هوش مصنوعی - تمرین تئوری سری چهارم

طاهر مظاهری - 98102346

سوال 1

الف)

$$\{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}\}_{mle} = \underset{\beta_0,\beta_1,\sigma}{\operatorname{argmax}} \prod_{n=1}^{N} p(\{x_n, y_n\} \mid \beta_1, \beta_0, \sigma) \underset{\epsilon_i \sim N(0,\sigma^2)}{=} \underset{\beta_0,\beta_1,\sigma}{\operatorname{argmax}} \prod_{n=1}^{N}$$

$$\left(N(y_n \mid \beta_0 + \beta_1 x_n, \sigma^2)\right) \xrightarrow{NLL} \underset{\beta_0,\beta_1,\sigma}{\operatorname{argmax}} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_n - \beta_1 x_n - \beta_0)^2}{2\sigma^2}\right) \xrightarrow{NLL}$$

(واریانس) (امید ریاضی توزیع)

$$\underset{\beta_0,\beta_1,\sigma}{\arg\min}\left(\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \beta_1 x_n - \beta_0)^2\right) \quad (*)$$

(ثابت نسبت به β_1, β_0) (ثابت)

$$\hat{\beta}_0: \frac{\partial}{\partial \beta_0} NLL = 0 = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \beta_1 x_n - \beta_0) \Rightarrow 0 = \sum_n y_n - N\hat{\beta}_0 - (\sum_n x_n)\times\hat{\beta}_1 \xrightarrow{/N}$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1: \frac{\partial}{\partial \beta_1} NLL = 0 = \frac{1}{\sigma^2}\sum_n (x_n y_n - \hat{\beta}_0 x_n - \beta_1 x_n^2) \rightarrow 0 = \sum_n x_n y_n - \hat{\beta}_0 \sum_n x_n - \hat{\beta}_1 \sum_n x_n^2$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \rightarrow 0 = \sum_n x_n y_n - (\bar{y} - \hat{\beta}_1 \bar{x})\sum_n x_n - \hat{\beta}_1 \sum_n x_n^2 \rightarrow \hat{\beta}_1 = \frac{\sum_n x_i y_i - \sum_n \bar{x}\bar{y}}{\sum_n x_n^2 - \sum_n \bar{x}^2}$$

(*): برای تخمین MLE باید NLL بدست آمده کمینه شود (برحسب β_0, β_1) که با توجه به NLL بدست آمده معادل کمینه کردن $\sum_n (y_n - \beta_1 x_n - \beta_0)^2$ است که همان مجموع مربعات خطا است

ب)

$$E\hat{\beta}_1 = E\left[\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\right] = E\left[\frac{\sum_i (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2}\right] = E\left[\frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_i (x_i - \bar{x})^2}\right]$$

$$*: \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i - \bar{y}\underbrace{\sum (x_i - \bar{x})}_{0} = \sum_i (x_i - \bar{x})y_i$$

جمله β_0 به دلیل صفر بودن حذف می‌شود

$$\rightarrow E[\hat{\beta}_1] = E\left[\frac{\sum (\beta_1 x_i + \varepsilon_i)(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right] = \frac{1}{\sum (x_i - \bar{x})^2}\left(\sum (\beta_1 x_i + E[\varepsilon_i]) \times (x_i - \bar{x})\right)$$

$E[\varepsilon_i] = 0$

$$= \frac{\beta_1 \sum x_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\beta_1 \sum (x_i^2 - x_i\bar{x})}{\sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)} = \beta_1 \Rightarrow \hat{\beta}_1 \text{ is unbiased}$$

$$E[\hat{\beta}_0] = E[\bar{y} - \hat{\beta}_1 \bar{x}] = E\bar{y} - E\hat{\beta}_1 \times \bar{x} = E[\beta_0 + \beta_1 x_i + E\varepsilon_i] - \beta_1 \bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} + 0$$

$$= \beta_0 \rightarrow \hat{\beta}_0 \text{ is unbiased}$$

$$VAR[\hat{\beta}_1] = VAR\left[\frac{\sum (x_i-\bar{x})(\beta_0+\beta_1 x_i+\varepsilon_i)}{\sum (x_i-\bar{x})^2}\right] \xrightarrow[\text{تصادفی نیستند و ارپانس صفر دارند حذف می‌شوند}]{\text{جملاتی } \beta_0+\beta_1 x_i \text{ چون}} VAR\left[\frac{\sum (x_i-\bar{x})\varepsilon_i}{\sum (x_i-\bar{x})^2}\right]$$

