Question 1

(1) True
(2) False
(3) True
(4) False
(5) False
(6) True
(7) False
(8) True
(9) False
(10) True
(11) True

Question 2

a) $\text{Posterior} = \dfrac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$

b) The posterior probability for a parameter is the prior probability of the parameter updated with information of the observations that has been incorporated into the likelihood.

c) Since my classifier performs poorly on the training set, I should first try to ensure that it performs well on the training set before trying to generalize to test sets. Thus, I should focus on improving the bias over the variance. Since boosting improves the bias and bagging does not, I should use boosting.

d) Boosting reduces the bias and increases the variance. Bagging leaves the bias unchanged and reduces the variance.

e) The assumption is given a class, the features in each pair of features are conditionally independent given the class.

f) The objective is to be minimized across all $\mathbf{m}_j$ and $r_j^{(i)}$ is

$$\sum_{i=1}^{N} \sum_{j=1}^{K} r_j^{(i)} ||\mathbf{m}_j - \mathbf{x}^{(i)}||^2$$

where N is the number of data points, K is the number of cluster centers, $\mathbf{m}_j$ is the jth cluster center, $\mathbf{x}^{(i)}$ is the ith data point, and $r_j^{(i)}$ is 1 is data point i is assigned to center $\mathbf{m}_j$ and 0 otherwise.

g)

$$Q^{\pi}(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} \mathcal{P}(s'|s,a) Q^{\pi}(s', \pi(s'))\, ds'$$

Question 3

a) k, the number of neighbours of each point used when voting for the most common label.

b) (i) number of layers in the network (ii) number of hidden units in any one of the layers.

c) λ, the ℓ_2 regularizer coefficient.

d) K, the number of dimensions of the subspace to project onto.

e) K, the number of cluster centers.

f) K, the number of weighted Gaussian distributions to include.

g) (i) α, the learning rate (ii) ϵ, the exploration probability.

Question 4

a) Let

$$\begin{aligned}
A_1 &= \text{the person has the disease} \\
A_2 &= \text{the person does not have the disease} \\
B_1 &= \text{the person tests positive} \\
B_2 &= \text{the person does not test positive}
\end{aligned}$$

Since 0.1% of the population has the disease, $P(A_1) = 0.001$ so $P(A_2) = 1 - 0.001 = 0.999$. Since the test is 99% accurate, $P(B_1|A_1) = 0.99$ and $P(B_2|A_2) = 0.99$ so $P(B_1|A_2) = 1 - 0.99 = 0.01$. The probability to be computed is $P(A_1|B_1)$. Using Bayes rule and law of total probability,

$$\begin{aligned}
P(A_1|B_1) &= \frac{P(B_1|A_1)P(A_1)}{P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2)} \\
&= \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999} \\
&= \frac{0.00099}{0.00099 + 0.00999} \\
&= \frac{0.00099}{0.01098} \\
&= \frac{99}{1098} \\
&= \frac{11}{122}
\end{aligned}$$

Question 4

b) Let

$$
\begin{aligned}
A_1 &= \text{the person has the disease} \\
A_2 &= \text{the person does not have the disease} \\
B_1 &= \text{the person tests positive the first time} \\
B_2 &= \text{the person tests positive the second time}
\end{aligned}
$$

Since 0.1% of the population has the disease, $P(A_1) = 0.001$ so $P(A_2) = 1 - 0.001 = 0.999$. Since the test is 99% accurate and the two runs of the test are independent given whether or not the person has the disease,

$$
\begin{aligned}
P(B_1|A_1) &= 0.99 \\
P(B_2|A_1) &= 0.99 \\
\Rightarrow P(B_1, B_2|A_1) &= P(B_1|A_1) \times P(B_2|A_1) \\
&= (0.99)^2 \qquad (1) \\
P(B_1^C|A_2) &= 0.99 \\
P(B_2^C|A_2) &= 0.99 \\
\Rightarrow P(B_1|A_2) &= 1 - 0.99 = 0.01 \\
\Rightarrow P(B_2|A_2) &= 1 - 0.99 = 0.01 \\
\Rightarrow P(B_1, B_2|A_2) &= (0.01)^2 \qquad (2)
\end{aligned}
$$

The probability to be computed is $P(A_1|B_1, B_2)$. Using Bayes rule, law of total probability, (1), and (2),

