دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳

علیرضا میررکنی - ۴۰۱۱۰۶۶۱۷

تمرین اول

پاسخ مسئله‌ی ۱.

الف) تابع فعال سازی **Softmax** اغلب در مسائل دسته‌بندی مورد استفاده قرار می گیرد، چرا که می توان مقادیر خروجی آن را به عنوان احتمال تعلق نمونه ورودی به هر یک از کلاس ها در نظر گرفت. تابع Softmax ورودی‌های خود را به مقادیر بین صفر و یک تبدیل می‌کند و مجموع خروجی‌های آن برابر با یک می باشد؛ بنابراین خروجی این تابع تمامی خصوصیات یک توزیع احتمالاتی معتبر بر روی کلاس های مختلف را دارا می باشد. این ویژگی باعث می‌شود که بتوان از آن برای پیش‌بینی احتمال تعلق یک نمونه به هر یک از کلاس‌ها استفاده کرد. همچنین Softmax یک تابع همواره مشتق پذیر است که این موضوع سبب می شود تا استفاده از آن به هنگام بهینه سازی تابع هزینه (هنگام آموزش مدل) مشکل ساز نباشد.

به‌طور خاص، در مدل‌های یادگیری عمیق مانند شبکه‌های عصبی، از این تابع فعال سازی در لایه آخر استفاده می‌شود تا پیش‌بینی‌های مدل به احتمالات برای کلاس‌های مختلف تبدیل شوند. اگر خروجی یک شبکه عصبی به صورت یک بردار باشد که هر عنصر آن نشان‌دهنده امتیاز (logit) برای یک کلاس است، Softmax این امتیازات را به احتمالات نرمال‌سازی می‌کند.

معادله تابع Softmax به شکل زیر است:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=۱}^{n} e^{z_j}}$$

که در آن z_i امتیاز یا logit کلاس i و n تعداد کل کلاس‌ها می باشد.

ب) واریانس بالا به این معنی است که مدل در هنگام تغییر داده‌های آموزشی، به شدت تغییر می‌کند. به عبارتی دیگر، مدل دچار بیش برازش (overfitting) شده است، یعنی مدل به خوبی به داده‌های آموزشی پاسخ می‌دهد، اما در مواجهه با داده‌های جدید (داده‌های آزمون) عملکرد ضعیفی دارد. در چنین حالتی، مدل بیش از حد به جزئیات و نویزهای موجود در داده‌های آموزشی توجه کرده و در تعمیم به داده‌های جدید (مثلا داده های موجود در مجموعه آزمون) دچار مشکل می‌شود.

یکی از روش‌های موثر برای کاهش واریانس مدل، استفاده از روش Regularization است. Regularization تکنیکی است که از پیچیدگی مدل کاسته و از بیش برازش آن جلوگیری می‌کند. یکی از رایج‌ترین و معروف ترین روش‌های Regularization، L2 regularization (یا Ridge Regression) نام دارد که یک جریمه بر روی مقادیر وزن‌های مدل اعمال می‌کند و آن‌ها را کوچک‌تر نگه می‌دارد تا مدل بیش از حد به داده‌های آموزشی وابسته نشود. یکی دیگر از روش های Regularization که به‌ویژه برای شبکه‌های عصبی به شکل گسترده ای مورد استفاده قرار می گیرد، تکنیک Early Stopping است. در این تکنیک، بهینه سازی مدل به شکل کامل انجام نمی شود و بعد از تعدادی مرحله فرآیند بهینه سازی را متوقف می کنیم تا از حفظ کردن جزئیات داده های آموزشی توسط مدل به منظور کاهش هر چه بیشتر تابع هزینه جلوگیری شود.

پ) در ابتدا به تفاوت میان این دو نوع رگرسیون دقت نمایید:

- **رگرسیون Ridge** از **L2 regularization** استفاده می‌کند که به تابع هزینه یک جریمه به شکل مجموع مربع وزن‌ها اضافه می‌کند:
$$\lambda \sum w_i^2$$
این جریمه وزن‌ها را کوچک می‌کند، اما معمولاً هیچ‌کدام از وزن‌ها را به طور کامل صفر نمی‌کند. در نتیجه همه ویژگی‌ها در مدل باقی می‌مانند ولی اثرگذاری کمتری خواهند داشت.

- **رگرسیون Lasso** از **L1 regularization** استفاده می‌کند و جریمه‌ای به شکل مجموع قدر مطلق وزن‌ها را به تابع هزینه اضافه می‌نماید:
$$\lambda \sum |w_i|$$
این جریمه باعث می‌شود که برخی از وزن‌ها دقیقاً به صفر برسند. بنابراین Lasso به طور خودکار ویژگی‌های غیرمهم را حذف می‌کند.

با توجه به توضیحات بالا، می توان گفت که اگر تمامی ویژگی‌های مدل با خروجی به‌طور قابل توجهی مرتبط باشند و حذف هیچ‌کدام از آن‌ها به کاهش خطای مدل کمک نکند، رگرسیون Ridge گزینه بهتری است. زیرا رگرسیون Lasso تمایل به صفر کردن وزن‌های برخی از ویژگی‌ها دارد که ممکن است منجر به حذف اطلاعات مهمی شود.

همچنین در بعضی موارد رگرسیون Ridge نسبت به رگرسیون Lasso پایدارتر است و در مواجهه با داده‌های جدید عملکرد بهتری دارد، به خصوص در مواقعی که داده‌ها به شدت پر نویز هستند. رگرسیون Ridge تمامی ویژگی‌ها را نگه می‌دارد و تاثیر همه را در نظر می‌گیرد، بنابراین ممکن است مدل نهایی تعمیم‌پذیری بهتری داشته باشد.

