CS188, Winter 2017
Problem Set 1:
Due 2/2/2017

Shannon Phu

1 Problem 1

(a)

$$P(x_i|\Theta) = \Theta^{x_i}(1-\Theta^{x_i}) \tag{1}$$

$$L(\Theta) = P(x_1, x_2, ..., x_N|\Theta) = \prod \Theta^{x_i}(1-\Theta^{x_i}) \tag{2}$$

The order of the elements doesn't matter.

(b)

$$l(\Theta) = log(L(\Theta)) = log(\prod \Theta^{x_i}(1-\Theta^{x_i})) \tag{3}$$

Because

$$log(ab) = log(a) + log(b) \tag{4}$$

$$l(\Theta) = \sum log(\Theta^{x_i}(1-\Theta^{x_i})) \tag{5}$$

Because

$$log(a^b) = blog(a0 \tag{6}$$

$$l(\Theta) = \sum [x_i log(\Theta) + (1-\Theta^{x_i})log(1-\Theta)))] \tag{7}$$

$$l(\Theta) = log(\Theta)\sum x_i + log(1-\Theta)\sum(1-x_i))) \tag{8}$$

$$l(\Theta) = log(\Theta)\bar{x}N + log(1-\Theta)\bar{x}N \tag{9}$$

Finding the first and second derivatives:

$$\frac{dl}{d\Theta} = \frac{\bar{x}N}{\Theta} - \frac{(1-\bar{x})N}{1-\Theta} \tag{10}$$

$$\frac{d^2l}{d\Theta^2} = -\frac{\bar{x}N}{\Theta^2} + \frac{(1-\bar{x})N}{(1-\Theta)^2} \tag{11}$$

I will set the first derivative to 0 and solve for theta to find the MLE:

$$\frac{\bar{x}N}{\Theta} = \frac{(1-\bar{x})N}{1-\Theta} \tag{12}$$

$$MLE = \Theta = \bar{x} \tag{13}$$

(c)

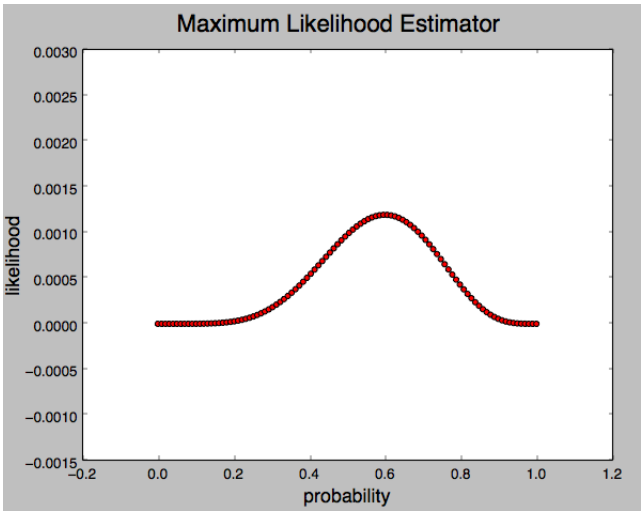

Figure 1: Based on the peak in this graph, the probability that results in the highest likelihood is 0.6. This matches what we theoretically computed using theory, which was also 0.6.

(d)

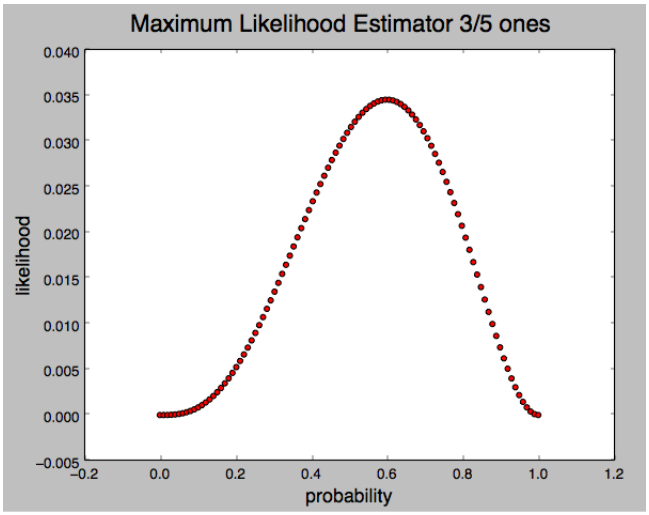

Figure 2: Based on the peak in this graph, the probability that results in the highest likelihood is 0.6. This matches what we theoretically computed using theory, which was also 0.6.

