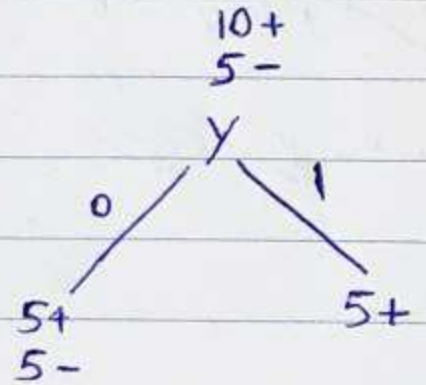
سوال ۱

الف) نادرست، نامساوی $H(X) \geq H(X|Y)$ همواره برقرار است ولی نامساوی $H(X|Y=y) \geq H(X)$ ممکن است صادق نباشد. مثال نقض:



در مثال روبرو، متغیر X می‌تواند + یا - باشد.

$$H(X) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$$

$$H(X|Y=0) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

$$H(X|Y=0) \not\geq H(X)$$

ب) درست، زیرا هیچ طبقه‌بند دیگری حضور ندارد که خطای آن کمتر از طبقه‌بند نیز باشد.

ج) درست، در یادگیری ماشین ما یک trade-off بین خطای بایاس و واریانس داریم و گاهی اوقات برای کاهش خطای واریانس، مقداری از بایاس را می‌پذیریم.

د) درست، زیرا با استفاده از روش دیگر نیز می‌توان دانش پیشین را وارد مدل کرد و فقط به داده‌ها متکی نبود.

هـ) درست، زیرا وقتی یک ویژگی حالت های زیادی داشته باشد، فرزندان آن شامل یک یا چند نمونه خواهند بود و لذا آنتروپی فرزندانش معمولاً نزدیک به صفر می‌شود و بیشترین information Gain را خواهد داشت ولی فروما این ویژگی، بهترین نیست و ممکن است دچار overfitting می‌شود.

سوال ۲

الف)

$$D = \{x_1, \cdots, x_n\}$$

$$\hat{\mu}_{MAP} = \arg\max_{\mu} P(\mu|D) = \arg\max_{\mu} P(D|\mu)\,P(\mu)$$

$$P(D|\mu) = \prod_{i=1}^{n} P(x_i|\mu) \Longrightarrow \ln P(D|\mu) = \sum_{i=1}^{n} \ln P(x_i|\mu)$$

$$= \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2}} = \underbrace{-\frac{n}{2}\ln 2\pi}_{\text{const.}} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \ln P(D|\mu) + \ln P(\mu)$$

$$= \arg\max_{\mu} \underbrace{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \ln\mu - \ln\sigma - \frac{\mu^2}{2\sigma^2}}_{g}$$

$$\frac{\partial g}{\partial \mu} = +\sum_{i=1}^{n}(x_i-\mu) + \frac{1}{\mu} - \frac{\mu}{\sigma^2} = 0$$

$$\Rightarrow +Z - n\mu + \frac{1}{\mu} - \frac{\mu}{\sigma^2} = 0$$

$$\Rightarrow R\mu^2 - Z\mu - 1 = 0$$

$$\Rightarrow \mu = \frac{+Z + \sqrt{Z^2+4R}}{2R}$$

$$\Rightarrow \boxed{\mu_{MAP} = \frac{Z}{2R}\left(1+\sqrt{1+\frac{4R}{Z^2}}\right)}$$

21 OCT

ب)

$$P(\mu|D) \propto P(D|\mu)\, P(\mu)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \quad \frac{\mu}{\sigma^2} e^{-\frac{\mu^2}{2\sigma^2}}$$

$$\propto \mu \exp\left(-\frac{\sum (x_i-\mu)^2 + \mu^2/\sigma^2}{2}\right)$$

همانطور که ملاحظه می‌شود، توزیع پسین دارای شکل متفاوتی از توزیع پیشین است. لذا conjugate نیست.

