq -۱, -x_۱ - x_۲ \geq -۲, x_۱ + x_۲ \geq -۲$$

با توجه به اینکه تابع فعال ساز هم پله می‌باشد، می‌توان جواب زیر را برای این سوال در نظر گرفت:



۴. (۲۰ نمره) در یک مسئله دسته‌بندی دو کلاسه (binary classification) از رگرسیون لاجستیک با تابع هزینه cross entropy استفاده کرده‌ایم. تابع هزینه برای یک نقطه از داده‌ها به صورت زیر تعریف می‌شود:

$$L = -\sum_{i=۱}^{n} y_i \log(\hat{y}_i)$$

که در آن y بردار برچسب‌های واقعی و $\hat{y}$ احتمالات پیش‌بینی‌شده با استفاده از تابع softmax به صورت زیر است:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=۱}^{n} e^{z_j}}$$

الف) مشتق تابع softmax را نسبت به z_k برای دو حالت $k = i$ و $k \neq i$ به دست آورید.

ب) با استفاده از مشتق تابع softmax، مشتق تابع هزینه cross entropy را نسبت به z_k محاسبه نمایید.

حل.

الف) برای محاسبه مشتق $\hat{y}_i$ نسبت به z_k دو حالت بررسی می‌کنیم:

حالت ۱: $k = i$ در این حالت، هدف ما محاسبه‌ی مشتق $\hat{y}_i$ نسبت به z_i (یعنی $\frac{\partial \hat{y}_i}{\partial z_i}$) است.

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=۱}^{n} e^{z_j}}$$

$$\frac{\partial \hat{y}_i}{\partial z_i} = \frac{\partial}{\partial z_i}\left(\frac{e^{z_i}}{\sum_{j=۱}^{n} e^{z_j}}\right)$$

با مشتق‌گیری از کسر داریم:

$$\frac{\partial \hat{y}_i}{\partial z_i} = \frac{e^{z_i}\left(\sum_{j=۱}^{n} e^{z_j}\right) - e^{z_i} e^{z_i}}{\left(\sum_{j=۱}^{n} e^{z_j}\right)^۲}$$

که ساده می‌شود به:

$$\frac{\partial \hat{y}_i}{\partial z_i} = \hat{y}_i(۱ - \hat{y}_i)$$

حالت ۲: $k \neq i$ در این حالت، هدف محاسبه مشتق $\hat{y}_i$ نسبت به z_k است (یعنی $\frac{\partial \hat{y}_i}{\partial z_k}$ برای $i \neq k$).

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=۱}^{n} e^{z_j}}$$

با مشتق‌گیری از تابع کسری، داریم:

$$\frac{\partial \hat{y}_i}{\partial z_k} = ۰ \cdot \left(\sum_{j=۱}^{n} e^{z_j}\right) - \frac{e^{z_i} e^{z_k}}{\left(\sum_{j=۱}^{n} e^{z_j}\right)^۲}$$

که نتیجه می‌دهد:

$$\frac{\partial \hat{y}_i}{\partial z_k} = -\hat{y}_i \hat{y}_k$$

ب) برای محاسبه مشتق تابع هزینه Cross-Entropy نسبت به z_k، از قاعده مشتق زنجیره‌ای استفاده می‌کنیم.
قاعده مشتق زنجیره‌ای:

$$\frac{\partial L}{\partial z_k} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_k}$$

با توجه به تعریف تابع هزینه:

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

مشتق نسبت به $\hat{y}_i$ به این صورت است:

$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

بنابراین، می‌توانیم رابطه زیر را بنویسیم:

$$\frac{\partial L}{\partial z_k} = \sum_{i=1}^{n} \left(-y_i \frac{1}{\hat{y}_i}\right) \cdot \frac{\partial \hat{y}_i}{\partial z_k}$$

اکنون از مشتق تابع Softmax که در بخش الف به دست آوردیم، استفاده می‌کنیم:
برای $i = k$ داریم:

$$\frac{\partial \hat{y}_k}{\partial z_k} = \hat{y}_k(1 - \hat{y}_k)$$

برای $i \neq k$ داریم:

$$\frac{\partial \hat{y}_i}{\partial z_k} = -\hat{y}_i \hat{y}_k$$

اکنون می‌توانیم دو قسمت را با هم ترکیب کنیم:

$$\frac{\partial L}{\partial z_k} = -y_k(1 - \hat{y}_k) + \sum_{i \neq k} y_i \hat{y}_k \qquad (1)$$

