CM146, Winter 2019
Problem Set 1: Decision trees
Due Jan 28, 2019

Jingjing Nie

1 Problem 1

(a) Problem 1a

Solution:

$$
\begin{aligned}
L(\theta) &= P_\theta(X_1, X_2, ..., X_n) \\
&= \prod_{i=1}^{n} P_\theta(X_1) \\
&= \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \\
&= \theta^{\sum_{i=1}^{n}\{X_i=1\}} (1-\theta)^{\sum_{i=1}^{n}\{X_i=0\}} \\
&= \theta^{\sum_{i=1}^{n}\{X_i\}} (1-\theta)^{n-\sum_{i=1}^{n}\{X_i\}}
\end{aligned}
$$

(b) Problem 1b

Solution: Take the log form of the function found in a,

$$\begin{aligned} log\ likelihood\ l(\theta) &= log(L(\theta)) \\ &= log(\theta^{\sum_{i=1}^{n}\{X_i\}}(1-\theta)^{n-\sum_{i=1}^{n}\{X_i\}}) \\ &= (\sum_{i=1}^{n}\{X_i\})log(\theta) + (n-\sum_{i=1}^{n}\{X_i\})log(1-\theta) \end{aligned}$$

Then we can take the first and second derivative of the function to find the closed-form formula for the MLE.

$$\frac{\partial l}{\partial \theta} = \frac{\sum_{i=1}^{n}\{X_i\}}{\theta} - \frac{n-\sum_{i=1}^{n}\{X_i\}}{1-\theta}$$

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{\sum_{i=1}^{n}\{X_i\}}{\theta^2} - \frac{n-\sum_{i=1}^{n}\{X_i\}}{(1-\theta)^2}$$

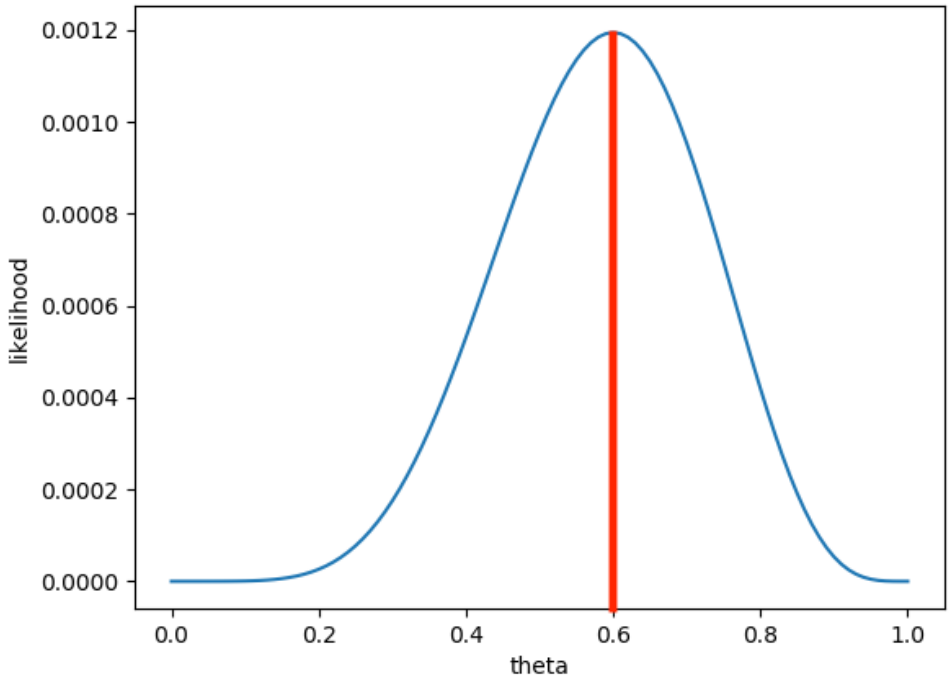
The second derivative of the function is always negative, so the θ value that can maximize the likelihood can be found by equating the first derivative form to 0 and find θ.

