# CS M146 Problem Set 1

a) $L(\theta) = \theta^{a}(1-\theta)^{n-a}$ where $a$ is the number of independent random variables where $X_i = 1$

b) $\ell(\theta) = a\log_g\theta + (n-a)\log_g(1-\theta)$

$$\ell'(\theta) = \frac{a}{\theta} + \frac{(n-a)}{(1-\theta)}(-1) = \frac{a}{\theta} - \frac{(n-a)}{(1-\theta)}$$

$$\ell''(\theta) = \frac{-a}{\theta^2} - (n-a)(-1)\frac{1}{(1-\theta)^2}(-1)$$

$$= -\frac{a}{\theta^2} - (n-a)\frac{1}{(1-\theta)^2} \quad -(2)$$

$$\frac{a}{\theta} = \frac{n-a}{1-\theta}$$

$$(1-\theta)a = \theta(n-a)$$

$$a - a\theta = \theta n - a\theta$$

$$\theta = \frac{a}{n}$$

Substituting $\theta = \frac{a}{n}$ into (2),

$$\ell''(\theta) = \frac{-a}{\left(\frac{a}{n}\right)^2} - (n-a)\frac{1}{\left(1-\frac{a}{n}\right)^2}$$

$$\ell''(\theta) = -\frac{n^2}{q} - (n-q)\frac{1}{\left(\frac{n-q}{n}\right)^2}$$

$$= -\frac{n^2}{q} - (n-q)\frac{(n^2)}{(n-q)^2}$$

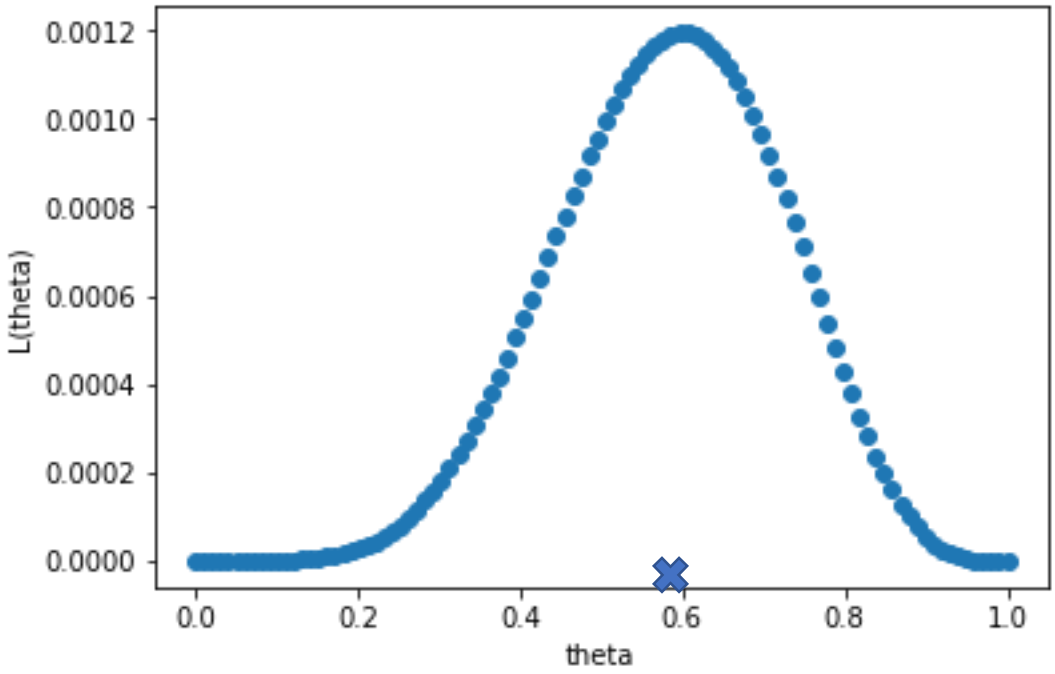
$$= -\frac{n^2}{q} - \frac{n^2}{(n-q)}$$

Since both $-\frac{n^2}{q}$ and $-\frac{n^2}{(n-q)}$ are $< 0$ for all $n > 1$,

$\ell''(\theta) < 0$ for $\theta = \frac{q}{n}$ and $\hat{\theta}_{MLE} = \frac{q}{n}$

c)