اگر فقط توضیح داده شود که conjugate نیست کفایت می‌کند و نمره کامل داده می‌شود

سوال ۳

هر دو کلاس نسبت‌های قرار دادن دو توزیع احتمال داده شده، برابر می‌باید:

$$\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) = \theta x \exp(-\theta x)$$

از دو طرف رابطه‌ی بالا $\ln$ می‌گیریم:

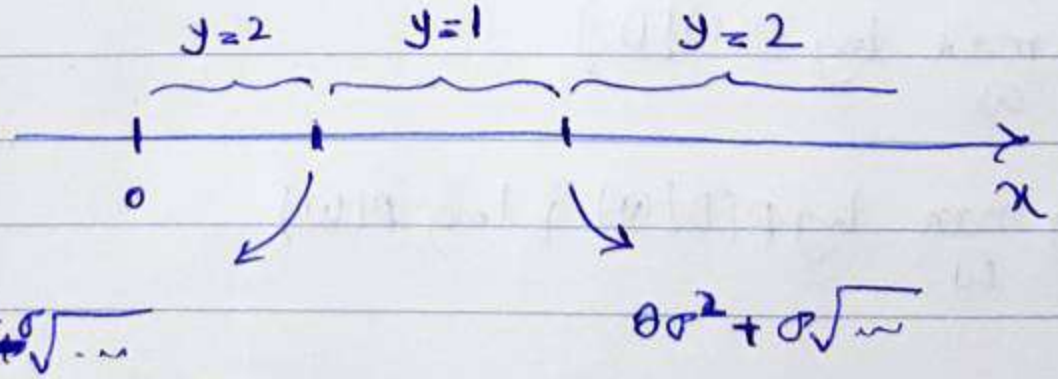
$$-2\ln\sigma - \frac{x^2}{2\sigma^2} = \ln\theta - \theta x$$

~~$\Rightarrow x^2 - 2\sigma^2\theta x + 2\sigma^2(\ln\theta\sigma^2) = 0$~~

$$\Rightarrow x^2 - 2\sigma^2\theta x + 2\sigma^2(\ln\theta\sigma^2) = 0$$

$$x = \sigma^2\theta \pm \sqrt{\sigma^4\theta^2 - 2\sigma^2\ln\theta\sigma^2}$$

$$\Rightarrow x = \theta\sigma^2 \pm \sigma\sqrt{\sigma^2\theta^2 - 2\ln\theta\sigma^2}$$



ممکن است براساس مقادیر θ و σ یکی از ریشه‌های معادله‌ی بالا منفی شود ولی نیازی به بررسی حالت‌های مختلف ریشه‌های معادله‌ی بالا نیست.

سوال ۴

با استفاده از GLM و رگرسیون خطی و یا با استفاده از لوکال رگرشن

راه دیگر این است که با استفاده از روش maximum likelihood مقدار

پارامتر w را تخمین بزنیم. هریک از روش‌های بالا اگر نوشته شود نمره کامل

داده می‌شود.

سوال ۵

$$\hat{w}_{MAP} = \arg\max_{w} \log P(w|D)$$

$$= \arg\max_{w} \log P(D|w) + \log P(w)$$

$$\log P(D|w) = \sum_{i=1}^{n} \log P(y_i|x_i,w) = \sum_{i=1}^{n} \underbrace{\log \frac{1}{\sqrt{2\pi}}}_{\text{const.}} - \frac{1}{2}(y_i - w^T x_i)^2$$

$$\Rightarrow \hat{w}_{MAP} = \arg\max_{w} -\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \underbrace{\log \frac{1}{(2b)^d}}_{\text{const.}} - \frac{|w|}{b}$$

$$\Rightarrow \hat{\omega}_{MAP} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - \omega^T x_i)^2 + \frac{2}{b} |\omega|$$

