1.

(a) $L(\theta) = p(X_1|\theta)\,p(X_2|\theta)\cdots p(X_n|\theta) = \prod_{i=1}^{n} p(X_i|\theta)$

$= p^{X_1}(1-p)^{1-X_1}\, p^{X_2}(1-p)^{1-X_2} \cdots p^{X_n}(1-p)^{1-X_n} = \prod_{i=1}^{n} p^{X_i}(1-p)^{X_i}$

$= p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}$

The likelihood function does not depend on the order of observing random variables.

(b) $\ell(\theta) = \log(L(\theta)) = \log p^{\sum X_i}(1-p)^{n-\sum X_i} = \sum X_i \log p + (n-\sum X_i)\log(1-p)$

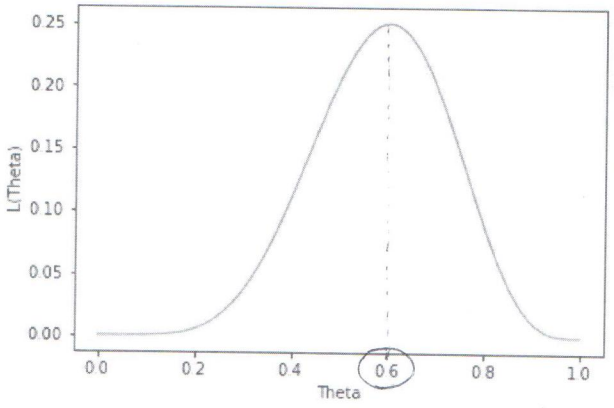
$$\frac{\partial \ell(\theta)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n-\sum_{i=1}^{n} X_i}{1-p}$$

The close form: $\frac{\partial \ell(\theta)}{\partial p} = 0$, $\frac{\sum X_i}{p} - \frac{n-\sum X_i}{1-p} = 0$

$(1-p)\sum X_i - np + p\sum X_i = 0$, $\sum X_i - np = 0$, $\Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$\frac{\partial^2 \ell(\theta)}{\partial p^2} = -\frac{\sum X_i}{p^2} - \frac{n-\sum X_i}{(1-p)^2}$$

(c)

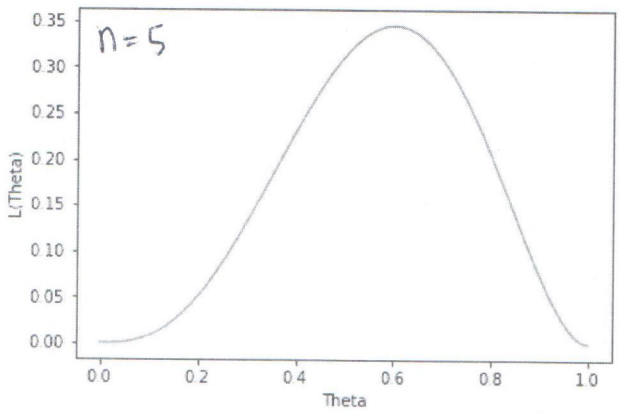


The close form answer:

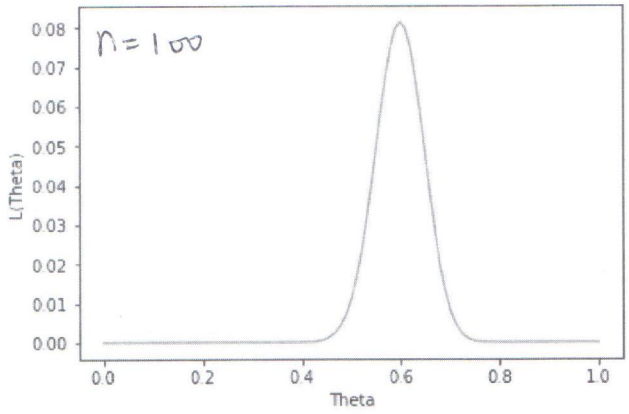
$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{6}{10} = 0.6$$

equal to the computational modeling answer

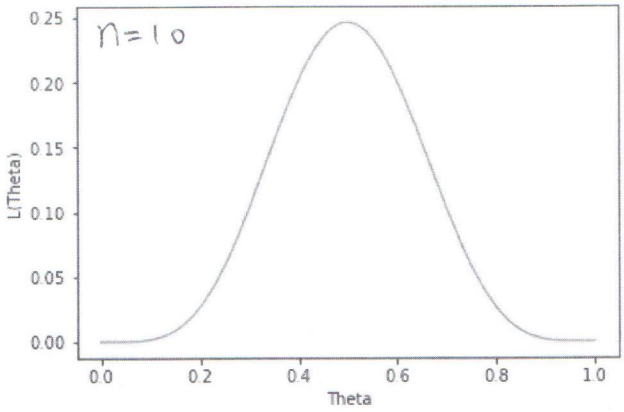
(d)



