View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Purdue E-Pubs

Purdue University
Purdue e-Pubs

ECE Technical Reports

Electrical and Computer Engineering

3-1-1997

# On the Qualitative Behavior of Impurity-Based Splitting Rules I: The Minima-Free Property

Craig W. Codrington
*Purdue University School of Electrical Engineering*

Carla E. Brodley
*Purdue University School of Electrical Engineering*

Follow this and additional works at: http://docs.lib.purdue.edu/ecetr

Codrington, Craig W. and Brodley, Carla E., "On the Qualitative Behavior of Impurity-Based Splitting Rules I: The Minima-Free Property" (1997). *ECE Technical Reports.* Paper 86.
http://docs.lib.purdue.edu/ecetr/86

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

# On the Qualitative Behavior of Impurity-Based Splitting Rules I: The Minima-Free Property

Craig W. Codrington
Carla E. Brodley

TR-ECE 97-5
March 1997

School of Electrical
and Computer Engineering
Purdue University
West Lafayette, Indiana 47907-1285

# On the Qualitative Behavior of Impurity-Based Splitting Rules I: The Minima-Free Property

Craig W. Codrington[1]
School of Electrical Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907-1285
email: codringt@ecti.purdue.edu

Carla E. Brodley
School of Electrical Engineering
1285 Electrical Engineering Building
Purdue University
West Lafayette, IN 47907-1285
email: brodley@ecn.purdue.edu
tel:(317) 494-0635

'This research was supported by the Purdue Research Foundation.

# Abstract

We show that all strictly convex $\cap$ impurity measures lead to splits at boundary points, and furthermore show that certain rational splitting rules, notably the information gain ratio, also have this property. A slightly weaker result is shown to hold for impurity measures that are only convex $\cap$, such as Inaccuracy.

# 1 Introduction

The classification problem can be stated roughly as: given a set of features (or attributes), find a way of dividing up the feature space into cells such that each cell reflects some specified property of the data mapped to that cell. Finding such a division is non-trivial; the infinitude of ways of dividing up the space effectively rules out direct enumeration. The complexity of the search can be limited by using a divide-and-conquer strategy, that is, rather than dividing the feature space all at once into many cells, just divide it into two cells; each cell can then be viewed as an instance of the original problem, and similarly can be recursively subdivided.

This strategy necessarily imposes a hierarchical structure on the partition, which can be viewed naturally as a tree (specifically, a Decision Tree (Morgan & Sonquist, 1963)). Each node in the tree represents a cell in a partition of feature space, and each branching represents a division of the space. The cells corresponding to the leaves constitute a partition of the space.

The divide-and-conquer approach to dividing up the feature space produces a sequence of finer and finer partitions; as each partition corresponds to a tree, we obtain a sequence of trees

$$T_0, T_1, .$$

$T_0$ consists of a single leaf node representing the entire feature space, and for each $i$, $T_{i+1}$ is obtained from $T_i$ by replacing a leaf node of $T_i$ with an internal (non-leaf) node branching into two new leaf nodes. This corresponds to splitting a cell in the partition represented by $T_i$ into two new cells separated by a *decision boundary.* The process terminates when all cells are pure, that is, consist of data from only one class. To reduce overfitting, a pruning phase often follows, but we will not be concerned with this here.

So far the tree representation and the feature space representation are entirely equivalent, but it turns out that abstracting the feature space into a tree has bought us something, namely the ability to handle non-ordinal (also called categorical) attributes. As the name implies, such attributes assume values that have no inherent ordering; it follows that a decision boundary in feature space cannot be used to discriminate different values of a non-ordinal attribute, as an ordering of values is required to know on which side of the boundary a given data point falls. Such attributes pose no problem for a tree, however, which can specify tests on the value on a non-ordinal attribute that. determine which branch to send a data point down. In this paper we will be exclusively concerned with ordinal (also called continuous) attributes, which assume values that. have a natural ordering.

To construct a decision tree, we must answer the following questions: Given a tree $T_i$, which leaf of $T_i$ should be split? and how?

The answer to the first question, which leaf to split, is: it doesn't matter. For if we assume that

- splitting continues until all leaf nodes are pure, and
- when splitting a leaf node t, the splitting rule doesn't look at any other leaves,

then the same tree is obtained regardless of the order in which leaves are split (Dietterich, Kearns, & Mansour, 1996).

