تمرین تئوری ۱ یادگیری ماشین

نویسنده:

سید احسان حسن بیگی - ۴۰۲۲۱۱۷۲۳

بخش۱

پرسش ۱

در الگوریتم یادگیری پرسپترون برای $\underbrace{\text{iteration}}_{\text{آپدیت وزن}}$ t اُم که به ازای sample i اُم رخ دهد، داریم:

$$\boxed{\begin{array}{l} w^t = w^{t-1} + y^{(i)} x^{(i)} \\ t \geq 1 \end{array}}$$ update rule

$$\boxed{w^0 = 0} \longrightarrow$$ بردار وزن‌ها را با تمام صفر initialize می‌کنیم

$$\xrightarrow[w^* \text{ در update rule}]{\text{ضرب داخلی طرفین}} w^t \cdot w^* = w^{t-1} \cdot w^* + \overbrace{y^{(i)} x^{(i)} \cdot w^*}^{y^{(i)} w^{*T} x^{(i)} \geq \gamma} \longrightarrow w^t \cdot w^* \geq w^{t-1} \cdot w^* + \gamma$$

طبق اثبات (I)

$$\circledast \; \boxed{\|w^t\| \geq t\gamma} \longleftarrow \begin{cases} w^t \cdot w^* \geq t\gamma \\ \|w^t\| \underbrace{\|w^*\|}_{1} \geq w^t \cdot w^* \quad \text{می‌دانیم} \end{cases}$$

$$\xrightarrow[\text{update rule}]{\text{گرفتن } norm^2 \text{ طرفین}} \|w^t\|^2 = \|w^{t-1} + y^{(i)} x^{(i)}\|^2 = \|w^{t-1}\|^2 + \overbrace{2 y^{(i)} \underbrace{x^{(i)} \cdot w^{t-1}}_{\hat{y}^{(i)}}}^{\ominus} + \overset{1}{\cancel{y^{(i)^2}}} \underbrace{\|x^{(i)}\|^2}_{\leq R^2}$$

$y^{(i)}$, $\hat{y}^{(i)}$ مختلف العلامت هستند زیرا باعث آپدیت وزن شده است

$$\longrightarrow \|w^t\|^2 \leq \|w^{t-1}\|^2 + R^2$$

$$\xrightarrow[\text{(II)}]{\text{طبق اثبات}} \boxed{\|w^t\|^2 \leq t R^2} \; \circledast\circledast$$

$$\xrightarrow{\circledast, \circledast\circledast} t^2\gamma^2 \leq \|w^t\|^2 \leq t R^2 \xrightarrow{t \geq 1} \boxed{t \leq \left(\frac{R}{\gamma}\right)^2}$$

اثبات I

با استقرا ثابت می‌کنیم که اگر $w^{t}.w^{*} \geqslant w^{t-1}.w^{*} + \gamma$ آنگاه $w^{t}.w^{*} \geqslant t\gamma$

$t \geqslant 1$

پایه $\xrightarrow{t=1} w^{1}.w^{*} \geqslant \underbrace{\underbrace{w^{0}}_{0}.w^{*}}_{0} + \gamma \longrightarrow w^{1}.w^{*} \geqslant \gamma$ ✓

$t \to t+1$

$$\left.\begin{array}{l} w^{t}.w^{*} \geqslant t\gamma \\ \left(y^{(i+k)}x^{(i+k)}\right).w^{*} \geqslant \gamma \end{array}\right\} \xrightarrow{+} \underbrace{\left(w^{t} + y^{(i+k)}x^{(i+k)}\right)}_{w^{t+1}}.w^{*} \geqslant (t+1)\gamma \quad ✓$$

اثبات II

با استقرا ثابت می‌کنیم که اگر $\|w^{t}\|^{2} \leqslant \|w^{t-1}\|^{2} + R^{2}$ آنگاه $\|w^{t}\|^{2} \leqslant tR^{2}$

$t \geqslant 1$

پایه $\xrightarrow{t=1} \|w^{1}\|^{2} \leqslant \underbrace{\|w^{0}\|^{2}}_{0} + R^{2} \longrightarrow \|w^{1}\|^{2} \leqslant R^{2}$ ✓

