یادگیری ماشین

نیم‌سال اول ۰۰-۰۱

مدرس: سید عباس حسینی

دانشکده مهندسی کامپیوتر

تمرین سوم

محمدجواد هزاره

شماره دانشجویی: ۹۸۱۰۱۰۷۴

سوال ۱

۱. **(c)**. زیاد بودن ضریب یک ویژگی نشان دهنده‌ی این است که با ثابت نگه داشتن سایر متغییرها و تغییر این متغییر، برچسب پیش‌بینی شده به مقدار زیادی تغییر می‌کند. در نگاه اول ممکن است این موضوع نشان دهنده‌ی این باشد که این ویژگی تاثیر زیادی در مدل دارد اما باید توجه کرد که این بزرگی ضریب می‌تواند ناشی از مقیاس آن ویژگی باشد. اگر مقیاس این ویژگی هماهنگ با سایر ویژگی‌ها نباشد، ضریب آن بسیار بزرگ یا بسیار کوچک خواهد شد. بنابراین بزرگ بودن ضریب لزوما به معنای مهم یا کم اهمیت بودن آن ویژگی نیست.

۲. آ) **درست**. با زیاد شدن داده‌ها مدل از حالت overfitting خارج می‌شود و پیچیدگی کنونی مدل نمی‌تواند تمام اطلاعاتی که داده‌ها در اختیارمان قرار می‌دهد را مدل کند. بنابراین زیاد کردن تعداد داده‌های آموزش، نه تنها باعث کاهش bias نمی‌شود، بلکه آن را افزایش می‌دهد.

ب) **غلط**. با کاهش خطا روی داده‌های آموزش، مدل روی این داده‌ها فیت می‌شود و قدرت تعمیم خود را از دست خواهد داد. بنابراین لزوما کاهش خطا روی داده‌های آموزش باعث کاهش خطا روی داده‌های تست نخواهد شد.

ج) **غلط**. افزایش پیچیدگی مدل، باعث کاهش خطای آموزش می‌شود اما همواره باعث کاهش خطای تست نخواهد شد. در ابتدا که پیچیدگی مدل کم است، خطای آموزش و تست هر دو زیاد است یا به عبارتی مدل دچار underfitting شده است. با افزایش پیچیدگی مدل، خطای آموزش به مرور کم می‌شود اما خطای تست تا جایی کم شده و پس از آن شروع به افزایش می‌کند؛ چرا که مدل به سمت فیت شدن روی داده‌های آموزش می‌رود و قدرت تعمیم خود را از دست خواهد داد.

د) **غلط**. می‌دانیم با افزایش پیچیدگی مدل، خطای آموزش کم می‌شود. از آن جایی که با مدل چندجمله‌ای درجه ۶ هنوز خطای آموزش به مقدار قابل توجهی زیاد است، کاهش دادن پیچیدگی مدل و استفاده از مدل خطی، نه تنها خطای آموزش را کاهش نداده بلکه باعث افزایش خطای آموزش نیز خواهد شد. برای کاهش خطای آموزش باید پیچیدگی مدل را افزایش داده یا به عبارتی از چندجمله‌ای‌های با درجه‌ی بالاتر استفاده کرد.

سوال ۲

آ) اگر از SSE برای هزینه استفاده کنیم، آن‌گاه هزینه برحسب $\boldsymbol{\omega}$ برابر خواهد بود با:

$$\begin{aligned} SSE(\boldsymbol{\omega}) &= \|Y - \hat{Y}\|^2 \\ &= \|Y - X\boldsymbol{\omega}\|^2 \\ &= (Y - X\boldsymbol{\omega})^T(Y - X\boldsymbol{\omega}) \\ &= Y^TY - 2Y^TX\boldsymbol{\omega} + \boldsymbol{\omega}^T(X^TX)\boldsymbol{\omega} \end{aligned}$$

برای پیدا کردن $\boldsymbol{\omega}$ بهینه کافی‌ست مشتق عبارت بالا را در $\boldsymbol{\omega}^*_{opt}$ برابر صفر قرار دهیم:

$$\begin{aligned} &\frac{d}{d\boldsymbol{\omega}} SSE(\boldsymbol{\omega}^*_{opt}) = -2(Y^TX)^T + 2(X^TX)\boldsymbol{\omega}^*_{\boldsymbol{opt}} = 0 \\ &\implies (X^TX)\boldsymbol{\omega}^*_{opt} = X^TY \\ &\implies \boxed{\boldsymbol{\omega}^*_{opt} = (X^TX)^{-1}X^TY} \end{aligned}$$