The second question, how to split, is still unresolved, judging from the extensive and growing literature on the subject (Breiman, Friedman, Olshen, & Stone, 1984; Brodley, 1995; Buntine & hiblett, 1992; Fayyad, 1994; Fayyad & Irani, 1990, 1992a, 1993; Goodman & Smyth, 1988; López de Mántaras, 1991; Lubinsky, 1995; Quinlan, 1990, 1996; Quinlan & Rivest, 1989). The accepted method is to select from a finite set of candidate splits that which minimizes a *splitting rule.*

The splitting rules we consider are based on an impurity function **I**, which measures the impurity of class labels in each leaf. For a leaf t with $N_i$ training samples from class $i$, for $i = 1, \ldots,$ m, the fraction of class $i$ labels is

$$p_i = \frac{N_i}{N}$$

where N $= N_1 + \ldots + N_m$. Then the impurity of leaf t is defined as

$$\begin{aligned} I(t) &= I(p_1, \ldots, p_m) \\ &= I(\vec{p}) \end{aligned}$$

where $\vec{p}$ is the class purity vector, defined by

$$\vec{p} = (p_1, \ldots, p_m)$$

We consider splitting rules of the form

$$\text{Choose the split that minimizes f} = \frac{|t_L|}{|t|} I(t_L) + \frac{|t_R|}{|t|} I(t_R) \tag{1}$$

where the minimization is carried out over a set of candidate splits of leaf t into leaves $t_L$ and $t_R$ ($|t|$ denotes the number of training samples in t and similarly for $|t_L|$ and $|t_R|$).

The splitting rule (1) can be justified on the basis of choosing the split that minimizes the expected impurity of the resulting tree with respect to the distribution of training data, which for a tree T is defined

$$\text{EII}\,|T] = \sum_{t \in \text{leaves}(T)} \frac{|t|}{N} I(t)$$

where $|t|$ is the number of training samples in leaf t, $I(t)$ is the impurity of this leaf, and leaves$(T)$ is the set of leaf nodes of T. If $T_{i+1}$ is the tree produced by splitting a leaf t into leaves $t_L$ and $t_R$, the expected impurity can be written

$$\begin{aligned} E[I\,|T_{i+1}] &= E[I\,|T_i] + \frac{|t_L|}{N} I(t_L) + \frac{|t_R|}{N} I(t_R) - \frac{|t|}{N} I(t) \\ &= E[I\,|T_i] + \frac{|t|}{N}\left(\frac{|t_L|}{|t|} I(t_L) + \frac{|t_R|}{|t|} I(t_R) - I(t)\right) \end{aligned}$$

Because $T_i$ and the leaf t to be split are fixed, the split that yields the tree $T_{i+1}$ of smallest average impurity is (1).

A variety of impurity functions have been proposed, including

- Inaccuracy, in which all samples other than those belonging to the majority class are counted as errors:

$$\text{Inacc}\,(p_1, \ldots, p_m) = 1 - \max_i p_i$$

- Gini (Breiman et al., 1984):

$$\text{Gini}\,(p_1, \ldots, p_m) = 1 - \sum_i p_i^2$$

- Entropy[2] (Lewis, 1962; Sethi & Sarvarayudu, 1982):

$$\text{Ent}\,(p_1,\ldots p_m) = \sum_{i=1}^{m} -p_i \log(p_i)$$

Entropy has become the impurity measure of choice, having found application in learning algorithms such as ID3 (Quinlan, 1983) and CN2 (Clark & Niblett, 1989). Representing the collection of training data by a tree introduces uncertainty in the sense that training samples mapped to the same leaf become indistinguishable from the point of view of the tree. Entropy is in some sense a measure of this uncertainty; if we use the tree to estimate the class label of a sample $\vec{X}$ drawn from the same distribution as the training data by outputting class label i with a probability equal to the proportion of class $i$ samples in the leaf containing $\vec{X}$, then the entropy of the tree is the number of bits it would take, on average, to correct the output of the tree.[3] Since partitioning a set reduces the number of possible arrangements, or equivalently, reduces its randomness, the decrease in entropy following a split reflects the additional information we have concerning the class of a sample; hence the amount by which the entropy is reduced following a split is called the information gain[4] (Quinlan, 1983):

$$\text{information gain} = \text{Ent}(t) - \left(\frac{|t_L|}{|t|}\text{Ent}(t_L) + \frac{|t_R|}{|t|}\text{Ent}(t_R)\right) \tag{2}$$