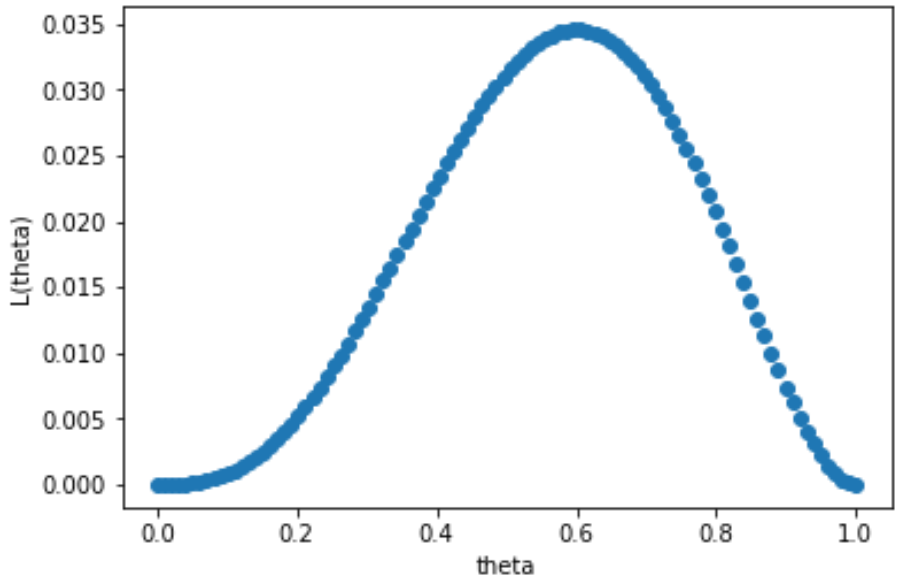
1c)



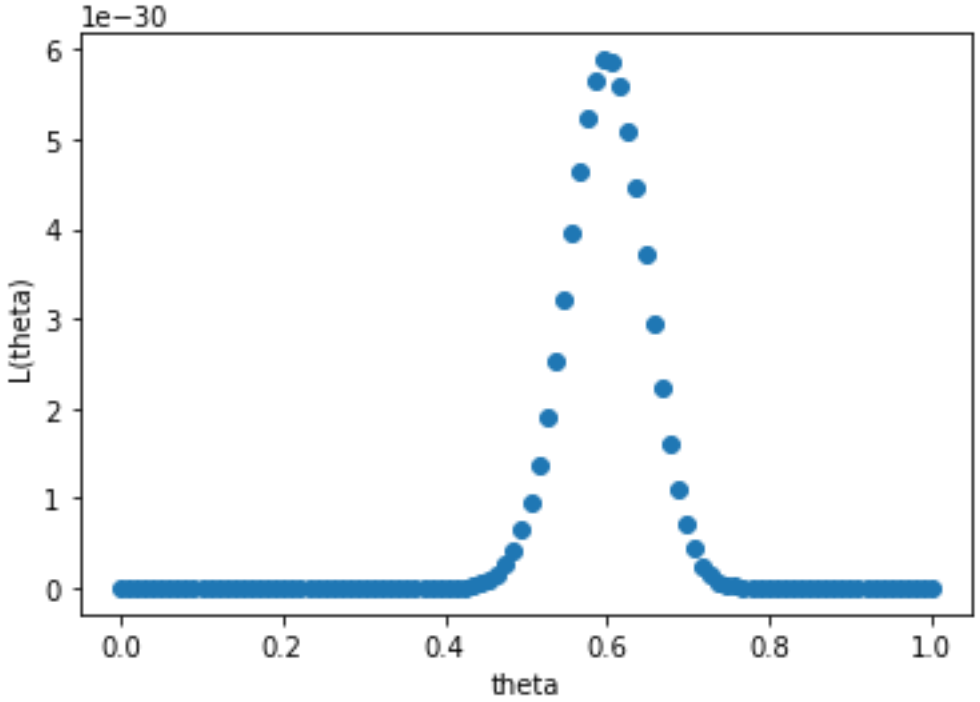

As seen from the plot, the value of θ that maximizes the likelihood is 0.6. This agrees with the closed form answer where $\theta = a/n = 6/10 = 0.6$

d)

