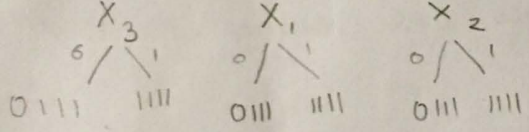
1a. The best 1-leaf decison tree classifies the data without spliting. In this case, all examples get classified as 1 given it has the most "votes". Since the classification truely only depends on the values of x_1, x_2 & x_3 and therefore 2^3 possible features. There are 2^n total features which can be used to classify the data, but the true function $f: X \to Y$ does not depend on 2^{n-3} of our features where $n \geq 4$. If data is classifies based on these 2^{n-3} features, it is an error therefore the 1-leaf tree makes $\frac{2^{n-3}}{2^n}$ mistakes or $2^{-3} = \frac{1}{8}$ of the time it errors.

1b. No, if we split at x_4 or any x_i where $i > 4$ we will still recieve at least one mistake. Spliting on x_4 infact gives $\frac{1}{8}$ mistake. Spliting on x_1, x_2, or x_3 alway will alway have at least 3/4 misclassified 1's in each case. Therefore, there is no split any better than the 1-leaf split here.

ex



1c. $H[Y] = -P(\text{error}) \log \text{err} - P(\text{correct}) \log \text{corr} = -\frac{1}{8}\log_{10}\left(\frac{1}{8}\right) - \frac{7}{8}\log_{10}\left(\frac{7}{8}\right) = 0.163/\log_{10}2 = \boxed{0.544}$

1d. Any split on x_1, x_2, or x_3 will reduce entropy a nonzero amount

1d. $H[Y|X_1] = \sum_x P(x_i = 1) \cdot H[Y|X_1] = \frac{1}{2} \cdot \left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} + \left(\frac{4}{4}\log\frac{4}{4} - 0\right)\right)$

$= 0.2442/\log_{10}2 = \boxed{0.811}$

this result is the same for $H[Y|X_2]$ & $H[Y|X_3]$ because they classify the same way as rats (see 6)

2a. Show that $0 \leq H[S] \leq 1$ & $H(S) = 1$ when $p = n$

when $p=n$ $H(S) = B\left(\frac{n}{2n}\right) = B\left(\frac{1}{2}\right) = -\frac{1}{2}\log_2\frac{1}{2} - \left(1-\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = -\frac{1}{2}\cdot -1 - \frac{1}{2}\cdot -1 = \frac{1}{2}+\frac{1}{2} = 1$ □

Since $p=n$ is the upper bound of $B(q)$ and $n=0$ is a lower bound (resulting in $B(1)=0 \doteq H(S)$ $H(S)$ must lie between $0 \leq H(S) \leq 1$

2b. $\frac{p_k}{p_k+n_k}$ ratio same of attributes. $Gain(S,A) = H[S] - H[S|X_j]$ where X_j is where we split, and $\frac{p_k}{p_k+n_k}$ for all k to j so

$H(S|X_j) = \sum_k H(S_k) \cdot \frac{p_k+n_k}{p+n} = H(S_k) \cdot \sum_{k=1}^{k} (p_k+n_k) \cdot \frac{1}{p+n} = H(S) \cdot \frac{p+n}{p+n} = H(S)$

$\therefore$ I Gain $= H[S] - H[S|X_j] = H[S] - H[S] = B\left(\frac{1}{1+n}\right) - B\left(\frac{p}{p+n}\right) = \boxed{0}$

3a Since $K=1$ is possible that is we let a point be its own neighbor, $K=1$ would minimize the training set error to zero. The training set is different from the test set. Since the model learns on the training set, a training set with value of $K=1$ could over fit the model to the training set and not allow the model to make any generalizations.

3b) A value of $K=5$ minimizes the LOOCV error. At $K=5$, the error rate is $\frac{2}{7}$ as $\frac{4}{14}$ points are incorrectly classified. Cross-Validation is a better measure of test set performance because it allows us to reuse data and test the learner on data it has never seen before leading to a lower bias in the errors.

