CS M146, Winter 2019
Problem Set 1: Decision Trees
Due Jan 28, 2019

Tian Ye

January 27, 2019

1 Maximum Likelihood Estimation

(a) **Solution:** The likelihood estimation is given by the following:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} P_\theta(X_i) && (1)\\
&= \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} && (2)\\
&= \theta^{x_1}(1-\theta)^{x_0} && (3)
\end{aligned}
$$

Where x_1 counts the number of cases which $X_i = 1$ and x_0 counts the number of cases which $X_i = 0$.
The order of the individual random variables X_i do not matter as they are independent from one another.

(b) **Solution:** Taking the log likelihood of the previous expression:

$$
\begin{aligned}
\ell(\theta) &= \log(\theta^{x_1}(1-\theta)^{x_0}) && (4)\\
&= x_1\log(\theta) + x_0\log(1-\theta) && (5)
\end{aligned}
$$

Taking the first and second derivatives of $\ell(\theta)$ with respect to θ:

$$
\begin{aligned}
\ell'(\theta) &= \frac{x_1}{\theta} - \frac{x_0}{1-\theta} && (6)\\
\ell''(\theta) &= -\frac{x_1}{\theta^2} - \frac{x_0}{(1-\theta)^2} && (7)\\
& && (8)
\end{aligned}
$$

Since $\ell''(\theta) < 0$, the function is always concave down.
We can therefore set $\ell'(\theta) = 0$ to solve for the MLE:

$$\theta_{MLE} = \frac{x_1}{x_1 + x_0} \quad (9)$$

(c) **Solution:**

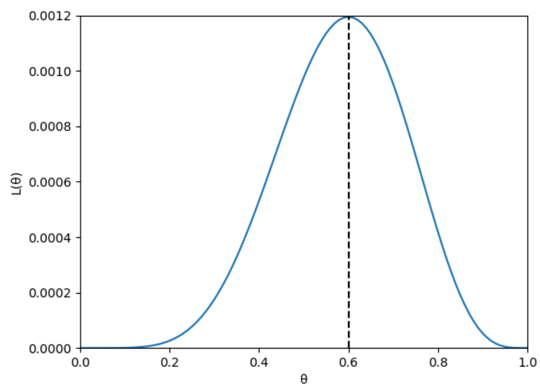

Figure 1: The figure above does agree with Equation 9 given in the previous section; we can see this as the maximum is at $\theta = 0.6$, which corresponds with $\frac{6}{4+6}$.

(d) **Solution:**

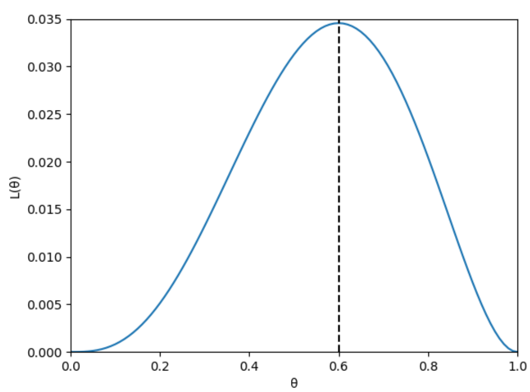

Figure 2: By decreasing the number of data points while maintaining the ratio of 1s to 0s, the likelihood plot keeps the same MLE while having a wider spread but higher likelihood.

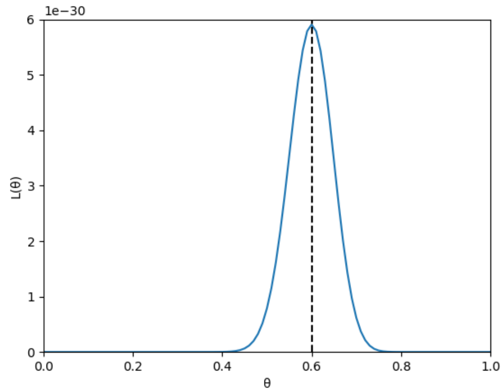

Figure 3: By increasing the number of data points while maintaining the ratio of 1s to 0s, the likelihood plot keeps the same MLE while having a narrower spread but a lower likelihood.

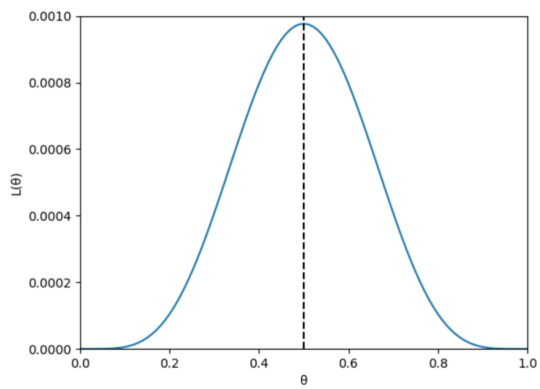

Figure 4: By maintaining the same number of data points while changing the ratio of 1s to 0s, the likelihood plot shifts the MLE while maintaining the spread and likelihood.

2 Splitting Heuristic for Decision Trees

(a) **Solution:** The best 1-leaf decision tree makes an error $\frac{1}{8}$ of the time. The decision tree is as follows: $Y = 0$ if and only if $X_1 = X_2 = X_3 = 0$. This leaves us with 2^{n-3} remaining binary vectors; hence the error is $\frac{2^{n-3}}{2^n} = \frac{1}{8}$.

(b) **Solution:** There does not exist a split that reduces the number of mistakes. If we split on X_1, X_2, or X_3, it will create a tree in which one leaf only contains 1s and the other leaf contains 1s with a proportion of $\frac{3}{4}$. Splitting on a value of $n \geq 4$ will create two leaves where the proportion of 1s is $\frac{7}{8}$. In both cases the tree will always predict 1, leaving an error rate of $\frac{1}{8}$.

(c) **Solution:** $\frac{1}{8}\log(8) + \frac{7}{8}\log(\frac{8}{5}) = 0.543$

(d) **Solution:** By splitting on X_1, X_2, or X_3, we can reduce the entropy on the output Y. The new entropy following the split is as follows:
$\frac{1}{2}[\frac{1}{4}\log(4) + \frac{3}{4}\log(\frac{4}{3})] = 0.406$
This reduces the entropy on Y by 0.137.

3 Entropy and Information

(a) **Solution:** We know that the entropy must fall within the range of $0 \leq H(S) \leq 1$ using the following:

$$B(q) = -q\log(q) - (1-q)\log(1-q) \tag{10}$$

For a set containing p positive examples and n negative examples, $H(S) = B(\frac{p}{n+p})$. We will show that $0 \leq H(S)$ by setting $p = 0$ and n to some value greater than 0:

$$B(\tfrac{0}{n+0}) = -q\log(q) - (1-q)\log(1-q) \tag{11}$$
$$B(0) = 0\log(0) - (1-0)\log(1-0) \tag{12}$$
$$H(S) = 0 \tag{13}$$

For any value of p such that $0 \leq p \leq n$, $H(S)$ increases and approaches 1 as q increases towards $\frac{1}{2}$. Once $p = n$, we yield the following:

$$B(\tfrac{n}{2n}) = -q\log(q) - (1-q)\log(1-q) \tag{14}$$
$$B(\tfrac{1}{2}) = -\tfrac{1}{2}\log(\tfrac{1}{2}) - (1-\tfrac{1}{2})\log(1-\tfrac{1}{2}) \tag{15}$$
$$H(S) = 1 \tag{16}$$

As p increases past n, the value q once again begins to decrease towards 0, and as such, $H(S)$ decreases back towards 0. Hence, $0 \leq H(S) \leq 1$.

(b) **Solution:**