در نهایت در رگرسیون Ridge به دلیل استفاده از جریمه $\lambda \sum w_i^2$، تمامی وزن‌ها به شکلی متعادل کوچک می‌شوند. این باعث می‌شود که مدل به طور هموارتر رفتار کند و پیش‌بینی خطی ملایم‌تری داشته باشد. در رگرسیون Lasso به دلیل حذف کامل برخی از ویژگی‌ها، تغییرات مدل می‌تواند ناگهانی‌تر باشد که ممکن است منجر به افزایش واریانس مدل و بیش برازش آن شود.

ت) رگولاریزیشن L_2 با اضافه کردن یک جریمه به‌شکل مضربی از مجموع مربعات وزن‌های مدل، پیچیدگی مدل را کاهش می‌دهد. تأثیر این کار بر تعادل بایاس–واریانس به شرح زیر است:

- **کاهش واریانس**: رگولاریزیشن L_2 به کاهش واریانس مدل کمک می‌کند. بدون رگولاریزیشن، یک مدل خطی ممکن است دچار بیش برازش شود، به این معنی که مدل بیش از حد به داده‌های آموزشی وابسته می‌شود و نمی‌تواند روی داده‌های جدید به خوبی عمل کند. با کنترل کردن اندازه وزن‌ها و جلوگیری از افزایش بیش از اندازه آنها، پیچیدگی مدل کاهش یافته و مدل کمتر دچار تغییرات شدید با تغییرات جزئی در داده‌های آموزشی می‌شود (اصطلاحا مدل هموارتر یا smooth تر می شود) و در نتیجه واریانس مدل کاهش می‌یابد.

- **افزایش بایاس**: اعمال رگولاریزیشن L_2 می‌تواند بایاس مدل را کمی افزایش دهد. زیرا محدود کردن وزن‌ها به این معنی است که مدل نمی‌تواند به اندازه کافی انعطاف‌پذیر باشد تا تمام جزئیات داده‌ها را فرا بگیرد.

در مجموع می توان گفت که رگولاریزیشن L_2 تعادلی بین بایاس و واریانس برقرار می‌کند. اگرچه بایاس کمی افزایش می‌یابد، اما کاهش واریانس معمولاً بهبود عملکرد مدل روی داده‌های جدید را به همراه دارد. این نوع رگولاریزیشن به طور کلی به بهبود تعمیم‌پذیری مدل کمک می‌کند.

پاسخ مسئله‌ی ۲.

الف) دقت کنید که اگر رگرسیون را فقط بر روی ویژگی jام انجام دهیم، هدف کمینه کردن تابع هزینه زیر خواهد بود:

$$\mathcal{L} = \sum_{i=1}^{n}(w_j^T x_j^{(i)} - y^{(i)})^2 = \sum_{i=1}^{n}(x_j^{(i)^T} w_j - y^{(i)})^2 = \|X_j^T w_j - y\|_2^2$$

چرا که در این صورت، مسئله با مسئله رگرسیون خطی زمانی که تنها یک ویژگی داریم (که در اینجا ویژگی jام در بردار ویژگی ها می باشد) معادل خواهد شد و می دانیم که در آنجا نیز از میانگین مربعات خطا به عنوان تابع هزینه استفاده می‌شود.

حال دقت کنید که می توان تابع هزینه را به شکل زیر ساده کرد:

$$\begin{aligned}
\mathcal{L} &= \|X_j^T w_j - y\|_2^2 \\
&= (X_j^T w_j - y)^T (X_j^T w_j - y) \\
&= \left((X_j^T w_j)^T - y^T\right)(X_j^T w_j - y) \\
&= (w_j^T X_j - y^T)(X_j^T w_j - y) \\
&= w_j^T X_j X_j^T w_j - w_j^T X_j y - y^T X_j^T w_j + y^T y
\end{aligned}$$

حال دقت کنید که داریم:

$$\begin{aligned}
&\langle X_j^T w_j, y\rangle = \langle y, X_j^T w_j\rangle \\
&\implies (X_j^T w_j)^T y = y^T (X_j^T w_j) \\
&\implies w_j^T X_j y = y^T X_j^T w_j
\end{aligned}$$

بنابراین خواهیم داشت:

$$\mathcal{L} = w_j^T X_j X_j^T w_j - 2 w_j^T X_j y + y^T y$$

برای پیدا کردن w_j بهینه که تابع $\mathcal{L}$ را کمینه کند، گرادیان $\mathcal{L}$ نسبت به w_j را محاسبه کرده و آن را برابر صفر قرار می دهیم:

$$\frac{\partial \mathcal{L}}{\partial w_j} = \left(X_j X_j^T + (X_j X_j^T)^T\right) w_j - 2 X_j y = 2 X_j X_j^T - 2 X_j y$$

پس می توان نوشت: (دقت کنید که X_j یک ماتریس سطری است، در نتیجه $X_j X_j^T$ یک عدد می باشد و می توان آن را به مخرج کسر منتقل نمود. همچنین فرض می کنیم که هیچ یک از ویژگی ها تماما صفر نیستند)

$$\frac{\partial \mathcal{L}}{\partial w_j} = 0 \implies 2 X_j X_j^T w_j - 2 X_j y = 0 \implies X_j X_j^T w_j = X_j y \implies w_j = \frac{X_j y}{X_j X_j^T}$$

که این همان حکم مسئله است.

