1.

(a) Suppose $\{y_i\}$ are sorted, and there are k number $< t$, i.e.,

$$\begin{cases} y_i < t, & i < k \\ y_i \geq t, & i \geq k \end{cases}$$

Then $S(t) = \sum_{i=1}^{k} (t - y_i) + \sum_{i=k+1}^{N} (y_i - t)$

$S'(t) = k - (N-k)$

$= N - 2k$

Let $S'(t) = 0$, we have $k = \frac{N}{2}$, which means there should be $\frac{N}{2}$ number smal then t, and $\frac{N}{2}$ number greater or equal than t, so t is the median number in $\{y_i\}$.

(b) $S(t) = \sum_{i=1}^{N} (t - y_i)^2$

$S'(t) = \sum_{i=1}^{N} 2(t - y_i) = 0 \Rightarrow t = \frac{\sum_{i=1}^{N} y_i}{N} = \text{mean}(\{y_i\})$.

(c) Again, assume there are k numbers in $\{y_i\} < t$ and $\{y_i\}$ are sorted. $S(t) = \sum_{i=1}^{k} (t - y_i)^R + \sum_{i=k+1}^{N} (y_i - t)^R$

$S'(t) = \sum_{i=1}^{k} R(t - y_i)^{R-1} + \sum_{i=k+1}^{N} R(y_i - t)^{R-1}$

Let $S'(t) = 0$, we have $\sum_{i=1}^{k} (t - y_i)^{R-1} = -\sum_{i=k+1}^{N} (y_i - t)^{R-1}$

Among the left terms, $(t - y_1)^{R-1}$ is the largest number, and among the right terms, $(y_N - t)^{R-1}$ is the largest number, we divide both side by $(t - y_1)^{R-1} (y_N - t)^{R-1}$,

For any numbers $< (t - y_1)$, $\lim_{R \to \infty} \left(\frac{t - y_i}{t - y_1}\right)^{R-1} \to 0$,

and any numbers $< (y_N - t)$, $\lim_{R \to \infty} \left(\frac{y_i - t}{y_N - t}\right)^{R-1} \to 0$

Suppose we have a numbers equal to y_1, b numbers equal to y_N. then we have equal to y_N. then we have

$$\frac{a}{(y_N-t)^{R-1}} = \frac{b}{(t-y_1)^{R-1}}$$

$$b(y_N-t)^{R-1} = a(t-y_1)^{R-1}$$

$$[b^{\frac{1}{R-1}}\cdot(y_N-t)]^{R-1} = [a^{\frac{1}{R-1}}\cdot(t-y_1)]^{R-1}$$

$$\Rightarrow b^{\frac{1}{R-1}}\cdot(y_N-t) = a^{\frac{1}{R-1}}\cdot(t-y_1)$$

Because $\lim_{R\to\infty} b^{\frac{1}{R-1}} = b^0 = 1$, $\lim_{R\to\infty} a^{\frac{1}{R-1}} = a^0 = 1$

we get $t = (y_1+y_N)/2$, i.e., $\frac{\min(\{y_i\})+\max(\{y_i\})}{2}$.

2. Assume that samples in D are i.i.d., then

$$P(D|\mu,\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$l(\mu,\theta) = \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

let $\frac{\partial l}{\partial \mu} = 0$. we have

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \Rightarrow \sum_{i=1}^{N} x_i = N\mu \Rightarrow \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

let $\frac{\partial l}{\partial \sigma} = 0$, we have

$$\frac{\partial l}{\partial \sigma} = \sum_{i=1}^{N}\left(\sqrt{2\pi}\sigma\cdot\frac{-1}{\sqrt{2\pi}\sigma^2} + \frac{2(x_i-\mu)^2}{2\sigma^3}\right) = \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^3} - \frac{N}{\sigma} = 0$$

$$\Rightarrow N\sigma^2 = \sum_{i=1}^{N}(x_i-\mu)^2 \Rightarrow \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

$e^{-\mu}\frac{\mu^x}{x!}$

3.

