

1. Entropy and Information Gain

(a) Let us consider

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

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

$$B''(\frac{p}{p+n}) = \frac{1}{(\frac{p}{p+n} - 1) \frac{p}{p+n}}$$

Since  $n = p$ ,

$$= \frac{1}{(0.5 - 1)0.5} < 0$$

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) Given

$$Gain(S, X_j) = H(S) - \sum_k \frac{|S_k|}{|S|} \cdot H(S_k)$$

we wish to find  $\frac{\partial^2 Gain(S, X_j)}{\partial S_k^2}$  in order to show that the function is convex. Then, we can argue that since  $Gain$  at the global minimum is positive and the function is convex, then  $Gain$  is strictly positive. So,

$$Gain'(S, X_j) = \frac{\partial}{\partial S_k} - \sum_k \frac{|S_k|}{S} H(S_k) = -\frac{p_k}{p+n} (\log(\frac{p_k}{p_k + n_k}) + \log(1 - \frac{p_k}{p_k + n_k})), \forall k$$

Setting this to 0 allows us to recognize that there exists a minimum where  $\frac{p_k}{p_k+n_k} = 0.5$ . We then realize that if (a) for  $k = 2$  where  $\frac{p_k}{p_k+n_k} = 0.5$  for all  $k$ , then  $Gain$  is positive. and (b) where  $k > 2$  and one such  $\frac{p_{ki}}{p_{ki}+n_{ki}} = 0.5$ , then  $Gain$  is positive since  $\sum_{k \neq ki} \frac{|S_k|}{|S|} \cdot H(S_k)$  is less than  $H(S)$ . Then, we show convexity:

$$\frac{\partial^2 Gain(S, X_j)}{\partial S_k^2} = -\frac{|S_k|}{|S|} \frac{1}{(\frac{p_k}{p_k+n_k} - 1)(\frac{p_k}{p_k+n_k})}, \forall k > 0$$

since  $\frac{p_k}{p_k+n_k} \leq 1$ . This shows that the function is convex with a global minimum greater than zero and thus strictly positive. Let us note that if all  $\frac{p_k}{p_k+n_k}$  are equal, then  $\sum_k \frac{|S_k|}{|S|} \cdot H(S_k)$  is equal to  $H(S)$  so  $Gain$  is 0.

## 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  $L_2$  loss of this node. Note:  $\text{class}(x)$  returns the value of the classification of  $x \in 0, 1$  and  $\text{AVG}$  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

To prove that the optimal split point always comes between examples with different Y-values, consider the following:

Suppose we have a split point between two classes of examples. If we move the split point to the left or to the right, we increase the size of one set  $S_1$  and decrease the size of the other  $S_2$ . Call this element that was added to  $S_1$  as  $i_1$  with class  $c_1$ . Consider the following cases (prior to moving  $c_1$ ):

- (a)  $MAJORITY(S_1) = c_1$   
Empirical loss does not change
- (b)  $MAJORITY(S_1) \neq c_1$ 
  - i.  $c_1$  becomes the majority after  $i_1$  is added  
Empirical loss does not change since there was a tie beforehand
  - ii.  $c_1$  is not majority after  $i_1$  is added  
Empirical loss increases by 1

At the same time, consider what happens to  $S_2$ :

- (a)  $MAJORITY(S_2) = c_1$ 
  - i.  $c_1$  stays the majority after  $i_1$  is removed  
Empirical loss does not change
  - ii.  $c_1$  is not longer the majority after  $i_1$  is removed  
Empirical loss increases by 1
- (b)  $MAJORITY(S_2) \neq c_1$   
Empirical loss decreases by 1

In all of the cases except one, empirical loss increases or stays the same. For these, this proves that the split point between elements of different Y-values are local minima.

For the case when empirical loss can decrease ( $S_2(b)$ ), consider what happens to  $S_1$ :

- (a)  $MAJORITY(S_1) = c_1$   
We are moving the item into a set where it already is of the majority, so the empirical loss in  $S_1$  remains the same. Consider the next element adjacent to  $i_1$  that we shall call  $i_2$ :
  - i. If  $i_2$  is not of class  $c_1$  we have found another split point that has a lower empirical loss than the first one and is between elements of different Y-values.
  - ii. If  $i_2$  is also of class  $c_1$ , we can almost move it into  $S_1$ , decreasing the overall empirical loss. We continue to do this until we reach (i). We will always reach (i) since  $c_1$  is not the majority class in  $S_2$ .
- (b)  $MAJORITY(S_1) \neq c_1$ 
  - i.  $c_1$  becomes the majority after  $i_1$  is added  
Same as (a) above. We will eventually find another split point with a lower empirical loss.

- ii.  $c_1$  is not majority after  $i_1$  is added  
The empirical loss changes cancel.

This proves that our empirical loss minima occur at split points. The absolute empirical loss is the minimum among the local minima so therefore it must occur at a split point between elements of different Y-values.

TL;DR Either a split point between elements with different Y-values is a local minima, or if by moving the split point you can decrease the empirical loss, you can continue moving this split point until you reach another local minima at a split point between elements with different Y-values.

#### 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 probability of those errors occurring. 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  $\epsilon$ . For example, consider the case when  $K = 3$  and each hypothesis makes 2 error per 10 guesses. An adversarial set of hypotheses can budget one error for a legitimate mistake and the next error for a coordinated mistake. The latter is used when another hypothesis makes a mistake, and you “support” its error and derive a majority by using your second error.

	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