تمرین چهارم یادگیری ماشین

آرمین سعادت
۹۴۶۵۸۲۹

سوال ۱)

الف: اگر a یک بردار باشد داریم:

$$\|a\|^2 = a^T a$$

$$\Rightarrow \|y - Xw\|^2 = (y - Xw)^T (y - Xw) = (y^T - w^T X^T)(y - Xw)$$

$$= y^T y - y^T X w - w^T X^T y + w^T X^T X w$$

$$\frac{\partial \|y - Xw\|^2}{\partial w} = -y^T X - (X^T y)^T + (X^T X w)^T + w^T X^T X = -2y^T X + 2w^T X^T X = 0$$

$$\Rightarrow 2w^T X^T X = 2y^T X \Rightarrow X^T X w = X^T y \Rightarrow w = (X^T X)^{-1} X^T y$$

ب:

$$\text{Ridge Regression}: \|y - Xw\|^2 + \lambda \|w\|^2$$

$$= (y^T y - y^T X w - w^T X^T y + w^T X^T X w) + (\lambda w^T w)$$

$$\frac{\partial\, \text{Ridge}}{\partial w} = \underbrace{(-2y^T X + 2w^T X^T X)}_{\text{از قسمت الف}} + 2\lambda w^T = 0 \Rightarrow w^T (X^T X + \lambda I) = y^T X$$

$$\Rightarrow w^T = y^T X (X^T X + \lambda I)^{-1} \Rightarrow w = (X^T X + \lambda I)^{-1} X^T y$$

سوال ۲)

الف : مانند اضافه کردن m داده جدید به مجموعه داده‌ها است که m تعداد ویژگی‌ها است.

مجموعه داده قبلی : $X_{n\times m}$ $\Rightarrow$ مجموعه داده جدید : $X' = \begin{pmatrix} X_{n\times m} \\ \sqrt{\lambda}\, I_{m\times m} \end{pmatrix}$

مجموعه تارگت قبلی : Y_n $\Rightarrow$ مجموعه تارگت جدید : $Y' = \begin{pmatrix} Y_n \\ 0_m \end{pmatrix}$

$$\underbrace{\|Y'-X'w\|^2}_{\text{LSE with adding some data points}} = \|Y'-Xw-\sqrt{\lambda}\, I_{m\times m} w\|^2 = \underbrace{\|Y-Xw\|^2+\lambda\|w\|^2}_{L_2 \text{ with regularization}}$$

ب) اگر Correlation بین پارامترهای مرتبط و نامرتبط زیاد باشد، L1 خیلی خوب عمل نمی‌کند. همچنین اگر وابستگی بین پارامترهای ورودی و خروجی زیاد باشد مشکل به وجود می‌آید.

در حقیقت با L1، تشخیص اینکه کدام پارامترها موثر هستند سخت‌تر است. پس اگر حجم داده کم باشد یا همبستگی پارامترها زیاد باشد، تشخیص حتی سخت‌تر هم می‌شود.

ج) کاملاً مشابه قسمت الف عمل می‌کنیم یعنی $\sqrt{\lambda_1}\, I_{m\times m}$ را اضافه کردیم.

$$X' = \begin{pmatrix} X_{n\times m} \\ \sqrt{\lambda_1}\, I_{m\times m} \end{pmatrix}, \quad Y' = \begin{pmatrix} Y_n \\ 0_m \end{pmatrix}$$

$$\|Y'-X'w\| + \lambda_2\|w\|_1 = \|Y'-Xw-\sqrt{\lambda_1}\, Iw\|^2 + \lambda_2\|w\|_1$$

$$= \|Y-Xw\|^2 + \lambda_1\|w\|_2^2 + \lambda_2\|w\|_1$$

سوال ۳)

الف:

$$\Phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \Rightarrow y = w^T \Phi(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\Rightarrow P(y|x_1, x_2) \sim \mathcal{N}(w^T \Phi(x), \sigma^2)$$

$P(y|x_1,x_2)$ همان $P(y|x)$ است که $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

ب: با فرض استقلال هر داده از داده‌ی دیگر داریم:

$$P(Y|X_1, X_2) = \prod_{i=1}^{n} P(y^i | x_1^i, x_2^i) = \prod_{i=1}^{n} \mathcal{N}(w^T \Phi(x^i), \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y^i - w^T \Phi(x^i))^2}{2\sigma^2}}$$

$$\Rightarrow \log(P(Y|X_1, X_2)) = \sum_{i=1}^{n} \log e^{\frac{-(y^i - w^T \Phi(x^i))^2}{2\sigma^2}} + \log \frac{1}{\sqrt{2\pi}\sigma}$$