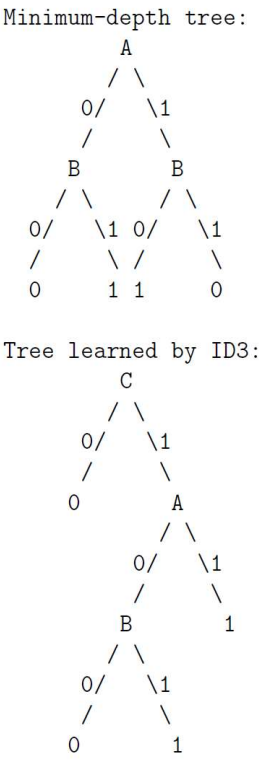
همانطور که ملاحظه می‌شود، λ با b رابطه‌ی عکس دارد.

سوال ۶ درخت تصمیم (۱۴ نمره)

فرض کنید که ویژگی باینری A، B و C به همراه برچسب کلاس هر داده بیان شده باشد (جدول زیر). می خواهیم درخت تصمیم با کمترین عمق را برای داده‌های داده شده بیابیم. آیا الگوریتم ID3 (بدون هرس کردن) درخت بهینه را می‌یابد؟ با رسم درخت‌های مربوطه جواب خود را شرح دهید.

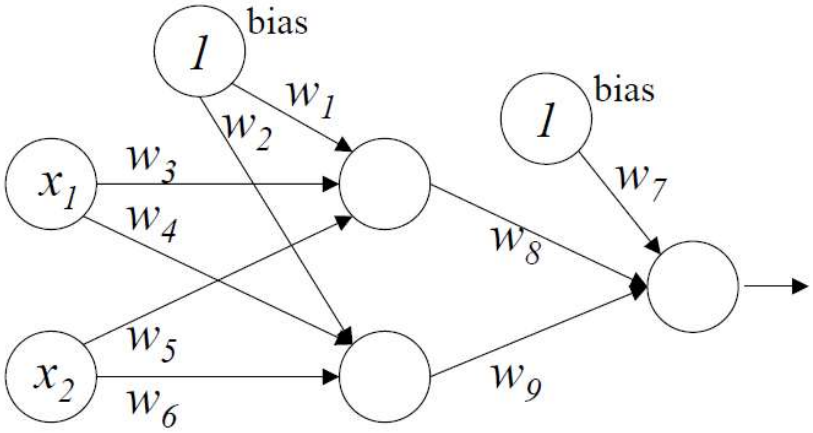
A	B	C	Class
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

جواب:



سوال ۷ شبکه عصبی (۱۲ نمره)

شبکه عصبی زیر را برای طبقه بندی دو کلاسه در نظر بگیرید. فرض کنید که لایه های میانی از تابع فعالسازی خطی ($h(z) = cz$) و تابع سیگموئید ($g(z) = \frac{1}{1+e^{-z}}$) در لایه خروجی استفاده می کند. این شبکه در صدد یادگیری $P(Y = 1|X, w)$ است که در آن $X = (x_1, x_2)$ و $W = (w_1, w_2, \dots, w_9)$ است.



الف) خروجی شبکه عصبی $P(Y = 1|X, w)$ را بر حسب پارامترهای شبکه (W ,x) و ثابت c نوشته و مرز تصمیم نهایی را به دست آورید

$$g(w_7 + w_8 h(w_1 + w_3 x_1 + w_5 x_2) + w_9 h(w_2 + w_4 x_1 + w_6 x_2))$$
$$= \frac{1}{1 + \exp(-(w_7 + cw_8 w_1 + cw_9 w_2 + (cw_8 w_3 + cw_9 w_4)x_1 + (cw_8 w_5 + cw_9 w_6)x_2))}$$

The classification boundary is :

$$w_7 + cw_8 w_1 + cw_9 w_2 + (cw_8 w_3 + cw_9 w_4)x_1 + (cw_8 w_5 + cw_9 w_6)x_2 = 0$$

ب) آیا می توان شبکه عصبی بدون لایه مخفی به دست آورد که معادل شبکه عصبی فوق باشد؟ در صورت وجود شبکه پیشنهادی را رسم کنید.