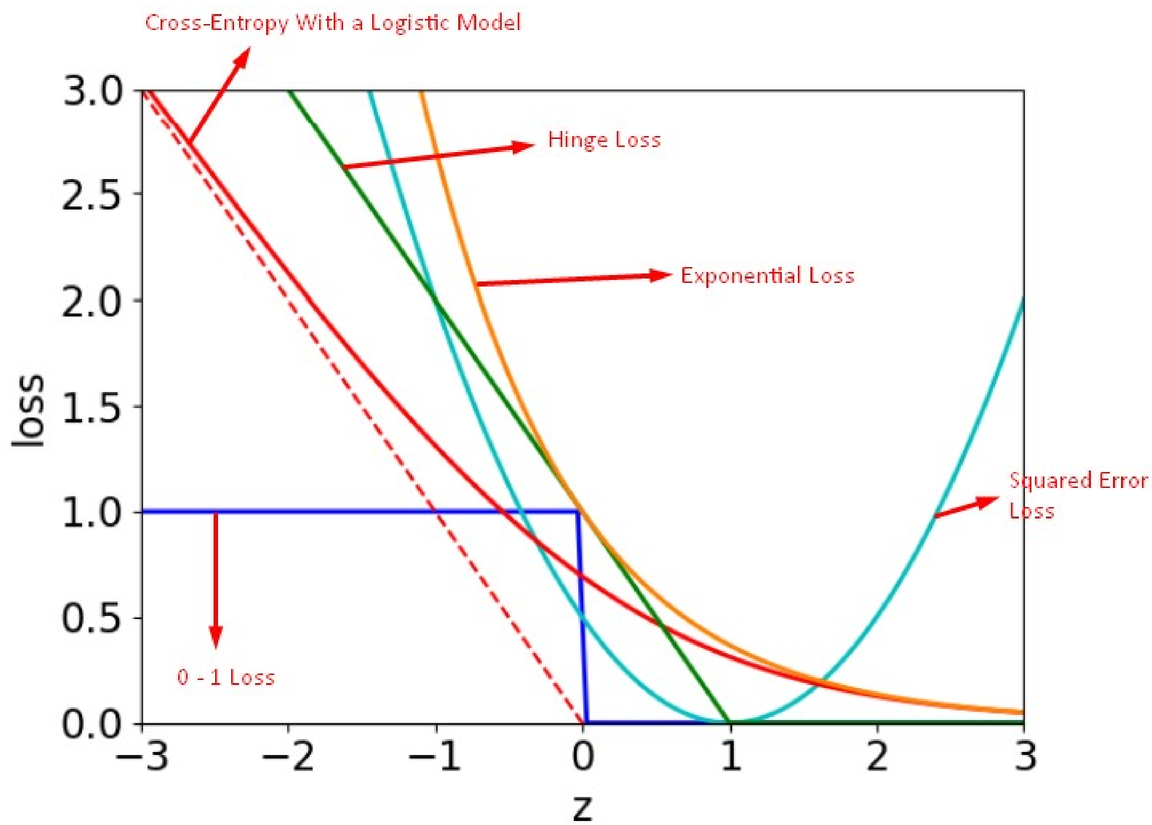
$$
\begin{aligned}
P(A_1|B_1, B_2) &= \frac{P(B_1, B_2|A_1)P(A_1)}{P(B_1, B_2|A_1)P(A_1) + P(B_1, B_2|A_2)P(A_2)} \\
&= \frac{(0.99)^2 \times 0.001}{(0.99)^2 \times 0.001 + (0.01)^2 \times 0.999} \\
&= \frac{0.9801 \times 0.001}{0.9801 \times 0.001 + 0.0001 \times 0.999} \\
&= \frac{0.0009801}{0.0009801 + 0.0000999} \\
&= \frac{0.0009801}{0.00108} \\
&= \frac{9801}{10800} \\
&= \frac{363}{400}
\end{aligned}
$$

Question 5

There will be 11 classes labels, which are $L_i = [0.1(i-1), 0.1i)$ for $i = 1, 2, 3, \ldots, 10$ and $L_{11} = \{1\}$. For a given data point $(\mathbf{x}, t) \in \mathcal{D}$, it will be transformed into $(\mathbf{x}, t') \in \mathcal{D}'$ where t' is the unique integer between 1 and 11 inclusively such that $t \in L_{t'}$. The multi-class classification problem is then to classify new data points with class labels L_i for $1 \leq i \leq 11$ given the data set $\mathcal{D}'$. If a point $(\mathbf{x}, t') \in \mathcal{D}'$ is classified using this new formulation, the prediction for the transformed point $(\mathbf{x}, t)$ from $\mathcal{D}$ is $0.1(t' - 1)$.

