(1)

a) $f(x|\theta) = \frac{x}{\theta^2} \exp\left\{\frac{-x^2}{2\theta^2}\right\} \quad x \geq 0$

$\ell(\theta|x) = \prod f(x|\theta) \rightarrow L(\theta|x) = \log \prod f(x|\theta) = \sum \log f(x|\theta)$

$L(\theta|x) = \sum \left(\log x - 2\log\theta - \frac{1}{2}\cdot\frac{x^2}{\theta^2}\right)$

$= \log\sum x - 2N\log\theta - \frac{1}{2\theta^2}\sum x^2$

MLE for $\theta$: $\frac{\partial L}{\partial \theta} = \frac{-2N}{\theta} - \left(\frac{\sum x^2}{2}\cdot(-2)\theta^{-3}\right) = \frac{-2N}{\theta} + \frac{\sum x^2}{\theta^3} = 0$

$\times\theta^3 \rightarrow -2N\theta^2 + \sum x^2 = 0 \rightarrow \theta^2 = \frac{\sum x^2}{2N} \rightarrow \theta = \sqrt{\frac{\sum x^2}{2N}}$

b) $f(x|\alpha,\theta) = \alpha\theta^{-\alpha}x^{\alpha-1}\exp\left\{-\left(\frac{x}{\theta}\right)^{\alpha}\right\} \quad x\geq 0, \alpha>0, \theta>0$

$\ell(\theta,\alpha|x) = \prod f(x|\alpha,\theta) \quad L(\alpha,\theta|x) = \log\prod f(x|\alpha,\theta) = \sum \log f(x|\alpha,\theta)$

$L(\alpha,\theta|x) = \sum \left(\log\alpha - \alpha\log\theta + (\alpha-1)\log x - \left(\frac{x}{\theta}\right)^{\alpha}\right)$

$= \sum \left(\log\alpha - \alpha\log\theta + \alpha\log x - \log x - \frac{x^{\alpha}}{\theta^{\alpha}}\right)$

$= N\log\alpha - \alpha N\log\theta + \alpha\log\sum x - \log\sum x - \frac{1}{\theta^{\alpha}}\sum x^{\alpha}$

MLE for $\theta$: $\frac{\partial L}{\partial\theta} = \frac{-\alpha N}{\theta} - \left(\sum x^{\alpha}\cdot(-\alpha)\cdot\frac{1}{\theta^{\alpha+1}}\right) = \frac{-\alpha N}{\theta} + \frac{\alpha\sum x^{\alpha}}{\theta^{\alpha+1}} = 0$

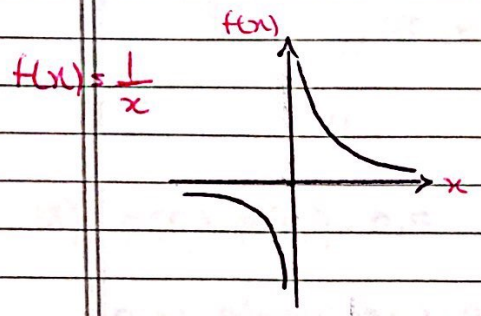
$*\theta^{\alpha+1} \rightarrow -\alpha N\theta^{\alpha} + \alpha\sum x^{\alpha} = 0 \rightarrow \theta^{\alpha} = \frac{\alpha\sum x^{\alpha}}{\alpha N}$

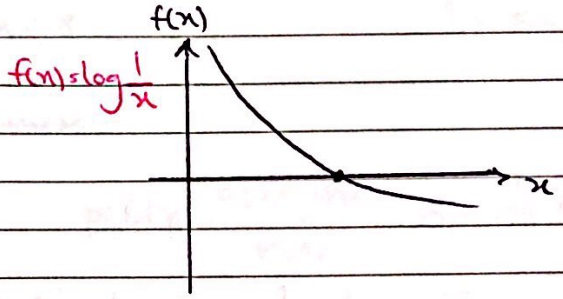
$\rightarrow \theta = \sqrt[\alpha]{\frac{\sum x^{\alpha}}{N}}$

(1)

c) $f(x|\theta) = \frac{1}{\theta} \quad 0 \leq x \leq \theta \;, \theta > 0$

$L(\theta|x) = \prod f(x|\theta) \qquad \ell(\theta|x) = \log \prod f(x|\theta) = \log \frac{1}{\theta^N}$




The goal is to maximize our likelihood which is maximizing $L(\theta|x)$. By looking at the right graph, MLE gets maximized when we go the most left side of horizontal axis. Then we can say that $\hat{\theta}$ should be minimized to get max MLE.

