1

Dhruv Chakraborty
UID: 204-962-098
COMSCI M146

HW #1

i) Splitting Heuristic for Decision Trees

(a) Of the given 2^n samples for $n \geq 4$, note that we will have 2^{n-3} samples with target (i.e. Y) = 0 and $2^n - 2^{n-3}$ samples with target 1. For $n \geq 4$, we have $2^n - 2^{n-3} > 2^{n-3}$, so the one leaf decision tree must clarify all samples as 1.

$\therefore$ 2^{n-3} mistakes are made by predicting 1 for samples with target 0.

(b) Since $Y = X_1 \vee X_2 \vee X_3$, it is independent of X_i for $i \geq 4$ and thus splitting on these will not reduce the num of mistakes. Moreover, splitting on X_j for $j = 1, 2, 3$ would lead to correct identification if $X_j = 1$ but when $X_j = 0$, the tree would still ~~classify~~ predict $Y = 1$ as $3/4$th of the samples down this branch still have target 1. So the predictive ~~accuracy~~ accuracy remains the same as with no splits.

$\therefore$ There is no single split that reduces the num of mistakes.

(c) Entropy $H[X] = -\sum_{k=1}^{k} P(X = a_k) \log P(X = a_k)$

$$= -\frac{2^{n-3}}{2^n} \log\left(\frac{2^{n-3}}{2^n}\right) - \left(1 - \frac{2^{n-3}}{2^n}\right) \log\left(1 - \frac{2^{n-3}}{2^n}\right)$$

$$= -\frac{1}{8} \log\left(\frac{1}{8}\right) - \frac{7}{8} \log\left(\frac{7}{8}\right)$$

$$H[X] \approx 0.543$$

(d) Yes, splitting on X_j for $j = 1, 2, 3$ reduces the entropy by a non-zero amt. As the tree is split in half, the entropy upon splitting on X_j is

$$H[X] = \frac{1}{2}(0) + \frac{1}{2}\left(-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right) = \frac{1}{2}(0.811)$$

$$H[X] = 0.406$$

2)

(a) $H(S) = B\left(\frac{p}{p+n}\right) = -\frac{p}{p+n}\log\left(\frac{p}{p+n}\right) - \left(1-\frac{p}{p+n}\right)\log\left(1-\frac{p}{p+n}\right)$

$$= -\frac{p}{p+n}\log\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log\left(\frac{n}{p+n}\right)$$

$$= \frac{-1}{p+n}\left(p\log\left(\frac{p}{p+n}\right) + n\log\left(\frac{n}{p+n}\right)\right)$$

Note that $\frac{p}{p+n}$ and $\frac{n}{p+n}$ are between 0 and 1 $\Rightarrow$

$-1 \leq \log\left(\frac{p}{p+n}\right), \log\left(\frac{n}{p+n}\right) \leq 0 \Rightarrow$

$$-(p+n) \leq p\log\left(\frac{p}{p+n}\right) + n\log\left(\frac{n}{p+n}\right) \leq 0$$

~~So~~

But $H(S) = \frac{-1}{p+n}\left(p\log\left(\frac{p}{p+n}\right) + n\log\left(\frac{n}{p+n}\right)\right)$

So $0 \leq H(S) \leq 1$.

Setting $p=n$, we get $H(S) = \frac{-1}{p+p}\left(p\log\left(\frac{p}{2p}\right) + p\log\left(\frac{n}{2n}\right)\right)$

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

$$= +\frac{2p}{2p} = 1$$

So $H(S) = 1$ when $p=n$.

(b) We know that the entropy prior to the split $H_{prior}[S] = B\left(\frac{p}{p+n}\right)$

Also, $\frac{p_1}{p_1+n_1} = \frac{p_2}{p_2+n_2} = \dots = \frac{p_k}{p_k+n_k}$

But we also have that $\sum_i p_k = p$ and $\sum n_k = n$

which means that $\frac{p_k}{n_k+p_k} = \frac{p}{p+n}$ $\forall k$

(2)

Dhruv C.
CS M146

2) (b) contd:

Entropy after the split $H[S]$

$$= \frac{p_1+n_1}{p+n} B\left(\frac{p_1}{p_1+n_1}\right) + \dots + \frac{p_k+n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right)$$

$$= \frac{p_1+n_1+\dots+p_k+n_k}{p+n} B\left(\frac{p_k}{p_k+n_k}\right) \quad \forall k$$

$$= \frac{p+n}{p+n} B\left(\frac{p_k}{p_k+n_k}\right) = B\left(\frac{p_k}{p_k+n_k}\right) \forall k.$$

So, Information Gain $= H_{prior}[S] - H[S]$

$$= B\left(\frac{p}{p+n}\right) - B\left(\frac{p_k}{p_k+n_k}\right)$$

$$= B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right) = \underline{\underline{0}}$$

3)

(a) In this dataset, $k=1$ minimizes the training set error, which would actually be 0% here.
Yet, this is not a good model and ~~leads~~ is not a reasonable estimate for training set error as it is highly likely to overfit the training data, choosing the same point continually and not generalizing well to the new data.

(b) $k=5$ minimizes LOOCV error for this data set, correctly identifying 10 points, giving us an error of $\frac{4}{14} \approx 28.6\%$.

Cross-validation provides a better measure of the models performance by ensuring that the model consistently performs ~~well~~ of different types of data that are left out from the training set. This allows the model to generalize well to unseen data and not overfit the training data.