Decision tree induction can be viewed as a process of driving the uncertainty in the class labels to zero. A reasonable splitting rule is thus to choose the split that maximizes the information gain, or equivalently (since t is fixed), minimizes

$$f = \frac{|t_L|}{|t|}\text{Ent}(t_L) + \frac{|t_R|}{|t|}\text{Ent}(t_R) \tag{3}$$

It has been observed that the information gain favors attributes that take on a larger set of values over the training set (Quinlan, 1990). To counteract this tendency, Quinlan normalized the information gain by dividing it by the information gained by knowing which of $t_L$ or $t_R$ contains a given sample, yielding a quantity he called the information gain ratio (Quinlan, 1990):

$$\text{information gain ratio} = \frac{\text{information gain}}{-\frac{|t_L|}{|t|}\log\left(\frac{|t_L|}{|t|}\right) - \frac{|t_R|}{|t|}\log\left(\frac{|t_R|}{|t|}\right)} \tag{4}$$

Fayyad and Irani showed that entropy always cuts at boundary points[5] [6] (Fayyad & Irani, 1992b). As this is a consequence of the splitting rule having no local minima over uniform sequences (a sequence of samples belonging to the same class), we refer to this as the minima-free property. We show that this property holds not only for the entropy impurity measure, but for all strictly convex $\cap$ impurity measures. This property is of interest for several reasons:

[2] Using the Entropy impurity measure in (1) is equivalent to choosing the split that maximizes the mutual information between the attributes and classes, which was the method used in (Lewis, 1962; Sethi & Sarvarayudu, 1982).

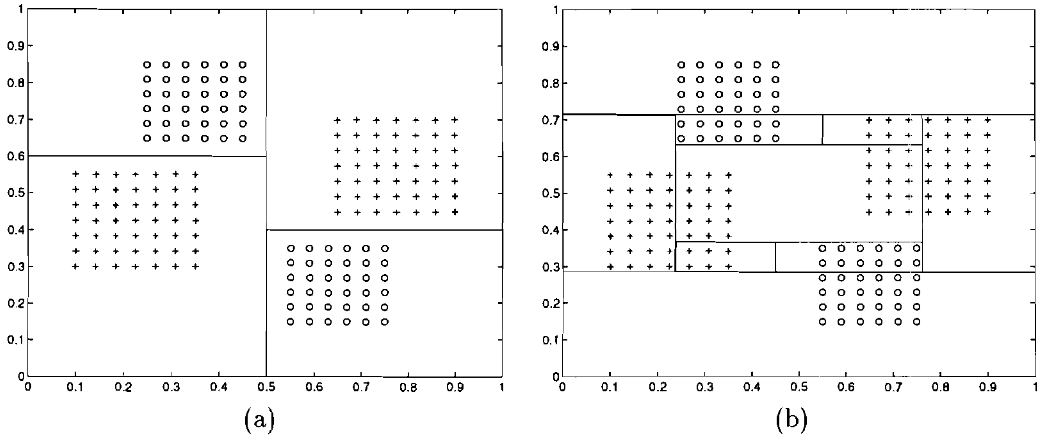
[3] This assumes that entropy is defined in terms of base 2 logs. We assume for the remainder of the paper that entropy is defined using base e logs, so that information is measured in *nats*, not *bits*. This is done without loss of generality to simplify differentiation.

[4] This is the reduction in entropy relative to the leaf t. The reduction in the average entropy of the tree is $\frac{|t|}{N}\left(\text{Ent}(t) - \left(\frac{|t_L|}{|t|}\text{Ent}(t_L) + \frac{|t_R|}{|t|}\text{Ent}(t_R)\right)\right) = \frac{|t|}{N}$ information gain.

[5] A rough definition of a boundary point is a point between samples belonging to different classes.