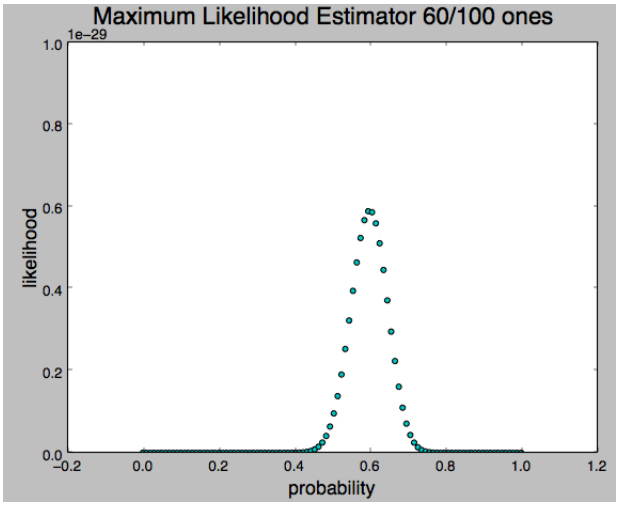

Figure 3: Based on the peak in this graph, the probability that results in the highest likelihood is 0.6. This matches what we theoretically computed using theory, which was also 0.6.

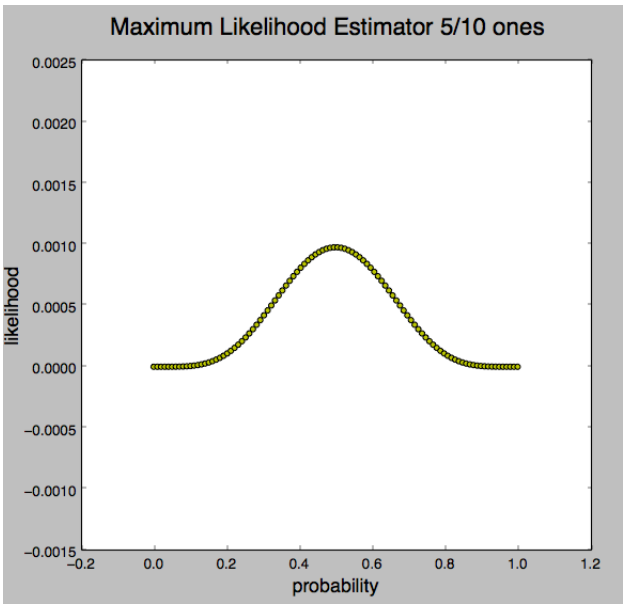

Figure 4: Based on the peak in this graph, the probability that results in the highest likelihood is 0.5. This matches what we theoretically computed using theory, which was also 0.5.

2 Problem 2

(a)

$$2^{n-3} mistakes \tag{14}$$

(b) No, all splits will lead to more error than if we made a single leaf decision tree. A single split would give the result of 0 or 1 both a 50/50 chance when the actual distribution leans heavily towards resulting in 1 and only in specific cases resulting in 0.

(c)

$$H[x] = -\sum p(x = a_k) log(p(x = a_k)) \tag{15}$$

$$= -(1-p) log(1-p) + p log(p) \tag{16}$$

$$= -\frac{2^n - 2^{n-3}}{2^n} log(\frac{2^n - 2^{n-3}}{2^n}) - \frac{2^{n-3}}{2^n} log(\frac{2^{n-3}}{2^n}) \tag{17}$$

$$= -(1 - 2^{-3}) log(1 - 2^{-3}) - 2^{-3} log(2^{-3}) \tag{18}$$

$$= 0.5436 \tag{19}$$

(d) Split on either x_1, x_2, or x_3. If Y = 0,

$$p = 0.5 * 0.5 = 0.25 \tag{20}$$

$$H[x] = -0.75 log(0.75) - 0.25 log(0.25) = 0.8113 \tag{21}$$

If Y = 1,

$$p = 1 \tag{22}$$

$$H[x] = -1 log(1) - 0 log(0) = 0 \tag{23}$$

Thus the conditional entropy is:

$$H[Y|X] = 0.5(0.8113) + 0.5(0) = 0.406 \tag{24}$$

3 Problem 3

(a) We are given that we have k subsets and the ratio

$$\frac{p_k}{p_k + n_k} \tag{25}$$

is equal for all subsets. To find information gain,

$$informationgain = H[Y] - H[Y|X] \tag{26}$$

Thus for information gain to be 0, H[Y] and H[YIX] must be equal. This can be proven since

$$H[Y|X] = \sum_{i=1}^{K} \frac{1}{K} H_i(S) \tag{27}$$

where

$$H_i[S] = -\frac{p_k}{p_k + n_k} log(\frac{p_k}{p_k + n_k}) - \frac{n}{p_k + n_k} log(\frac{n}{p_k + n_k}) \tag{28}$$

Because

$$H_1(S) = H_2(S) = ... = H_K(S) \tag{29}$$

then Equation 27 means that

$$H[Y|X] = H_K(S) = H(S) \tag{30}$$

Thus because the two are equal, information gain is 0.