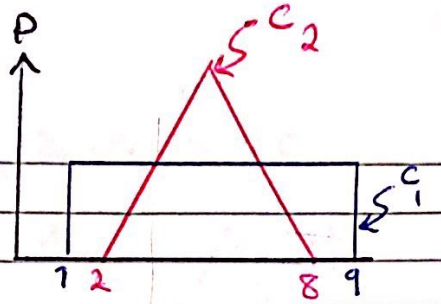
$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{\theta^N} \quad 0 \leq x \leq \theta$

Since $0 \leq x \leq \theta$, then we can conclude that $\hat{\theta}$ will be minimized if $x = \theta$

$\hat{\theta} = \max\{x_1, x_2, \ldots, x_n\} \quad 0 \leq x \leq \theta \rightarrow \max x = \theta$

(2)

$$P(x|C_1) = \begin{cases} 1/8 & 1 \le x < 9 \\ 0 & \text{otherwise} \end{cases}$$




$$P(x|C_2) = \begin{cases} \frac{1}{9}(x-2) & 2 \le x < 5 \\ \frac{1}{9}(8-x) & 5 \le x < 8 \\ 0 & \text{otherwise} \end{cases} \qquad 1 \le x < 9$$

a) $P(C_1) = P(C_2) = 0.5 \quad x = 4 \qquad P(C|x) = \frac{P(x|C)P(C)}{P(x)} \qquad P(x) = \sum_i P(x|C_i)P(C_i)$

$$P(x=4) = P(x=4|C_1)P(C_1) + P(x=4|C_2)P(C_2) = \frac{1}{8}\cdot\frac{1}{2} + \frac{2}{9}\cdot\frac{1}{2} = \frac{17}{144}$$

$$P(C_1|x=4) = \frac{P(x=4|C_1)P(C_1)}{P(x=4)} = \frac{\frac{1}{8}\cdot\frac{1}{2}}{\frac{17}{144}} = \frac{144}{272} = \frac{9}{17}$$

$$P(C_2|x=4) = \frac{P(x=4|C_2)P(C_2)}{P(x=4)} = \frac{\frac{2}{9}\cdot\frac{1}{2}}{\frac{17}{144}} = \frac{144}{153} = \frac{16}{17}$$

$P(C_2|x=4) > P(C_1|x=4) \Rightarrow x=4$ with equal priors is classified as $C_2$

b) $P(C_1) = 0.7 \quad P(C_2) = 0.3 \quad x = 6$

$$P(x=6) = P(x=6|C_1)P(C_1) + P(x=6|C_2)P(C_2) = \frac{7}{10}\cdot\frac{1}{8} + \frac{2}{9}\cdot\frac{3}{10} = \frac{111}{720}$$

$$P(C_1|x=6) = \frac{P(x=6|C_1)P(C_1)}{P(x=6)} = \frac{\frac{1}{8}\cdot\frac{7}{10}}{\frac{111}{720}} = \frac{21}{37}$$

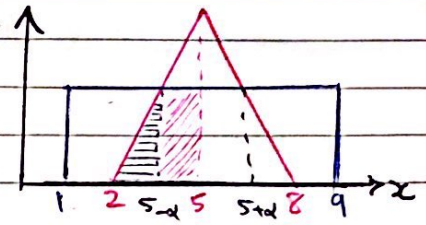
$$P(C_2|x=6) = \frac{P(x=6|C_2)P(C_2)}{P(x=6)} = \frac{\frac{2}{9}\cdot\frac{3}{10}}{\frac{111}{720}} = \frac{16}{37}$$

$P(C_1|x=6) > P(C_2|x=6) \Rightarrow x=6$ with different priors is classified as $C_1$

(2)

c) $\phi(x) = |x-5| - \alpha \quad 0 \le \alpha \le 2 \quad$ if $\phi(x) < 0 \rightarrow C_2$ & $\phi(x) \ge 0 \rightarrow C_1$