ب) اگر جمله‌ی منظم ساز L2 را به تابع هزینه اضافه کنیم، آن‌گاه:

$$E(\boldsymbol{\omega}) = SSE(\boldsymbol{\omega}) + \lambda\boldsymbol{\omega}^T\boldsymbol{\omega}$$

با صفر قرار دادن مشتق رابطه‌ی بالا به ازای $\boldsymbol{\omega}^*$ خواهیم داشت:

$$\begin{aligned} &\frac{d}{d\boldsymbol{\omega}} E(\boldsymbol{\omega}^*) = -2(Y^TX)^T + 2(X^TX)\boldsymbol{\omega}^* + 2\lambda\boldsymbol{\omega}^* = 0 \\ &\implies -X^TY + (X^TX)\boldsymbol{\omega}^* + \lambda\boldsymbol{\omega}^* = 0 \\ &\implies (X^TX + \lambda\boldsymbol{I})\boldsymbol{\omega}^* = X^TY \\ &\implies \boxed{\boldsymbol{\omega}^* = (X^TX + \lambda\boldsymbol{I})^{-1}X^TY} \end{aligned}$$

ج) **اگر**: برای این قسمت می‌دانیم رابطه $\Sigma X = XF$ را داریم. هم‌چنین از آنجایی که Σ ماتریس کوواریانس است،

ماتریسی متقارن خواهد بود. پس برای $\boldsymbol{\omega}^*_{new}$ خواهیم داشت:

$$
\begin{aligned}
\boldsymbol{\omega}^*_{new} &= \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}Y \\
&= \left(X^TXF^{-1}\right)^{-1}X^T\Sigma^{-1}Y && (\Sigma^{-1}X = XF^{-1}) \\
&= \left(X^TXF^{-1}\right)^{-1}(F^T)^{-1}X^TY && (X^T\Sigma^{-1} = (F^T)^{-1}X^T) \\
&= \left(F^T(X^TX)F^{-1}\right)^{-1}X^TY && \left((AB)^{-1} = B^{-1}A^{-1}\right) \\
&= \left((F^TX^T)(XF^{-1})\right)^{-1}X^TY \\
&= \left(X^T\Sigma\Sigma^{-1}X\right)^{-1}X^TY \\
&= \left(X^TX\right)^{-1}X^TY \\
&= \boldsymbol{\omega}^*_{opt}
\end{aligned}
$$

فقط اگر: برای این قسمت می‌دانیم $\boldsymbol{\omega}^*_{new} = \boldsymbol{\omega}^*_{opt}$ ، بنابراین خواهیم داشت:

$$
\begin{aligned}
&\boldsymbol{\omega}^*_{new} = \boldsymbol{\omega}^*_{opt} \\
&\quad\Longrightarrow\ \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}Y = \left(X^TX\right)^{-1}X^TY \\
&\quad\Longrightarrow\ \left(X^T\Sigma^{-1}XF\right)^{-1}X^T\Sigma^{-1} = \left(X^TX\right)^{-1}X^T \\
&\quad\Longrightarrow\ X^T = \left(X^T\Sigma^{-1}X\right)\left(X^TX\right)^{-1}X^T\Sigma \\
&\quad\Longrightarrow\ X = \Sigma X\left(X^TX\right)^{-1}\left(X^T\Sigma^{-1}X\right) \\
&\quad\Longrightarrow\ X\left(X^T\Sigma^{-1}X\right)^{-1}\left(X^TX\right) = \Sigma X
\end{aligned}
$$

بنابراین اگر تعریف کنیم $F = \left(X^T\Sigma^{-1}X\right)^{-1}\left(X^TX\right)$ ، خواسته‌ی مسئله اثبات می‌شود. با توجه به وارون‌پذیر بودن X^TX، ماتریس F نیز وارون‌پذیر خواهد بود و داریم $\Sigma X = XF$.

