Franklin Hu, Sunil Pedapudi
Assignment 3

1. Entropy and Information Gain

   (a) Let us consider

$$
\begin{aligned}
B(q) &= -qlog(q) - (1-q)log(1-q) \\
\frac{dB}{dq} &= log(1-q) - log(q) \\
\frac{d^2B}{d^2q} &= \frac{1}{(q-1)q}
\end{aligned}
$$

   Then, let $q = \frac{p}{p+n}$. We wish to find a maxima in order to demonstrate $H(S) = B(\frac{p}{p+n}) \leq 1$. Then,

$$
\begin{aligned}
B'(\frac{p}{p+n}) &= log(1 - \frac{p}{p+n}) - log(\frac{p}{p+n}) \\
&= log(\frac{n}{p+n}) - log(\frac{p}{p+n}) \\
&= log(\frac{\frac{n}{p+n}}{\frac{p}{p+n}}) \\
&= log(\frac{n}{p}) = 0
\end{aligned}
$$

   This shows that there exists an optima where $n = p$ and we can verify that this point is a maximum by

$$
\begin{aligned}
B''(\frac{p}{p+n}) &= \frac{1}{(\frac{p}{p+n} - 1)\frac{p}{p+n}} \\
\text{Since } n = p, \\
&= \frac{1}{(0.5-1)0.5} < 0
\end{aligned}
$$

   Therefore, there exists a maximum when $n = p$. Note that in this scenario,

$$
\begin{aligned}
H(S) = B(\frac{p}{p+p}) &= B(0.5) \\
&= -0.5 \cdot log(0.5) - 0.5 \cdot log(0.5) \\
&= -log(0.5) \\
&= 1
\end{aligned}
$$

   which shows that the equality is achieved under said constraint.

   (b) In event where the ratio $\frac{p_k}{p_k+n_k}$ is the same for all k then the weighted sum would be equal to the overall entropy $H(S)$:

$$
\begin{aligned}
\text{Gain}(S, X_j) =& H(S) - \sum_k \frac{|S_k|}{|S|} \cdot H(S) \\
=& 0
\end{aligned}
$$

For all other ratios, the gain will be positive. Since $H(S) \leq 1$, for the gain to be positive:

$$\begin{aligned}
0 <& H(S) - \sum_k \frac{|S_k|}{|S|} \cdot H(S_k) \\
\sum_k \frac{|S_k|}{|S|} \cdot H(S_k) <& H(S) \\
\sum_k \frac{p_k + n_k}{p + n} \cdot &
\end{aligned}$$

$$\begin{aligned}
\text{Gain}(S, X_j) =& H(S) - \sum_k \frac{|S_k|}{|S|} H(S_k) \\
=& B(\frac{p}{p+n}) - \sum_k \frac{p_k + n_k}{p+n} B(\frac{p_k}{p_k + n_k}) \\
\text{Gain}'(S, X_j) =& \\
\text{Gain}''(S, X_j) =& \frac{1}{\frac{p}{p+n} \cdot (\frac{p}{p+n} - 1)} - \sum_k \frac{|S_k|}{|S|} \cdot \frac{1}{\frac{p_k}{p_k+n_k} \cdot (\frac{p_k}{p_k+n_k} - 1)} \\
=& \frac{p+n}{p \cdot (\frac{-n}{p+n})} - \sum_k \frac{|S_k|}{|S|} \cdot \frac{p_k + n_k}{p_k \cdot (\frac{-n_k}{p_k+n_k})} \\
=& -\frac{(p+n)^2}{np} - \sum_k \frac{p_k + n_k}{p+n} \cdot \frac{(p_k + n_k)^2}{-n_k p_k}
\end{aligned}$$

2. Empirical Loss and Splits
   Discrete attributes – 0/1 loss
   Without loss of generality, let us examine a node with $m + n$ examples that we wish to split over an arbitrary attribute. This node contains m correctly classified examples and n incorrectly classified examples. We recognize that the empirical 0/1 loss for this node is $\frac{n}{m+n}$. After splitting this node, we observe two children: one with $m' + n'$ examples and another with $m'' + n''$ examples where m', m'' represent the count of correctly classified examples in each child and n', n'' represent the count of incorrectly classified examples in each child. We wish to show that the empirical loss across both these children is no worse than the empirical loss of the original node. Thus,

$$\frac{m' + n'}{m + n} \frac{n'}{m' + n'} + \frac{m'' + n''}{m + n} \frac{n''}{m'' + n''} = \frac{n' + n''}{m + n} \leq \frac{n}{m + n}$$