$P(C_1) = P(C_2) = 0.5$




$P(\text{missclassification}) = \int \left( P(C_1^*, C_2, x^t) + P(C_2^*, C_1, x^t) \right) dx$

$P(C_1^*, C_2, x^t)$ means the prediction is that $x^t$ belongs to $C_1$, but it actually belongs to $C_2$

$P(C_2^*, C_1, x^t)$ means the prediction is that $x^t$ belongs to $C_2$, but it actually belongs to $C_1$

$$P(C_1^*, C_2, x^t) = P(C_1^* | C_2, x^t) P(C_2, x^t) = P(C_1^* | x^t) P(x^t | C_2) P(C_2)$$

$$P(C_1^* | x^t) = \begin{cases} 1 & x \in [2, 5-\alpha] \ \& \ [5+\alpha, 8] \\ 0 & \text{otherwise} \end{cases}$$

$$P(C_2^* | x^t) = \begin{cases} 1 & x \in [5-\alpha, 5+\alpha] \\ 0 & \text{otherwise} \end{cases}$$

for minimizing the $\alpha$, we have to calculate the area of missclassification, Set it to zero, and Calculate $\alpha$.
Because of the symmetry around $x=5$, we can calculate one side and multiply it by 2.

$$P(\text{misclassification}) = \left( \int_2^{5-\alpha} P(x|C_2) P(C_2) \, dx + \int_{5-\alpha}^{5} P(x|C_1) P(C_1) \right) * 2$$

($P(C_2) = P(C_1) = 1/2$)

$$= 2 * \left[ \frac{1}{2} \int_2^{5-\alpha} \frac{1}{9}(x-2) \, dx + \frac{1}{2} \int_{5-\alpha}^{5} \frac{1}{8} \, dx \right] = \frac{1}{9} \left( \frac{x^2}{2} - 2x \right) \Big|_2^{5-\alpha} + \frac{1}{8} x \Big|_{5-\alpha}^{5}$$

Simplifying →

$$= \frac{1}{9} \left( \frac{(5-\alpha)^2}{2} - 2(5-\alpha) - 2 + 4 \right) + \frac{1}{8} (5 - 5 + \alpha) = \frac{1}{9} \left( \frac{25 - 10\alpha + \alpha^2 + 4\alpha - 16}{2} \right) + \frac{\alpha}{8}$$

$$= \frac{1}{9} \left( \frac{\alpha^2 - 6\alpha + 9}{2} \right) + \frac{\alpha}{8}$$

taking derivative $\frac{\partial}{\partial \alpha} \left[ \frac{1}{9} \left( \frac{\alpha^2 - 6\alpha + 9}{2} \right) + \frac{\alpha}{8} \right] = \frac{1}{9} \alpha - \frac{1}{3} + \frac{1}{8} = 0 \Rightarrow \alpha = \frac{15}{8} = 1.875$

(2)

C) Cont'd

I understood that $\alpha = 1.875$ is actually equals to the intersection of classes 1 and 2

$$\rightarrow \frac{1}{8} = \frac{1}{9}(x'-2) \rightarrow x'-2 = \frac{9}{8}, \quad x' = \frac{9}{8} + 2 = \frac{25}{8} = 3.125$$

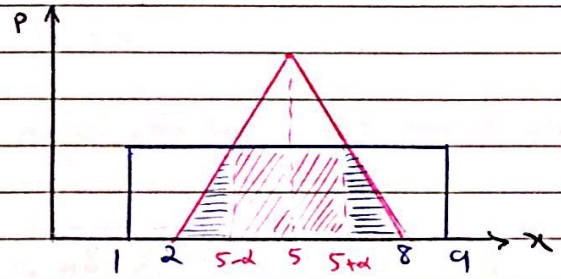
and if $5-\alpha = 5-1.875 = 3.125$

So in conclusion, best $\alpha$ for minimizing the probability of missclasification, is when $5-\alpha = x'$

$$p(\text{misclassification}) = 2 * \left( \int_2^{5-\alpha} \frac{1}{9}(x-2) \cdot \frac{1}{2}\, dx + \int_{5-\alpha}^{5} \frac{1}{8} \cdot \frac{1}{2}\, dx \right)$$

$$= \frac{1}{9}\left(\frac{(\alpha-3)^2}{2}\right) + \frac{\alpha}{8} = \frac{1}{18}\left(\frac{15}{8}-3\right)^2 + \frac{15}{8 \cdot 8} = \frac{39}{128}$$

for calculating misclassification I basically calculated the area below & multiplied it by 2.




