دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

پاسخ مسائل تشریحی تکلیف دوم

استاد:دکتر ادیبی

پوریا صامتی
۹۹۳۶۲۳۰۲۹

پائیز ۱۴۰۲

۱- پاسخ الف، ب، پ

likelihood

$$\ell(w) = P(D \mid w)$$

$$= P\left((x^1,t^1), (x^2,t^2), (x^3,t^3), \ldots, (x^N,t^N) \mid w\right) \Rightarrow$$

چون xها در توضیحات مسئله معلوم هستند، پس در سمت دیگر شرط می‌توانند تعریف شوند

$$\rightarrow P\left(t^1, t^2, t^3, \ldots, t^N \mid x^1, x^2, x^3, \ldots, x^N, w\right)$$

با فرض مستقل بودن داده‌ها چون همگی مستقل هستند

$$\Rightarrow P(t^1 \mid x^1, w) \times P(t^2 \mid x^2, w) \times P(t^3 \mid x^3, w) \times \cdots \times P(t^N \mid x^N, w)$$

$$\Rightarrow \ell(w) = \prod_{i=1}^{N} P(t^i \mid x^i, w)$$

$$-\ln(\ell_w) = -\ln\left(\prod_{i=1}^{N} \left(1 - P(C=0 \mid x^i)\right)^{t^i} \cdot \left(P(C=0 \mid x^i)\right)^{1-t^i}\right)$$

$$\Rightarrow -\ln(\ell_w) = -\left[\ln\left(\sum_{i=1}^{N} t^i \left[1 - P(C=0 \mid x^i)\right] + (1-t^i)\left[P(C=0 \mid x^i)\right]\right)\right]$$

$$\Rightarrow -\ln(\ell_w) = -\left[\sum_{i=1}^{N} t^i \ln \underbrace{\left[1 - P(C=0 \mid x^i)\right]}_{P(C=1 \mid x^i)} + (1-t_i) \ln P(C=0 \mid x^i)\right]$$

$$\Rightarrow -\ln(\ell_w) = -\sum_{i=1}^{N} t^i \ln P(C=1 \mid x^i) - \sum_{i=1}^{N} (1-t^i) \ln P(C=0 \mid x^i)$$

$$\Rightarrow -\ln(\ell_w) = -\sum_{i=1}^{N} t^i \ln\left[\frac{\exp(-z^i)}{1+\exp(-z^i)}\right] - \sum_{i=1}^{N} (1-t^i) \ln\left[\frac{1}{1+\exp(-z^i)}\right]$$

$$\Rightarrow -\ln(\ell_w) = -\sum_{i=1}^{N} -z^i \times t^i + \cancel{\sum_{i=1}^{N} t^i (1+\exp(-z^i))}$$

$$+ \sum_{i=1}^{N} \ln(1+\exp(-z^i)) - \cancel{\sum_{i=1}^{N} t^i \ln(1+\exp(-z^i))}$$

$$\Rightarrow \boxed{-\ln(\ell_w) = \sum_{i=1}^{N} z^i t^i + \sum_{i=1}^{N} \ln(1+\exp(-z^i))}$$

$$\nabla l_w = \frac{\partial l}{\partial w} = \frac{\partial l}{\partial z} \cdot \frac{\partial z}{\partial w}$$

$$\Rightarrow \begin{cases} \nabla l_w = \left(\sum_{i=1}^{N} t^i + \sum_{i=1}^{N} \frac{-\exp(-z^i)}{1+\exp(-z^i)} \right) \cdot \left(x^i \right) \\ z = w^T x = \sum_{i=1}^{N} w \cdot x^i \end{cases}$$

$$\Rightarrow \nabla l_w = \sum_{i=1}^{N} \left(t^i + \frac{-\exp(-z^i)}{1+\exp(-z^i)} \right) \cdot x^i$$

$$\Rightarrow \frac{\partial l_w}{\partial w} = \sum_{i=1}^{N} x^i \left(t^i + \frac{-\exp(-z^i)}{1+\exp(-z^i)} \right)$$

$$= \sum_{i=1}^{N} x^i \left(t^i - P(C=1 \mid x^i) \right)$$

$$w^{t+1} = w^t - \eta_t \sum_{i=1}^{N} x^i \left(t^i - P(c=1 \mid x^i) \right)$$

پاسخ ۲

پاسخ 2)

الف) خیر هیچ تفاوتی نمی‌کند. زیرا مدل اول bias ندارد.

$$\begin{cases} \sigma(w_1(0) + w_2(0)) = \sigma(0) \\ \sigma(0) = \frac{1}{1+e^{(0)}} = 1/2 \end{cases}$$

ب) باز هم به خاطر اینکه این مدل bias ندارد، در حین آموزش نیز فقط در $x=[0,0]^T$ ضرب می‌شود که باعث می‌شود دسته‌بندیان در اثر مشاهده این داده تغییری در وزن‌ها رخ ندهد.