[6] In this paper we shall sometimes blur the distinction between splitting rules and impurity measures, so that the phrase "entropy always cuts at boundary points" should be interpreted as "the splitting rule that uses entropy as an impurity measure always cuts at boundary points".

- **Efficiency** – if it can be shown that the splitting rule being used cuts only at boundary points, then only boundary points need to be considered as potential cut points. This can represent a substantial savings in computation (Fayyad & Irani, 1992b).
- **Accuracy** – Intuition suggests that splitting rules that cut at interior points (i.e. non-boundary points) result in a finer partition of the space, and thus a larger tree, than splitting rules that disallow such cuts; such overfitting may lead to a reduction in predictive accuracy. This intuition is sometimes wrong; as Figure 1 shows, splitting between samples of the same class can occasionally lead to a *smaller* tree. However, this example does not change the fact that it is difficult to justify, on the basis of a one dimensional projection of the data along some attribute, cutting at an interior point rather than a boundary point, as we have no way of distinguishing the situation depicted in Figure 1, where it is a good idea to cut at an interior point, from the myriad other situations in which it is a bad idea.
- **Insight** – Knowing something about where a splitting rule splits may lead to a deeper understanding of its weaknesses, laying the groundwork for new splitting rules that address these weaknesses.



(a) (b)

Figure 1: An example showing that cutting between samples belonging to the same class is not always a bad idea. (a) Hand-drawn partition that cuts between samples of the same class. (b) Entropy-guided partition that does not.

The organization of this paper is as follows: Sections 2 (mathematical preliminaries) and 3 (notation) provide a framework in which to discuss the theorems and proofs presented in Section 4; the theory is then applied to specific splitting rules in Section 5.

## 2 Mathematical Preliminaries

The following is intended as a brief review.

**Definition 1** A *function $f$ defined on a convex set $\Omega$ is* <u>*convex* $\cap$</u> *if for all* $5,\vec{y} \in \Omega$ *and for all* $\lambda \in [0,1]$

$$f(\lambda\vec{x} + (1-\lambda)\vec{y}) \geq \lambda f(\vec{x}) + (1-\lambda)f(\vec{y})$$

Definition 2 A *function $f$ defined on a convex set $\Omega$ is* <u>*strictly convex* $\cap$</u> *if for all $\vec{x}, \vec{y} \in \Omega$, $\vec{x}$ # $\vec{y}$, and for all $\lambda \in (0,1)$*

$$f(\lambda\vec{x} + (1 - \lambda)\vec{y}) > \lambda f(\vec{x}) + (1 - \lambda)f(\vec{y})$$

Definition 3 *The* <u>*Hessian matrix*</u> V$^2$1 *of an impurity function $I(p_1, \ldots, p_m)$ is the matrix of second derivatives dejined by*

$$\nabla^2 I = \begin{bmatrix} \frac{\partial^2 I}{\partial p_1 \partial p_1} & \frac{\partial^2 I}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_1 \partial p_m} \\ \frac{\partial^2 I}{\partial p_2 \partial p_1} & \frac{\partial^2 I}{\partial p_2 \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_2 \partial p_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 I}{\partial p_m \partial p_1} & \frac{\partial^2 I}{\partial p_m \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_m \partial p_m} \end{bmatrix}$$

Fact 1 *(Negative semidefinite matrices (Strang, 1988, p. 339))*
*The following are equivalent:*

1. *The matrix M is negative semidefinite.*
2. *All eigenvalues of M are $\leq 0$.*
3. *$\vec{x}^T M \vec{x} \leq 0$ for all $\vec{x}$.*

Fact 2 *(Negative definite matrices (Strung, 1988, p. 331))*
*The following are equivalent:*

1. *The matrix M is negative definite.*
2. *All eigenvalues of M are $<$ 0.*
3. *$\vec{x}^T M \vec{x} < 0$ for all $\vec{x} \neq 0$.*

Fact 3 *(Relation between strict convexity and negative definzte Hessian (Peressini, Sullivan, & Uhl, 1988, p 54))*
*For a function $f$ dejined on a convex set $\Omega$, if the Hessian off is negative definite, then $f$ is strictly convex* n.