Question 6

a)



Cross-Entropy With a Logistic Model
Hinge Loss
Exponential Loss
Squared Error Loss
0 - 1 Loss
loss
z
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3

Question 6

b) The partial derivative of the 0-1 loss with any of the weights in the linear equation is 0 everywhere that it is defined. Thus, this loss function cannot be minimized using gradient descent because changing the weights by by a small amount proportional to the gradient would not change the loss whenever the gradient is defined.

c) Support vector machines uses hinge loss.

d) Adaboost can be interpreted as using an exponential loss.

Question 7

a) Substituting $z = w_1 x$ into $y = w_2 z$ gives $y = w_2(w_1 x) \Rightarrow y = (w_1 w_2)x$, which is the relation of the y and x. The one-layer NN consists of an input x and and a layer that transforms x and outputs $y = (w_1 w_2)x$.

b) Using the chain rule,

$$\begin{aligned}\frac{dl}{dw_2} &= \frac{dl}{dy}\frac{dy}{w_2} \\ &= \frac{1}{2}(2)(y-t)z \\ &= (y-t)z\end{aligned}$$

$$\begin{aligned}\frac{dl}{dw_1} &= \frac{dl}{dy}\frac{dy}{dw_1} \\ &= \frac{dl}{dy}\frac{dy}{dz}\frac{dz}{dw_1} \\ &= \frac{1}{2}(2)(y-t)w_2 x \\ &= (y-t)w_2 x\end{aligned}$$

c) The loss function with respect to w_1 and w_2 is found by substituting $y = w_1 w_2 x$ and $y = w_2 z$ into $\frac{1}{2}(y-t)^2$ to get $l_1(w_1) = \frac{1}{2}(w_1 w_2 x - t)^2$ and $l_2(w_2) = \frac{1}{2}(w_2 z - t)^2$. To test if $l_1(w_1)$ is convex, it is sufficient to check if $\frac{d^2 l_1}{dw_1^2} \geq 0$ for all w_1.

$$\begin{aligned}\frac{d^2 l_1}{dw_1^2} &= \frac{d^2}{dw_1^2}\frac{1}{2}(w_1 w_2 x - t)^2 \\ &= \frac{d}{dw_1}\frac{1}{2}(2)(w_1 w_2 x - t)(w_2 x) \\ &= (w_2 x)(w_2 x) \\ &= (w_2 x)^2 \geq 0\end{aligned}$$

Thus, the loss function with respect to w_1 is convex. To test if $l_2(w_2)$ is convex, it is sufficient to check if $\frac{d^2 l_2}{dw_2^2} \geq 0$ for all w_2.

$$\begin{aligned}\frac{d^2 l_2}{dw_2^2} &= \frac{d^2}{dw_2^2}\frac{1}{2}(w_2 z - t)^2 \\ &= \frac{d}{dw_2}\frac{1}{2}(2)(w_2 z - t)(z) \\ &= (z)(z) \\ &= z^2 \geq 0\end{aligned}$$

Thus, the loss function with respect to w_2 is convex.

Question 8

a) Since t is sampled from $\{0, 1\}$ with equal probability of $t = 0$ or $t = 1$, $P(t = 0) = \frac{1}{2}$ and $P(t = 1) = \frac{1}{2}$.

If $t = 0$, then x is sampled uniformly from $[0, 1]$ so the density of x given $t = 0$ is $P(x|t = 0) = \frac{1}{1-0} = 1$ for all $x \in [0, 1]$ and $P(x|t = 0) = 0$ for all $x \notin [0, 1]$.

If $t = 1$, then x is sampled uniformly from $[0, 2]$ so the density of x given $t = 1$ is $P(x|t = 1) = \frac{1}{2-0} = \frac{1}{2}$ for all $x \in [0, 2]$ and $P(x|t = 1) = 0$ for all $x \notin [0, 2]$.