پ)

$$\sum_i \log\left(y^{(i)} \mid x^{(i)}, \underline{w}\right) - \frac{\lambda}{2} \|\underline{w}\|^2 = \underbrace{\sum_i \log \sigma\left(y^{(i)} w^T x^{(i)}\right)}_{\frac{1}{2} \sum_{i=0}^{N} y^{(i)} w^T x^{(i)}} - \frac{\lambda}{2} \|\underline{w}\|^2$$

$$\Rightarrow \log[l(w)] = \frac{1}{2} \sum_{i=0}^{N} y^{(i)} \underline{w}^T x^{(i)} - \frac{\lambda}{2} \|\underline{w}\|^2$$

$$\Rightarrow \frac{\partial \log[l(w)]}{\partial \underline{w}} = \frac{1}{2} \sum_{i=0}^{N} y^{(i)} x^{(i)} - \frac{\lambda}{2} \cdot 2\underline{w}$$

$$\Rightarrow \frac{\partial \log[l(w)]}{\partial \underline{w}} = \frac{1}{2} \sum_{i=0}^{N} y^{(i)} x^{(i)} - \lambda \underline{w} = 0 \Rightarrow$$

$$\Rightarrow \boxed{\underline{w} = \frac{\frac{1}{2} \sum_{i=1}^{N} y^{(i)} x^{(i)}}{\lambda}}$$

$\Rightarrow$ همانطور که مشخص است از رابطه بالا w با ضریب λ می‌توان دریافت که این دو با همدیگر نسبت معکوس دارند. افزایش λ موجب کاهش $\underline{w}$ می‌شود و بالعکس.

پاسخ ۳)

پاسخ 3) الف)

$$P(y|x,a,b,c) = \mathcal{N}(ax^2+bx+c, \sigma^2)$$

$$\Rightarrow P(y|x,a,b,c) = \prod_{i=1}^{N} P(y^{(i)}|x^{(i)},a,b,c)$$

$$\xrightarrow{\log} \log(P(y|x,a,b,c)) = \log\left(\prod_{i=1}^{N} P(y^{(i)}|x^{(i)},a,b,c)\right) = \log\left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)\right)$$

$$= \log\left(\underbrace{\frac{1}{\sigma\sqrt{2\pi}}}_{(\sigma^2 2\pi)^{\frac{1}{2}}}\right)^n - \sum_{i=1}^{N}\left(\frac{1}{2\sigma^2}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)\right)^2$$

$$= \frac{n}{2}\log[\sigma^2 2\pi] - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)^2$$

$$\frac{\partial l}{\partial a} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)\cdot[-2]\cdot[x^{(i)}]^2$$

$$= +\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)[x^{(i)}]^2$$

$$\frac{\partial l}{\partial b} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)\cdot[2]\cdot[-x^{(i)}]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)x^{(i)}$$

$$\frac{\partial l}{\partial c} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)\cdot 2\cdot[-1]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)$$

$$\frac{\partial l}{\partial x} = 0 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)^2\cdot 2\cdot[0 - 2ax^{(i)} - b - 0]$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)^2(2ax^{(i)}+b)$$

$$\frac{\partial l}{\partial \sigma^2} = \frac{n}{2}\cdot\left[\frac{1}{\frac{1}{\sigma^2 2\pi}}\right] - \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)$$

$$= n\pi\sigma^2 - \frac{1}{2\sigma^4}\sum_{i=1}^{N}\left(y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right)$$

$$\alpha = \frac{\sum_{i=1}^{N} x(i)^2 y(i)}{\sum_{i=1}^{N} x(i)^4} \Rightarrow a = \frac{[(5^2)\times 8] + [1^2 \times 4] + [3^2 \times (-2)] + [0^2 \cdot (-8)]}{5^4 + 1^4 + 3^4 + 0^4} \simeq 0.26$$