$$\frac{\sum_{i=1}^{n}\{X_i\}}{\theta} - \frac{n-\sum_{i=1}^{n}\{X_i\}}{1-\theta} = 0$$

$$\theta = \frac{\sum_{i=1}^{n}\{X_i\}}{n}$$

(c) Problem 1c

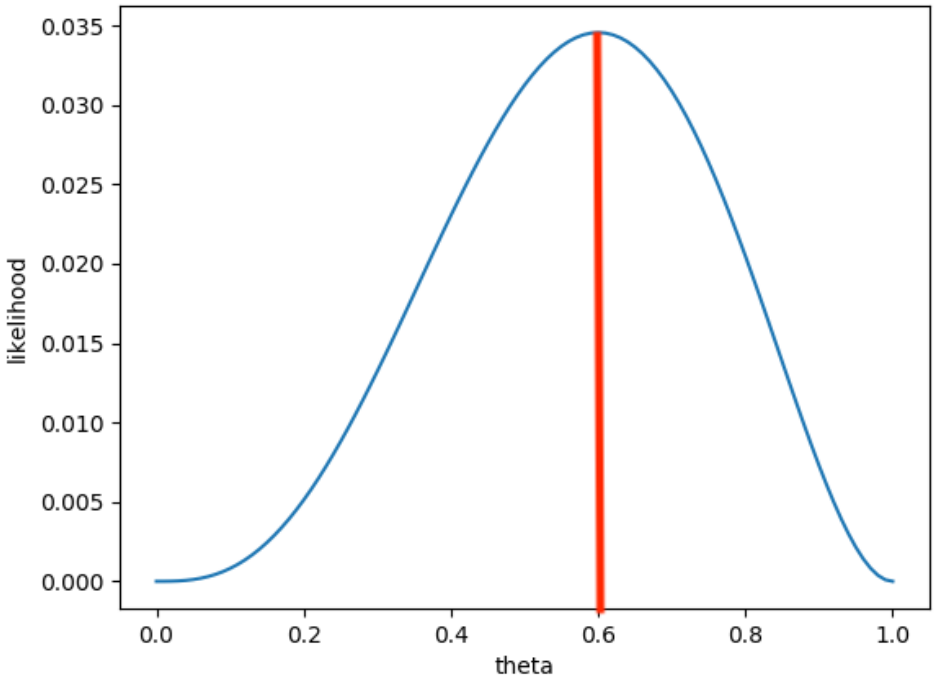
Solution:



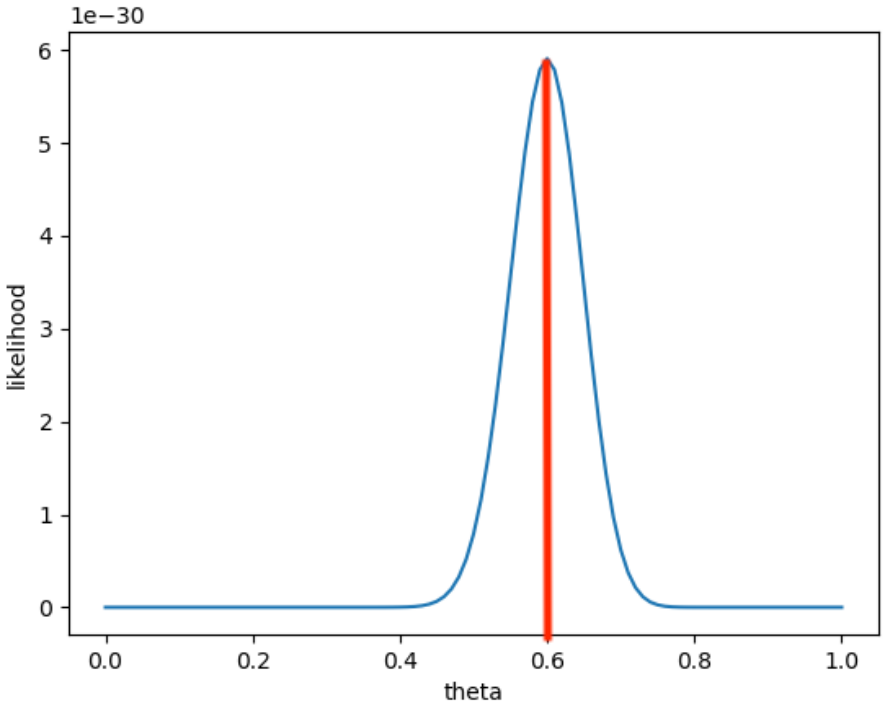
According to the plot, the value of θ_{MLE} is approximately 0.6.
According to the closed-form formula, since n is 10, and we have 6 ones, so θ, is 0.6.
Thus, the two values found by plot and closed-form formula agree with each other.

