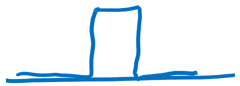
سوال ۱

الف) درست، اضافه کردن ترم منظم ساز باعث ساده تر شدن مدل می شود و لذا خطای بایاس افزایش می یابد. از طرفی، مدل های ساده تر دارای خطای واریانس کمتری هستند.

ب) غلط، روش کاهش گرادیان در صورت همگرایی، به مینیمم محلی می رسد و لذا اگر تابع convex نباشد لزوما به global min نمی رسد.

ج) غلط، میزان جستجو در ساختار داده ی KD-tree بستگی به پراکندگی داده ها دارد و لذا هزینه جستجو ثابت نیست.

د) درست، استفاده از پارامترهای مشترک باعث کاهش تعداد کل پارامترها می شود.

ه) درست، با استفاده از شبکه RBF می توان تابعی به شکل  را تقریب زد و لذا از کنار هم قرار گرفتن نورون های RBF می توان هر تابعی را تقریب زد. به عبارت دیگر، شبکه های RBF جزء universal func. approx هستند.

و) غلط، از آنجایی که شبکه های عصبی universal func. approx. هستند، خطای بایاس کمی دارند.

سوال ۲

مرز دو کلاس در نقاطی است که احتمال پسین دو کلاس مساوی باشد:

$$P(y=1|x) = P(y=2|x)$$

$$\Rightarrow \frac{P(x|y=1)\,\cancel{P(y=1)}}{\cancel{P(x)}} = \frac{P(x|y=2)\,\cancel{P(y=2)}}{\cancel{P(x)}} \Rightarrow P(x|y=1) = P(x|y=2)$$

$$\Rightarrow \frac{1}{\cancel{2\pi}\sqrt{\frac{1}{9\times 8}}} \exp\left\{-\frac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}9 & 0\\ 0 & 8\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right\} = \frac{1}{\cancel{2\pi}\sqrt{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right\}$$

$$\xrightarrow{\log} +\log\sqrt{72} - \frac{9}{2}x_1^2 - 4x_2^2 = +\log\sqrt{2} - x_1^2 - \frac{1}{2}x_2^2 \Rightarrow \frac{7}{2}x_1^2 + \frac{7}{2}x_2^2 = \log\sqrt{\frac{72}{2}}$$

$$\Rightarrow x_1^2 + x_2^2 = \frac{\cancel{2}}{\cancel{7}}\log 6 \longrightarrow$$

مرز جداکننده یک دایره به شعاع $\sqrt{\frac{\cancel{2}}{\cancel{7}}\log 6}$ است

سوال ۳

الف)

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log P(D|\theta) = \arg\max_{\theta} \log P(x_1,\dots,x_n|\theta) \overset{i.i.d}{=} \arg\max_{\theta} \log \prod_{i=1}^{n} P(x_i|\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \left(\log\theta + \theta\log b - (\theta+1)\log x_i\right)$$

$$= \arg\max_{\theta} \underbrace{n\log\theta + n\theta\log b - (\theta+1)\sum_{i=1}^{n}\log x_i}_{L(\theta)}$$

$$\frac{\partial L(\theta)}{\partial\theta} = \frac{n}{\theta} + n\log b - \sum_{i=1}^{n}\log x_i = 0 \implies \frac{n}{\theta} = \sum_{i=1}^{n}\log x_i - n\log b$$

$$\implies \boxed{\hat{\theta}_{ML} = \frac{n}{\sum_{i=1}^{n}\log x_i - n\log b}}$$

ب)

$$P(\theta|D) \propto P(D|\theta)\,P(\theta) \overset{i.i.d.}{=} \prod_{i=1}^{n} P(x_i|\theta)\,P(\theta) = \prod_{i=1}^{n}\left(\frac{\theta\, b^{\theta}}{x_i^{\theta+1}}\right) c\,\theta^{\alpha-1}\, e^{-\beta\theta}$$

$$= \frac{\theta^{n}\, b^{n\theta}}{\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}}\, c\,\theta^{\alpha-1}\, e^{-\beta\theta} = c\,\frac{\theta^{n+\alpha-1}\, e^{n\theta\ln b}\, e^{-\beta\theta}}{e^{(\theta+1)\ln\prod_{i=1}^{n} x_i}}$$

$$= \underbrace{\frac{c}{e^{\sum \ln x_i}}}_{c'}\, \theta^{n+\alpha-1}\, e^{-\theta\left(\beta - n\ln b + \sum_{i=1}^{n} \ln x_i\right)}$$

$$= \text{Gamma}\Big(\theta \,\Big|\, \underbrace{n+\alpha}_{\alpha_{new}},\ \underbrace{\beta - n\ln b + \sum_{i=1}^{n} \ln x_i}_{\beta_{new}}\Big)$$