البته باید به این نکته توجه نمایید که چون تابع $\mathcal{L}$ یک تابع محدب است، در نقطه ای که گرادیانش صفر می شود به کمینه سراسری خود خواهد رسید.

ب) در ابتدا دقت کنید که تابع هزینه در این قسمت به شکل زیر می باشد:

$$\mathcal{L} = \sum_{i=۱}^{n} (w^T x^{(i)} - y^{(i)})^۲ = \sum_{i=۱}^{n} (x^{(i)^T} w - y^{(i)})^۲ = \|X^T w - y\|_۲^۲$$

در ادامه مشابه قسمت الف تابع هزینه را ساده می کنیم:

$$\begin{aligned}
\mathcal{L} &= \|X^T w - y\|_۲^۲ \\
&= (X^T w - y)^T (X^T w - y) \\
&= \left((X^T w)^T - y^T\right)(X^T w - y) \\
&= (w^T X - y^T)(X^T w - y) \\
&= w^T X_j X_j^T w - w^T X y - y^T X^T w + y^T y
\end{aligned}$$

حال دقت کنید که داریم:

$$\begin{aligned}
&\langle X^T w, y\rangle = \langle y, X^T w\rangle \\
&\Longrightarrow (X^T w)^T y = y^T (X^T w) \\
&\Longrightarrow w^T X y = y^T X^T w
\end{aligned}$$

بنابراین خواهیم داشت:

$$\mathcal{L} = w^T X X^T w - ۲ w^T X y + y^T y$$

برای پیدا کردن w بهینه که تابع $\mathcal{L}$ را کمینه کند، گرادیان $\mathcal{L}$ نسبت به w را محاسبه کرده و آن را برابر صفر قرار می دهیم:

$$\frac{\partial \mathcal{L}}{\partial w} = \left(X X^T + (X X^T)^T\right) w - ۲ X y = ۲ X X^T - ۲ X y$$

$$\frac{\partial \mathcal{L}}{\partial w} = ۰ \Longrightarrow ۲ X X^T w - ۲ X y = ۰ \Longrightarrow X X^T w = X y$$

چون در صورت سوال فرض شده است که ویژگی ها دو به دو بر هم عمودند، برای هر $۱ \leqslant i \neq j \leqslant L$ تساوی $X_i X_j^T = ۰$ برقرار می باشد. در نتیجه می توان نوشت:

$$X X^T = \begin{bmatrix} X_۱ X_۱^T & X_۱ X_۲^T & \dots & X_۱ X_L^T \\ X_۲ X_۱^T & X_۲ X_۲^T & \dots & X_۲ X_L^T \\ \vdots & \vdots & \ddots & \vdots \\ X_L X_۱^T & X_L X_۲^T & \dots & X_L X_L^T \end{bmatrix} = \begin{bmatrix} X_۱ X_۱^T & ۰ & \dots & ۰ \\ ۰ & X_۲ X_۲^T & \dots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \dots & X_L X_L^T \end{bmatrix}$$

در نتیجه با فرض اینکه هیچ کدام از ویژگی ها تماما صفر نیستند، $X X^T$ یک ماتریس قطری با درایه های قطری ناصفر بوده و وارون پذیر می باشد. پس خواهیم داشت:

$$X X^T w = X y \Longrightarrow w = (X X^T)^{-۱} X y$$

از طرفی دقت کنید که داریم:

$$(XX^T)^{-۱} = \begin{bmatrix} X_۱X_۱^T & ۰ & \dots & ۰ \\ ۰ & X_۲X_۲^T & \dots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \dots & X_LX_L^T \end{bmatrix}^{-۱} = \begin{bmatrix} \frac{۱}{X_۱X_۱^T} & ۰ & \dots & ۰ \\ ۰ & \frac{۱}{X_۲X_۲^T} & \dots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \dots & \frac{۱}{X_LX_L^T} \end{bmatrix}$$

همچنین می توان نوشت:

$$Xy = \begin{bmatrix} X_۱ \\ X_۲ \\ \vdots \\ X_L \end{bmatrix} y = \begin{bmatrix} X_۱y \\ X_۲y \\ \vdots \\ X_Ly \end{bmatrix}$$

در نهایت با جایگذاری تساوی های بالا در عبارت به دست آمده برای بردار w، خواهیم داشت:

$$w = (XX^T)^{-۱}Xy = \begin{bmatrix} \frac{۱}{X_۱X_۱^T} & ۰ & \dots & ۰ \\ ۰ & \frac{۱}{X_۲X_۲^T} & \dots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ ۰ & ۰ & \dots & \frac{۱}{X_LX_L^T} \end{bmatrix} \begin{bmatrix} X_۱y \\ X_۲y \\ \vdots \\ X_Ly \end{bmatrix} = \begin{bmatrix} \frac{X_۱y}{X_۱X_۱^T} \\ \frac{X_۲y}{X_۲X_۲^T} \\ \vdots \\ \frac{X_Ly}{X_LX_L^T} \end{bmatrix}$$

$$\Longrightarrow \forall ۱ \leqslant j \leqslant L;\ w_j = \frac{X_jy}{X_jX_j^T}$$

و در نتیجه حکم مسئله درست می باشد.

پ) در این قسمت، تابع هزینه به شکل زیر می باشد:

$$\mathcal{L} = \sum_{i=۱}^{n}(w_j^Tx_j^{(i)} + w_۰ - y^{(i)})^۲ = \sum_{i=۱}^{n}(x_j^{(i)^T}w_j + w_۰ - y^{(i)})^۲ = \|X_j^Tw_j + w_۰\mathcal{I} - y\|_۲^۲$$

به طوری که $\mathcal{I}$ برداری n بعدی است که تمامی مولفه‌های آن برابر ۱ هستند.