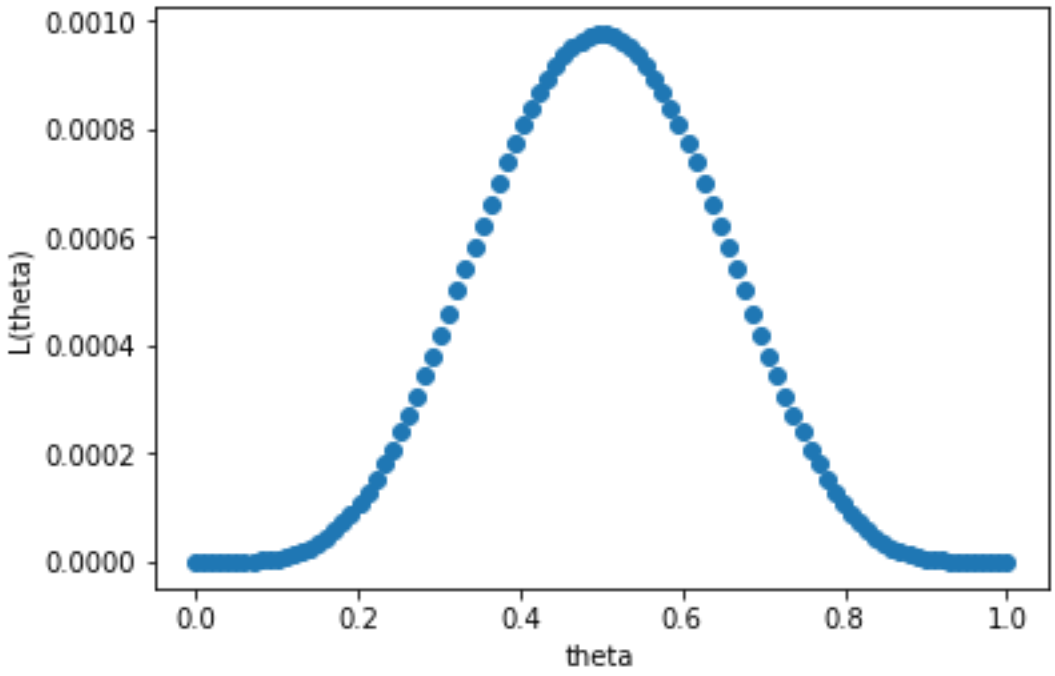
3 1s and 2 0s:



60 1s and 40 0s:




5 1s and 5 0s:



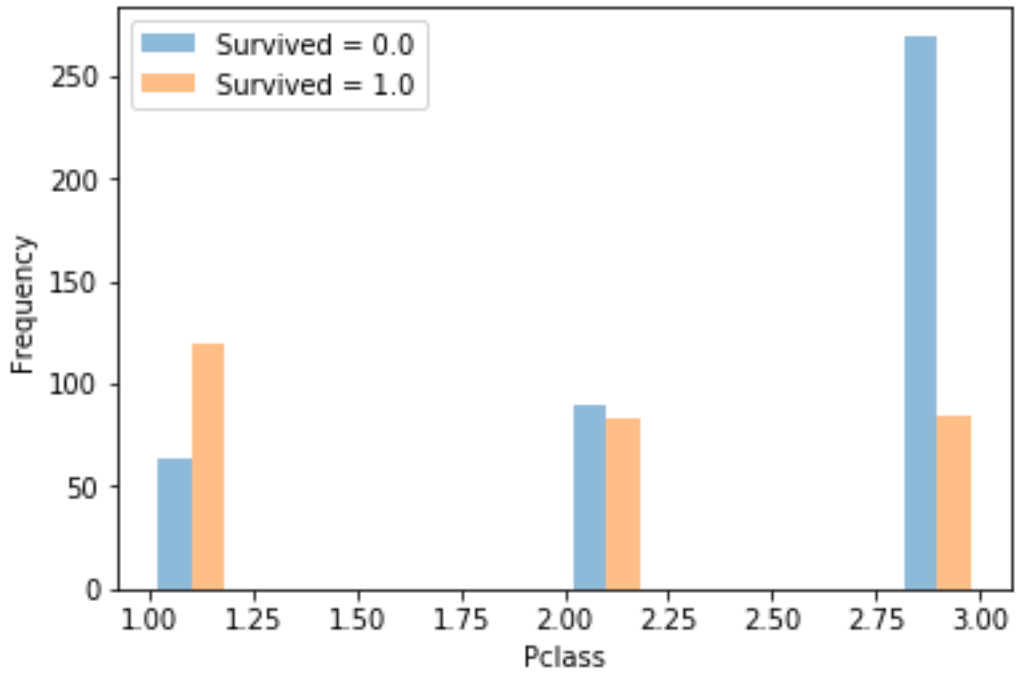

The maximum likelihood estimates for the first and second plot is the same as the ratio of a/n is the same, this can also be observed from the plot. The maximum likelihood estimate for the

third plot is lower (0.5) as also can be observed from the plot where the peak aligns with the middle of the x axis.