Question 8

b) Using the formula for the posterior probability and (a),

$$\begin{aligned} P(t=0|x) &= \frac{P(x|t=0)P(t=0)}{P(x)} \\ &= \frac{P(x|t=0)}{2P(x)} \end{aligned} \tag{3}$$

Using the law of total probability,

$$\begin{aligned} P(x) &= P(x|t=0)P(t=0) + P(x|t=1)P(t=1) \\ \Rightarrow P(x) &= \frac{1}{2}P(x|t=0) + \frac{1}{2}P(x|t=1) \end{aligned} \tag{4}$$

If $x \in [0,1]$, then using a), (3), and (4) gives

$$\begin{aligned} P(x) &= \frac{1}{2}(1) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \\ &= \frac{1}{2} + \frac{1}{4} \\ &= \frac{3}{4} \\ \Rightarrow P(t=0|x) &= \frac{1}{2(3/4)} \\ &= \frac{2}{3} \end{aligned}$$

If $x \in [1,2]$, then using a) and (4) gives

$$\begin{aligned} P(x) &= \frac{1}{2}(0) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \\ &= 0 + \frac{1}{4} \\ &= \frac{1}{4} \\ \Rightarrow P(t=0|x) &= \frac{0}{2(1/4)} \\ &= 0 \end{aligned}$$

If $x \notin\in [0,1]$, then using a) and (4) gives

$$\begin{aligned} P(x) &= \frac{1}{2}(0) + \frac{1}{2}(0) \\ &= 0 + 0 \\ &= 0 \\ \Rightarrow P(t=0|x) &= \frac{0}{2(0)} \\ \Rightarrow P(t=0|x) &\text{ is undefined} \end{aligned}$$

Thus, $P(t=0|x)$ is $\frac{2}{3}$ if $x \in [0,1]$, is 0 if $x \in [1,2]$, and is undefined if $x \notin [0,2]$.

Question 9

a) Let the two classes be t_1 associated with Σ_1 and t_2 associated with Σ_2. Using the formula for $\log p(t_k|\mathbf{x})$ on slide 44/52 of the lecture 7 slides, the equation is

$$\log p(t_1|\mathbf{x}) = \log p(t_2|\mathbf{x})$$
$$\Rightarrow -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \log p(t_1) - \log p(\mathbf{x}) =$$
$$-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_2^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2) + \log p(t_2) - \log p(\mathbf{x})$$
$$\Rightarrow -\frac{1}{2}\log|\Sigma_1^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1) + \log p(t_1) = -\frac{1}{2}\log|\Sigma_2^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2) + \log p(t_2)$$

b) Replacing Σ_1 and Σ_2 with Σ in the equation in a), the decision boundary is

$$\Rightarrow -\frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1) + \log p(t_1) = -\frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\mu_2) + \log p(t_2)$$
$$\Rightarrow -\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1) + \log p(t_1) = -\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma^{-1}(\mathbf{x}-\mu_2) + \log p(t_2)$$

Using the expansion of $(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)$ on slide 44/52 of the lecture 7 slides, the decision boundary reduces to

$$-\frac{1}{2}(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mu_1^T\Sigma^{-1}\mathbf{x}) + \log p(t_1) = -\frac{1}{2}(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\mu_2^T\Sigma^{-1}\mathbf{x}) + \log p(t_2)$$
$$\Rightarrow -\frac{1}{2}(-2\mu_1^T\Sigma^{-1}\mathbf{x}) + \log p(t_1) = -\frac{1}{2}(-2\mu_2^T\Sigma^{-1}\mathbf{x}) + \log p(t_2)$$
$$\Rightarrow \mu_1^T\Sigma^{-1}\mathbf{x} + \log p(t_1) = \mu_2^T\Sigma^{-1}\mathbf{x} + \log p(t_2)$$
$$\Rightarrow (\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1})\mathbf{x} + (\log p(t_1) - \log p(t_2)) = 0$$

Since $(\mu_1^T\Sigma^{-1} - \mu_2^T\Sigma^{-1})$ is a dimension D vector and $(\log p(t_1) - \log p(t_2))$ is constant, the decision boundary is linear in $\mathbf{x}$.