مشابه قسمت های قبل، در ابتدا تابع هزینه را ساده می کنیم:

$$\begin{aligned}
\mathcal{L} &= \|X_j^Tw_j + w_۰\mathcal{I} - y\|_۲^۲ \\
&= (X_j^Tw_j + w_۰\mathcal{I} - y)^T(X_j^Tw_j + w_۰\mathcal{I} - y) \\
&= (w_j^TX_j + \mathcal{I}^Tw_۰^T - y^T)(X_j^Tw_j + w_۰\mathcal{I} - y) \\
&= w_j^TX_jX_j^Tw_j + w_j^TX_jw_۰\mathcal{I} - w_j^TX_jy + \mathcal{I}^Tw_۰^TX_j^Tw_j + \mathcal{I}^Tw_۰^Tw_۰\mathcal{I} \\
&\qquad - \mathcal{I}^Tw_۰^Ty - y^TX_j^Tw_j - y^Tw_۰\mathcal{I} + y^Ty
\end{aligned}$$

حال دقت کنید که مشابه انچه که پیش تر نشان دادیم، در اینجا نیز داریم:

$$\langle X_j^T w_j, w_۰\mathcal{I}\rangle = \langle w_۰\mathcal{I}, X_j^T w_j\rangle \Longrightarrow w_j^T X_j w_۰\mathcal{I} = \mathcal{I}^T w_۰^T X_j^T w_j$$

$$\langle X_j^T w_j, y\rangle = \langle y, X_j^T w_j\rangle \Longrightarrow w_j^T X_j y = y^T X_j^T w_j$$

$$\langle w_۰\mathcal{I}, y\rangle = \langle y, w_۰\mathcal{I}\rangle \Longrightarrow \mathcal{I}^T w_۰^T y = y^T w_۰\mathcal{I}$$

بنابراین خواهیم داشت:

$$\mathcal{L} = w_j^T X_j X_j^T w_j + \mathcal{I}^T w_۰^T w_۰\mathcal{I} + y^T y + ۲w_j^T X_j w_۰\mathcal{I} - ۲w_j^T X_j y - ۲\mathcal{I}^T w_۰^T y$$

برای پیدا کردن $w_j, w_۰$ بهینه که تابع $\mathcal{L}$ را کمینه کنند، گرادیان $\mathcal{L}$ نسبت به هر کدام از این دو متغیر را محاسبه کرده و آنها را برابر صفر قرار می دهیم: (دقت کنید که چون $w_۰$ و w_j عدد هستند، می توان آن ها را در ضرب با ماتریس ها جابجا کرد و همچنین تساوی های $w_۰^T = w_۰$ و $w_j^T = w_j$ برقرار می باشند)

$$\frac{\partial \mathcal{L}}{\partial w_j} = ۲w_j X_j X_j^T + ۲w_۰ X_j\mathcal{I} - ۲X_j y$$
$$\frac{\partial \mathcal{L}}{\partial w_۰} = ۲w_۰\mathcal{I}^T\mathcal{I} + ۲w_j X_j\mathcal{I} - ۲\mathcal{I}^T y$$

$$\frac{\partial \mathcal{L}}{\partial w_j} = ۰ \Longrightarrow w_j X_j X_j^T + w_۰ X_j\mathcal{I} = X_j y$$
$$\frac{\partial \mathcal{L}}{\partial w_۰} = ۰ \Longrightarrow w_۰\mathcal{I}^T\mathcal{I} + w_j X_j\mathcal{I} = \mathcal{I}^T y$$

در نهایت با حل دستگاه دو معادله دو مجهول بالا، خواهیم داشت:

$$\mathbf{w_۰} = \frac{\mathbf{d_۲} - \frac{\mathbf{Bd_۱}}{\mathbf{A}}}{\mathbf{C} - \frac{\mathbf{B^۲}}{\mathbf{A}}} = \frac{\mathbf{d_۲A} - \mathbf{d_۱B}}{\mathbf{CA} - \mathbf{B^۲}}$$

$$\mathbf{w_j} = \frac{\mathbf{d_۱} - \mathbf{B}\left(\frac{\mathbf{d_۲} - \frac{\mathbf{Bd_۱}}{\mathbf{A}}}{\mathbf{C} - \frac{\mathbf{B^۲}}{\mathbf{A}}}\right)}{\mathbf{A}} = \frac{\mathbf{d_۱C} - \mathbf{d_۲B}}{\mathbf{CA} - \mathbf{B^۲}}$$

به‌طوری که در عبارات بالا داریم:

$$A = X_j X_j^T, \quad B = X_j\mathcal{I}, \quad C = \mathcal{I}^T\mathcal{I}, \quad d_۱ = X_j y, \quad d_۲ = \mathcal{I}^T y$$

پاسخ مسئله‌ی ۳.