د) اگر داده‌های جدید $\widetilde{X}$ با برچسب $\widetilde{Y}$ را به داده‌های قبلی اضافه کنیم، آن‌گاه ماتریس برچسب‌ها و ماتریس داده‌ها به صورت زیر خواهد بود:

$$
\begin{cases}
\bar{Y} = \begin{bmatrix} Y \\ \widetilde{Y} \end{bmatrix} \\
\bar{X} = \begin{bmatrix} X \\ \widetilde{X} \end{bmatrix}
\end{cases}
$$

بنابراین برای خطای L_1 داده‌های جدید داریم:

$$L(\omega, \lambda) = \|\bar{Y} - \bar{X}\omega\|^2 + \lambda\|\omega\|_1$$

اگر بخواهیم این تابع با تابع خطای داده شده برابر باشد، ضریب λ تابع بالا را همان ضریب خطای L_1 در تابع داده شده در نظر گرفته و داریم:

$$\begin{aligned}
\|\bar{Y} - \bar{X}\omega\|^2 + \lambda_2\|\omega\|_1 = &\quad \|Y - X\omega\|^2 + \lambda_1\|\omega\|_2^2 + \lambda_2\|\omega\|_1 \\
\|Y - X\omega\|^2 + \|\widetilde{Y} - \widetilde{X}\omega\|^2 = &\quad \|Y - X\omega\|^2 + \lambda_1\|\omega\|_2^2 \\
\|\widetilde{Y} - \widetilde{X}\omega\|^2 = &\quad \lambda_1\|\omega\|_2^2 \\
\widetilde{Y}^T\widetilde{Y} - 2\widetilde{Y}^T\widetilde{X}\omega + \omega^T(\widetilde{X}^T\widetilde{X})\omega = &\quad \lambda_1\omega^T\omega
\end{aligned}$$

بنابراین بین داده‌های جدید و برچسب‌هایشان باید رابطه‌ی زیر برقرار باشد تا تابع خطاهای مذکور با یکدیگر برابر شوند:

$$\widetilde{Y}^T\left(\widetilde{Y} - 2\widetilde{X}\omega\right) = \omega^T\left(\lambda_1\boldsymbol{I} - \widetilde{X}^T\widetilde{X}\right)\omega$$

اگر برچسب داده‌های جدید یعنی $\widetilde{Y}$ را ماتریس صفر در نظر بگیریم، سمت چپ تساوی بالا صفر شده و اگر ماتریس داده‌های جدید یعنی $\widetilde{X}$ را برابر $\sqrt{\lambda_1}\boldsymbol{I}$ در نظر بگیریم، سمت راست نیز صفر خواهد شد. بنابراین در این صورت خطای L_1 داده‌های $\bar{X}$ و $\bar{Y}$ برابر با خطای داده شده روی X و Y خواهد بود.

سوال ۳

با توجه به واگرایی Kullback-Liebler برای توزیع‌های $p(\boldsymbol{x})$ و $q(\boldsymbol{x})$ که $\boldsymbol{x}$ یک بردار n بعدی است داریم:

$$KL(p||q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} = \int p(\boldsymbol{x})\big[\log(p(\boldsymbol{x})) - \log(q(\boldsymbol{x}))\big] d\boldsymbol{x}$$

با توجه به اینکه توزیع $q(\boldsymbol{x})$ یک توزیع نمایی با میانگین $\boldsymbol{\mu}$ و ماتریس کوواریانس Σ است داریم:

$$\begin{aligned}
\log(q(\boldsymbol{x})) &= \log\left((2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})
\end{aligned}$$

بنابراین برای واگرایی KL که تابعی از $\boldsymbol{\mu}$ و Σ خواهد بود داریم:

$$KL(p||q) = f(\boldsymbol{\mu}, \Sigma) = \int p(\boldsymbol{x}) \left(\log(p(\boldsymbol{x})) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(|\Sigma|) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) d\boldsymbol{x}$$

برای پیدا کردن کمینه‌ی این تابع باید گرادیان f را محاسبه کرده و برابر صفر قرار دهیم. برای محاسبه‌ی گرادیان، با توجه به‌این که انتگرال روی متغییر $\boldsymbol{x}$ گرفته می‌شود، می‌توان نخست از تابع زیر انتگرال مشتق گرفت و سپس انتگرال را محاسبه کرد. هم‌چنین مبنای لگاریتم را عدد نپر در نظر می‌گیریم و در نتیجه بجای $\log$ داریم $\ln$. بنابراین برای گرادیان f داریم:

$$\begin{cases} \dfrac{\partial}{\partial \boldsymbol{\mu}} f(\boldsymbol{\mu}, \Sigma) = \displaystyle\int p(\boldsymbol{x}) \left(\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) d\boldsymbol{x} \\ \dfrac{\partial}{\partial \Sigma} f(\boldsymbol{\mu}, \Sigma) = \displaystyle\int p(\boldsymbol{x}) \left(\frac{1}{2|\Sigma|}(|\Sigma|\Sigma^{-1}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}\Sigma^{-1} \right) d\boldsymbol{x} \end{cases}$$