for $x^t$ in:

→ blue areas are supposed to be classified as C1

→ Pink areas are supposed to be classified as C2

(3)

Assumption 1: S1 & S2 are learned from each class, shared cov = False

discriminant function $g_i(x) = -\frac{1}{2}\log|S_i| - \frac{1}{2}(x^T S_i^{-1} x - 2x^T S_i^{-1} m_i + m_i^T S_i^{-1} m_i) + \log \hat{P}(C_i)$

$= x^T W_i x + w_i^T x + w_{i0}$

$W_i = -\frac{1}{2} S_i^{-1} \quad w_i = S_i^{-1} m_i \quad w_{i0} = -\frac{1}{2} m_i^T S_i^{-1} m_i - \frac{1}{2}\log|S_i| + \log \hat{P}(C_i)$

C = My Confusion Matrix = $\begin{bmatrix} 23 & 7 \\ 7 & 63 \end{bmatrix}$ r is the result from test data

$C_{11}$ 23 : if prediction = 1 & r = 1

$C_{12}$ 7 : if prediction = 1 & r = 2

$C_{21}$ 7 if prediction = 2 & r = 1

$C_{22}$ 63 : if prediction = 2 & r = 2

for Class 1 we predicted 23 correctly out of 30 & for Class 2 we predicted 63 correctly out of 70. The total should be 100 = test data.

Assumption 2: S1 = S2 are learned from both classes, shared cov = True

Discriminant function $g_i(x) = -\frac{1}{2}(x - m_i)^T S^{-1}(x - m_i) + \log \hat{P}(C_i) = w_i^T x + w_{i0}$

$w_i = S^{-1} m_i \quad w_{i0} = -\frac{1}{2} m_i^T S^{-1} m_i + \log \hat{P}(C_i)$

S = Common Covariance Matrix = $\sum_i \hat{P}(C_i) S_i$ we have Si from each class

C = My Confusion Matrix = $\begin{bmatrix} 28 & 10 \\ 2 & 60 \end{bmatrix}$

$C_{11}$ 28 if prediction = 1 & r = 1

$C_{12}$ 10 if prediction = 1 & r = 2

$C_{21}$ 2 if prediction = 2 & r = 1

$C_{22}$ 60 if prediction = 2 & r = 2

for Class 1 we predicted 28 tests correctly out of 30 which is slightly better than assumption (1)

for Class 2 we predicted 60 correctly and 10 wrong which is slightly worse than assumption (1)

(3)

Assumption 3: $S_1 = S_2$ Covariance matrix is a diagonal matrix

Discriminant function

$$g_i(x) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j^t - m_{ij}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

We will be classifying based on weighted Euclidean distance to the nearest mean.

C = My confusion Matrix = $\begin{bmatrix} 24 & 6 \\ 6 & 64 \end{bmatrix}$

$C_{11}$ 24: if prediction = 1 & r = 1
$C_{12}$ 6: if prediction = 1 & r = 2
$C_{21}$ 6: if prediction = 2 & r = 1
$C_{22}$ 64: if prediction = 2 & r = 2

for Class 1 we predicted 24 tests correctly and 6 wrong
for Class 2 we predicted 64 tests correctly and 6 wrong.

Overall it is slightly better than assumption ① for both classes.
for class 1, better than assumption 1, worse than assumption 2
for class 2, better than assumptions 1 & 2

→ We can conclude that if our data points don't have any covariance meaning $cov(X,Y) = 0$ and they are independent, then we only get the variance for each data point & see how far they are from mean, and at the end we get a better prediction for our model. Our model won't be overfitted and will result in a better model.

Problem with assumption 1 is that we are having more parameters and also we are overfitting our model.

Assumption 2, we are having less parameters than assumption 1 so it resulted in a better model.

Assumption 3, not only we have the least amount of parameters, but also assumed data points are independent so overall we got the best prediction there.

→ With same training set, we reduced the model complexity and also reduced overfitting so we got better results!