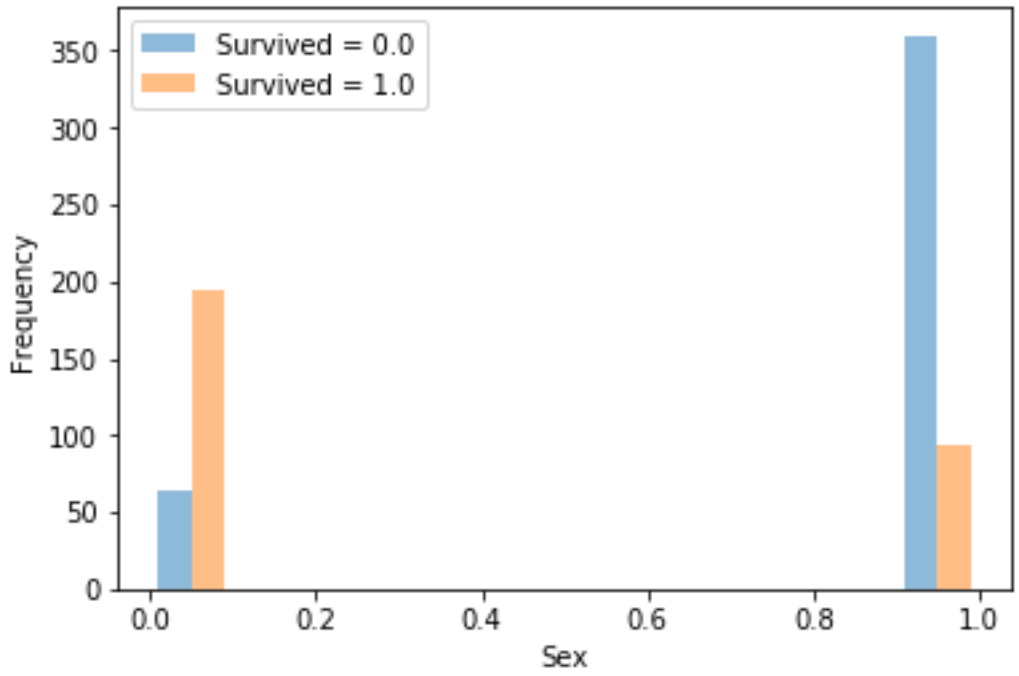
The likelihood function for the first plot is smoother compared to the third plot, with the second plot being the least smooth among the three i.e. it has the steepest change as theta gets closer to the maximum likelihood estimate.

5

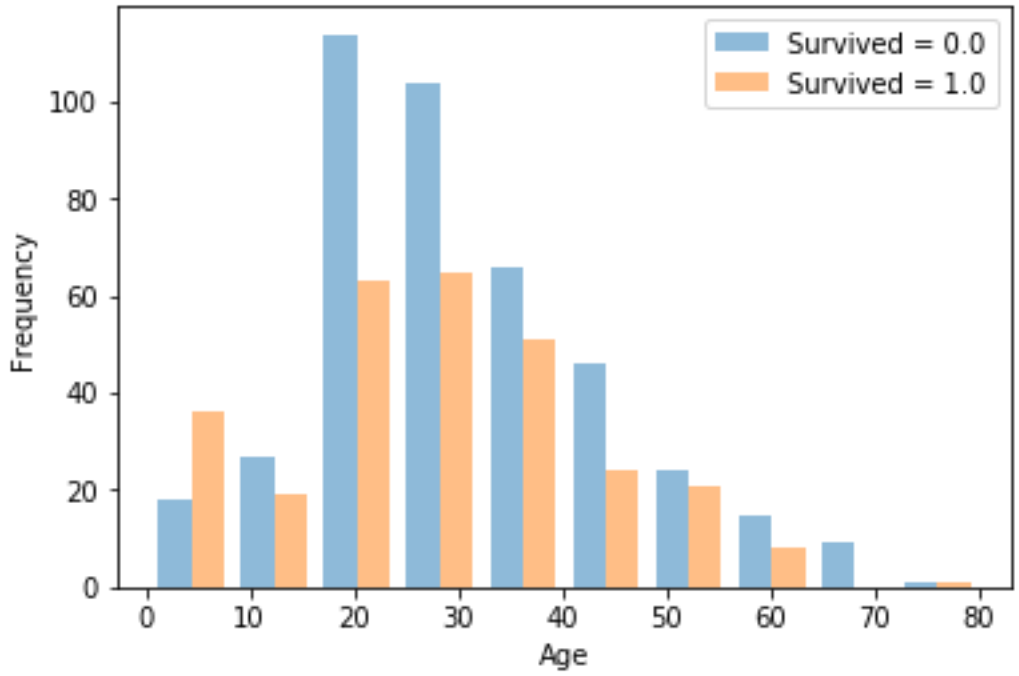
a) Survival rate = survived to not survived ratio within the class