4 Problem 4

(a)
i. Pclass: upper class passengers survived at a higher frequency than lower class passengers
ii. Sex: females were more likely to survive than men
iii. Age: children below the age of 10 survived at a significantly higher frequency than any other age group
iv. SibSp: having 1 sibling correlated with higher frequency of survival
v. Parch: smaller family sizes with 1 or 2 children had a higher frequency of survival
vi. Fare: passengers who paid higher fare had a higher frequency of survival
vii. Embarked: passengers who embarked at Cherbourg had a higher frequency of survival than other embarking locations

(b) An error of 0.485 was exactly obtained.

(c) Using the DecisionTreeClassifier we obtain an error of 0.014.

(d)
i. MajorityVoteClassifier
- train error: 0.40377855887521963
- test error: 0.40734265734265768

ii. RandomClassifier
- train error: 0.48901581722319881
- test error: 0.48657342657342667

iii. DecisionTreeClassifier
- train error: 0.011528998242530775
- test error: 0.2397902097902099

(e)

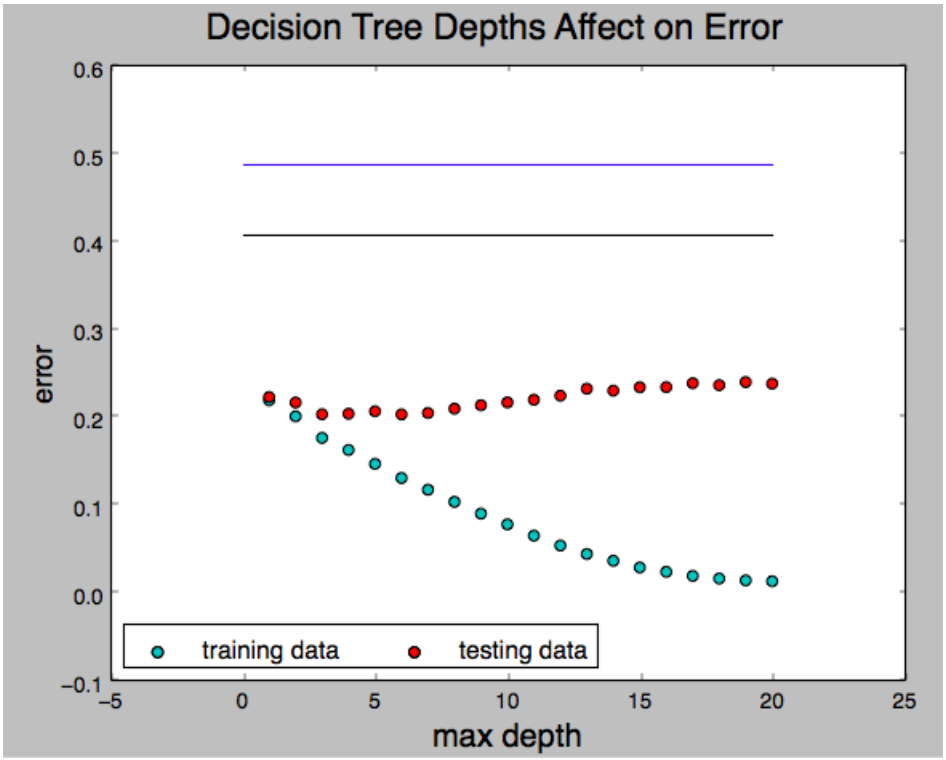

Figure 5: The blue line at top represents the average RandomClassifier error and the black line below that represents the MajorityClassifier. Both these models' error isnt affected by max depth. The scatter plot underneath tells us that the best max depth for our model is at 3. We see overfitting occur as max depth is over 3 when the error starts to increase again after reaching a minimum.

(f)

Figure 6: As we have more training data, our training error should increase because our model does not generalize well enough. But with more training data, the model can form better results on our testing data thus decreasing the original testing error.