(c) For the lowest i.e. k=1 the error is 10/14 ≈ 71.4%
whereas for the highest i.e. k=13 the error is 14/14 ≈ 100%.
Using too large a value of k is bad as it is prone to underfitting whereas too small a value leads to overfitting the training set.

4.1 Visualization

(a)

Pclass: Passengers with higher ticket classes have a much lower chance of survival than those with a lower ticket class. This is especially evident for passengers travelling i.e. with a 3rd class ticket.

Sex: Women have a much higher chance of survival than men

Age: People aged 20-40 seem to have the lowest rates of survival whereas children (particulary 0-10) have the highest rates of survival.

SibSp: Those with no siblings or spouse on board have bad survival rates, whereas those with only one on board seem to do the best. Those with more than 1 do not do as well, with 5 siblings/spouse having it the worst.

Parch: Those with no children or parents on board fare badly, with 1/2 parents + children being on board seems to be the best case.

Fare: Passengers that paid a higher fare have much higher chances of survival

Embarked: Passengers boarding at Cherbourg fare much better than those embarking at Queenstown or Southhampton.

<u>4. Applying Decision Trees</u>

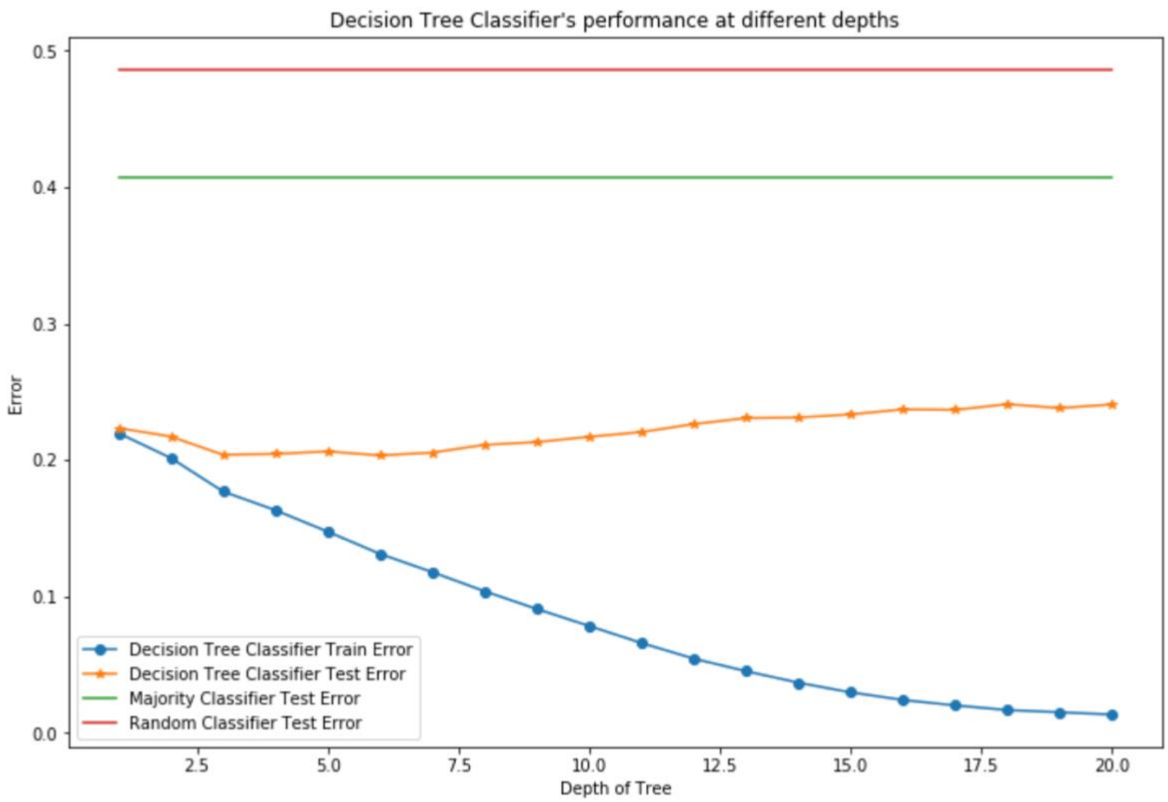
4.2: Evaluation

(b) The training error when classifying using majority vote is 0.404 i.e. 40.4% and when classifying using a random classifier is 0.485 i.e. 48.5%

(c) The training error of the decision tree classifier is 0.014 i.e. 1.4%.

(d) The average training and test errors using cross validation for each classifier are:
Majority Vote Classifier:
train_error: 0.403778558875
test_error: 0.407342657343
Random Classifier:
train_error: 0.489015817223
test_error: 0.486573426573
Decision Tree Classifier:
train_error: 0.0115289982425
test_error: 0.240839160839

(e) A depth limit of 3 is the best for the Decision Tree Classifier as it has both low training and test error. This leads to generalized results since the model doesn't overfit. This evidence is seen clearly as although training error continues to decrease for trees with depth greater than 3, the test error stays the same or even increases.



(f) The key observation to be made from this graph is that as the model is fed more and more training data, it is less likely to overfit the data. This is seen clearly as although the training error starts off really small, the test error is quite large as the model is unable to generalize. As the amount of data the model is trained on is increased, we see that the training and test errors seem to converge, indicating that the model is not overfitting the training data.