$VAR[aX] = a^2 VAR X$

$$\longrightarrow VAR[\hat{\beta}_1] = \left(\frac{\sum (x_i-\bar{x})^2}{\left(\sum (x_i-\bar{x})^2\right)^2}\right) VAR[\varepsilon_i] = \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}$$

ε_i are independent

$$VAR[\hat{\beta}_0] = VAR\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = VAR\left[\frac{1}{n}\left(\beta_1(\sum x_i) + n\beta_0 + (\sum \varepsilon_i) - \hat{\beta}_1 \sum x_i\right)\right]$$

deterministic ← deterministic → VAR = 0

$$= VAR\left[\frac{1}{n}\left(\sum_i \varepsilon_i - (\sum_i x_i)\hat{\beta}_1\right)\right] = \frac{1}{n^2}\left(n\sigma^2 + (n\bar{x})^2 \frac{\sigma^2}{\sum (x_i-\bar{x})^2} - 2n\bar{x}\frac{Cov(\sum y_i, \sum (x_j-\bar{x})\beta_1)}{\sum (x_i-\bar{x})^2}\right)$$

$$= \frac{\sigma^2}{n} + \bar{x}\sigma^2_{\hat{\beta}_1} - 2\bar{x}\frac{\sum (x_i-\bar{x}) \overset{0}{\cancel{\text{Cov}}} \times VAR\, y_1}{n\sum (x_i-\bar{x})^2} \qquad (\sigma^2_{\bar{y}} \nearrow \quad \sigma^2_{\hat{\beta}_1} \nearrow)$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i-\bar{x})^2} = \frac{\sigma^2\left(n\bar{x}^2 + \sum (x_i-\bar{x})\right)}{n\sum (x_i-\bar{x})^2}$$

$$\longrightarrow \frac{\sigma^2 \sum x_i^2}{n\sum (x_i-\bar{x})^2}$$

ب)

$$\hat{\beta}_1 = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} = \frac{\sum_i (x_i-\bar{x}) y_i - \overset{0}{\cancel{\sum (x_i-\bar{x})\bar{y}}}}{\sum_i (x_i-\bar{x}) x_i - \overset{0}{\cancel{\sum_i (x_i-\bar{x})\bar{x}}}} = \frac{\sum_i (x_i-\bar{x}) y_i}{\sum_i (x_i-\bar{x}) x_i}$$

$\sum_i x_i - \bar{x} \overset{?}{=} 0$ ✓ $\Longrightarrow Y_i = (x_i-\bar{x})$ بله طبق رابطه بالا به فرم مشابه رسید؛

پ)

$$E\tilde{\beta}_1 = E\frac{\sum Y_i y_i}{\sum Y_i x_i} = E\frac{\sum Y_i(\beta_0+\beta_1 x_i+\varepsilon_i)}{\sum Y_i x_i} = E\frac{\overset{0}{\cancel{\beta_0 \sum Y_i}} + \beta_1 \sum Y_i x_i + \sum Y_i \varepsilon_i}{\sum_i Y_i x_i}$$

β ثابت $\sum Y_i$ چون

غیر تصادفی اند، بین Y_i و x_i هم غیر تصادفی اند

$$\longrightarrow E\tilde{\beta} = \frac{\beta_1 \sum Y_i x_i + \overset{0}{\cancel{\sum Y_i E\varepsilon_i}}}{\sum_i Y_i x_i} = \beta_1 \qquad (VAR(\gamma x + \beta_0) = 0)$$