دقت کنید که برای قرارگیری نقطه به مختصات (x_1, x_2) در داخل و یا روی مرز ناحیه مشخص شده، تمامی نابرابری های زیر باید به صورت همزمان برقرار باشند:

$$x_2 \leqslant 1 \qquad\qquad x_2 \geqslant -1$$

$$x_2 \leqslant 2 - x_1 \qquad\qquad x_2 \leqslant 2 + x_1$$

$$x_2 \geqslant -2 + x_1 \qquad\qquad x_2 \geqslant -2 - x_1$$

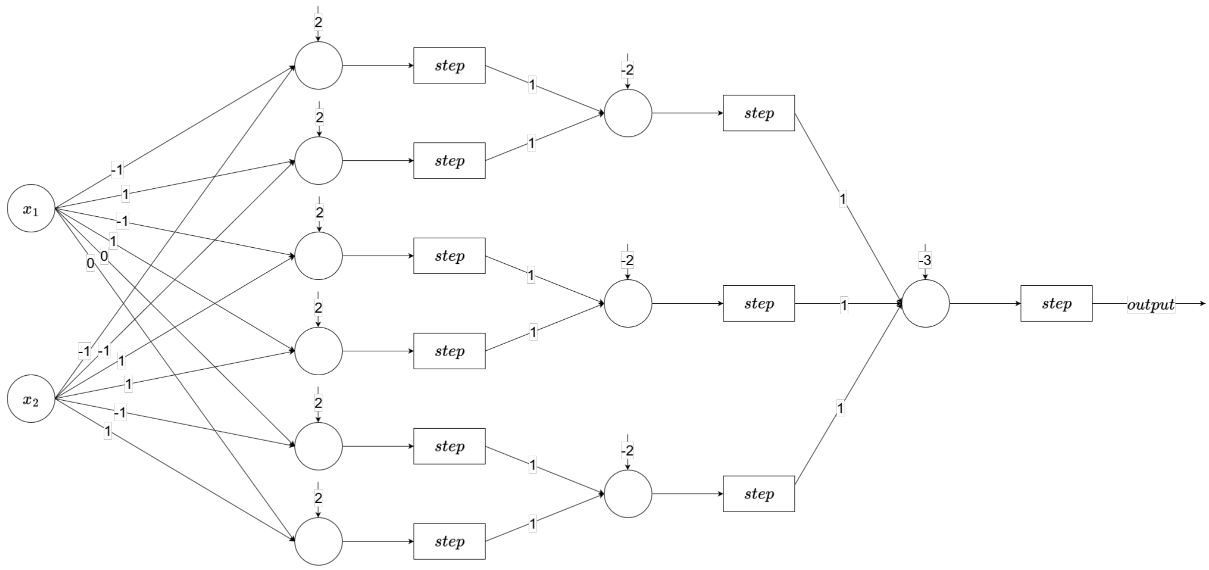
یا به طور معادل باید داشته باشیم:

$$1 - x_2 \geqslant 0 \qquad\qquad x_2 + 1 \geqslant 0$$

$$2 - x_1 - x_2 \geqslant 0 \qquad\qquad 2 + x_1 - x_2 \geqslant 0$$

$$x_2 + 2 - x_1 \geqslant 0 \qquad\qquad x_2 + 2 + x_1 \geqslant 0$$

خروجی شبکه عصبی زیر که دارای دو لایه مخفی می باشد، تنها زمانی یک می شود که مختصات نقطه ورودی در تمامی نابرابری های بالا صدق کند و در غیر این صورت خروجی آن صفر خواهد شد.



شکل ۱: شبکه عصبی با دولایه مخفی

دقت کنید که در طراحی شبکه بالا، تعریف تابع پله به شکل زیر در نظر گرفته شده است:

$$step(x) = \begin{cases} 1 & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

پاسخ مسئله‌ی ۴.

الف) در دو حالت گفته شده، مشتق تابع softmax را نسبت به z_k به‌دست می‌آوریم.

- **حالت** $k = i$:

$$
\begin{aligned}
\frac{\partial \hat{y}_k}{\partial z_k} &= \frac{\partial \frac{e^{z_k}}{\sum_{j=1}^{n} e^{z_j}}}{\partial z_k} \\
&= \left[\frac{\partial e^{z_k}}{\partial z_k} \cdot \sum_{j=1}^{n} e^{z_j} - \frac{\partial \left(\sum_{j=1}^{n} e^{z_j} \right)}{\partial z_k} \cdot e^{z_k} \right] \cdot \frac{1}{\left(\sum_{j=1}^{n} e^{z_j} \right)^2} \\
&= \left[e^{z_k} \cdot \sum_{j=1}^{n} e^{z_j} - e^{z_k} \cdot e^{z_k} \right] \cdot \frac{1}{\left(\sum_{j=1}^{n} e^{z_j} \right)^2} \\
&= \frac{e^{z_k}}{\sum_{j=1}^{n} e^{z_j}} - \frac{e^{2z_k}}{\left(\sum_{j=1}^{n} e^{z_j} \right)^2} \\
&= \frac{e^{z_k}}{\sum_{j=1}^{n} e^{z_j}} - \left(\frac{e^{z_k}}{\sum_{j=1}^{n} e^{z_j}} \right)^2 \\
&= \hat{y}_k - \hat{y}_k^2 \\
&= \hat{y}_k (1 - \hat{y}_k)
\end{aligned}
$$

- **حالت** $k \neq i$:

$$
\begin{aligned}
\frac{\partial \hat{y}_i}{\partial z_k} &= \frac{\partial \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}}{\partial z_k} \\
&= \left[\frac{\partial e^{z_i}}{\partial z_k} \cdot \sum_{j=1}^{n} e^{z_j} - \frac{\partial \left(\sum_{j=1}^{n} e^{z_j} \right)}{\partial z_k} \cdot e^{z_i} \right] \cdot \frac{1}{\left(\sum_{j=1}^{n} e^{z_j} \right)^2} \\
&= \left[-e^{z_k} \cdot e^{z_i} \right] \cdot \frac{1}{\left(\sum_{j=1}^{n} e^{z_j} \right)^2} \\
&= -\frac{e^{z_i}}{\left(\sum_{j=1}^{n} e^{z_j} \right)} \cdot \frac{e^{z_k}}{\left(\sum_{j=1}^{n} e^{z_j} \right)} \\
&= -\hat{y}_i \hat{y}_k
\end{aligned}
$$

