درس هوش مصنوعی **تمرین تئوری چهارم**

متین باقری **۴۰۲۱۰۵۷۲۷**

---

۱) آ) نادرست. الزاما *همواره* درست نیست. ممکن است با داده های تست نیز تطابق داشته باشد و خطا در تست نیز کاهش یابد. درواقع ممکن است داده های آموزش، نویزی در تضاد با داده های تست نداشته باشد.

ب) درست. اگر bias مدل زیاد باشد، یعنی مدل بسیار ساده است و نمی‌تواند الگوهای پیچیده‌ی موجود در داده‌ها را به درستی یاد بگیرد و افزایش تعداد داده‌های آموزشی تأثیر چندانی در کاهش bias ندارد، چون مدل اساساً نمی‌تواند روابط پیچیده را یاد بگیرد؛ حتی اگر داده‌های بیشتری داشته باشد، باز هم به همان صورت ساده آن‌ها را مدل‌سازی می‌کند.

ج) نادرست. طبقه بندی حالت خاصی از رگرسیون نیست.

د) نادرست. الزاما اینطور نیست و ممکن است مجددا برچسب غلط پیش بینی شود. مثال: ($y_i = -1$)

---

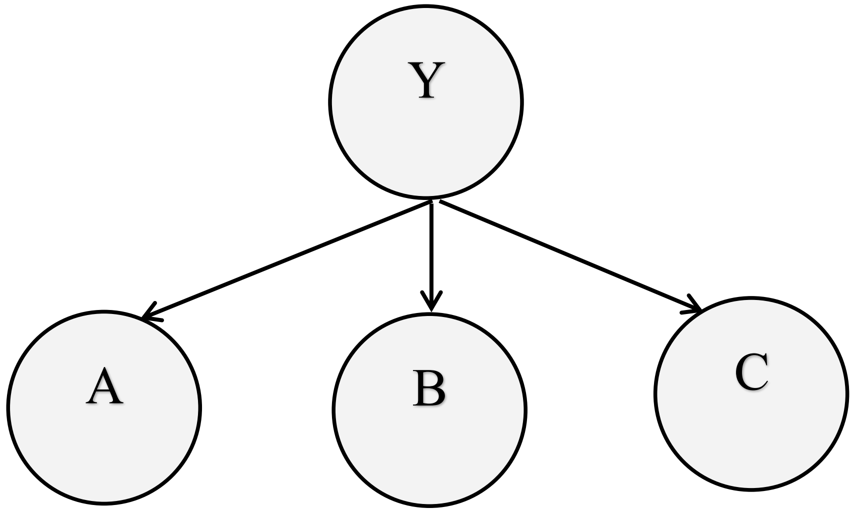
۲) آ) بله، به صورتی خطی جدا پذیر اند. چون δ آن مثبت است. مثال خط جدا کننده: $f_7 + f_8 - 3$

ب) در اولین مشاهده حاصل 1 است که بزرگتر از 0 است و A خروجی میدهیم که صحیح است پس بردار وزن تغییری نمیکند. پس از اولین مشاهده: (bias: 1, f_8: 0, f_7: 0)

در دومین مشاهده حاصل 1 است که بزرگتر از 0 است و A خروجی میدهیم که صحیح نیست و بردار وزن با (-1, -1, -1) جمع میشود. پس از دومین مشاهده: (bias: 0, f_8: -1, f_7: -1)

ج) – بله: (f_8: 1, f_7: 1, bias: -7)   – خیر   – خیر   – خیر

۳)



A \ Y	0	1
0	¼	4/6
1	¾	2/6

B \ Y	0	1
0	2/4	1/6
1	2/4	5/6

C \ Y	0	1
0	0/4	3/6
1	4/4	3/6

آ)

Y = 0 ~: 4/10 * 3/4 * 2/4 * 0/4 = 0

Y = 1 ~: 6/10 * 2/6 * 5/6 * 3/6 = 1/12

تناسب احتمال مشاهده چنین رکوردی در صورت نادرست بودن Y برابر صفر، و در صورت درست بودن Y برابر $\frac{1}{12}$ است.

بنابراین در طبقه بندی **درست** قرار می‌گیرد.