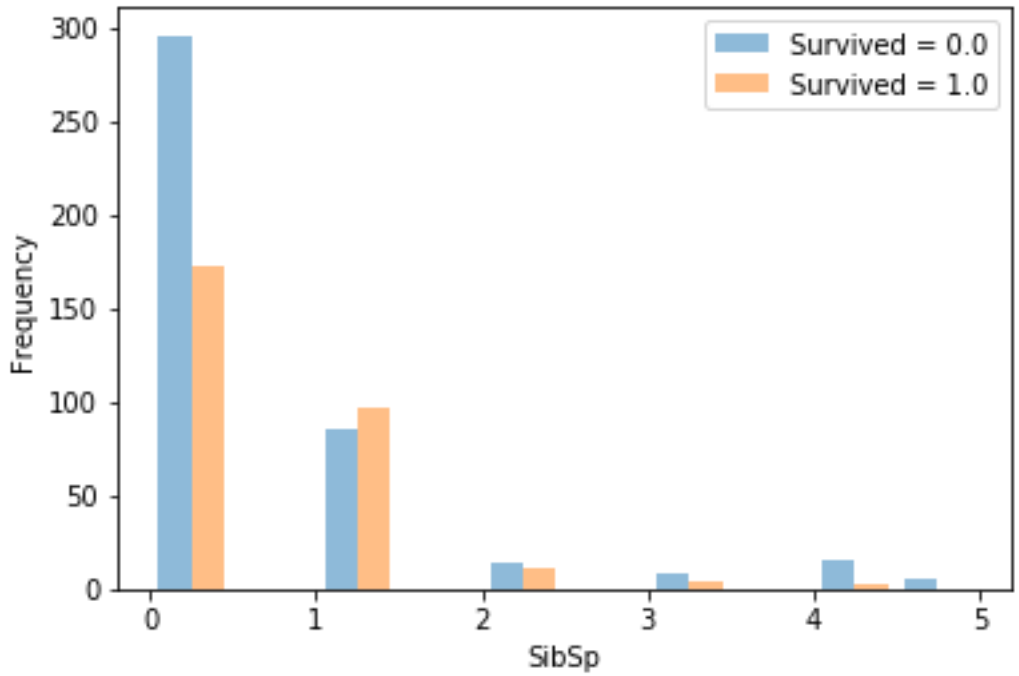

Higher PClass had higher survival rate (where PClass = 1 is higher than PClass = 3)