   We recognize that $n' + n'' = n$ and thus obtain $\frac{n}{m+n}$ which is the empirical loss of the original parent node.
   Continuous attributes – L_2 loss
   Without loss of generality, let us examine a node with $m + n = |E|$ examples that we wish to split over an arbitrary attribute. This node contains m correctly classified examples and n incorrectly classified examples which belong to the set E . We associate a value of 0 with each correctly classified example and a value of 1 with each incorrectly classified example. Then, we wish to find the L2 loss of this node. Note: class(x) returns the value of the classification of x $\in 0, 1$ and AVG returns the average

over the values of the classification of the examples.

$$\begin{aligned}
Loss &= \sum_{x \in E} (\text{class}(x) - \text{AVG}(E))^2 \\
&= \sum_{x \in E} (\text{class}(x) - \frac{n}{m+n})^2 \\
&= m\left(0 - \frac{n}{m+n}\right)^2 + n\left(1 - \frac{n}{m+n}\right)^2 \\
&= m\left(\frac{n}{m+n}\right)^2 + n\left(\frac{m}{m+n}\right)^2 \\
&= \frac{mn^2 + nm^2}{(m+n)^2} \\
&= \frac{mn(m+n)}{(m+n)^2} \\
&= \frac{mn}{m+n}
\end{aligned}$$

Then, after splitting this node, we observe two children: one with $m' + n'$ examples and another with $m'' + n''$ examples where m', m'' represent the count of correctly classified examples in each child and n', n'' represent the count of incorrectly classified examples in each child. We wish to show that the empirical loss across both these children is no worse than the empirical loss of the original node. Thus,

$$\begin{aligned}
\frac{mn}{m+n} &\geq \frac{m'+n'}{m+n}\frac{m'n'}{m'+n'} + \frac{m''+n''}{m+n}\frac{m''n''}{m''+n''} \\
mn &\geq m'n' + m''n'' \\
mn &\geq m'n' + (m-m')(n-n') \\
mn &\geq m'n' + (mn - mn' - m'n + m'n') \\
mn &\geq mn + 2m'n' - mn' - m'n \\
0 = 2m'n' - 2m'n' &\geq 2m'n' - mn' - m'n
\end{aligned}$$

which we obtain from the observation that $mn' > m'n'$ and $m'n > m'n'$ since $m > m', n > n'$

3. Splitting continuous attributes

4. Majority voting

   (a) Since the errors made by each hypothesis are independent, the error of the ensemble algorithm is simply the sum of the probabilities of the combinations of getting a majority multiplied by the probablity of those errors occuring. If we have K hypotheses, this error is:

$$\text{Error(ensemble)} = \sum_{i=\lfloor \frac{K}{2} \rfloor + 1}^{K} \binom{K}{i} \cdot \epsilon^i \cdot (1-\epsilon)^{K-i}$$

   (b) If the independent assumption is removed, the error of the ensemble algorithm can be worse than ϵ. For example, consider the case when having $K = 3, \epsilon = \frac{2}{10}$. If te hypothesis are adversarial in attempting to make the overall algorithm produce more errors, they can orchestrate their answers in the following way (X's are incorrect results):

	H1	H2	H3
1			
2	X	X	
3			
4		X	X
5			
6	X		X
7			
8			
9			
10			

In this example, each hypothesis has an error rate of $\frac{2}{10}$ but the overall ensemble algorithm has an $\epsilon = \frac{3}{10}$.

5. Programming question

The below accuracies were obtained by running the decision tree classifiers on the entire set of training data (trainingData.csv) and the subset that just contained the samples from that station.

(a) One decision tree

Station Tree	trainingData.csv	trainingData_N.csv
tree_12	0.651816	0.696116
tree_19	0.617914	0.788136
tree_24	0.660420	0.803245
tree_3	0.655570	0.855061
tree_38	0.681941	0.944010
tree_63	0.581799	0.797180
tree_65	0.272412	0.068442
tree_7	0.689539	0.968118
tree_74	0.579006	0.780132
tree_80	0.687383	0.955301

(b) Bagged accuracy

Station Tree	trainingData.csv	trainingData_N.csv
tree_12	0.690356	0.893257
tree_19	0.690881	0.638418
tree_24	0.690367	0.720573
tree_3	0.690322	0.788569
tree_38	0.690322	0.925191
tree_63	0.690222	0.552053
tree_65	0.159342	0.545128
tree_7	0.690322	0.957833
tree_74	0.690065	0.568212
tree_80	0.690322	0.942247