ب)

Y = 0 ~: 4/10 * 1/4 * 2/4 * 4/4 = 1/20

Y = 1 ~: 6/10 * 4/6 * 1/6 * 3/6 = 1/30

1/30 > 1/20 بنابراین: **نادرست**

ج)

$$Y = 0 \sim: \quad \frac{4+1}{10+2} \times \frac{1+1}{4+2} \times \frac{2+1}{4+2} \times \frac{0+1}{4+2} = \frac{5}{432}$$

$$Y = 1 \sim: \quad \frac{6+1}{10+2} \times \frac{4+1}{6+2} \times \frac{1+1}{6+2} \times \frac{3+1}{6+2} = \frac{35}{768}$$

درست

د) طبقه بندی فعلی: **نادرست**

$$Y = 0 \sim: \quad \frac{4}{10} \times \frac{3}{4} \times \frac{2}{4} \times \frac{4}{4} = \frac{3}{20}$$

$$Y = 1 \sim: \quad \frac{6}{10} \times \frac{2}{6} \times \frac{1}{6} \times \frac{3}{6} = \frac{1}{60}$$

برای رسیدن به این هدف، بهترین رکوردی که میتوانیم اضافه کنیم، همین مقادیر a, b, c را دارد و مقدار Y آن درست است.

در آن صورت خواهیم داشت:

$$Y = 0 \sim: \quad \frac{4}{11} \times \frac{3}{4} \times \frac{2}{4} \times \frac{4}{4} = \frac{3}{22}$$

$$Y = 1 \sim: \quad \frac{7}{11} \times \frac{3}{7} \times \frac{2}{7} \times \frac{4}{7} = \frac{24}{539}$$

حتی در این صورت هم جواب هنوز نادرست است. بنابراین نمیتوان با افزودن تنها یک رکورد، طبقه بندی را تغییر داد.

۴) آ) بیشترین IG معادل است با کمترین H(Y|X):

$$H(Y|A) = -\left(\frac{5}{10}\left(\frac{4}{5}\log\frac{4}{5}+\frac{1}{5}\log\frac{1}{5}\right)+\frac{5}{10}\left(\frac{3}{5}\log\frac{3}{5}+\frac{2}{5}\log\frac{2}{5}\right)\right) \cong 0.8464$$

$$H(Y|A) = \frac{1}{2}(0.72+0.97) = 0.845$$

$$H(Y|B) = -(\frac{7}{10}\left(\frac{5}{7}\log\frac{5}{7}+\frac{2}{7}\log\frac{2}{7}\right)+\frac{3}{10}\left(\frac{2}{3}\log\frac{2}{3}+\frac{1}{3}\log\frac{1}{3}\right)) \cong 0.8796$$

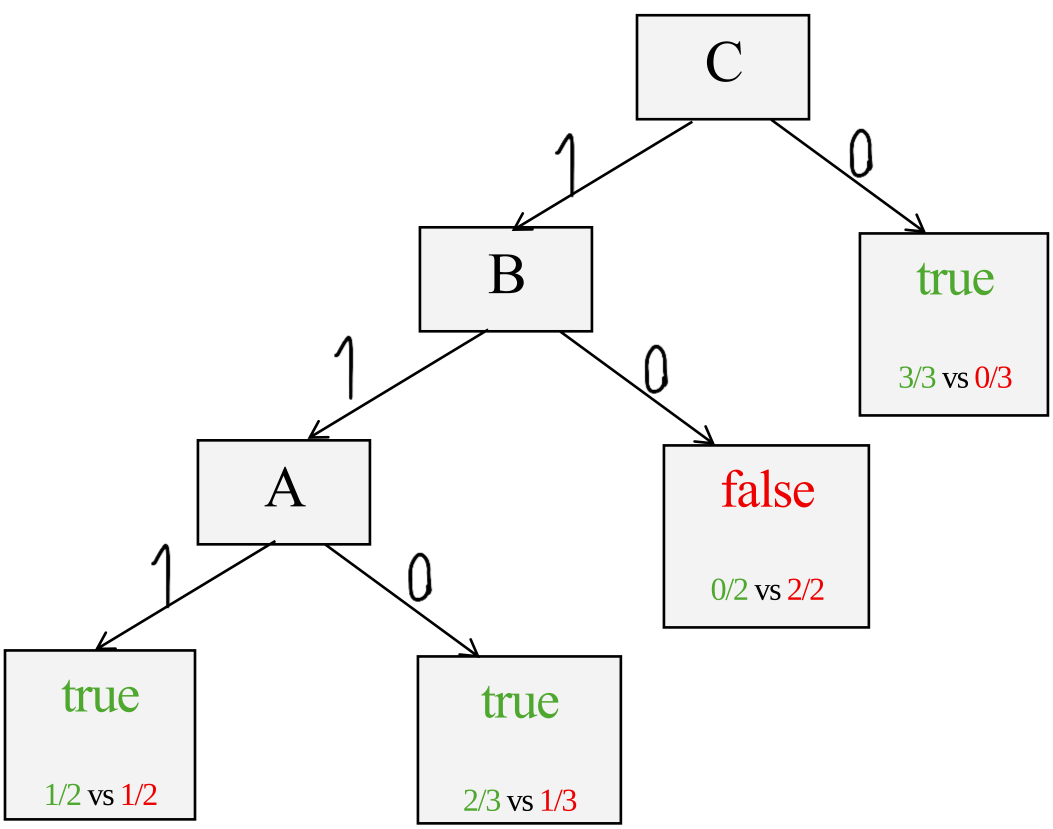
$$H(Y|B) = \frac{3}{10}*0.92+\frac{7}{10}*0.86 = 0.878$$