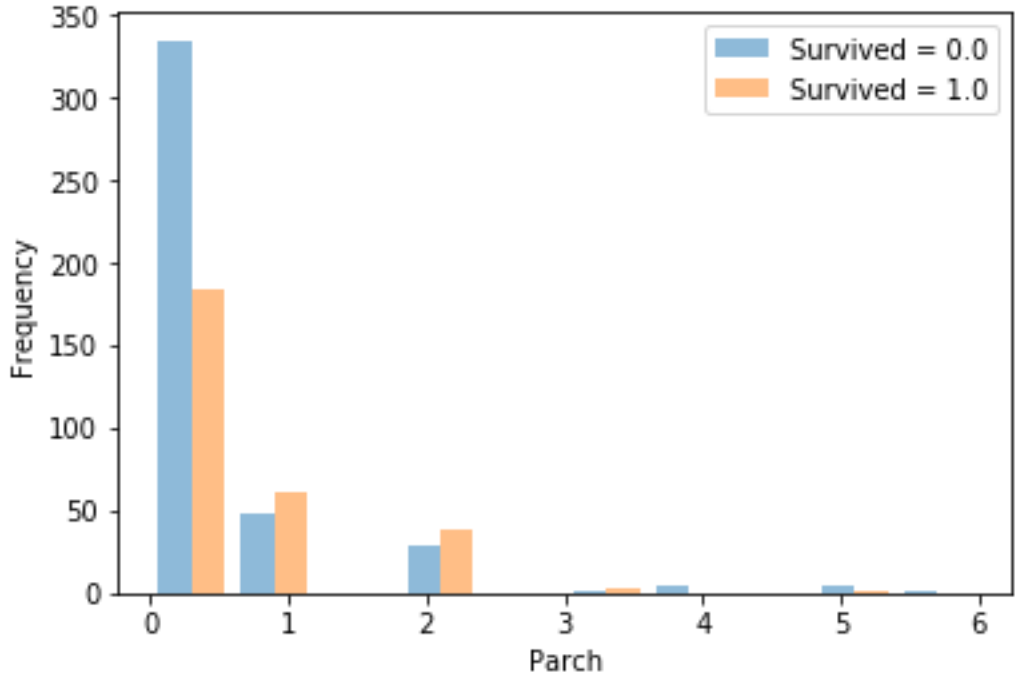

Females had higher survival rate than males



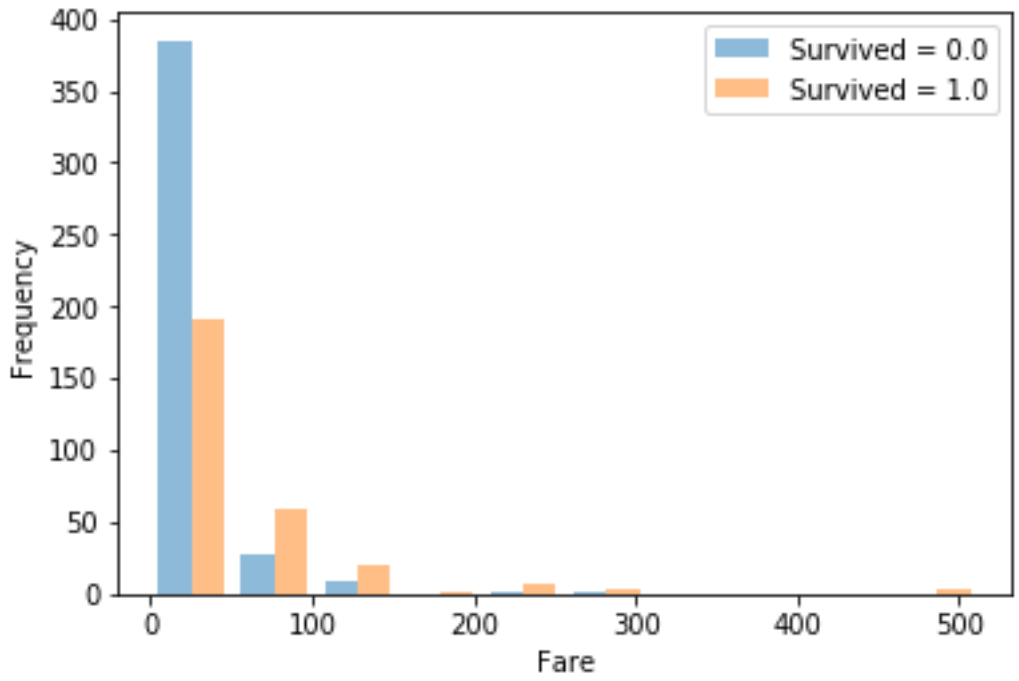

People aged 20-40 have the lowest survival rate but also consist of the majority of people and have more people who survived in absolute terms.



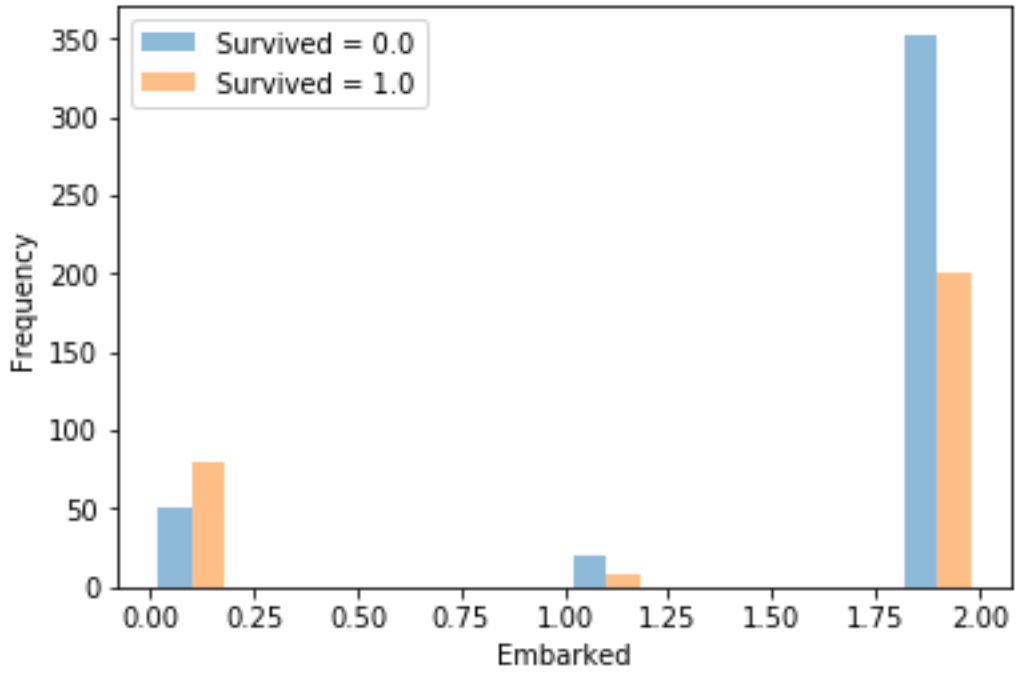

More people with 0 or 1 SibSp survived but people with 1 or 2 SibSp have highest survival rate.