که با صفر قرار دادن روابط بالا به ازای $\boldsymbol{\mu}^*$ و Σ^* داریم:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}} f(\boldsymbol{\mu}^*, \Sigma^*) = {\Sigma^*}^{-1} \int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu}^*)d\boldsymbol{x} = 0 \\ \implies \boldsymbol{\mu}^* \int p(\boldsymbol{x})d\boldsymbol{x} = \int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x} \\ \implies \boxed{\boldsymbol{\mu}^* = \mathbb{E}_p[\boldsymbol{x}]} \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \Sigma} f(\boldsymbol{\mu}^*, \Sigma^*) = \frac{1}{2}\left(\int p(\boldsymbol{x}) \left(\boldsymbol{I} - (\boldsymbol{x}-\boldsymbol{\mu}^*)(\boldsymbol{x}-\boldsymbol{\mu}^*)^T{\Sigma^*}^{-1} \right) d\boldsymbol{x} \right) {\Sigma^*}^{-1} = 0 \\ \implies \boldsymbol{I} \int p(\boldsymbol{x})d\boldsymbol{x} = \left(\int (\boldsymbol{x}-\boldsymbol{\mu}^*)(\boldsymbol{x}-\boldsymbol{\mu}^*)^T p(\boldsymbol{x})d\boldsymbol{x} \right) {\Sigma^*}^{-1} \\ \implies \boldsymbol{I} = \mathrm{Var}_p(\boldsymbol{x}){\Sigma^*}^{-1} \\ \implies \boxed{\Sigma^* = \mathrm{Var}_p(\boldsymbol{x})} \end{aligned}$$

بنابراین بهترین تخمینی که از توزیع $p(\boldsymbol{x})$ با استفاده از توزیع‌های نرمال می‌توان داشت، توزیع نرمالی با میانگین $\boldsymbol{\mu}^*$ و ماتریس کوواریانس Σ^* است.

سوال ۴

آ) اگر برای مدل داشته باشیم $\hat{y}^{(i)} = w_j x_j^{(i)} + \epsilon$ ، و برای تابع هزینه از MSE استفاده کنیم، آنگاه داریم:

$$\begin{aligned} MSE(w_j) &= \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - w_j x_j^{(i)}\right)^2 \end{aligned}$$

که برای کمینه‌کردن خطا داریم:

$$\begin{aligned} \frac{d}{dw_j}MSE(w_j^*) = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - w_j^* x_j^{(i)}\right)x_j^{(i)} = 0 \\ \implies w_j^*\sum_{i=1}^{n} x_j^{(i)} x_j^{(i)} = \sum_{i=1}^{n} y^{(i)} x_j^{(i)} \\ \implies w_j^*(\boldsymbol{x_j^T x_j}) = \boldsymbol{x_j^T y} \\ \implies \boxed{w_j^* = \frac{\boldsymbol{x_j^T y}}{\boldsymbol{x_j^T x_j}}} \end{aligned}$$

ب) می‌دانیم جواب بهینه برای بردار $\boldsymbol{w}$ اگر همه‌ی ویژگی‌ها را لحاظ کنیم، از رابطه‌ی زیر بدست می‌آید:

$$\boldsymbol{w}^* = (X^T X)^{-1} X^T \boldsymbol{y}$$

حال اگر ستون‌های X را با بردارهای ستونی $\boldsymbol{x_i}$ نشان دهیم، آنگاه برای ماتریس $X^T X$ داریم:

$$(X^T X)_{ij} = \boldsymbol{x_i^T x_j}$$

که اگر بدانیم ستون‌های X متعامد هستند، آنگاه ماتریس $X^T X$ برابر خواهد بود با:

$$(X^T X)_{ij} = \begin{cases} 0 & i \neq j \\ \boldsymbol{x_i^T x_i} & i = j \end{cases}$$

حال برای حاصل‌ضرب وارون این ماتریس در ماتریس X^T خواهیم داشت:

$$X^{\dagger} = \begin{pmatrix} \frac{1}{\boldsymbol{x_1^T x_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\boldsymbol{x_2^T x_2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\boldsymbol{x_m^T x_m}} \end{pmatrix}_{m\times m} \times \begin{pmatrix} \boldsymbol{x_1^T} \\ \boldsymbol{x_2^T} \\ \vdots \\ \boldsymbol{x_m^T} \end{pmatrix}_{m\times n} = \begin{pmatrix} \frac{\boldsymbol{x_1^T}}{\boldsymbol{x_1^T x_1}} \\ \frac{\boldsymbol{x_2^T}}{\boldsymbol{x_2^T x_2}} \\ \vdots \\ \frac{\boldsymbol{x_m^T}}{\boldsymbol{x_m^T x_m}} \end{pmatrix}_{m\times n}$$