$$H(Y|C) = -(\frac{7}{10}\left(\frac{4}{7}\log\frac{4}{7}+\frac{3}{7}\log\frac{3}{7}\right)+\frac{3}{10}\left(\frac{3}{3}\log\frac{3}{3}\right)) \cong 0.6894$$

$$H(Y|C) = \frac{7}{10}*\ 0.99+\frac{3}{10}*0.00 = 0.693$$

بنابراین C بیشترین IG را دارد.

ب)

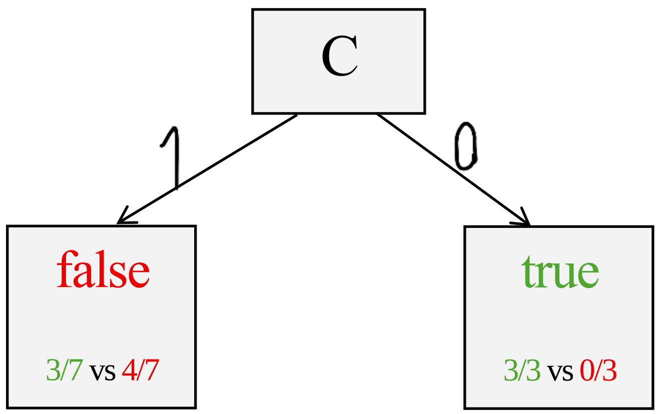


میتوان به جای گره A، true گذاشت.

ج) ۱) false

۲) true

د) درخت به این شکل می‌شود:



۱) false

۲) true

۵) آ)

$1*1 + 0*0 + 0*-1 + 1*1 + 1 = 3 \Rightarrow 1$

ب) ۱) اگر در مورد تصمیم گیری درون یک برگ صحبت کنیم، درست است. اما اگر تصمیم گیری در حالت کلی را بررسی کنیم، نادرست است. چون ۴ برگ داریم، در حالت کلی به کمک ۴ خط مختلف تصمیم گیری می شود و تصمیم گیری کلی مانند مدل پرسپترون صرفا یک خط نیست.

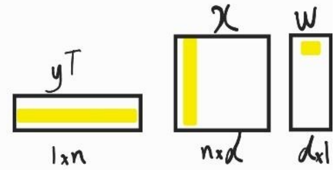
۲) نادرست، تصمیم گیری در برگ ها در درخت پرسپترون نسبت به درخت تصمیم دقیق تر است و احتمال underfit شدن آن کمتر است.

ج)

- نادرست، اولی به کمک آنتروپی انتخاب میشود و دومی با training error rate و نتیجه الزاما یکسان نیست.
- نادرست، از آنجا که راس های متفاوتی میتوانند انتخاب شوند، رسیدن به Training Error صفر میتواند با مسیر هایی با طول متفاوت انجام شود.
- درست، چون نهایتا Training Error هر دو درخت صفر است.
- نادرست، اگر ورودی از training data ما نباشد، چنین تضمینی وجود ندارد و از آنجا که مدل ها متفاوت اند، میتوانند به ازای ورودی یکسان، خروجی متفاوت داشته باشند.
- نادرست، الزامی به انجام این کار وجود ندارد.