People with 1 or 2 parent/child have higher survival rate but most people do not have any parent/child and consist of the majority of people who survived.




People with higher fare have higher survival rate but most people paid no fare.



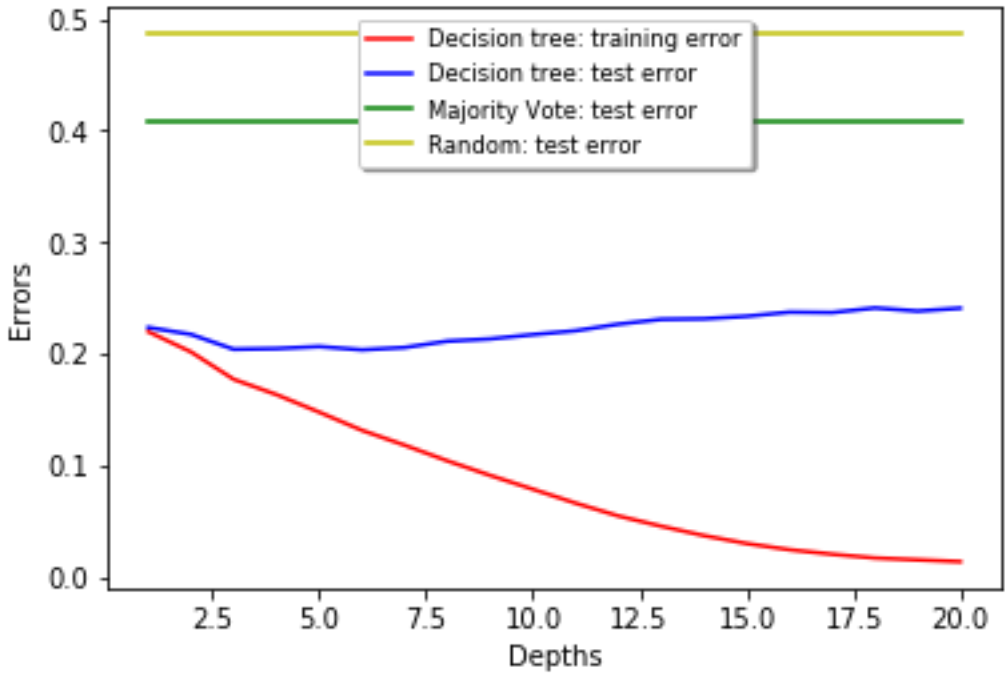

People who embarked at Cherbourg (Embarked= 0) have the highest survival rate but most people embarked at Southampton (Embarked = 2).

b) The training error is as expected, 0.485.

c) The training error is 0.014

d) Decision Tree error: training error = 0.012 test error = 0.240
Majority Vote error: training error = 0.404 test error = 0.407
Random error: training error = 0.489 test error = 0.487

e) The best depth is 3 which leads to lowest test error value of 0.204 as seen from the blue line on the plot.




Overfitting can be seen from the increasing divergence between the test and training errors (the blue and red line) as the depth increases. Even though the training error decreases at higher depths, the test error increases due to overfitting.

f)


Test error is always higher than training error. As the percentage of training data is increased, test error decreases while training error increases and hence the gap between them reduces.

2a) Number of mistakes = number of possible combinations

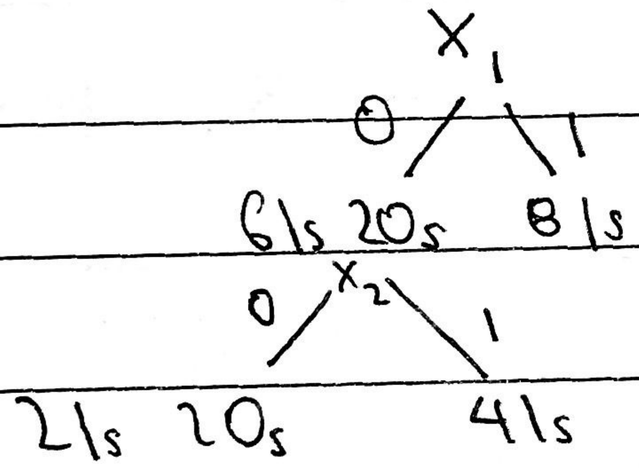
where $X_1 = 0, X_2 = 0, X_3 = 0$

for $n \geq 4$

$= 2^{n-3}$

b) Splitting on $X_i$ where $i > 3$ will not help reduce the number of mistakes as Y does not depend on it