$t \to t+1$

$$\left.\begin{array}{l} \|w^{t}\|^{2} \leqslant tR^{2} \\ \|y^{(i+k)}x^{(i+k)}\|^{2} \leqslant R^{2} \end{array}\right\} \xrightarrow{+} \underbrace{\|w^{t} + y^{(i+k)}x^{(i+k)}\|^{2}}_{\|w^{t+1}\|^{2}} \leqslant \|w^{t}\|^{2} + \|y^{(i+k)}x^{(i+k)}\|^{2} \leqslant (t+1)R^{2} \quad ✓$$

پرسش ۲

نرم ماتریسی

$$\boxed{1} \quad J(w) = \sum_{i=1}^{n} \left(y^{(i)} - w^T x^{(i)}\right)^2 = \| Xw - y \|^2$$

$$X = \begin{bmatrix} - & x^{(1)^T} & - \\ & \vdots & \\ - & x^{(n)^T} & - \end{bmatrix}_{n \times (d+1)} \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}_{n \times 1}$$

$$\xrightarrow[J(w)]{\text{minimizing}} \nabla_w J(w) = 2X^T (Xw - y) = 0$$

$$\longrightarrow X^T X w = X^T y \longrightarrow \hat{w} = \underbrace{(X^T X)^{-1} X^T}_{X^{\dagger} \text{ pseudo inverse}} y$$

$\boxed{2}$

- برای استفاده از این فرمول باید وارون ماتریس $X^T X$ را حساب کنیم که یک ماتریس $(d+1) \times (d+1)$ بعدی است و بسته به تعداد فیچرها می‌تواند هزینه بر باشد $\xleftarrow{\text{راه حل}}$ استفاده از روش های iterative مانند gradient descent

- از نظر تئوری احتمال این وجود دارد که ماتریس $X^T X$ اصلاً وارون پذیر نباشد چون $\det(X^T X) = 0$ و البته اگر تعداد فیچرها از تعداد سمپل ها بیشتر باشد که قطعاً وارون پذیر نخواهد بود $\xleftarrow{\text{راه حل}}$ اگر از regularization استفاده کنیم عبارتی متناسب با λ در قطر اصلی ماتریس $X^T X$ اضافه می‌شود که کمک می کند $\det(X^T X) \neq 0$ و وارون پذیر شود

- ممکن است شرایط مسئله به صورتی باشد که تمام دیتا را در اول کار نداشته باشیم و stream سمپل ها در حین یادگیری به دست ما برسد (online learning) $\xleftarrow{\text{راه حل}}$ در این حالت نیز استفاده از روش های iterative مانند gradient descent به کار می آید

فرم ماتریسی

X دل مانند بخش $\boxed{1}$

$$F = \begin{bmatrix} F_1 & 0 & \cdots & 0 \\ 0 & F_2 & 0 \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & F_n \end{bmatrix}_{n \times n}$$

$\boxed{3}$

$$J(w) = \sum_{i=1}^{n} F_i \left(y^{(i)} - w^T x^{(i)}\right)^2 = \underbrace{\left(Xw - y\right)^T}\ \underbrace{F\left(Xw - y\right)}$$

$$\begin{bmatrix} x^{(1)} \cdot w - y^{(1)} & \cdots & x^{(n)} \cdot w - y^{(n)} \end{bmatrix}_{1 \times n} \begin{bmatrix} F_1\left(x^{(1)} \cdot w - y^{(1)}\right) \\ \vdots \\ F_n\left(x^{(n)} \cdot w - y^{(n)}\right) \end{bmatrix}_{n \times 1}$$

$$\nabla_w J(w) = \left(Xw - y\right)^T \underbrace{\left(F + F^T\right)}_{2F} \left(X\right) = 0$$

$$\xrightarrow{T} X^T F^T \left(Xw - y\right) = 0 \longrightarrow X^T F^T X w = X^T F^T y \longrightarrow \hat{w} = \left(X^T F^T X\right)^{-1} X^T F^T y$$

$$= \boxed{\left(X^T F X\right)^{-1} X^T F y}$$

$\boxed{4}$

$$E_{x,y}\left[\left(y - w^T x\right)^2\right] = E_{x,y}\left[y^2 - 2 w^T x y + \left(w^T x\right)^2\right]$$

$$= E_y\left[y^2\right] - 2 E_{x,y}\left[w^T x y\right] + E_x\left[\left(w^T x\right)\left(x^T w\right)\right]$$

$$= E_y\left[y^2\right] - 2 w^T \underbrace{E_{x,y}\left[x y\right]}_{C} + w^T \underbrace{E_x\left[x x^T\right]}_{R} w = E_y\left[y^2\right] - 2 w^T C + w^T R w$$