Question 10

a) Since the rounds are independent, the likelihood is the product of the individual round likelihoods.

$$\begin{aligned} P(\mathcal{D}_N|\theta) &= \prod_{i=1}^{N} P(K = K_i|\theta) \\ &= \prod_{i=1}^{N} \theta(1-\theta)^{K_i} \\ &= \theta^N (1-\theta)^{\sum_{i=1}^{N} K_i} \end{aligned}$$

b)

$$\begin{aligned} \log P(\mathcal{D}_N|\theta) &= \log(\theta^N (1-\theta)^{\sum_{i=1}^{N} K_i}) \\ &= \log(\theta^N) + \log((1-\theta)^{\sum_{i=1}^{N} K_i}) \\ &= N \log(\theta) + \left(\sum_{i=1}^{N} K_i\right) \log(1-\theta) \end{aligned}$$

Question 10

c) The MLE occurs at a point where derivative of the log-likelihood is 0. The derivative of the log-likelihood is

$$\begin{aligned}
\frac{d}{d\theta}\log P(\mathcal{D}_N|\theta) &= \frac{d}{d\theta}\left(N\log(\theta) + \left(\sum_{i=1}^{N} K_i\right)\log(1-\theta)\right) \\
&= N\left(\frac{d}{d\theta}\log(\theta)\right) + \left(\sum_{i=1}^{N} K_i\right)\left(\frac{d}{d\theta}\log(1-\theta)\right) \\
&= \frac{N}{\theta} + \left(\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}(-1) \\
&= \frac{N}{\theta} - \left(\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}
\end{aligned}$$

Setting the derivative to 0 gives

$$\begin{aligned}
\frac{N}{\theta} - \left(\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta} &= 0 \\
\Rightarrow \frac{N}{\theta} &= \left(\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta} \\
\Rightarrow \frac{1-\theta}{\theta} &= \frac{\sum_{i=1}^{N} K_i}{N} \\
\Rightarrow \frac{1}{\theta} - 1 &= \frac{\sum_{i=1}^{N} K_i}{N} \\
\Rightarrow \frac{1}{\theta} &= \frac{N + \sum_{i=1}^{N} K_i}{N} \\
\Rightarrow \theta &= \frac{N}{N + \sum_{i=1}^{N} K_i}
\end{aligned}$$

To determine if this is a local maximum, the second derivative will be computed.

$$\begin{aligned}
\frac{d^2}{d\theta^2}\log P(\mathcal{D}_N|\theta) &= \frac{d}{d\theta}\left(\frac{N}{\theta} - \left(\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}\right) \\
&= -\frac{N}{\theta^2} - \left(\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2}(-1)(1-\theta)' \\
&= -\frac{N}{\theta^2} + \left(\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2}(-1) \\
&= -\frac{N}{\theta^2} - \left(\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2} < 0
\end{aligned}$$

The θ computed is thus a local maximum. Since it is the only local maximum, it is a global maximum so the MLE is $\dfrac{N}{N + \sum_{i=1}^{N} K_i}$.

Question 10

d) The prior belief can be rephrased as that the expected value of θ is less than $\frac{1}{1+r}$. Since θ follows a Beta(a, b) distribution, the expected value of θ is $\frac{a}{a+b}$. The condition then becomes

$$\begin{aligned} \frac{a}{a+b} &< \frac{1}{1+r} \\ \Rightarrow \frac{a+b}{a} &< 1+r \\ \Rightarrow 1+\frac{b}{a} &< 1+r \\ \Rightarrow \frac{b}{a} &< r \end{aligned}$$

One possible choice is $a = 1$ and $b = r/2$.

Question 10

e) Using 28/52 of the week 7 slides, the MAP estimator maximizes $\log p(\theta) + \log p(\mathcal{D}|\theta)$. Using the Beta$(a, b)$ prior and the log-likelihood found in b), this can be written as

$$\begin{aligned}
&\log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right) + N\log(\theta) + \left(\sum_{i=1}^{N} K_i\right)\log(1-\theta) \\
=&\log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) + \log(\theta^{a-1}) + \log(\theta^{b-1}) + N\log(\theta) + \left(\sum_{i=1}^{N} K_i\right)\log(1-\theta) \\
=&\log\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right) + (N+a-1)\log\theta + \left(b-1+\sum_{i=1}^{N} K_i\right)\log(1-\theta)
\end{aligned}$$

The maximum of the above expression occurs when its derivative is 0. Taking the derivative of the above expression gives

$$\begin{aligned}
\frac{d}{d\theta}\left(\log p(\theta) + \log p(\mathcal{D}|\theta)\right) =& \frac{N+a-1}{\theta} + \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}(-1) \\
=& \frac{N+a-1}{\theta} - \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}
\end{aligned}$$