Fact 4 *(Relation between convexity and negative semidefinite Hessian (Peressini et al., 1988, p. 54))*
*For a function $f$ defined on a convex set $\Omega$, $f$ is convex* n *if and only if its Hessian is negative semidefinite.*

Fact 5 *(Conditions for a Minimum (Peressini et al., 1988, p. 3))*
*A point $x^*$ is a strict local minimum of a function $f(x)$ if $f'(x^*) = 0$ and $f''(x^*) > 0$.*

# 3 Notation

The following strategy for partitioning a leaf $t$ into leaves $t_L$ and $t_R$ was used in (Breirnan et al., *1984):*

1. For each attribute $A$, project the training samples down to the A-axis and evaluate the splitting rule $f$ at all points that are halfway between adjacent training samples. Let $f(c(A))$ denote the minimum of the splitting rule over all such points, where $c(A)$ is the point that achieves the minimum.

2. Split along the attribute $A_0$ that minimizes $f(c(A))$ (i.e. $A_0 = \arg\min_A f(c(A))$) at the cut point $c(A_0)$. This defines a decision boundary (actually a plane) in feature space that splits $t$ into $t_L$ and $t_R$.

The process of projecting the training data down to the A-axis produces a *sequence:*

**Definition 4** A *sequence is an ordering of the samples contained in some interval after projecting the training data onto a feature space axis corresponding to some attribute* A, *or more generally, onto an arbitrary vector in feature space.*

We will be particularly concerned with sequences of samples that all belong to the same class:

**Definition 5 A** *uniform sequence is a sequence that contains only samples belonging to the same class.*

We assume we are at the stage of the above algorithm where we are selecting the best cut point for some ,tribute A. Let N denote the number of training samples in the node to be split, and let the variable denote the position of the cut point along A. Since the splitting rules we consider depend only on the number of samples of each class on each side of the cut point, and not on their actual positions, without loss of generality we can assume that, after sorting the N training samples in order of increasing A, the first sample occurs at n = 0.5, the second at n = 1.5, and so on. Then the set of cuts at n = 1,..., N – 1 exhaust all possible ways of cutting between adjacent samples.

We have assumed here that the values produced by projecting the data onto the A axis are unique, that is, there are no repeated values. We will develop the theory under this simplifying assumption, and then show in Section **4.6** that the theory remains valid even when repeated values are allowed.

Many of the results in this paper relate to boundary points, which, for the restricted case we are considering (no repeated values), can be defined as follows:

**Definition 6** *Given a sequence of training samples, a boundary point is a point between samples belonging to different classes.*

Later, in Section **4.6,** we shall give a more precise definition which also applies to the case of repeated values.

We assume that each training sample belongs to one of m classes, numbered from 1 to m. For a given cut point n, we define, for $i = 1, \ldots$ m

$$\begin{aligned}
L_i(n) &\equiv \text{the number of class } i \text{ samples to the left of n.}\\
R_i(n) &\equiv \text{the number of class } i \text{ samples to the right of n.}\\
L(n) &\equiv \text{the total number of samples to the left of n.}\\
R(n) &\equiv \text{the total number of samples to the right of n.}\\
\ell_i(n) &\equiv \text{the fraction of class i samples to the left of n.}\\
&= \frac{L_i(n)}{L(n)}\\
r_i(n) &\equiv \text{the fraction of class i samples to the left of n.}\\
&= \frac{R_i(n)}{R(n)}
\end{aligned}$$

Figure 2 may help to make this notation more concrete.

Collect the fractions $\ell_i(n)$ into a vector $\vec{\ell}(n)$:

$$\vec{\ell}(n) = [\ell_1(n)\ \ell_2(n)\ \ldots\ \ell_m(n)]^T$$

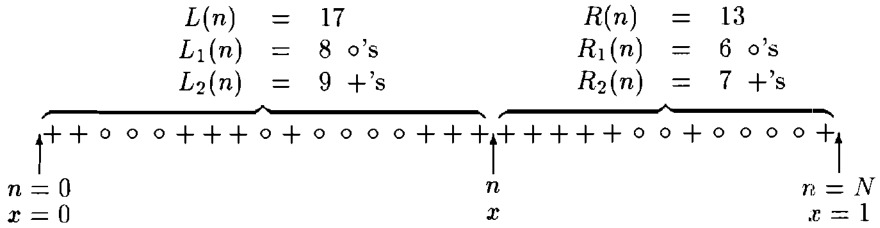


Figure 2: Notation used for model.