(d) Problem 1d

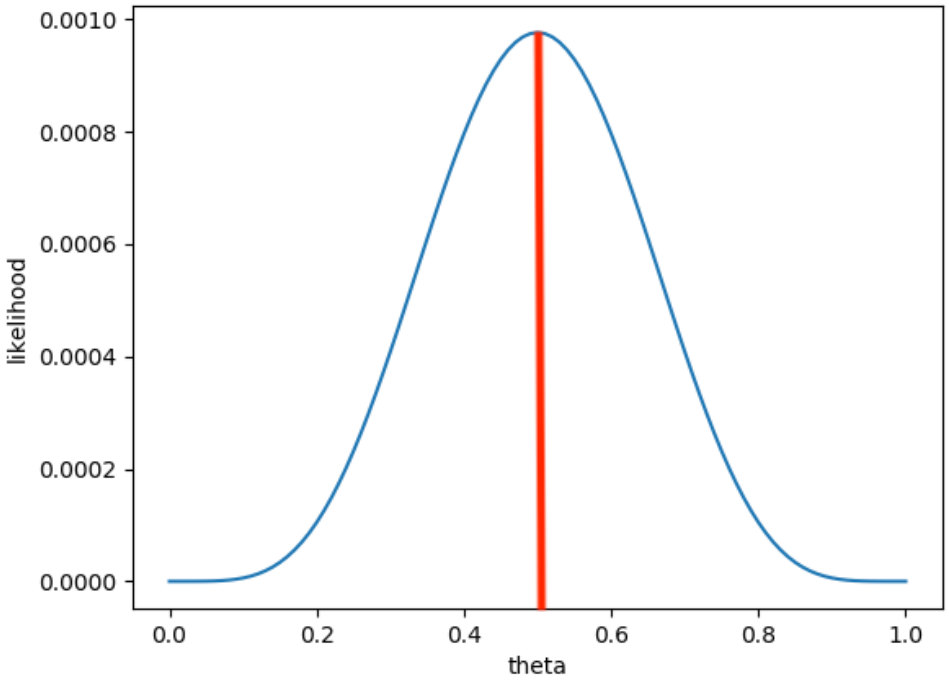
Solution:



n = 5, three 1s and two 0s:



n = 100, sixty 1s and forty 0s:



n = 10, five 1s and five 0s:

According to the likelihood function, the likelihood values are dependent on the proportion of the variables that are equal to 1. In the plots, the one in part c, and the first two ones in part d all have the same proportion values for "1" variables, 0.6, so the MLE in all those cases are approximately 0.6. In the last case, since the proportion changed to 0.5, MLE also changes to 0.5. Also, as the number n increases, the plots are packed more closely to the maximum value.

2 Problem 2

(a) Problem 2a

Solution:

There are n boolean features, and Y is 1 when $X_1 or X_2 or X_3$ is 1. Therefore, only when $X_1 and X_2 and X_3$ are all 0 can Y be 0, and the other (n-3) features can be either 1 or 0. So in total, the best 1-leaf decision tree can predict all inputs to be 1, and thus will make mistakes of $2^{(n-3)}$.

$$2^{(n-3)}/2^n = 1/8$$

So 1/8 of the time.

(b) Problem 2b

Solution:

No, there is not. If we want to split on $X_1, X_2, or X_3$, then the data will be split into two leaves, where one contains only 1s, and the other contains only a quarter of 0s. In both cases, the predictions are both 1. If we want to split on the other X_i for $i >= 4$, then the data will be split into two leaves, and both contain 1/8 0s and 7/8 1s. In both cases, the predictions are also both 1. Thus, whether we split it or not will not affect the error rate. It will remain as 1/8.

(c) Problem 2c

Solution: $H(X) = -\frac{1}{8}log(\frac{1}{8}) - \frac{7}{8}log(\frac{7}{8}) = 0.543$

(d) Problem 2d

Solution:

To reduces the entropy of the output Y by a non-zero amount, we can split on either $X_1, X_2, or X_3$. The resulting conditional entropy will be:
$H(X) = -\frac{1}{2} \times 0 - \frac{1}{2} \times (\frac{1}{4}log(\frac{1}{4}) + \frac{3}{4}log(\frac{3}{4})) = 0.406$