بنابراین برای $\boldsymbol{w}$ داریم:

$$\boldsymbol{w}^* = X^{\dagger}\boldsymbol{y} = \begin{pmatrix} \frac{\boldsymbol{x_1^T}}{\boldsymbol{x_1^T x_1}} \\ \frac{\boldsymbol{x_2^T}}{\boldsymbol{x_2^T x_2}} \\ \vdots \\ \frac{\boldsymbol{x_m^T}}{\boldsymbol{x_m^T x_m}} \end{pmatrix}_{m\times n} \times \boldsymbol{y}_{n\times 1} = \begin{pmatrix} \frac{\boldsymbol{x_1^T y}}{\boldsymbol{x_1^T x_1}} \\ \frac{\boldsymbol{x_2^T y}}{\boldsymbol{x_2^T x_2}} \\ \vdots \\ \frac{\boldsymbol{x_m^T y}}{\boldsymbol{x_m^T x_m}} \end{pmatrix}_{m\times 1}$$

بنابراین همانطور که دیده می‌شود، w_j^* برابر با حالتی است که فقط از ویژگی jام برای پیداکردن مقدار بهینه w_j استفاده کرده بودیم.

ج) اگر از یکی از ویژگی‌ها و بایاس استفاده کرده باشیم، مدل ما به صورت $\hat{y}^{(i)} = w_0 + w_j x_j^{(i)}$ خواهد بود. بنابراین برای MSE داریم:

$$MSE(w_0, w_j) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_j x_j^{(i)}\right)^2$$

برای پیدا کردن مقدار بهینه w_0 و w_j باید گرادیان MSE در نقاط بهینه صفر شود، بنابراین:

$$\begin{cases} \dfrac{\partial}{\partial w_0} MSE(w_0, w_j) = -\dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_j x_j^{(i)}\right) \\ \dfrac{\partial}{\partial w_j} MSE(w_0, w_j) = -\dfrac{2}{n}\displaystyle\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_j x_j^{(i)}\right) x_j^{(i)} \end{cases}$$

با صفر قراردادن روابط بالا برای w_0^* و w_j^* داریم:

$$\frac{\partial}{\partial w_0} MSE(w_0^*, w_j^*) = 0$$

$$\implies \frac{\sum_{i=1}^n y^{(i)}}{n} - w_0^* \frac{\sum_{i=1}^n 1}{n} - w_j^* \frac{\sum_{i=1}^n x_j^{(i)}}{n} = 0$$

$$\implies w_0^* = \mathbb{E}[y] - w_j^* \mathbb{E}[x_j] \quad (\star)$$

و با استفاده از مقدار بدست آمده برای w_0^* داریم:

$$\frac{\partial}{\partial w_j} MSE(w_0^*, w_j^*) = 0$$

$$\implies \frac{\sum_{i=1}^n y^{(i)} x_j^{(i)}}{n} - w_0^* \frac{\sum_{i=1}^n x_j^{(i)}}{n} - w_j^* \frac{\sum_{i=1}^n x_j^{(i)} x_j^{(i)}}{n} = 0$$

$$\implies \mathbb{E}[yx_j] - w_0^* \mathbb{E}[x_j] - w_j^* \mathbb{E}[x_j^2] = 0$$

$$\overset{(\star)}{\implies} \mathbb{E}[yx_j] - \mathbb{E}[y]\mathbb{E}[x_j] + w_j^* \mathbb{E}[x_j]\mathbb{E}[x_j] - w_j^* \mathbb{E}[x_j^2] = 0$$

$$\implies w_j^* = \frac{\mathbb{E}[yx_j] - \mathbb{E}[y]\mathbb{E}[x_j]}{\mathbb{E}[x_j^2] - \mathbb{E}[x_j]^2} = \frac{\mathrm{Cov}(y, x_j)}{\mathrm{Var}(x_j)}$$

بنابراین $\boxed{w_j^* = \frac{\mathrm{Cov}(y, x_j)}{\mathrm{Var}(x_j)}}$ و $\boxed{w_0^* = \mathbb{E}[y] - w_j^* \mathbb{E}[x_j]}$.