$\hat{p} = \frac{3}{5} = 0.6$



$\hat{p} = \frac{60}{100} = 0.6$



$\hat{p} = \frac{5}{10} = 0.5$

We can see that according to the likelihood function, the p can be estimated.

The comparison between first plot and the second plot is that the n of the second plot is much more larger than the first plot, which means that the number of multiplication is much more larger. And that's the reason cause the sharper shape of the second one comparing with the first one.

2. (a)

mistake probability: assuming all the $Y = 1$, then

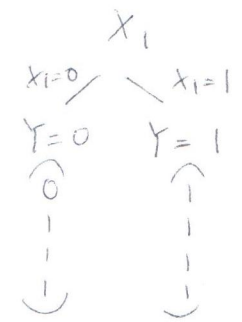
probability of mistake $= \frac{1}{2^3} = \frac{1}{8}$

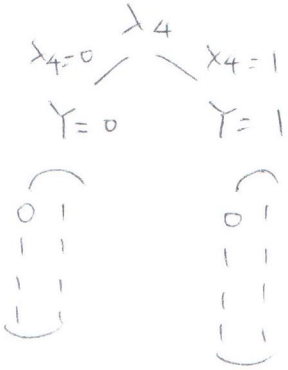
Since there is one condition mistake will happen when $X_1 = 0, X_2 = 0, X_3 = 0$

(b)

No. For example, if we split by X_1 or X_2 or X_3, then the mistake will be $\frac{3}{8}$; if we split by X_4, X_5, ..., then the mistake will be $\frac{7}{16} + \frac{1}{16} = \frac{1}{2}$

Both are larger than $\frac{1}{8}$.





(c) $H[X] = -\left[\frac{7}{8}\log_2\frac{7}{8} + \frac{1}{8}\log_2\frac{1}{8}\right]$

$= 0.5435$

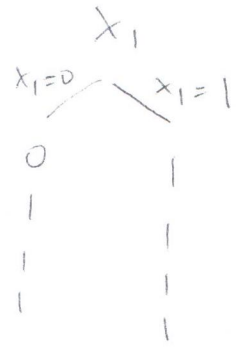
(d) If we split by X_1, then the entropy will be:

$$H[X|X_1] = -\left[\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right] \times \frac{1}{2} - \left[\frac{4}{4}\log\frac{4}{4} + \frac{0}{4}\log\frac{0}{4}\right] \times \frac{1}{2}$$

$= 0.4056$

$H[X] - H[X|X_1] = 0.5435 - 0.4056$

$= 0.1379$



reduce the entropy by splitting using X_1.

3. (a)

Before split, the entropy $H(S) = -\left[\frac{p}{p+n}\log\frac{p}{p+n} + \frac{n}{p+n}\log\frac{n}{p+n}\right]$

$$H[S|\lambda_j] = -\sum_{i=1}^{k}\left(\frac{p_i}{p_i+n_i}\log\frac{p_i}{p_i+n_i} + \frac{n_i}{p_i+n_i}\log\frac{n_i}{p_i+n_i}\right)p(S_i)$$

Since $\frac{p_k}{p_k+n_k}$ are all the same,

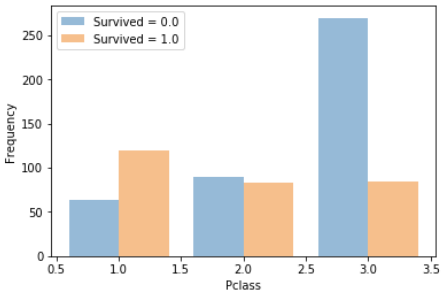
$$H[S|\lambda_j] = -\left(\frac{p}{p+n}\log\frac{p}{p+n} + \frac{n}{p+n}\log\frac{n}{p+n}\right)\sum_{i=1}^{k}p(S_i)$$