ت)

$$VAR\,\tilde{\beta}_1 = VAR\left[\frac{\sum Y_i y_i}{\sum Y_i x_i}\right] = VAR\left[\frac{\sum Y_i \cancel{\beta}(\beta_1 x_i + \beta_0 + \varepsilon_i)}{\sum Y_i x_i}\right] = VAR\left[\frac{\sum Y_i \varepsilon_i}{\sum Y_i x_i}\right]$$

deterministic

$$= \frac{\sum Y_i^2 VAR\,\varepsilon_i}{(\sum_i Y_i x_i)^2} = \sigma^2 \frac{\sum Y_i^2}{(\sum Y_i x_i)^2}$$

$$\frac{VAR\,\hat{\beta}_1}{VAR\,\tilde{\beta}_1} = \frac{\frac{\sigma^2}{\sum (x_i-\bar{x})x_i}}{\frac{\sigma^2 \sum Y_i^2}{(\sum Y_i x_i)^2}} =$$

$$\frac{(\sum Y_i x_i)^2}{(\sum Y_i^2)(\sum (x_i-\bar{x})x_i)} \leq 1 \Longrightarrow VAR\,\hat{\beta} \leq VAR\,\tilde{\beta}$$

$\sum Y_i = 0$ چون Y_i ها ⇒

$$\begin{cases} \sum Y_i^2 \geq \sum Y_i x_i \\ \sum (x_i-\bar{x})x_i \geq \sum Y_i x_i \end{cases}$$

سوال 2

$$\min_{w} \lambda w^T w + \|Xw - Y\|_2^2$$

الف)

- یک بردار مقدار اولیه برای w انتخاب می‌کنیم و مقدار η (ضریب یادگیری) را انتخاب می‌کنیم
- تا وقتی که $F(w)$ از سطح آستانه‌ای ما بزرگتر است، کار زیر را انجام می‌دهیم:
- از بین نمونه‌ها (n تا نمونه‌ای دیتاست) دسته‌ای را به صورت تصادفی انتخاب می‌کنیم (دسته‌ای به اندازه n)
- برای نمونه‌های انتخابی در مرحله قبل، کار زیر را انجام می‌دهیم:

$$w \Leftarrow w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla F_i(w)$$

ب)

w is vector so, $\|w\|_2 = \underbrace{w^T w}$ → adding this factor to arg min argument, large $\|w\|_2$ would be penalized by (tikhonov regularization) so, $\|w_2\|_2 < \|w_1\|_2$

به بیان دیگر اضافه کردن $\lambda w^T w$ باعث کوچک شدن فضای جواب w برای کمینه کردن $L(w)$ است که به این معنی است که $L(w_1) < L(w_2)$، حال اگر $\|w_1\| > \|w_2\|$ باشد بدین معنی است که w_2 پاسخ مناسبی برای مسئله دوم نیست (چرا که w_1 پاسخ بهینه‌تری است) که خلاف فرض است w_2 پاسخ مسئله دوم باشد پس $\|w_1\| \not< \|w_2\|_2$ است

سوال 3

این ویژگی در نگاه اول تاثیر بسیار زیادی روی مدل ما داشته و باید حفظ شود چرا که x_i ها را بشدت تحت تاثیر قرار می‌دهد و y_i ها اثرپذیری زیادی از این ویژگی دارند؛ اما در حقیقت بزرگی ضرایب ویژگی‌ها به تنهایی موجب اثرگذاری خیلی زیاد این پارامتر نیست چرا که این ضریب بسیار بزرگ می‌تواند به موجب واحد کوچک آن ویژگی باشد (مثلا برحسب MB بودن بجای برحسب KB بودن) وی‌توان B بودن در دست داشتن اطلاعات بیشتر در مورد این ویژگی نظر داد (این اطلاعات بیشتر می‌تواند اندازه standard deviation باشد) و پارامتر مناسب می‌تواند حاصل ضرب σ_{x_i} در ضریب محاسبه شده باشد