Splitting on $X_i$ where $i \leq 3$ :




It will still leave with the same number of mistakes as values of $X_1, X_2$, and $X_3$ have to be known to make a correct guess

Typo

c) Entropy of $Y = -\sum_{k=1}^{k} P(Y=q_k) \log P(Y=q_k)$

$$= -\frac{2^{n-3}}{2^n} \log \frac{2^{n-3}}{2^n}$$

$$-\frac{2^n - 2^{n-3}}{2^n} \log \frac{2^n - 2^{n-3}}{2^n}$$

$$= -\frac{1}{8}\log\frac{1}{8} - \frac{7}{8}\log\frac{7}{8}$$

$$= 0.544$$

d) Yes, if we split at a given $X_i$ where $i \leq 3$

$$H[Y|X] = \left(\frac{1}{2}\right)\left(-\frac{2}{8}\log\frac{1}{4} - \frac{6}{8}\log\frac{3}{4}\right)$$

$$+ \left(\frac{1}{2}\right)(-0\log 0 - 1\log 1)$$

$$= 0.406$$

3) a) $H(S) = B\left(\frac{p}{p+n}\right)$

$0 \leq \frac{p}{p+n} \leq 1$

$B'(q) = -\frac{q}{q} - \frac{(1-q)}{(1-q)}(-1) - \log q - (-1)\log(1-q)$

$= -\log q - 1 + \log(1-q) + 1$

$= \log(1-q) - \log(q)$

$= \log\left(\frac{1-q}{q}\right)$

$\log\left(\frac{1-q}{q}\right) = 0$

$\frac{1-q}{q} = 1$

$q = \frac{1}{2}$

$B(q)$ is maximized when $q = \frac{1}{2}$

Hence $H(S) = B\left(\frac{p}{p+n}\right)$ is maximized when $\frac{p}{p+n} = \frac{1}{2}$

i.e. when $p = n$.

When $q = \frac{1}{2}$, $B(q) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = -1\log\left(\frac{1}{2}\right) = 1$

When $q = 0$ (or 1), $B(q) = 0$

$\therefore$ $0 \leq H(S) \leq 1$ and $H(S) = 1$ when $p = n$.

b) If the ratio $\frac{p_k}{p_k+n_k}$ is the same for all K subsets, the ratio of each subset = ratio of whole set.

i.e.

$$\frac{p}{p+n} = \frac{p_1+p_2+p_3+\cdots+p_k}{p_1+p_2+\cdots+p_k+n_1+n_2+\cdots+n_k}$$

$$\frac{p_k}{p_k+n_k} = \frac{\frac{1}{k}p}{\frac{1}{k}(p+n)} = \frac{p}{p+n}$$

Since $\frac{p_k}{p_k+n_k} = \frac{p}{p+n}$,

$$H[Y] = H[Y|X_j]$$

So $GAIN = H[Y] - H[Y|X_j] = 0$

$H[Y] = H[Y_j] = B\left(\frac{p}{p+n}\right)$ where $Y_j$ corresponds to one of the K subsets

$$H[Y|X_j] = H[Y_1]\frac{p_1+n_1}{p+n} + H[Y_2]\frac{p_2+n_2}{p+n} + \cdots + H[Y_k]\frac{p_k+n_k}{p+n}$$

$$= H[Y]\left(\frac{p_1+n_1+p_2+n_2+\cdots+p_k+n_k}{p+n}\right)$$

$$= H[Y]$$

4 a) $k=1$. The resulting training set error is 0 as the nearest neighbor is the point itself.

This is not a reasonable estimate of test set error as it will only be 0 if the test set is exactly the same as the training set and encourages overfitting.

b) & c) LOOCV errors:

$k=1$, error $= \frac{0+1+1+1+1+1+0+0+1+1+1+1+1+0}{14}$

$= \frac{10}{14} = 0.714$

$k=3$, error $= \frac{6}{14} = 0.4286$

$k=5$, error $= \frac{4}{14} = 0.2857$

$k=7$, error $= \frac{4}{14} = 0.2857$

$k=9$, error $= \frac{6}{14}$

$k=13$, error $= \frac{14}{14} = 1$

$k=5/7$ minimizes the LOOCV error, the resulting error is 0.2857

Typo

Cross validation is a better measure of test set performance as performance is measured based on unseen & test set data and is averaged based on different sets of test data. It prevents overfitting.

The LOOCV for K=1 is 0.714 and for K=13 is 1. Too small a value of K may lead to greater susceptibility to anomalies and too large a value misclassification due to taking into account unrelated data points and classifying based on data further away