۶) آ)

$$J_\lambda = \frac{1}{2}(y - Xw)^T(y - Xw) + \lambda|w|$$



$$= \frac{1}{2}y^Ty + \frac{1}{2}w^T\underbrace{X^TX}_{I}w - \frac{1}{2}\underbrace{w^TX^Ty}_{1\times 1} - \frac{1}{2}\underbrace{y^TXw}_{1\times 1} + \lambda|w|$$

$$\Rightarrow g(y) = \frac{1}{2}y^Ty \quad , \quad f = \frac{1}{2}w_i^2 + \lambda|w_i| - (y^TX_{:,i})w_i$$

مقدار $g(y)$ که فارغ از X و w مشخص است و در کمینه کردن تابع نقشی ندارد. در بخش تابع f درایه i ام بردار w تنها با ویژگی i ام ارتباط دارد. (چون در $y^TX_{:,i}$ ضرب می‌شود و ستون i ام ماتریس X تنها شامل ویژگی i ام است.)

ب)

$$\text{if } w_i > 0 \rightarrow$$

$$w_i = \arg\min_{w_i} \underbrace{w_i\left(\frac{w_i}{2} + \lambda - y^TX_{:,i}\right)}_{h(w_i)}$$

$$h(w_i) = w_i\left(\frac{w_i}{2} - w_i\right) = -\frac{w_i^2}{2} < 0 \leftarrow (w_i > 0)$$

$$(w_i = 0) \Rightarrow h(w_i) = 0$$

$$\frac{d}{dw_i}h(w_i) = w_i + \lambda - y^TX_{:,i} = 0 \rightarrow w_i = y^TX_{:,i} - \lambda$$

$$w_i = \max\left(0, y^TX_{:,i} - \lambda\right)$$

ج)

$$w_i = \arg\min_{w_i} \underbrace{\frac{w_i^2}{2} - \lambda w_i - (y^T x_{:,i})\, w_i}_{h(w_i)}$$

$$\frac{d}{dw_i} h(w_i) = w_i - \lambda - y^T x_{:,i} = 0 \rightarrow w_i = \lambda + y^T x_{:,i}$$

$$w_i = \min(\lambda + y^T x_{:,i} \,,\, -\epsilon)$$

مقدار بسیار کوچک منفی

د)

در صورتی که $\lambda > y^T x_{:,i}$ باشد

یا $\lambda + y^T x_{:,i} > 0$ باشد

یعنی $\lambda > -y^T x_{:,i}$

اعمال: $\lambda > |y^T x_{:,i}|$

ه)

$$\frac{d}{dw_i} f = w_i(\lambda+1) - y^T X_{:,i} = 0$$

$$w_i = \frac{y^T X_{:,i}}{\lambda+1} \quad , \quad f = \frac{3}{2} w_i^2 (\lambda+1)$$

$$\left(\lim_{\lambda \to \pm\infty} w_i = 0 \right)$$

$$y^T X_{:,i} = 0 \Longrightarrow w_i = 0$$

در حالت قبل میتوانستیم با تنظیم مقدار λ وزن یک ویژگی را صفر مطلق کنیم. در حالی که در اینجا میتوان با بردن λ به سمت $\pm\infty$ وزن را به سمت صفر میل داد.

۷)

$$f(w,x)=\frac{1}{1+e^{-(-۲+۶-۳)}}=\frac{1}{1+e^{-1}}=\frac{e}{1+e}\simeq ۰٫۷۳۱$$

خروجی گره ها را به ترتیب از چپ به راست شماره گذاری می کنیم.