با حذف مقادیر Constant داریم:

$$\log(P(Y|X_1, X_2)) \propto \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y^i - w^T \Phi(x^i))^2$$

ج: با توجه به منفی در عبارت بالا، بیشینه شدن $P(Y|X_1,X_2)$ معادل با کمینه شدن حاصل جمع بالا است. پس داریم:

$$f(w_0, w_1, w_2, w_3, w_4) = \sum_{i=1}^{n} (y^i - w^T \Phi(x^i))^2$$

Scanned with CamScanner

د:

$$\nabla_w f(w) = \nabla_w \left(\sum_{i=1}^{n} \left(y^i - w^T\phi(x^i)\right)^2 \right) = -2 \sum_{i=1}^{n} \left(y^i - w^T\phi(x^i)\right) \phi(x^i)^T$$

هـ:

$$w^{i+1} = w^i - a\nabla_w f(w) = w^i + a\left(y_i - w^T\phi(x^i)\right)\phi(x^i)^T$$

که a ضریب یادگیری همان rate است.

سؤال ۴)

الف:

$$Y = Xw + \epsilon \Rightarrow X^TY = X^TXw + X^T\epsilon \Rightarrow X^TXw = X^TY - X^T\epsilon$$

$$\Rightarrow w = (X^TX)^{-1}X^TY - (X^TX)^{-1}X^T\epsilon = (X^TX)^{-1}X^T(Xw+\epsilon) - (X^TX)^{-1}X^T\epsilon$$

$$(X^TX)^{-1}X^TY = w^* \Rightarrow w = w^* - (X^TX)^{-1}X^T\epsilon$$

$$\epsilon \sim N(0, I) \Rightarrow w \sim \left(w^*, \|(X^TX)^{-1}X^T\|^2\right)$$

$$\|(X^TX)^{-1}X^T\|^2 = (X^TX)^{-1}X^T\left((X^TX)^{-1}X^T\right)^T = (X^TX)^{-1}X^TX(X^TX)^{-T}$$

$$= \underbrace{\left((X^TX)^{-1}(X^TX)\right)}_{I}\left((X^TX)^T\right)^{-1} = (X^TX)^{-1}$$

$$\Rightarrow \hat{w} \sim N\left(w^*, (X^TX)^{-1}\right)$$

ب)

$$Var(Y) = Var(X\hat{w} + \epsilon) = Var(X\hat{w}) + Var(\epsilon)$$

$$= X\,Var(\hat{w})X^T + Var(\epsilon) = X(X^TX)^{-1}X^T + I$$

نکته: $Var(a+b) = Var(a) + Var(b)$

نکته: $Var(Xw) = X\,Var(w)X^T$

ج:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \Rightarrow A_n = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \longrightarrow A_n = f(x, n)$$

پس A_n به صورت تابعی از n و x_i نوشته شد.

$$\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T = \sum_{i=1}^{n} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = A_n$$

$$\forall s: \quad s^T A_n s = \sum_{i=1}^{n} s_i^T \phi(x_i)\phi(x_i)^T s_i = \sum \left(s_i^T\phi(x_i)\right)\left(s_i^T\phi(x_i)\right)^T$$

$$= \sum_{i=1}^{n} \| s_i^T \phi(x_i) \|_2^2 \geq 0 \Rightarrow$$ نیمه معین مثبت

ج:

$$J(X) = E_{x\sim Q}\left(Var(y(x))\right) = E_{x\sim Q}\left[x^T (X^T X)^{-1} x\right]$$

$$X^T X = A_n = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \Rightarrow (X^T X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$\Rightarrow E\left[x^T (X^T X)^{-1} x\right] = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \times \left(\sum x_i^2 - 2E(x)\sum x_i + E[x^2]n\right)$$

$$E[x] = 0, \quad Var(x) = V^2$$

$$\Rightarrow J(X) = \frac{\sum x_i^2 + nV^2}{n\sum x_i^2 - (\sum x_i)^2}$$

سؤال ۵)