از آنجایی که $\sum_{i \neq k} y_i = 1 - y_k$، معادله به شکل زیر ساده می‌شود:

$$\frac{\partial L}{\partial z_k} = -y_k(1 - \hat{y}_k) + \hat{y}_k(1 - y_k) \qquad (2)$$

که در نهایت ساده می‌شود به:

$$\frac{\partial L}{\partial z_k} = \hat{y}_k - y_k \tag{۳}$$

نتیجه نهایی: این نشان می‌دهد که مشتق تابع هزینه Cross-Entropy به سادگی برابر است با تفاوت احتمال پیش‌بینی‌شده ($\hat{y}_k$) و برچسب واقعی (y_k):

$$\frac{\partial L}{\partial z_k} = \hat{y}_k - y_k \tag{۴}$$

۵. (۲۰ نمره) به سوالات زیر در مورد رگرسیون Ridge پاسخ دهید:

الف) نشان دهید که به ازای مقادیر $\lambda > ۰$ واریانس ضرایب Ridge از واریانس ضرایب رگرسیون خطی کوچکتر است:

$$\text{Var}(\hat{\beta}^{LS}) > \text{Var}(\hat{\beta}^{Ridge}(\lambda))$$

ب) نشان دهید به ازای مقادیر $\lambda > ۰$ رابطه زیر برقرار است:

$$\text{tr}\left\{\text{Var}[\hat{Y}(\lambda)]\right\} = \sigma^۲ \sum_{j=۱}^{p} (D_x)_{jj}^۴ \left[(D_x)_{jj}^۲ + \lambda\right]^{-۲}$$

در این رابطه D_x یک ماتریس قطری است که شامل مقادیر تکین X است.

حل. الف) از تجزیه **SVD** ماتریس X استفاده می‌کنیم.

فرض: $\epsilon \sim N(۰, \sigma^۲ I)$، $\quad y = X\beta + \epsilon$

$$\beta_{LS} = (X^T X)^{-۱} X^T y \quad \Rightarrow \quad \text{Cov}(\hat{\beta}_{LS}) = E\left[(\hat{\beta}_{LS} - E(\hat{\beta}))(\hat{\beta}_{LS} - E(\hat{\beta}))^T\right]$$

$$\beta_{ls} = (X^T X)^{-۱} X^T (X\beta + \epsilon) \Rightarrow \beta_{ls} = \beta + (X^T X)^{-۱} X^T \epsilon$$
$$\text{Cov}(\beta_{ls}) = \text{Cov}((X^T X)^{-۱} X^T \epsilon)$$

$$\text{Cov}(\beta_{ls}) = E\left[((X^T X)^{-۱} X^T \epsilon - \underbrace{E\left[(X^T X)^{-۱} Z^T \epsilon\right]}_{۰})((X^T X)^{-۱} X^T \epsilon - \underbrace{E\left[(X^T X)^{-۱} X^T \epsilon\right]}_{۰})^T\right]$$

$$\text{Cov}(\beta_{ls}) = E\left[((X^T X)^{-۱} X^T \epsilon)(\epsilon^T X ((X^T X)^{-۱})^T)\right]$$

$$\xrightarrow{X=UDV^T} cov(\vec{\beta_{ls}}) = E[(UD^۲V^T)^{-۱} V D U^T \epsilon\epsilon^T U D V^T (V D^۲ V^T)^{-۱}] \to E[\epsilon\epsilon^T] = \sigma^۲ I$$

$$\Rightarrow \sigma^۲ ((V D^۲ V^T)^{-۱})^T = \sigma^۲ \sum_{j=۱}^{p} D_{jj}^{-۲} V_j V_j^T$$

$$\beta_{ridge} = (X^T X + \lambda I)^{-۱} X^T y \to cov(\vec{\beta_{ridge}}) = \sigma^۲ (V D^۲ V^T + \lambda I)^{-۱} V D^۲ V^T ((V D^۲ V^T + \lambda I)^{-۱})^T$$

$$= \sigma^۲ \sum_{i=۱}^{p} \frac{D_{jj}^۲}{(D_{jj}^۲ + \lambda)^۲} V_j V_j^T$$

با توجه به اینکه واریانس هریک از درایه‌های بردار β برابر با درایه متناظر روی قطر $cov(\hat{\beta})$ است داریم:

$$var(\beta_{j_{ls}}) = \frac{\sigma^۲}{D_{jj}^۲}$$

$$var(\beta_{j_{ridge}}) = \sigma^۲ \frac{D_{jj}^۲}{(D_{jj}^۲ + \lambda)^۲}$$

$$\xrightarrow{\lambda > ۰} (D_{jj}^۲ + \lambda)^۲ = D_{jj}^۴ + \lambda^۲ + ۲\lambda D_{jj} > D_{jj}^۲$$

ب)

$$\beta = (X^T X + \lambda I)^{-۱} X^T y$$

$$\hat{Y} = X\beta + \epsilon$$

$$X = UDV^T$$

$$cov(\epsilon\epsilon^T) = \sigma^۲ I$$

$$var[\hat{Y}(\lambda)] = E[(\hat{Y} - E\hat{Y})(\hat{Y} - E\hat{Y})^T] \Rightarrow E[UD(D^۲ + \lambda I)^{-۱} DU^T \epsilon\epsilon^T UD((D^۲ + \lambda I)^{-۱})^T DU^T]$$

$$= \sigma^۲ UD(D^۲ + \lambda I)^{-۱} D^۲ ((D^۲ + \lambda I)^{-۱})^T DU^T$$

$$(D^۲ + \lambda I)^{-۱} = \begin{bmatrix} \frac{۱}{(D_۱^۲ + \lambda)} & ۰ & ۰ & ۰ \\ ۰ & \frac{۱}{(D_۲^۲ + \lambda)} & ۰ & ۰ \\ ۰ & ۰ & \ddots & ۰ \\ ۰ & ۰ & ۰ & \frac{۱}{(D_P^۲ + \lambda)} \end{bmatrix}_{P*P}$$

$$D^۲ = \begin{bmatrix} D_۱^۲ & ۰ & ۰ & ۰ \\ ۰ & D_۲^۲ & ۰ & ۰ \\ ۰ & ۰ & \ddots & ۰ \\ ۰ & ۰ & ۰ & D_P^۲ \end{bmatrix}_{P*P}$$

$$var(\hat{Y}) = \sigma^۲ U \begin{bmatrix} \frac{D_۱^۲}{(D_۱^۲ + \lambda)^۲} & ۰ & ۰ & ۰ \\ ۰ & \frac{D_۲^۲}{(D_۲^۲ + \lambda)^۲} & ۰ & ۰ \\ ۰ & ۰ & \ddots & ۰ \\ ۰ & ۰ & ۰ & \frac{D_P^۲}{(D_P^۲ + \lambda)^۲} \end{bmatrix}_{P*P} U^T$$

$$\rightarrow tr(var(\hat{Y})) = \sigma^۲ \sum_{j=۱}^{P} \frac{D_j^۴}{(D_j^۲ + \lambda)^۲}$$

سوالات عملی (۱۰۰ نمره)

---

۱. (۱۰۰ نمره) برای حل سوالات به $notebook$ های ضمیمه شده مراجعه کنید.

(۱) (۶۰ نمره) برای پاسخ به تمرین عملی اول ابتدا فایل نوت‌بوک قرار گرفته را باز کنید و سپس مراحل را مطابق آن‌چه که از شما خواسته شده انجام دهید. در نهایت، مقادیر پیش‌بینی شده برای دیتاست InsuranceData_test.csv را در یک فایل به نام submission.csv که شامل یک ستون به نام $charges$ میباشد، ذخیره کنید. فایل خروجی و فایل نوتبوک را در یک فایل zip قرار دهید و آن را به فرمت $HW۱\_P۱\_[STD\_ID].zip$ نامگذاری کرده و آپلود کنید.
توجه بفرمایید این سوال دارای داوری خودکار می‌باشد و ۱۵ نمره از ۶۰ نمره به این قسمت تعلق دارد.

(۲) (۴۰ نمره) برای پاسخ به تمرین عملی دوم تنها کافی است نوتبوک $Perceptron.ipynb$ را تکمیل کرده و سپس مطابق با فرمت ذکر شده آپلود کنید.

حل.
notebook های حل‌شده ضمیمه شده‌اند.