(a) $P(\mu=1|D) = e^{-1}\cdot\frac{1^2}{2!}\cdot e^{-1}\frac{1^3}{3!}\cdot e^{-1}\frac{1^2}{2!}$

$$l(\mu) = -1+\ln(\tfrac{1}{2})-1+\ln(\tfrac{1}{6})-1+\ln(\tfrac{1}{2})$$

$$= 2\ln(\tfrac{1}{2})+\ln(\tfrac{1}{6})-3$$

(b) $l(\mu) = 2\ln(\frac{\mu^2}{2})+\ln(\frac{\mu^3}{6})-3\mu$

$$l'(\mu) = 2\ln(\mu^2)-2\ln(2)+\ln(\mu^3)-\ln(6)-3\mu$$

$$= 4\ln(\mu)+3\ln(\mu)-3\mu-\ln(24)$$

$$\xRightarrow{l'(\mu)=0} 7\ln(\mu)-3\mu = \ln(24)$$

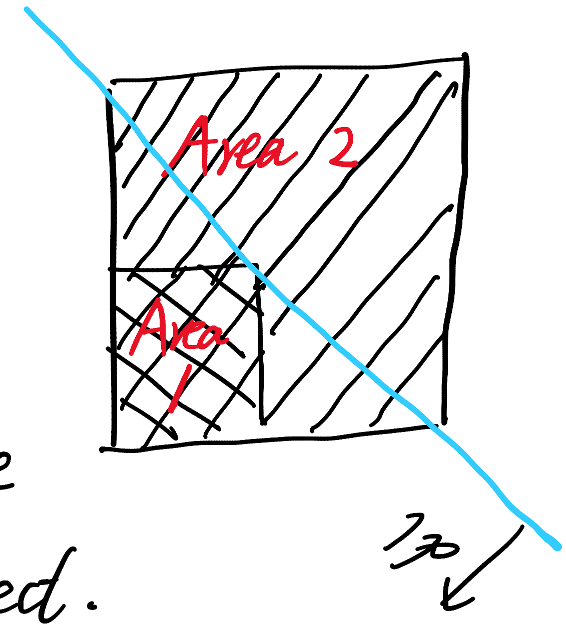
Don't know how to solve this equation.

4.

Please checkout the Jupyter notebook.

5.

(a). For any points in the area 1, the best guess is class 1, and any points in area 2, the best guess is class 2. thus the accuracy would be the chance that the guess is correct.



$$\text{Best Accuracy} = \frac{750}{2000} + \frac{1250}{2000} \cdot \frac{1000}{1250}$$

$$= 0.875$$

(b) For points in area 2. kNN will alway correct, for points in area 1, the correct chance will be

$$\frac{1000}{2000} \cdot \frac{1000}{1250} + \frac{250}{2000} \cdot \frac{250}{1250} = 0.4 + 0.025 = 0.425$$

So, the over all accuracy for kNN $= 0.425 + \frac{750}{2000} = 0.8$

(C) The first question to consider is, does the best linear separator cross Area 1? And if yes, where is the best cross position?

Denote the linear classifier as l. If l does not cross Area 1, then the best accuracy it can achieve is $1 - \frac{250}{1000} = 0.75$, because the prediction on half of the points of class two would be wrong. Now, if we want to achive better accuracy, apparently, l has to move towards the bottom left vertex of the square, as shown in the picture. Suppose the length of the big squre is 1, and l has moved x distance along the direction starting from the center of the big square, then we have

$$\mathrm{dArea(x)} = \sqrt{2}\mathrm{x} - \mathrm{x}^2$$
$$dArea1(x) = x^2$$
$$dArea2(x) = dArea(x) - dArea1(x) = \sqrt{2}x - 2x^2$$
$$0 \le x \le \frac{\sqrt{2}}{2}$$

Where $\mathrm{dArea(x)}$ is the area that l has swiped at x, $dArea1(x)$ is the swiped area that belongs to Area 1 and $dArea2(x)$ is the area that belongs to area 2, so we can calculate the accuracy at x

$$acc(x) = 0.75 - dArea1(x) \times \frac{1000}{2000} + dArea2(x) \times \frac{4000}{2000} = 0.75 - 2dArea1(x) + 0.5dArea2(x)$$

Let $acc'(x) = 0$, we get $x = \frac{0.5\sqrt{2}}{6}$. Thus the best accuracy for a linear classifier is

$$acc\left(\frac{0.5\sqrt{2}}{6}\right) = 0.75 - 2dArea1\left(\frac{0.5\sqrt{2}}{6}\right) + 0.5dArea2\left(\frac{0.5\sqrt{2}}{6}\right) = 0.792$$