الف :

$$Y = f(x,w) + \epsilon \,,\quad \epsilon \sim N(0,\beta^{-1})$$

$$\Rightarrow P(y|x,w,\beta) \sim N(f(x,w), \beta^{-1})$$

با در نظر گرفتن استقلال بین n داده داریم :

$$P(Y|X,w,\beta) = \prod_{i=1}^{n} N(f(x_i,w),\beta^{-1})$$

$$= \prod_{i=1}^{n} \left(\frac{\sqrt{\beta}}{2\pi} e^{-\frac{(y_i - f(x_i,w))^2}{2\beta^{-2}}}\right) \xrightarrow{\log} \frac{-N}{2}\log(2\pi) + \frac{N}{2}\log(\beta) - \frac{\beta}{2}\sum_{i=1}^{n} (y_i - f(x_i,w))^2$$

پس بیشینه شدن $P(Y|X,w,\beta)$ معادل شد با کمینه شدن $\sum_{i=1}^{n} (y_i - f(x_i,w))^2$ که همان SSE است.

ب :

$$P(w|Y) \propto P(Y|w)\,P(w)$$

پس برای بیشینه کردن $P(w|Y)$، باید $P(Y|w)P(w)$ را بهینه کنیم که با log گرفتن از ۲ طرف، نتیجه می‌شود که $\log P(Y|w) + \log P(w)$ باید بهینه شود.

$$P(w) \sim N(0,\alpha^{-1}) = \frac{1}{\sqrt{2\pi^k \Sigma}} e^{-\frac{1}{2} w^T \alpha I w} \Rightarrow \log P(w) \propto -\sum_{i=1}^{k} w_i^2$$

از الف هم به دست آوردیم که $\log P(Y|w) \propto -\sum_{i=1}^{n} (y_i - f(x_i,w))^2$

پس بیشینه کردن $P(w|Y)$، معادل است با کمینه کردن عبارت زیر:

$$\frac{\beta}{2}\sum_{i=1}^{n} (y_i-f(x_i,w))^2 + \frac{\alpha}{2}\sum_{i=1}^{k} w_i^2 = \frac{\beta}{2} SSE + \frac{\alpha}{2}\|w\|_2^2 \rightarrow SSE + \text{Ridge\_Regularized}$$

ج:

$$P(w|Y) \propto P(Y|w)P(w)$$

$$P(w) \propto \frac{\alpha}{2} e^{-\alpha|w|} \Rightarrow \log P(w) \sim -\alpha|w|$$

از این جا به بعد مشابه قسمت ب عمل می کنیم.

$$MAP \sim -\frac{\beta}{2}\sum_{i=1}^{n}(y_i - f(x_i,w))^2 - \alpha\|w\|_1$$

پس برای بیشینه شدن $P(w|Y)$، باید عبارت زیر کمینه شود:

$$\frac{\beta}{2}\sum_{i=1}^{n}(y_i - f(x_i,w))^2 + \alpha\|w\|_1 \longrightarrow \text{SSE with Lasso regulolized}$$

سوال ۵)

الف:

$$P(w) \propto e^{-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)} = e^{-\frac{1}{2}w^T S_0^{-1} + m_0^T S_0^{-1} w \underbrace{- \frac{1}{2} m_0^T S_0^T m_0}_{\text{چون مستقل از w است، عملا ضریب می شود.}}}$$

$$\Rightarrow P(w) \propto \exp(J^T w - \tfrac{1}{2} w^T P w), \quad J = S_0^{-1} m_0, \; P = S_0^{-1}$$

$$P(w|D) \propto P(D|w)P(w) = \prod_{i=1}^{n} P(y_i|x_i,w) \times P(w)$$

$$= \left(\prod_{i=1}^{n} \exp\left(-\frac{\beta}{2}(y_i - w^T x_i)^2\right)\right) \times \exp\left(J^T w - \tfrac{1}{2} w^T P w\right)$$

قسمت های مستقل از w را در نظر نمی گیریم چون عملا ضریب می شوند.

$$\Rightarrow P(w|D) \propto \exp\left\{(J + \beta\sum y_i x_i)^T w - \tfrac{1}{2} w^T (P + \beta \sum x_i x_i^T) w\right\}$$

$$P_N = S_N^{-1} = P + \beta\sum x_i x_i^T = S_0^{-1} + \beta X^T X \Longrightarrow S_N^{-1} = S_0^{-1} + \beta X^T X$$ ☑

$$J_N = S_N^{-1} m_N = (J + \beta\sum y_i x_i)^T = S_0^{-1} m_0 + \beta X^T Y \Rightarrow m_N = S_N(S_0^{-1} m_0 + \beta X^T Y)$$ ☑