$\nabla_w E_{x,y}\left[\left(y - w^T x\right)^2\right] = R\hat{w} - C = 0 \longrightarrow$ به جوابی برای $\hat{w}$ نمی‌رسیم

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \big(y^{(i)} - \underbrace{\hat{y}}_{w^T x^{(i)}}\big)^2$$

$$w^* = \arg\min_{w} E_{x,y}\Big[(y - \underbrace{\hat{y}}_{w^T x})^2\Big] \xrightarrow{\partial/\partial w} \frac{\partial}{\partial w} E_{x,y}\Big[(y - w^{*T} x)^2\Big]$$

$$= E_{x,y}\Big[(y - w^{*T} x) \cdot x\Big] = 0 \quad ⊛$$

$\hat{w}$ وزن هایی است که باتوجه به دیتاست محدود مان به آن رسیده‌ایم

w^* وزن هایی است که اگر توزیع $\underbrace{p(Y|X)}_{\text{target distribution}}$ را داشته باشیم به آن خواهیم رسید

خطایی که به واسطه تقریب زدن خط، براساس دیتاست محدود داریم و اگر می دانستیم به ازای هر X چه لیبل هایی باید داشته باشیم (داشتن p(Y|X)) خط بهتری می کشیدیم

$$\text{approximation error} = E_{x,y}\Big[(y^* - \hat{y})^2\Big] = E_{x,y}\Big[(w^{*T} x - \hat{w}^T x)^2\Big]$$

$$\text{structural error} = E_{x,y}\Big[(y - y^*)^2\Big] = E_{x,y}\Big[(y - w^{*T} x)^2\Big]$$

خطایی که به واسطه ساختار انتخاب شده (خط) داریم و اگر ساختار بهینه ای پیدا می کردیم می توانستیم در پیش بینی به صفر برسیم

$$\text{error} = E_{x,y}\Big[(y - \hat{y})^2\Big] = E_{x,y}\Big[(y - \hat{w}^T x)^2\Big] = E_{x,y}\Big[(y - w^{*T} x + w^{*T} x - \hat{w}^T x)^2\Big]$$

$$= \underbrace{E_{x,y}\Big[(y - w^{*T} x)^2\Big]}_{\text{structural error}} + \underbrace{E_{x,y}\Big[(w^{*T} x - \hat{w}^T x)^2\Big]}_{\text{approxiamation error}} + 2E_{x,y}\Big[(y - w^{*T} x)\underbrace{(w^{*T} x - \hat{w}^T x)}_{\underbrace{(w^{*T} - \hat{w}^T)x}_{x^T(w^* - \hat{w})}}\Big]$$

چون عدد است می توانیم T بگیریم

$$= \text{structural error} + \text{approximation error} + \underbrace{2\underbrace{E_{x,y}\Big[(y - w^{*T} x)x^T\Big]}_{0\ ⊛}(w^* - \hat{w})}_{0}$$

پرسش ۳

⊛ چون x بردار است داریم:

1

$$\hat{w} = (\underbrace{x^T x}_{\text{عدد}})^{-1} x^T y = \frac{1}{\underbrace{x^T x}_{x \cdot x}} \overbrace{x^T y}^{x \cdot y} = \frac{\langle x, y \rangle}{\langle x, x \rangle}$$