Setting the above derivative to 0 gives

$$\begin{aligned}
\frac{N+a-1}{\theta} - \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta} &= 0 \\
\Rightarrow \frac{N+a-1}{\theta} &= \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta} \\
\Rightarrow \frac{1-\theta}{\theta} &= \frac{b-1+\sum_{i=1}^{N} K_i}{N+a-1} \\
\Rightarrow \frac{1}{\theta} - 1 &= \frac{b-1+\sum_{i=1}^{N} K_i}{N+a-1} \\
\Rightarrow \frac{1}{\theta} &= \frac{N+a-1+b-1+\sum_{i=1}^{N} K_i}{N+a-1} \\
\Rightarrow \theta &= \frac{N+a-1}{N+a+b-2+\sum_{i=1}^{N} K_i}
\end{aligned}$$

To determine if this is a local maximum, the second derivative will be computed.

$$
\begin{aligned}
\frac{d^2}{d\theta^2}\left(\log p(\theta) + \log p(\mathcal{D}|\theta)\right) &= \frac{d}{d\theta}\left(\frac{N+a-1}{\theta} - \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{1-\theta}\right) \\
&= -\frac{N+a-1}{\theta^2} - \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2}(-1)(1-\theta)' \\
&= -\frac{N}{\theta^2} + \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2}(-1) \\
&= -\frac{N}{\theta^2} - \left(b-1+\sum_{i=1}^{N} K_i\right)\frac{1}{(1-\theta)^2} < 0
\end{aligned}
$$

The θ computed is thus a local maximum. Since it is the only local maximum, it is a global maximum so the MAP estimate for θ is $\dfrac{N+a-1}{N+a+b-2+\sum_{i=1}^{N} K_i}$.

Question 10

f) By the formula for the posterior,

$$\begin{aligned} P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \\ &= \frac{P(\mathcal{D}|\theta)P(\theta)}{\int_{-\infty}^{\infty} P(\mathcal{D}|\theta')P(\theta')\,\theta'} \end{aligned}$$

Substituting the expression for the likelihood $P(\mathcal{D}|\theta)$ and the expression for the Beta(a,b) prior for $P(\theta)$ gives

$$\begin{aligned} P(\theta|\mathcal{D}) &= \frac{\left(\theta^N(1-\theta)^{\sum_{i=1}^N K_i}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right)}{\int_{-\infty}^{\infty}\left(\theta'^N(1-\theta')^{\sum_{i=1}^N K_i}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta'^{a-1}(1-\theta')^{b-1}\right)\,\theta'} \\ &= \frac{\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{N+a-1}(1-\theta)^{b-1+\sum_{i=1}^N K_i}\right)}{\int_{-\infty}^{\infty}\left(\theta'^N(1-\theta')^{\sum_{i=1}^N K_i}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta'^{a-1}(1-\theta')^{b-1}\right)\,\theta'} \end{aligned}$$

Since the posterior distribution is a Beta distribution, the exponents on the θ and $1-\theta$ indicate that the distribution is Beta$\left(N+a, b+\sum_{i=1}^N K_i\right)$ so

$$P(\theta|\mathcal{D}) = \frac{\Gamma(N+a+b+\sum_{i=1}^N K_i)}{\Gamma(N+a)\Gamma(b+\sum_{i=1}^N K_i)}\theta^{N+a-1}(1-\theta)^{b-1+\sum_{i=1}^N K_i}$$

g) The mean of a Beta(x,y) distribution is $\dfrac{x}{x+y}$ so the expected value of θ is found by substituting $x = N+a-1$ and $y = b-1+\sum_{i=1}^N K_i$ into the formula to get $\dfrac{N+a-1}{N+a-1+b-1+\sum_{i=1}^N K_i} = \dfrac{N+a-1}{N+a+b-2+\sum_{i=1}^N K_i}$.

Question 10

h) **MLE**: An advantage is that it is an optimization problem so it can be solved easily using gradient descent and the gradient operation is implemented by many software packages. A disadvantage it can give inaccurate results when there is little data.

MAP: An advantage is that it enables belief about a parameter to be incorporated into the prior which can give a more accurate estimate of the parameter. A disadvantage is that the information required to get a prior distribution can be scarce or wrong, both of which will reduce the accuracy of the parameter estimate.

Bayesian Estimation: An advantage is it works well when there is little data. A disadvantage is that it requires integration to compute the posterior mean of the parameter and it is more difficult to perform this using software tools compared to the MLE.