ب)

$$b = \frac{\sum_{i=1}^{N} x^{(i)} y^{(i)}}{\sum_{i=1}^{N} x^{(i)2}} \Rightarrow b = \frac{[5\times 8] + [1\times 4] + [3\times(-2)] + [0\times(-8)]}{5^2 + 1^2 + 3^2 + 0} \simeq 1.08$$

$$C = \frac{\sum_{i=1}^{N} \left[y^{(i)} - a[x^{(i)}]^2 - bx^{(i)}\right]}{N} = \frac{[8 - a(5)^2 - b(5)] + [4 - a(1)^2 - b(1)] + [-2 - a(3)^2 - b(3)] + [-8 - a(0)^2 - b(0)]}{4} \simeq -4.2$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} \left[y^{(i)} - a[x^{(i)}]^2 - bx^{(i)} - c\right]^2}{N} = \frac{[8 - a(5)^2 - b(5) - c]^2 + [4 - a(1)^2 - b(1) - c]^2 + [-2 - a(3)^2 - b(3) - c]^2 + [-8 - a(0)^2 - b(0) - c]}{4}$$

$$\simeq 15.43$$

پاسخ ۴)

پاسخ 4)

$$P(y|x,w) = N(xw, \sigma^2 I) = \prod_{i=1}^{|D|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

↳ likelihood distribution

$$P(w) = N(0, \sigma_0^2 I) = \prod_{j=1}^{|w|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(w_j)^2}{2\sigma_0^2}\right) \Rightarrow \text{Perior Distribution}$$

$$MAP = P(w, x|y) = P(y|x,w) \cdot P(w)$$

$$\overset{\log}{\Longrightarrow} \underbrace{\log P(w,x|y)}_{loss} = \log P(y|x,w) \cdot P(w) \Rightarrow \log P(w,x|y) = \log\left(\prod_{i=1}^{|D|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x)^2}{2\sigma^2}\right)\right) \times \log\left(\prod_{j=1}^{|D|} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(w_j)^2}{2\sigma_0^2}\right)\right)$$

$$\xrightarrow[w]{\text{مشتق برحسب}} \frac{\partial\, loss}{\partial w} = \frac{1}{\sigma^2} \sum_{i=1}^{|D|} (y_i - w^T x_i) x_i - \frac{1}{\sigma_0^2} \sum_{j=1}^{|w|} w_j$$

$$\Rightarrow \frac{\partial\, loss}{\partial w} = 0 \Rightarrow w = \frac{\sum_{j=1}^{|D|} y_i x_i}{\sum_{i=1}^{|D|} x_i x_i^T + \left(\frac{\sigma}{\sigma_0}\right)^2 I^T}$$

CamScanner

پاسخ ۵)

پاسخ 5) الف) درست است، زیرا اگر مشتق را در رابطه مقابل می‌توانیم برحسب a نیز حساب کنیم؛

$$y_i = \log(x_1^{a_1} e^{a_2}) + \varepsilon_i = \log x_1^{a_1} + \log e^{a_2} + \varepsilon_i = a_1 \log x_1 + a_2 \log x_2 + \varepsilon_i$$

ب) غلط است. زیرا نمی‌توانیم رابطه $y_i = x_1^{a_1} e^{a_2}$ را بصورت رابطه خطی بین a و y تعریف کنیم.

پاسخ ۶)

پاسخ 6) طبق یک قانون سرانگشتی می دانیم که اگر درجه چند جمله ای برابر m باشد، تعداد داده مورد نیاز برای آموزش مدل برابر است با [5m - 10m]. یعنی 5 برابر تا 10 برابر m باید نمونه داشته باشیم.

پس محدودیتی از نظر تعداد داده نداریم. بهتر است مدل B بهتر است اتفاقی که رخ می دهد مدل در حین آموزش، ضریب پایینی را به جمله x^5 اختصاص می دهد. بنابراین این مدل ها شبیه مدل درجه 4 رفتار خواهد کرد (چون داده ها با یک رابطه درجه 4 تولید شده اند)

مدل A از آنجا که درجه 2 است، برای یک مجموعه داده که می دانیم از درجات بالاتر تولید شده، بسیار ساده است و باعث underfitting می شود.