2

columns of X are orthogonal $\longrightarrow$

$$X^T X = \begin{bmatrix} N & 0 & \cdots & 0 \\ 0 & \langle x_1, x_1 \rangle & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle x_d, x_d \rangle \end{bmatrix}$$

$$\longrightarrow (X^T X)^{-1} X^T = \begin{bmatrix} \frac{1}{N} & & & \\ & \frac{1}{\langle x_1, x_1 \rangle} & & \\ & & \ddots & \\ & & & \frac{1}{\langle x_d, x_d \rangle} \end{bmatrix} \begin{bmatrix} 1 & 1 & & 1 \\ x_1^{(1)} & x_1^{(2)} & & x_1^{(N)} \\ \vdots & \vdots & & \vdots \\ x_d^{(1)} & x_d^{(2)} & & x_d^{(N)} \end{bmatrix} = \begin{bmatrix} \frac{1}{N} & \cdots & \frac{1}{N} \\ \frac{x_1^{(1)}}{\langle x_1, x_1 \rangle} & \cdots & \frac{x_1^{(N)}}{\langle x_1, x_1 \rangle} \\ \vdots & & \vdots \\ \frac{x_d^{(1)}}{\langle x_d, x_d \rangle} & \cdots & \frac{x_d^{(N)}}{\langle x_d, x_d \rangle} \end{bmatrix}$$

$$\longrightarrow (X^T X)^{-1} X^T y = \hat{w} = \begin{bmatrix} \frac{\sum_i y^{(i)}}{N} \\ \frac{\langle x_1, y \rangle}{\langle x_1, x_1 \rangle} \\ \vdots \\ \frac{\langle x_d, y \rangle}{\langle x_d, x_d \rangle} \end{bmatrix}$$

⊛ مشاهده می‌شود که به برداری از ضرایب رگرسیون تک بعدی رسیدیم

3

$$y = w_1 x + w_0 \longrightarrow E[y] = E[w_1 x + w_0] = w_1 E[x] + w_0 \longrightarrow \boxed{w_0 = E[y] - w_1 E[x]}$$

$$\mathrm{cov}(x, y) = \mathrm{cov}(x, w_1 x + w_0) = \mathrm{cov}(x, w_1 x) + \mathrm{cov}(x, w_0)$$

$$= w_1 \underbrace{\mathrm{cov}(x, x)}_{\mathrm{var}(x)} + w_0 \underbrace{\mathrm{cov}(x, 1)}_{0} \longrightarrow \mathrm{cov}(x, y) = w_1\, \mathrm{var}(x) \longrightarrow \boxed{w_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)}}$$

بخش۲ (MLE, MAP)

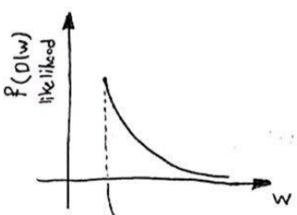
پرسش ۱

$$D = \{X^{(1)}, \dots, X^{(N)}\}$$

likelihood for continuous random vars $= f(D|\theta)$

$X^{(i)}$s are iid $U(-w, w) \rightarrow$ $w \geq \max(x^{(1)}, \dots, x^{(N)})$, $-w \leq \min(x^{(1)}, \dots, x^{(N)})$ $\rightarrow$ $w \geq \max(|x^{(1)}|, \dots, |x^{(N)}|)$

$$f(D|w) = \underbrace{\prod_{i=1}^{N} f_X(x^{(i)}|w)}_{\left(\frac{1}{2w}\right)^N} =$$



$\max(|x^{(1)}|, \dots, |x^{(N)}|)$

⊛ مشاهده می‌شود که تابع likelihood نزولی است و اگر بخواهیم آنرا بیشینه کنیم $\hat{w} = \max(|x^{(1)}|, \dots, |x^{(N)}|)$

پرسش ۲

$\boxed{1}$

$$\text{likelihood} = \ell(\mu) = p(D|\mu) = \prod_{i=1}^{N} f_x(x^{(i)}|\mu) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^{(i)}-\mu)^2}{2\sigma^2}}$$

$$\rightarrow L(\mu) = \ln \ell(\mu) = \sum_{i=1}^{N} -\ln(\sigma\sqrt{2\pi}) - \frac{(x^{(i)}-\mu)^2}{2\sigma^2}$$

$$\rightarrow \frac{\partial L(\mu)}{\partial \mu} = 0 \rightarrow \sum_{i=1}^{N} \frac{x^{(i)}-\mu}{\sigma^2} = 0 \rightarrow \sum_{i=1}^{N} x^{(i)} = N\mu \rightarrow \boxed{\hat{\mu} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}}$$