ب) با توجه به مشتق های به دست آمده در قسمت قبل، خواهیم داشت:

$$\begin{aligned}
\frac{\partial L}{\partial z_k} &= -\frac{\partial \left(\sum_{i=۱}^{n} y_i \log \hat{y}_i\right)}{\partial z_k} \\
&= -\frac{\partial(y_k \log \hat{y}_k)}{\partial z_k} - \frac{\partial \left(\sum_{i\neq k} y_i \log \hat{y}_i\right)}{\partial z_k} \\
&= -\frac{\partial(y_k \log \hat{y}_k)}{\partial z_k} - \sum_{i\neq k} \frac{\partial(y_i \log \hat{y}_i)}{\partial z_k} \\
&= -y_k \cdot \frac{\partial \log \hat{y}_k}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_k} - \sum_{i\neq k} y_i \cdot \frac{\partial \log \hat{y}_i}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_k} \\
&= -y_k \cdot \frac{۱}{\hat{y}_k} \cdot \hat{y}_k(۱ - \hat{y}_k) + \sum_{i\neq k} y_i \cdot \frac{۱}{\hat{y}_i} \cdot \hat{y}_k \hat{y}_i \\
&= y_k(\hat{y}_k - ۱) + \sum_{i\neq k} y_i \hat{y}_k \\
&= \sum_{i=۱}^{n} y_i \hat{y}_k - \hat{y}_k
\end{aligned}$$

پاسخ مسئله‌ی ۵.

الف) در ابتدا تابع هزینه را برای رگرسیون Ridge و رگرسیون خطی می نویسیم و با محاسبه گرادیان این توابع و برابر قرار دادن گرادیان آنها با صفر، پارامتر بهینه را برای هرکدام به دست می‌آوریم. دقت کنید که فرض می‌کنیم ماتریس X و بردار های y و β به شکل زیر تعریف شده اند:

$$y = X\beta + e, \quad e \sim \mathcal{N}(۰, \sigma^۲ I), \quad x_i \in \mathbb{R}^L, \quad y \in \mathbb{R}^n$$

$$X = [x_۱, x_۲, \cdots, x_n]^T, \quad y = [y_۱, y_۲, \cdots, y_n]^T, \quad \beta = [\beta_۱, \beta_۲, \cdots, \beta_L]^T$$

همچنین فرض می کنیم که ویژگی ها مستقل از یکدیگرند (چرا که در غیر این صورت، رگرسیون خطی پاسخ نخواهد داشت).

- **رگرسیون خطی:**

$$\begin{aligned}
\mathcal{L}_{LS} &= \|X\hat{\beta}_{LS} - y\|_۲^۲ \\
&= \left(X\hat{\beta}_{LS} - y\right)^T (X\hat{\beta}_{LS} - y) \\
&= \hat{\beta}_{LS}^T X^T X \hat{\beta}_{LS} - \hat{\beta}_{LS}^T X^T y - y^T X \hat{\beta}_{LS} + y^T y \\
&= \hat{\beta}_{LS}^T X^T X \hat{\beta}_{LS} - ۲\hat{\beta}_{LS}^T X^T y + y^T y
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_{LS}}{\partial \hat{\beta}_{LS}} = ۰ &\Longrightarrow ۲X^T X \hat{\beta}_{LS} - ۲X^T y = ۰ \\
&\Longrightarrow X^T X \hat{\beta}_{LS} = X^T y \\
&\Longrightarrow \hat{\beta}_{LS} = (X^T X)^{-۱} X^T y
\end{aligned}$$

- **رگرسیون Ridge:**

$$\begin{aligned}
\mathcal{L}_{Ridge} &= \|X\hat{\beta}_{Ridge} - y\|_۲^۲ + \lambda \|\hat{\beta}_{Ridge}\|_۲^۲ \\
&= \left(X\hat{\beta}_{Ridge} - y\right)^T (X\hat{\beta}_{Ridge} - y) + \lambda \hat{\beta}_{Ridge}^T \hat{\beta}_{Ridge} \\
&= \hat{\beta}_{Ridge}^T X^T X \hat{\beta}_{Ridge} - \hat{\beta}_{Ridge}^T X^T y - y^T X \hat{\beta}_{Ridge} + y^T y + \lambda \hat{\beta}_{Ridge}^T \hat{\beta}_{Ridge} \\
&= \hat{\beta}_{Ridge}^T X^T X \hat{\beta}_{Ridge} - ۲\hat{\beta}_{Ridge}^T X^T y + y^T y + ۲\lambda \hat{\beta}_{Ridge}^T \hat{\beta}_{Ridge}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_{Ridge}}{\partial \hat{\beta}_{Ridge}} = ۰ &\Longrightarrow ۲X^T X \hat{\beta}_{Ridge} - ۲X^T y + ۲\lambda \hat{\beta}_{Ridge} = ۰ \\
&\Longrightarrow X^T X \hat{\beta}_{Ridge} + \lambda \hat{\beta}_{Ridge} = X^T y \\
&\Longrightarrow (XX^T + \lambda I)\hat{\beta}_{Ridge} = X^T y \\
&\Longrightarrow \hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-۱} X^T y
\end{aligned}$$

در ادامه، واریانس هر کدام از این ضرایب را محاسبه خواهیم کرد.