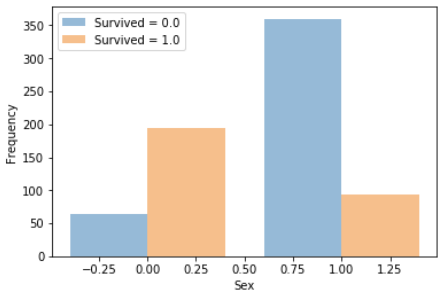
$$= -\left[\frac{p}{p+n}\log\frac{p}{p+n} + \frac{n}{p+n}\log\frac{n}{p+n}\right]\cdot 1$$

$$= -\left[\frac{p}{p+n}\log\frac{p}{p+n} + \frac{n}{p+n}\log\frac{n}{p+n}\right]$$

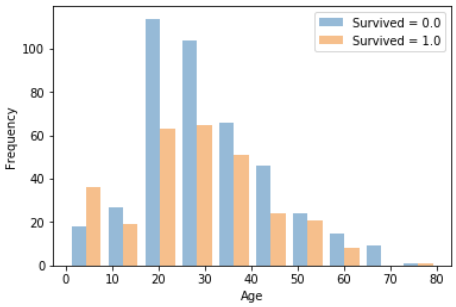
$$= H[S]$$

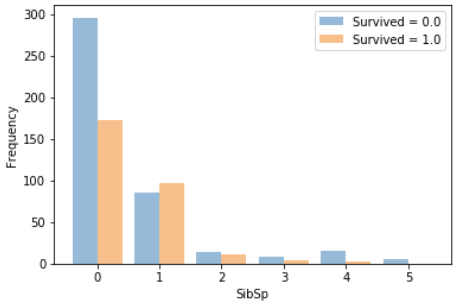
$$\Rightarrow H[S] - H[S|\lambda_j] = 0$$



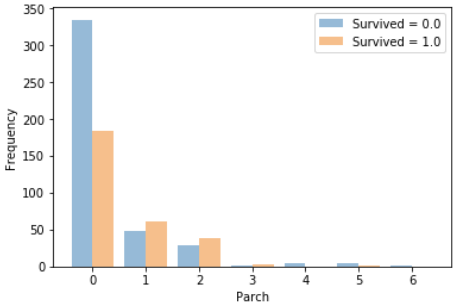


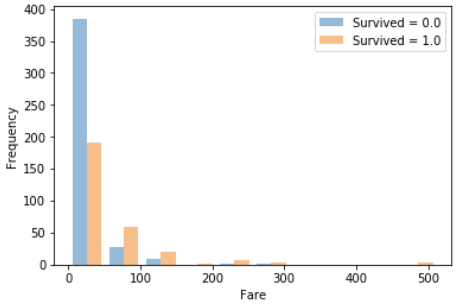
People in the upper class has the highest survival rate, while people in the lower class has the lowest survival rate. The survival rate of female (sex = 0) is higher than the survival rate of male (sex = 1).



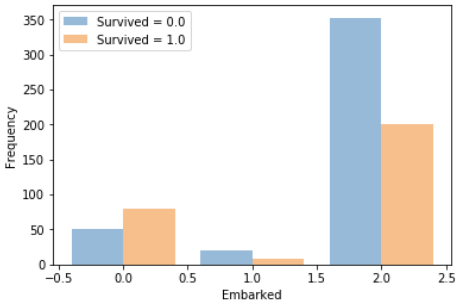


People under 10 years old has the highest survival rate, while people between 20 to 30 years old has the lowest survival rate. People without sibling and spouse has the lowest survival rate, while people with one sibling or spouse has the highest survival rate.





People without parent and children has the lowest survival rate, while people with one to three parent or children has higher survival rate. People with higher fare fee has higher survival rate, while people with lower fare has the lowest survival rate.



People embarked at Southampton has the lowest survival rate, while people embarked at Cherbourg has the highest survival rate.