ج) بله، همان طور که در قسمت قبل محاسبه شد، توزیع احتمال پسین از همان نوع توزیع احتمال پیشین شده است.

د)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D) = \frac{\alpha_{new}-1}{\beta_{new}} = \frac{n+\alpha-1}{\beta - n\log b + \sum_{i=1}^{n} \log x_i}$$

ه) بله، زیرا:

$$\lim_{n\to\infty} \hat{\theta}_{MAP} = \lim_{n\to\infty} \frac{n+\alpha-1}{\beta - n\log b + \sum_{i=1}^{n} \log x_i} = \lim_{n\to\infty} \frac{n}{-n\log b + \sum_{i=1}^{n} \log x_i} = \lim_{n\to\infty} \hat{\theta}_{ML}$$

سوال ۴

الف)

$$y_i = w^T x_i + \epsilon_i \Rightarrow P(y_i|x_i) = \text{laplace}(y_i \mid w^T x_i, 1) = \frac{1}{2} e^{-|y_i - w^T x_i|}$$

$$\hat{w}_{ML} = \arg\max_w \log P(D|w) = \arg\max_w \log \prod_{i=1}^{n} P(y_i|x_i) = \arg\max_w \sum_{i=1}^{n} \log P(y_i|x_i)$$

$$= \arg\max_w \sum_{i=1}^{n} \Big(\underbrace{-\log 2}_{\text{const}} - |y_i - w^T x_i| \Big) = \arg\max_w - \sum_{i=1}^{n} |y_i - w^T x_i|$$

$$\boxed{= \arg\min_w \sum_{i=1}^{n} |y_i - w^T x_i|}$$

ب)

$$\hat{w}_{MAP} = \arg\max_{w} \overset{\log}{} P(w|D) = \arg\max_{w} \overset{\log}{} P(D|w)\,P(w) = \arg\max_{w} \underbrace{\log P(D|w)}_{\text{در قسمت قبل حساب شد}} + \log P(w)$$

$$= \arg\max_{w} \; -\sum_{i=1}^{n} \Big(\underbrace{\log 2}_{\substack{\text{const}\\ \text{with respect to } w}} - |y_i - w^T x_i| \Big) - \underbrace{d \log 2b}_{\substack{\text{const}\\ \text{with respect to } w}} - \frac{\|w\|_1}{b}$$

$$= \arg\min_{w} \sum_{i=1}^{n} |y_i - w^T x_i| + \frac{\|w\|_1}{b}$$

لذا λ و b رابطه‌ی عکس دارند.

سوال ۵

1. [4 points] What is the entropy H(Passed)?

★ **ANSWER:**

$$H(Passed) = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6})$$
$$H(Passed) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$$
$$H(Passed) = \boxed{\log_2 3 - \frac{2}{3} \approx 0.92}$$

2. [4 points] What is the entropy H(Passed | GPA)?

★ **ANSWER:**

$$H(Passed|GPA) = -\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(1\log_2 1)$$
$$H(Passed|GPA) = \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0)$$
$$H(Passed|GPA) = \boxed{\frac{2}{3} \approx 0.66}$$

★ **ANSWER:**

$$H(Passed|Studied) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(1\log_2 1)$$
$$H(Passed|Studied) = \frac{1}{2}(\log_2 3 - \frac{2}{3}$$
$$H(Passed|Studied) = \boxed{\frac{1}{2}\log_2 3 - \frac{1}{3} \approx 0.46}$$

4. [4 points] Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

★ **ANSWER:** We want to split first on the variable which maximizes the information gain $H(Passed) - H(Passed|A)$. This is equivalent to minimizing $H(Passed|A)$, so we should split on "Studied?" first.