Then the impurity of the samples to the left of $n$ is

$$I(\ell_1(n), \ell_2(n), \ \ldots \ \ell_m(n))$$

which we condense to

$$I(\vec{\ell}(n))$$

Collect the fractions $r_i(n)$ into a vector $\vec{r}(n)$:

$$\vec{r}(n) = [r_1(n) r_2(n) \ \ldots \ r_m(n)]_T$$

Then the impurity of the samples to the right of $n$ is

$$I(r_1(n), r_2(n), \ \ldots \ r_m(n))$$

which we condense to

$$I(\vec{r}(n))$$

The splitting rule *(1)* becomes

$$f(n) = \frac{n}{N} I(\vec{\ell}(n)) + (1 - \frac{n}{N}) I(\vec{r}(n)) \tag{5}$$

We introduce the continuous variable $x = n/N$ so that we can use calculus to investigate the necessary and sufficient conditions for $f$ to have a minimum. Substituting $n = Nx$ into (5), we have

$$f(Nx) = \frac{Nx}{N} I(\vec{\ell}(Nx)) + \frac{N - Nx}{N} I(\vec{r}(Nx)) \tag{6}$$

which becomes

$$f(x) = xI(\vec{\ell}(x)) + (1 - x) I(\vec{r}(x)) \tag{7}$$

under the following identifications:

$$\begin{array}{lcl} f(X) & & f(n) \text{ evaluated at } n = Nx. \\ \vec{\ell}(x) & \equiv & \vec{\ell}(n) \text{ evaluated at } n = Nx. \\ \vec{r}(x) & \equiv & \vec{r}(n) \text{ evaluated at } n = Nx. \end{array}$$

Note that:

- As x varies from 0 to 1, $n$ varies from 0 to N.
- $\vec{\ell}(n)$ and $\vec{r}(n)$ (and hence $\vec{\ell}(x)$ and $\vec{r}(x)$) take on values in the "purity space"

$$\Omega = \left\{ (p_1, \ldots, p_m) \,:\, 0 \leq p_i \leq 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m} p_i = 1 \right\}$$

We shall sometimes write $f(x;I)$ in place of $f(x)$ to make explicit which impurity measure is being used

# 4 Theorems and Proofs

In this section we show that:

- Splitting rules based on strictly convex ∩ impurity measures cut only at boundary points.
- Splitting rules based on strictly convex ∩ impurity measures are strictly decreasing over the first uniform sequence and strictly increasing over the last.
- The set of global minimizers of splitting rules based on convex ∩ impurity measures includes at least one boundary point, but may also include all interior and boundary points of one or more uniform sequences (see Figure 3).
- Splitting rules based on convex ∩ impurity measures are non-increasing over the first uniform sequence and non-decreasing over the last.
- Splitting rules of the form
$$F(x) = \frac{g(x)}{h(x)}$$
cut only at boundary points, under certain conditions on g and $h$.
- The above results hold even when repeated values of an attribute occur over the training data.

Below we develop a model for how impurity-based splitting rules behave over uniform sequences; this model forms the basis for many of the results in this paper.

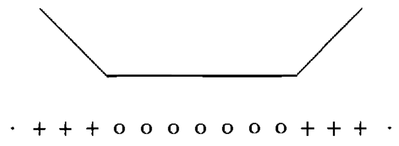


Figure 3: Hypothetical behavior at minimum of splitting rule $f(x;I)$ for a convex ∩ innpurity measure $I$

## 4.1 A Model for Impurity-Based Splitting Rules over Uniform Sequences

We are interested in studying how the splitting rule behaves as the cut point is varied within a sequence of samples belonging to the same class. We choose one such sequence, assumed to extend from $n_a$ to $n_b$ and to consist entirely of samples from class k. We fix a reference point at some position *:no* within the chosen sequence, and define, for $1 \leq$ i $\leq$ m:

$L_i$ the number of class i samples to the left of *no.*
$R_i$ $\equiv$ the number of class i samples to the right of *no.*
$L$ $\equiv$ the total number of samples to the left of $n_0$.
$R$ $\equiv$ the total number of samples to the right of $n_0$.