$n_1 = w_0\, x_0 = -۲\ ,\ n_۲ = w_1\, x_1 = ۶\ ,\ n_۳ = n_1 + n_۲ = ۴\ ,$

$n_۴ = n_۳ + w_۲ = 1\ ,\ n_۵ = -1 \cdot n_۴ = -1\ ,\ n_۶ = e^{n_۵} = ۰٫۳۶۷۹$

$n_۷ = 1 + n_۶ = ۱٫۳۶۷۹\ ,\ f = \frac{1}{n_۷} = ۰٫۷۳۱۱$

back prop:

$\frac{df}{dn_۷} = \frac{-1}{n_۷^۲} \simeq -۰٫۵۳۴\ ,\ \frac{dn_۷}{dn_۶} = 1\ ,\ \frac{dn_۶}{dn_۵} = e^{-1}\ ,$

$\frac{dn_۵}{dn_۴} = -1\ ,\ \frac{dn_۴}{dn_۳} = 1\ ,\ \frac{dn_۳}{dn_۲} = 1\ ,\ \frac{dn_۳}{dn_1} = 1$

به کمک ضرب جملات بالا داریم:

$$\frac{df}{dx_0} = \frac{df}{dn_1} \cdot \frac{dn_1}{dx_0} = ۰٫۳۹۳۴$$

$$\frac{df}{dw_0} = \frac{df}{dn_1} \cdot \frac{dn_1}{dw_0} = -۰٫۱۹۶۷$$

$$\frac{df}{dx_1} = \frac{df}{dn_۲} \cdot \frac{dn_۲}{dx_1} = -۰٫۵۹۰۱$$

$$\frac{df}{dw_1} = \frac{df}{dn_۲} \cdot \frac{dn_۲}{dw_1} = -۰٫۳۹۳۴$$

$$\frac{df}{dw_۲} = \frac{df}{dn_۴} \cdot \frac{dn_۴}{dw_۲} = ۰٫۱۹۶۷$$

۸) آ)

I. 2

II. 3

III. 4

IV. 5

V. 4

VI. 3

output. 5

ب)

۸) ب) ۱)

$$\frac{\partial y}{\partial a} = \frac{\partial y}{\partial s_1} \times \frac{\partial s_1}{\partial a} =$$

$$1 \times \frac{r_1 \cdot e^{-a \cdot r_1}}{(1+e^{-a \cdot r_1})^2} = \frac{2e^{-2}}{(1+e^{-2})^2} \simeq 0,21$$

$$\frac{\partial y}{\partial b} = \frac{\partial y}{\partial s_2} \times \frac{\partial s_2}{\partial b} = \frac{r_2 e^{-b r_2}}{(1+e^{-b r_2})^2} = 0$$

۲)

$$\frac{\partial r_1}{\partial c} = x_1 = 1 \quad , \quad \frac{\partial r_1}{\partial e} = x_2 = -1$$

۳)

$$\frac{\partial r_2}{\partial d} = 0 \quad , \quad \frac{\partial r_2}{\partial f} = 0$$

چون $max\ x$ صفر را انتخاب کرده.

۴)

$$\frac{\partial y}{\partial r_1} = \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial r_1} = 1 \times$$

$$= \frac{e^{-2}}{(1+e^{-2})^2} \simeq 0,105$$

$$\frac{\partial y}{\partial r_2} = \frac{\partial y}{\partial s_2} \cdot \frac{\partial s_2}{\partial r_2} = 1 \times \frac{b e^{-b r_2}}{(1+e^{-b r_2})^2} = \frac{e^{0}}{(1+e^{0})^2}$$

$$= \frac{1}{4}$$

۵)

$$\frac{\partial y}{\partial c} = \frac{\partial y}{\partial r_1} \cdot \frac{\partial r_1}{\partial c} = \frac{e^{-2}}{(1+e^{-2})^2} \simeq 0,105$$

$$\frac{\partial y}{\partial e} = \frac{\partial y}{\partial r_1} \cdot \frac{\partial r_1}{\partial e} = \frac{-e^{-2}}{(1+e^{-2})^2} \simeq -0,105$$

۶)

$$\frac{\partial y}{\partial d} = \frac{\partial y}{\partial r_2} \cdot \frac{\partial r_2}{\partial d} = 0$$

$$\frac{\partial y}{\partial f} = \frac{\partial y}{\partial r_2} \cdot \frac{\partial r_2}{\partial f} = 0$$

۷) $\eta = 1$

$a = 1 - 0,21 = 0,79$

$b = 1$

$c = 4 - 0,105 = 3,895$

$d = 1$

$e = 2 - (-0,105) = 2,105$

$f = 2$