• **رگرسیون خطی:**

$$\begin{aligned}\hat{\beta}_{LS} &= (X^TX)^{-1}X^Ty\\ &= (X^TX)^{-1}X^T(X\beta+e)\\ &= (X^TX)^{-1}(X^TX)\beta+(X^TX)^{-1}X^Te\\ &= \beta+(X^TX)^{-1}X^Te\end{aligned}$$

$$\begin{aligned}\Longrightarrow \mathrm{Var}(\hat{\beta}_{LS}) &= \mathrm{Var}\left(\beta+(X^TX)^{-1}X^Te\right)\\ &= \mathrm{Var}\left((X^TX)^{-1}X^Te\right)\\ &= (X^TX)^{-1}X^T\,\mathrm{Var}(e)\left[(X^TX)^{-1}X^T\right]^T\\ &= (X^TX)^{-1}X^T\,\mathrm{Var}(e)X(X^TX)^{-1}\\ &= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1}\\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

• **رگرسیون Ridge:** در ابتدا دقت کنید که می توان نوشت:

$$\begin{aligned}\hat{\beta}_{Ridge} &= (X^TX+\lambda I)^{-1}X^Ty\\ &= (X^TX+\lambda I)^{-1}(X^TX)(X^TX)^{-1}X^Ty\\ &= (X^TX+\lambda I)^{-1}X^TX\hat{\beta}_{LS}\end{aligned}$$

در نتیجه خواهیم داشت:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{Ridge}) &= \mathrm{Var}\left((X^TX+\lambda I)^{-1}X^TX\hat{\beta}_{LS}\right)\\ &= (X^TX+\lambda I)^{-1}X^TX\,\mathrm{Var}(\hat{\beta}_{LS})\left[(X^TX+\lambda I)^{-1}X^TX\right]^T\\ &= (X^TX+\lambda I)^{-1}X^TX\,\mathrm{Var}(\hat{\beta}_{LS})X^TX(X^TX+\lambda I)^{-1}\\ &= (X^TX+\lambda I)^{-1}X^TX\sigma^2(X^TX)^{-1}X^TX(X^TX+\lambda I)^{-1}\\ &= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\end{aligned}$$

حال نشان می‌دهیم که $\mathrm{Var}(\hat{\beta}_{LS})-\mathrm{Var}(\hat{\beta}_{Ridge})$ یک ماتریس مثبت معین می‌باشد.

در ابتدا تعریف می‌کنیم

$$W = X^TX(X^TX+\lambda I)^{-1}$$

بنابراین به‌دست می‌آید:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{Ridge}) &= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\\ &= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX)^{-1}X^TX(X^TX+\lambda I)^{-1}\\ &= \sigma^2W^T(XX^T)^{-1}W\end{aligned}$$

حال می‌توان نوشت:

$$\begin{aligned}
&\mathrm{Var}(\hat{\beta}_{LS}) - \mathrm{Var}(\hat{\beta}_{Ridge})\\
&= \sigma^۲ (X^TX)^{-۱} - \sigma^۲ W^T (X^TX)^{-۱} W\\
&= \sigma^۲ \left\{ W^T (W^T)^{-۱} (X^TX)^{-۱} W^{-۱} W - W^T (X^TX)^{-۱} W \right\}\\
&= \sigma^۲ W^T \left\{ (W^T)^{-۱} (X^TX)^{-۱} W^{-۱} - (X^TX)^{-۱} \right\} W\\
&= \sigma^۲ W^T \left\{ (X^TX)^{-۱} (X^TX+\lambda I)(X^TX)^{-۱}(X^TX+\lambda I)(X^TX)^{-۱} - (X^TX)^{-۱} \right\} W\\
&= \sigma^۲ W^T \left\{ (I+\lambda (X^TX)^{-۱})(X^TX)^{-۱}(I+\lambda(X^TX)^{-۱}) - (X^TX)^{-۱} \right\} W\\
&= \sigma^۲ W^T \left\{ \left((X^TX)^{-۱} + \lambda (X^TX)^{-۲} \right) (I+\lambda(X^TX)^{-۱}) - (X^TX)^{-۱} \right\} W\\
&= \sigma^۲ W^T \left\{ (X^TX)^{-۱} + \lambda (X^TX)^{-۲} + \lambda (X^TX)^{-۲} + \lambda^۲ (X^TX)^{-۳} - (X^TX)^{-۱} \right\} W\\
&= \sigma^۲ W^T \left\{ ۲\lambda (X^TX)^{-۲} + \lambda^۲ (X^TX)^{-۳} \right\} W\\
&= \sigma^۲ (X^TX+\lambda I)^{-۱} X^TX \left\{ ۲\lambda (X^TX)^{-۲} + \lambda^۲ (X^TX)^{-۳} \right\} X^TX (X^TX+\lambda I)^{-۱}\\
&= \sigma^۲ (X^TX+\lambda I)^{-۱} \left\{ ۲\lambda I + \lambda^۲ (X^TX)^{-۱} \right\} (X^TX+\lambda I)^{-۱}
\end{aligned}$$

دقت کنید که می دانیم:

$$\forall v \neq ۰ :\ u = (X^TX+\lambda I)^{-۱} v \neq ۰$$

و در نتیجه خواهیم داشت:

$$\begin{aligned}
v^T \left[\mathrm{Var}(\hat{\beta}_{LS}) - \mathrm{Var}(\hat{\beta}_{Ridge}) \right] v &= \sigma^۲ u^T \left\{ ۲\lambda I + \lambda^۲ (X^TX)^{-۱} \right\} u\\
&= \sigma^۲ \lambda u^T u + \sigma^۲ \lambda^۲ u^T (X^TX)^{-۱} u > ۰
\end{aligned}$$

بنابراین $\mathrm{Var}(\hat{\beta}_{LS}) - \mathrm{Var}(\hat{\beta}_{Ridge})$ یک ماتریس مثبت معین می‌باشد و حکم درست است.