Figure 4 may help to make this discussion more concrete. For an arbitrary position n within the uniform

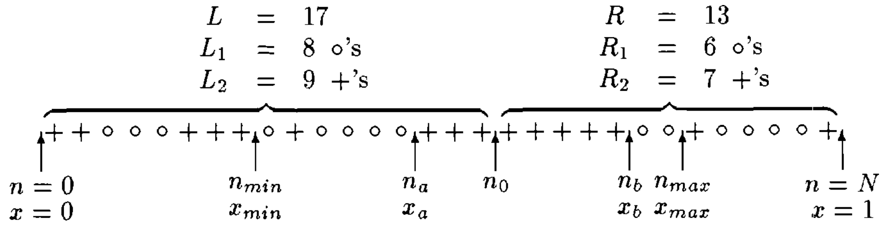


Figure 4: Notation used for model.

sequence $[n_a, n_b]$, the number of class k samples to the left and right of $n$ are respectively $n - L + L_k$ and $L - n + R_k$, while the number of class i samples to the left and right of $n$ are respectively $L_i$ and $R_i$, for $i \neq$ k. The number of samples to the left and right of $n$ are respectively $n$ and $N - n$. The fraction of the samples to the left of $n$ that belong to class i is

$$\ell_i(n) = \begin{cases} \frac{L_i}{n} & i \neq k \\ \frac{n-L+L_k}{n} & i = k \end{cases}$$

Collect these fractions into a vector $\vec{\ell}(n)$:

$$\vec{\ell}(n) = [\ell_1(n)\ell_2(n) \ \ldots \ \ell_m(n)]^T$$

Then the impurity of the samples to the left of $n$ is

$$I(\ell_1(n), \ell_2(n), \ \ldots \ \ell_m(n))$$

which we condense to

$$I(\vec{\ell}(n))$$

The fraction of the samples to the right of $n$ that belong to class i is

$$r_i(n) = \begin{cases} \frac{R_i}{N-n} & i \neq k \\ \frac{L-n+R_k}{N-n} & i = k \end{cases}$$

Collect these fractions into a vector $\vec{r}(n)$:

$$\vec{r}(n) = [r_1(n)\ r_2(n)\ \ldots\ r_m(n)]^T$$

Then the impurity of the samples to the right of $n$ is

$$I(r_1(n), r_2(n),\ \ldots\ r_m(n))$$

which we condense to

$$I(\vec{r}(n))$$

Then our discrete model of how the splitting rule (5) varies as a function of the cut point $n$ is

$$f(n) = \frac{n}{N} I(\vec{\ell}(n)) + (1 - \frac{n}{N}) I(\vec{r}(n)) \tag{8}$$

where $\vec{\ell}(n)$ and $\vec{r}(n)$ have components

$$\begin{aligned} \ell_i(n) &= \begin{cases} \frac{L_i}{n} & i \neq k \\ \frac{n-L+L_k}{n} & i = k \end{cases} \\ &= \begin{cases} \frac{L_i}{n} & i \neq k \\ \frac{n-n_{min}}{n} & i = k \end{cases} \end{aligned}$$

and

$$\begin{aligned} r_i(n) &= \begin{cases} \frac{R_i}{N-n} & i\ f\ k \\ \frac{L+R_k-n}{N-n} & i = k \end{cases} \\ &= \begin{cases} \frac{R_i}{N-n} & i\ f\ k \\ \frac{n_{max}-n}{N-n} & i = k \end{cases} \end{aligned}$$

The model is valid for $n = n_a, \ldots, n_b$, since $n$, and $n_b$ mark the boundaries of the uniform sequence of interest. The quantities $n_{min}$ and $n_{max}$ appearing above are defined by

$$\begin{aligned} n_{min} &= L - L_k \\ n_{max} &= L + R_k \end{aligned}$$

$n_{min}$ is where the sequence would begin if all class $k$ points to the left of the sequence were contiguous with it, and thus represents a "worst case[7]" starting point for the sequence. Similarly, $n_{max}$ is where the sequence would end if all class $k$ points to the right of the sequence were contiguous with it, and thus represents a "worst case" end point for the sequence.