(c)

Classifying using Decision Tree...

training error: 0.014

(d)

Classifying using Majority Vote... -- average training error: 0.397

-- average testing error: 0.434

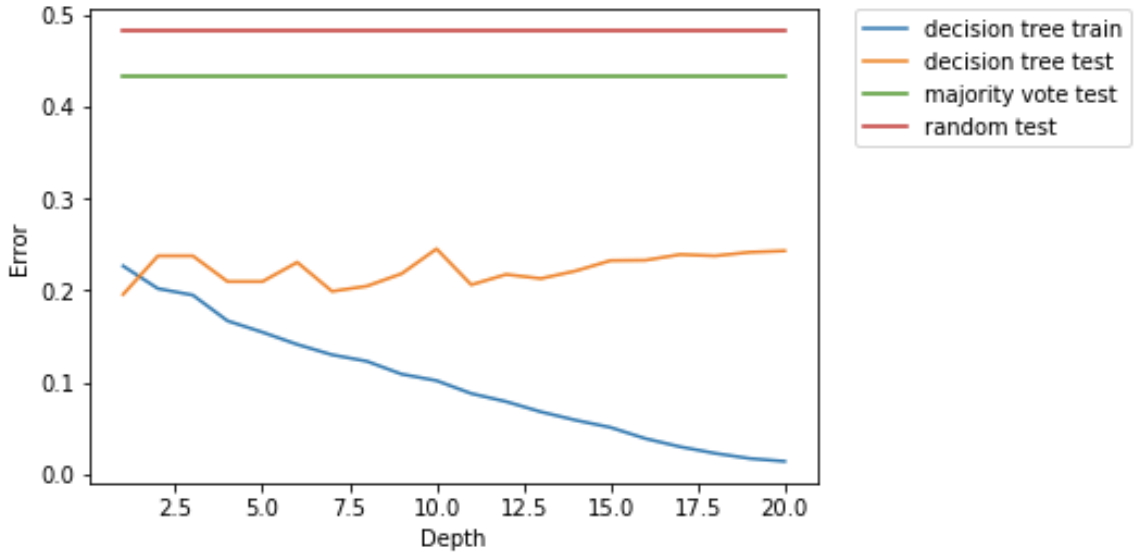
Classifying using Random... -- average training error: 0.517

-- average testing error: 0.483

Classifying using Decision Tree... -- average training error: 0.012

-- average testing error: 0.241

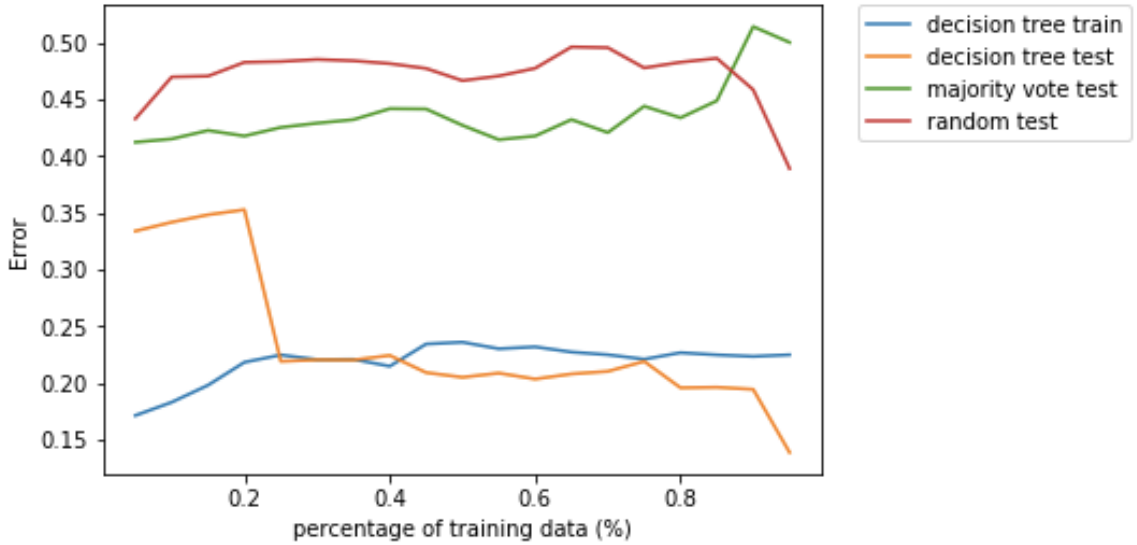
(e)



The best depth should be the depth with the smallest error of the decision tree testing set, which equals to one in the figure.

The overfitting happened, since we can see the validating curve doesn't decrease with the increasing depth as the training curve. Since the error of training set decrease when the depth increase, but the error of validating (testing) set doesn't, we could say that the overfitting happened when training the model.

(f)



We can see that while the percentage of training data increase, by using the decision tree as our model, the training error increase, but the testing error decrease. The best split is 0.95. The reason is at 95 % of training data, the model has the smallest testing error.