$\boxed{2}$

فرض می کنیم که $prior \sim N(\mu_0, \beta^2)$

$$p(D|\mu)\,p(\mu) = \prod_{i=1}^{N} f_x(x^{(i)}|\mu) \times f_\mu(\mu)$$

$$= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^{(i)}-\mu)^2}{2\sigma^2}} \times \frac{1}{\beta\sqrt{2\pi}} e^{\frac{-(\mu-\mu_0)^2}{2\beta^2}}$$

$$\xrightarrow{\ln} L(\mu) - \frac{(\mu-\mu_0)^2}{2\beta^2} - \ln(\beta\sqrt{2\pi})$$

$$\xrightarrow{\frac{\partial}{\partial\mu}=0} \sum_{i=1}^{N} \frac{x^{(i)}-\mu}{\sigma^2} - \frac{1}{\beta^2}(\mu-\mu_0) = 0$$

$$\rightarrow \frac{N\mu}{\sigma^2} + \frac{\mu}{\beta^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N} x^{(i)} + \frac{\mu_0}{\beta^2} \rightarrow (\mu)\left(\frac{N\beta^2+\sigma^2}{\cancel{\sigma^2\beta^2}}\right) = \frac{\beta^2\sum x^{(i)} + \sigma^2\mu_0}{\cancel{\sigma^2\beta^2}}$$

$$\rightarrow \hat{\mu} = \frac{\beta^2\sum x^{(i)} + \sigma^2\mu_0}{N\beta^2+\sigma^2} = \boxed{\frac{\frac{\beta^2}{\sigma^2}\sum_{i=1}^{N} x^{(i)} + \mu_0}{\frac{\beta^2}{\sigma^2}N + 1}}$$

$\boxed{3}$

⊛ وقتی $N \rightarrow \infty$ در فرمول $\hat{\mu}_{MAP}$ ترم های μ_0 در صورت و 1 در مخرج بی اهمیت می شوند و به فرمول $\hat{\mu}_{MLE}$ می رسیم

یعنی وقتی تعداد سمپل ها بسیار زیاد باشند، هر دو تخمین یک جواب را می دهند و همگرا می شوند

پرسش ۳

$D = \{X^{(1)}, \dots, X^{(N)}\}$

⊛ چون α را می‌دانیم $p(D|\alpha,\beta) \rightarrow p(D|\beta)$

$$p(\beta|D) \propto \overbrace{p(D|\beta)}^{\text{iid}\ \prod_{i=1}^{N} p(x^{(i)}|\beta)} \; p(\beta)$$

$p(D|\beta) \longrightarrow \sim gamma(\overset{\text{known}}{\alpha}, \beta)$

$p(\beta) \longrightarrow \sim gamma(\alpha_0, \beta_0)$

$$p(D|\beta) = \prod_{i=1}^{N}\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} {x^{(i)}}^{\alpha-1} e^{-\beta x^{(i)}}\right) = \underbrace{\frac{1}{\Gamma^{N}(\alpha)}\beta^{\alpha N} e^{-\beta\sum_{i=1}^{N} x^{(i)}}}_{⊛}\left(\prod_{i=1}^{N} x^{(i)}\right)^{\alpha-1} 1_{x^{(i)}>0}$$

$$p(\beta) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \underbrace{\beta^{\alpha_0-1} e^{-\beta_0\beta}}_{⊛⊛} 1_{\beta>0}$$

$$\xrightarrow{⊛\ ⊛⊛} p(\beta|D) \propto \beta^{\alpha N+\alpha_0-1} e^{-\beta\left(\beta_0+\sum_{i=1}^{N} x^{(i)}\right)} 1_{\beta>0,\ x^{(i)}>0}$$

که متناسب است با $gamma\left(\alpha_0+\alpha N,\ \beta_0+\sum_{i=1}^{N} x^{(i)}\right)$

⊛ بنابراین توزیع prior, posterior از یک جنس شد ، توزیع گاما برای توزیع گاما با α مشخص conjugate prior می‌باشد