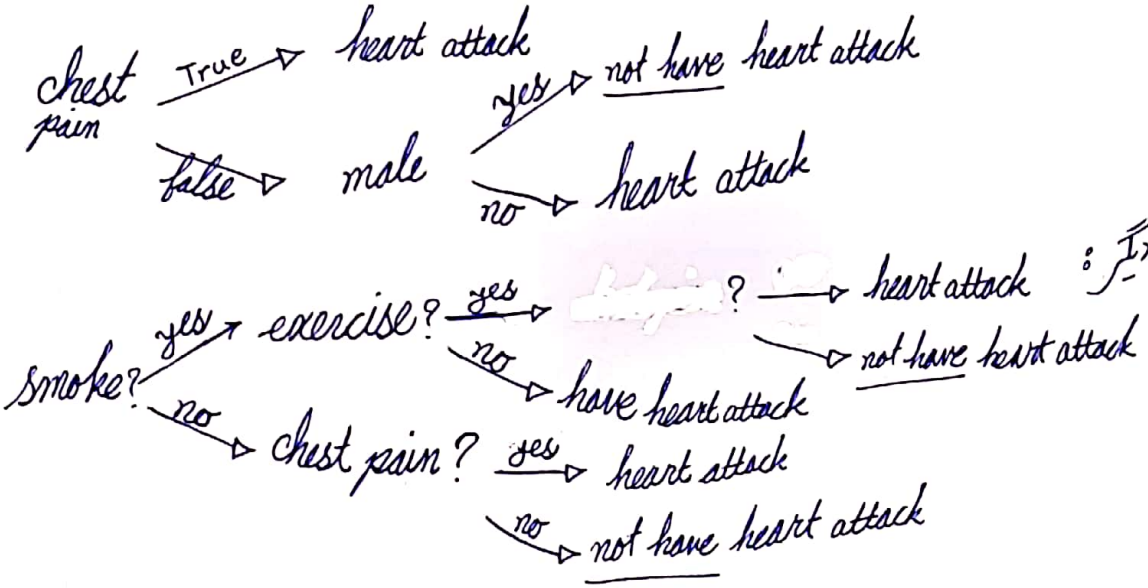
$x_{i,p}$ ← نمونه‌ها
← ویژگی iام

سوال 4

- غلط؛ گاها در محاسبه bias، به ازای برحسب تعداد نمونه‌ها می‌رسیم که نشاندهنده اثرپذیری bias از تعداد نمونه‌ها است.
- غلط؛ بدلیل وجود نویز در داده‌های آموزش و همچنین عدم تطابق رفتار مدل با رفتار واقعی نمی‌آیند، کم کردن زیاد خطا منجر به overfit می‌شود که می‌تواند منجر به این شود داده‌های تست خطای زیادی داشته باشند (نمونه‌ای از overfit در اسلاید 16 صفحه 8 وجود دارد)
- به طور کلی غلط است؛ افزایش پیچیدگی تا چندین مرحله محدود می‌تواند باعث بهبود عملکرد مدل هم در training set و هم در داده تست شود اما افزایش پیچیدگی بیش از اندازه موجب افزایش خطا در داده تست شود (اصولا علت وجود داده تست برانداز کردن هایپرپارامترها (مانند پیچیدگی مدل) است)

سوال 5

الف)



درخت تصمیم گیری دیگر:

ب)

از درخت بالایی استفاده می‌کنیم

- اگر درد سینه دارند، احتمال حمله قلبی وجود دارد
- اگر درد سینه ندارند و مرد هستند، احتمال حمله قلبی وجود ندارد
- " و زن هستند، احتمال حمله قلبی وجود دارد

سوال 6

می‌توان هر رشته بیت دودویی را با رشته بایزی به طول d ساخت، همچنین با پرسش $x_i = 0$؟ دیگر ام مشخص می‌شود که بیت iام صفر است یا نه؛ به این روش در برگ‌های این درخت دودویی کامل همه رشته‌های بایزی ممکن به طول d بیت ساخته می‌شود که هر رشته در تناظر یک به یک با یکی از اعضای مجموعه رشته بندهای $\{0,1\}^d$ خواهد بود، پس درخت تصمیم گیری حداکثر تا عمق d تا گره دارد (که یعنی عمق $d+1$ دارد) (در صورتی که 0 درایه‌ها قابل ساده شدن باشند این امکان وجود دارد که حداکثر عمق این درخت کمتر باشد.