ب:

$$P(w) \sim N(0, \alpha^{-1} I) \Rightarrow M_0 = 0,\ S_0 = \alpha^{-1} I$$

$$P(Y|w) \sim N(w^T x, \beta^{-1})$$

طبق الف خواهیم داشت:

$$S_N^{-1} = S_0^{-1} + \beta X^T X = \alpha I + \beta X^T X$$

$$M_N = S_N(0 + \beta X^T Y) = S_N \times \beta X^T Y$$

$$\Rightarrow P(w|Y) \sim N(M_N, S_N),\quad S_N^{-1} = \alpha I + \beta X^T X,\quad M_N = \beta S_N X^T Y$$

ج: برای به دست آوردن S_N و M_N از Prior استفاده کردیم که M_0 و S_0 بود. خب همین کار را برای این قسمت هم انجام می دهیم. و چون Σ تقارنی است، پس می توانیم از هر x استفاده کنیم.

$$S_{N+1}^{-1} = S_N^{-1} + \beta x_{n+1} x_{n+1}^T$$

$$M_{N+1} = S_{N+1}(S_N^{-1} M_N + \beta x_{n+1} y_{n+1})$$

پس همان سؤال الف است با این تفاوت که در مرحلهٔ $N+1$ ، X همان x_{n+1}^T می شود، Y همان y_{n+1} می شود.
و به جای S_0 و M_0 از S_N و M_N استفاده می کنیم چون Prior به روز رسانی می شود.
خب برای ادامه از رابطهٔ به دست آمده استفاده می کنیم:

$$S_{N+1}^{-1} = S_0^{-1} + \beta \sum_{i=1}^{n+1} x_i x_i^T$$

$$M_{N+1} = S_{N+1}(S_N^{-1} M_N + \beta x_{n+1} y_{n+1}) = S_{N+1}(S_n^{-1}(S_n(S_{n-1}^{-1} M_{n-1} + \beta x_n y_n) + \beta x_{n+1} y_{n+1}$$

$$\Rightarrow M_{N+1} = S_{N+1}\left(S_0^{-1} M_0 + \beta \sum_{i=1}^{n+1} x_i y_i\right)$$

پس مرحله به مرحله این چنین می شود.

از آن جا که $X^T Y = \sum x_i y_i$ و $X^T X = \sum x_i x_i^T$ ، پس جواب معادل این است که همه داده ها یکجا تحلیل شود.

د :

$$S_{N+1}^{-1} = S_N^{-1} + \beta x_{n+1} x_{n+1}^T$$

$$\Rightarrow S_{N+1} = \left(S_N^{-1} + \beta x_{n+1} x_{n+1}^T\right)^{-1} = \left(S_N^{-1} + \sqrt{\beta}\, x_{n+1} \sqrt{\beta}\, x_{n+1}^T\right)^{-1}$$

$$= S_N - \frac{S_N (\sqrt{\beta}\, x_{n+1}) (\sqrt{\beta}\, x_{n+1}^T) S_N}{1 + (\sqrt{\beta}\, x_{n+1}^T) S_N (\sqrt{\beta}\, X_{n+1})} = S_N - \frac{\beta S_N x_{n+1} x_{n+1}^T S_N}{1 + \beta x_{n+1}^T S_N x_{n+1}}$$

$$\Rightarrow S_N - S_{N+1} = \frac{\beta S_N x_{n+1} x_{n+1}^T S_N}{1 + \beta x_{n+1}^T S_N x_{n+1}}$$

$$x^T (S_N - S_{N+1}) x = \frac{x^T S_N x_{n+1} x_{n+1}^T S_N x}{\frac{1}{\beta} + x_{n+1}^T S_N x_{n+1}} = \frac{(x^T S_N x_{n+1})^2}{\frac{1}{\beta} + x_{n+1}^T S_N x_{n+1}}$$

چون S_N مثبت معین است، پس مخرج همیشه مثبت است. صورت هم که مربع کامل شد.

پس داریم:

$$x^T (S_N - S_{N+1}) x \geqslant 0 \Rightarrow \sigma_N^2 - \sigma_{N+1}^2 \geqslant 0$$

$$\Rightarrow \sigma_{N+1}^2 - \sigma_N^2 \leq 0 \Rightarrow$$ واریانس کاهش می‌یابد.