3 Problem 3

(a) Problem 3a

Solution:

$$B(q) = -qlogq - (1-q)log(1-q)$$

Since q is between 0 and 1, (1-q) is also between 0 and 1. Thus, log(q) and log(1-q) are smaller than 0, and the term B(q), with negative signs, will be greater than 0.
Then, we could find the extrema of this term by taking the derivatives:

$$\frac{\partial B(q)}{\partial q} = -log(q) - 1 - (1-q)log(1-q) + 1 = 0$$
$$log(\frac{1-q}{q}) = 0$$
$$\frac{1-q}{q} = 1$$
$$q = \frac{1}{2}$$
$$\frac{\partial^2 B(q)}{\partial q^2} = -\frac{1}{q(1-q)}$$

Since q and (1-q) are greater than 0, the second derivative term is smaller than 0, suggesting that the result we have found for first derivative is the local maximum.

Plug in q=0.5

$$B(q) = -0.5log(0.5) - (1-0.5)log(1-0.5) = -log(0.5) = 1$$

Thus,

$$0 \leq H(s) = B(\frac{p}{p+n}) \leq 1$$
$$0 \leq H(s) \leq 1$$

$$
\begin{aligned}
&when\ p = n, \\
&\frac{p}{p+n} = \frac{1}{2} \\
H(s) &= H(\frac{1}{2}) \\
&= -\frac{1}{2}log(\frac{1}{2})(1\frac{1}{2})log(1\frac{1}{2}) \\
&= -\frac{1}{2}log(\frac{1}{2}) - \frac{1}{2}log(\frac{1}{2}) \\
&= -log(\frac{1}{2}) \\
&= log(2) \\
&= 1
\end{aligned}
$$

(b) Problem 3b

Solution:

Since the sample is split into k subsets and they all have the same ratio, then $\frac{p_k}{p_k+n_k} = \frac{p}{p+n}$. Thus, the entropy for S is also B(q). So the information gain can be calculated as:

$$
\begin{aligned}
Gain &= B(\frac{p}{p+n}) - \sum_k (\frac{p_k+n_k}{p+n}) B(\frac{p}{p+n}) \\
&= B(\frac{p}{p+n}) - B(\frac{p}{p+n}) \\
&= 0
\end{aligned}
$$

4 Problem 4

(a) Problem 4a

Solution:

When k is 1, the training set error will be minimized, since the point can be its own neighbor. The resulting training error is 0 as it can always correctly classify itself.
Using the training set error, especially for this k value, will take in all the details and noises in the training data. This will lead to overfitting, which will then harm the performance of the built model for new data. So in this case, even if the training set error is 0, the test set error could be much bigger than 0. Thus, the training set error is not a reasonable estimate of test set error.

(b) Problem 4b

Solution:

When k is 5 or 7, the leave-one-out cross-validation error is 4/14.

(c) Problem 4c

Solution:

The LOOCV errors for the lowest k, 0, and highest k, 13, for this data set are 0 and 1. Using too small value for k will overfit the training data and thus lead to high error when taking new data. Using too large value for k will misclassify all training data points, and thus will also lead to high error when taking new data.

5 Problem 5

(a) Problem 5.1

Solution:

Pclass: First class has the highest survival rate, and the third class has the lowest.
sex: Female has higher survival rates than male.
age: People at age smaller than 10 (children) have the highest survival rate, and people at age between 20 and 40 have the lowest survival rate.
SibSp: People with siblings have higher survival rate than people with no siblings.
Parch: People with parent or children have higher survival rate than people who do not have parent or children.
Fare: People who pay more have higher survival rate than people who pay little.
Embark: People who embarked on 0 port have higher survival rates than the others.

(b) Problem 5.2b

Solution:

Classifying using Majority Vote...
– training error: 0.404
Classifying using Random...
– training error: 0.485

(c) Problem 5.2c

Solution:

Classifying using Decision Tree...
– training error: 0.014

(d) Problem 5.2d

Solution:

Classifying using Majority Vote...
– training error: 0.404

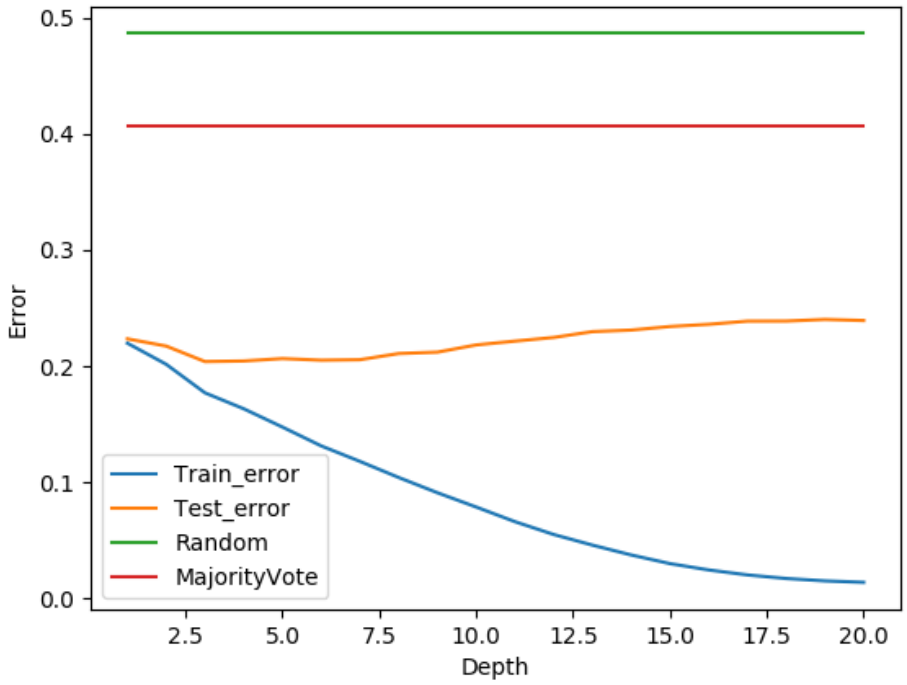
– testing error: 0.407

Classifying using Random Classifier...
– training error: 0.489
– testing error: 0.487

Classifying using Decision Tree...
– training error: 0.012
– testing error: 0.241

(e) Problem 5.2e

Solution:

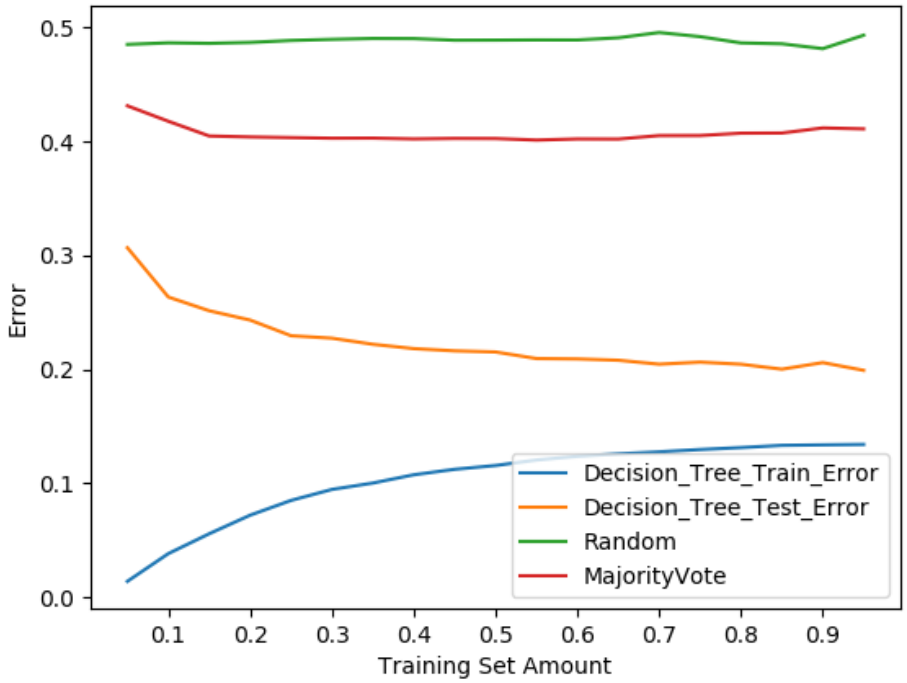


According to the plot, the best depth limit to use for this data is 6. There is over-fitting, because as the value of depth increases, whereas the training error decreases, the test error decreases slightly and then increases. Thus, although more depth has helped decrease the training

error, it harms the test error due to over-fitting.

(f) Problem 5.2f

Solution:



According to the plot, as the amount of training set increases, the training error increases, and the test error slightly decreases. Thus, with a larger training set, the gap between the two types of errors decreases. So more data can be beneficial.