دقت کنید که $\mathrm{trace}\{\mathrm{Var}(\hat{\beta}_{LS}) - \mathrm{Var}(\hat{\beta}_{Ridge}) = \sum_i \lambda_i \left(\mathrm{Var}(\hat{\beta}_{LS}) - \mathrm{Var}\,\hat{\beta}_{Ridge} \right\}$، که برابر مجموع اختلاف واریانس ها می‌باشد، نیز مثبت است؛ چرا که بخاطر مثبت معین بودن این ماتریس تمامی مقادیر ویژه آن مثبت‌اند.

ب) در ابتدا دقت کنید که داریم:

$$\hat{Y}(\lambda) = X\hat{\beta}_{Ridge}$$

بنابراین با توجه به واریانس به‌دست آمده برای $\hat{\beta}_{Ridge}$ در قسمت قبل، می‌توان نوشت:

$$\begin{aligned}
\mathrm{Var}\left[\hat{Y}(\lambda) \right] &= \mathrm{Var}(X\hat{\beta}_{Ridge})\\
&= X\, \mathrm{Var}(\hat{\beta}_{Ridge}) X^T\\
&= \sigma^۲ X (X^TX+\lambda I)^{-۱} X^TX (X^TX+\lambda I)^{-۱} X^T
\end{aligned}$$

حال تجزیه SVD ماتریس X را به شکل زیر در نظر بگیرید:

$$X = U\Sigma V^T$$

به‌طوری که U و V ماتریس هایی متعامدند (به عبارت دیگر تساوی های $U^T = U^{-1}$ و $V^T = V^{-1}$ برقرار هستند) و Σ یک ماتریس قطری است که درایه‌های واقع بر روی قطر اصلی در آن، مقادیر تکین X می‌باشند.

بنابراین داریم:

$$X^TX = \left(U\Sigma V^T\right)^T\left(U\Sigma V^T\right) = V\Sigma(U^TU)\Sigma V^T = V\Sigma^2V^T$$

در ادامه دقت کنید که با توجه به خاصیت تغییرناپذیری trace با شیفت چرخشی در ضرب ماتریس‌ها، خواهیم داشت:

$$\begin{aligned}
\operatorname{trace}\left\{\operatorname{Var}\left[\hat{Y}(\lambda)\right]\right\} &= \operatorname{trace}\left\{\sigma^2X(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}X^T\right\}\\
&= \sigma^2\operatorname{trace}\left\{X(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}X^T\right\}\\
&= \sigma^2\operatorname{trace}\left\{X^TX(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1}\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(X^TX(X^TX+\lambda I)^{-1}\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(V\Sigma^2V^T(V\Sigma^2V^T+\lambda I)^{-1}\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(V\Sigma^2V^T(V\Sigma^2V^T+\lambda VV^T)^{-1}\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(V\Sigma^2V^T\left(V(\Sigma^2+\lambda I)V^T\right)^{-1}\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(V\Sigma^2(V^TV)(\Sigma^2+\lambda I)^{-1}V^T\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{\left(V\Sigma^2(\Sigma^2+\lambda I)^{-1}V^T\right)^2\right\}\\
&= \sigma^2\operatorname{trace}\left\{V\Sigma^2(\Sigma^2+\lambda I)^{-1}(V^TV)\Sigma^2(\Sigma^2+\lambda I)^{-1}V^T\right\}\\
&= \sigma^2\operatorname{trace}\left\{V\Sigma^2(\Sigma^2+\lambda I)^{-1}\Sigma^2(\Sigma^2+\lambda I)^{-1}V^T\right\}\\
&= \sigma^2\operatorname{trace}\left\{V^TV\Sigma^2(\Sigma^2+\lambda I)^{-1}\Sigma^2(\Sigma^2+\lambda I)^{-1}\right\}\\
&= \sigma^2\operatorname{trace}\left\{\Sigma^2(\Sigma^2+\lambda I)^{-1}\Sigma^2(\Sigma^2+\lambda I)^{-1}\right\}\\
&= \sigma^2\operatorname{trace}\left\{\Sigma^4(\Sigma^2+\lambda I)^{-2}\right\} \qquad (*)\\
&= \sigma^2\operatorname{trace}\left\{\begin{bmatrix}\sigma_1(X)^4 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \sigma_p(X)^4\end{bmatrix}\begin{bmatrix}\sigma_1(X)^2+\lambda & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \sigma_p(X)^2+\lambda\end{bmatrix}^{-2}\right\}\\
&= \sigma^2\operatorname{trace}\left\{\begin{bmatrix}\sigma_1(X)^4 & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \sigma_p(X)^4\end{bmatrix}\begin{bmatrix}\frac{1}{(\sigma_1(X)^2+\lambda)^2} & \dots & 0\\ \vdots & \ddots & \vdots\\ 0 & \dots & \frac{1}{(\sigma_p(X)^2+\lambda)^2}\end{bmatrix}\right\}\\
&= \sigma^2\sum_{i=1}^{p}\sigma_i(X)^4[\sigma_i(X)^2+\lambda]^{-2}
\end{aligned}$$

که این همان حکم مسئله می‌باشد (البته دقت کنید که در صورت سوال، باید به‌جای مقادیر ویژه، مقادیر تکین قرار بگیرد).