3c) The lowest $K=1$ has an error of $\frac{10}{14}$ because it would only correctly classify 4 points specifically the points on the diagonal ends. The highest $K=13$ places every point as a neighbor which would miss-classify the left-out point everytime because the majority of points would belong to the other class, thus this error would be $14/14$. Too large K value can lead to underfitting while too small a K can cause overfitting.

4a) PClass: Members of a lower pclass (first class) correspond to higher survival rate while higher values correspond to lower. Third class had the lowest survival rate.

Sex: Women had a higher survial rate than men.

Age: Children had the highest survival rate while most adults, particularly ages 20-30 had a lower survival.

Sibsp: Those with no siblings had a lower survival rate than those with 1-3 siblings. Beyond 3 siblings and survival rate appeared to fall.

Parch: People without children or parents had a lower survival rate than those with children and parents in general.

Fair: Those who paid higher fairs had a higher survival rate than those who paid lower fairs.

Embarked: Those who embarked from Cherbourg had the highest survival rate and our of those who left from Queenstown and Southampton only a third survived.

4b) My RandomClassifier had a training error of 0.485

4c) My DecisionTreeClassifier training error was 0.014

4d) For each of my classifiers:

MajorityVoteClassifier:

Training Error: 0.404

Testing Error: 0.407

RandomClassifier:

Training Error: 0.486

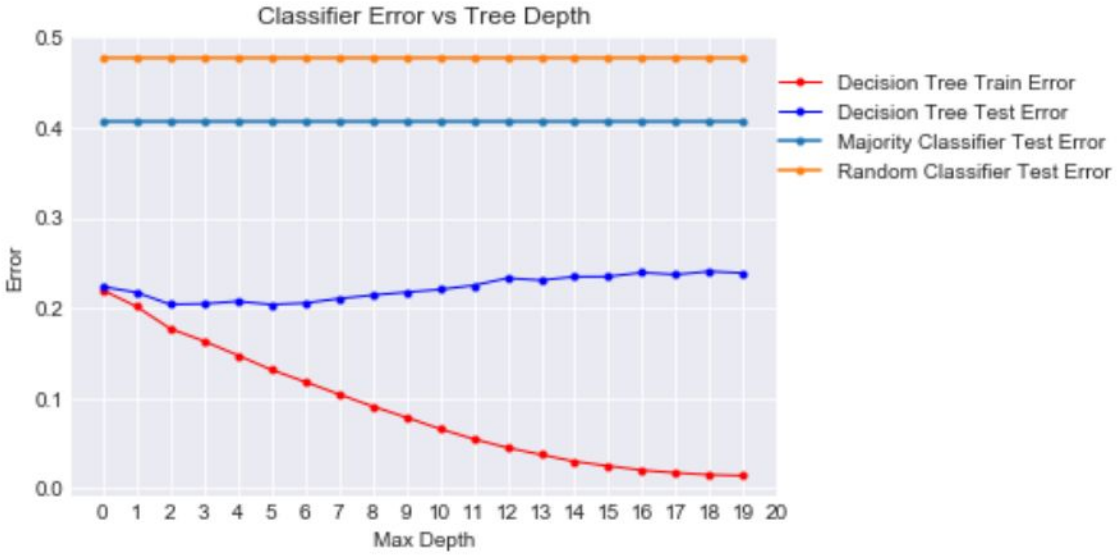
Testing Error: 0.478

DecisionTreeClassifier:

Training Error: 0.012

Testing Error: 0.241

4e)



A depth of 6 is the best to use for the DecisionTreeClassifier. At max_depth = 6, the classifier has the lowest testing *and* training error for the data set. While the testing error for both max_depth = 3 and 6 is 0.204, max_depth = 6 has the lowest training error. This can be inferred from the plot.
Yes, I do see overfitting as the testing error decreases drastically and training error begins to climb.

4f) As the proportion of data used for training increases, the training error rises and the testing error falls. Training set size matters for the decision tree classifier because of overfitting The two errors both converse as the amount of samples used in training increase. Since the random and majority classifiers classify based on randomness, probabilities, and majority, they aren't susceptible to overfitting.