As before, to get a handle on what is happening between the samples we introduce the continuous variable $x = n/N$ and substitute $n = Nx$ into the above expression to get

$$f(Nx) = \frac{Nx}{N} I(\vec{\ell}(Nx)) + \frac{N - Nx}{N} I(\vec{r}(Nx)) \tag{9}$$

[7] Or best case, depending on how you look at it.

where $\vec{\ell}(Nx)$ and $\vec{r}(Nx)$ have components

$$\begin{aligned}
\ell_i(Nx) &= \begin{cases} \frac{L_i}{Nx} & i \neq k \\ \frac{Nx-L+L_k}{Nx} & i = k \end{cases} \\
&= \begin{cases} \frac{L_i}{Nx} & i \neq k \\ 1 - \frac{L-L_k}{Nx} & i = k \end{cases} \\
&= \begin{cases} \frac{L_i}{Nx} & i \neq k \\ 1 - \frac{x_{min}}{x} & i = k \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
r_i(Nx) &= \begin{cases} \frac{R_i}{N-Nx} & i \neq k \\ \frac{L-Nx+R_k}{N-Nx} & i = k \end{cases} \\
&= \begin{cases} \frac{R_i}{N-Nx} & i \neq k \\ \frac{L+R_k-N+N-Nx}{N-Nx} & i = k \end{cases} \\
&= \begin{cases} \frac{R_i}{N(1-x)} & i \neq k \\ \frac{L+R_k-N}{N(1-x)} + 1 & i = k \end{cases} \\
&= \begin{cases} \frac{R_i}{N(1-x)} & i \neq k \\ \frac{x_{max}-1}{1-x} + 1 & i = k \end{cases}
\end{aligned}$$

The quantites $x_{min}$ and $x_{max}$ appearing above are defined by

$$\begin{aligned}
x_{min} &= \frac{L-L_k}{N} \\
x_{max} &= \frac{L+R_k}{N}
\end{aligned}$$

$x_{min}$ is where the sequence would begin if all class k points to the left of the sequence were contiguous with it, and thus represents a "worst case" starting point for the sequence. Similarly, $x_{max}$ is where the sequence would end if all class k points to the right of the sequence were contiguous with it, and thus represents a "worst case" end point for the sequence.

Under the identifications

$$\begin{aligned}
f(x) & \quad f(n) \text{ evaluated at } n = Nx. \\
\vec{\ell}(x) &\equiv \vec{\ell}(n) \text{ evaluated at } n = Nx. \\
\vec{r}(x) &\equiv \vec{r}(n) \text{ evaluated at } n = Nx.
\end{aligned}$$

our continuous model of how the average impurity varies as a function of the cut point $x$ becomes

$$f(x) = xI(\vec{\ell}(x)) + (1-x)I(\vec{r}(x)) \tag{10}$$

where $\vec{\ell}(x)$ and $\vec{r}(x)$ have components

$$\ell_i(x) = \begin{cases} \frac{L_i}{Nx} & i \neq k \\ 1 - \frac{x_{min}}{x} & i = k \end{cases}$$

and

$$r_i(x) = \begin{cases} \frac{R_i}{N(1-x)} & i \neq k \\ \frac{x_{max}-1}{1-x} + 1 & i = k \end{cases}$$

The model is valid for x between x, $= \frac{n_a}{N}$ and $x_b = \frac{n_b}{N}$, since x, and $x_b$ mark the boundaries of the sequence of interest in x-coordinates.

## 4.2 Strictly Convex n Impurity Measures

The main result of this section is

**Theorem 4.1** *If the impurity measure* I *is strictly convex* $\cap$ *on*

$$\Omega = \{(p_1, \ldots, p_m) : 0 \leq p_i \leq 1 \textit{ foralli, and } \sum_{i=1}^{m} p_i = 1)$$

*then the set of global minima of the splitting rule*

$$f(n; \mathrm{I}) = \frac{\mathrm{n}}{N} I(\vec{\ell}(n)) + (1 - \frac{\mathrm{n}}{N}) I(\vec{r}(n))$$

*consists entirely of boundary points.*

**Proof:** The proof consists of the following steps:

- We first show that f(x), the continuous analog of f(n), is strictly convex n over uniform sequences (Lemma 4.1).
- We then show that the convexity off (x) over a uniform sequence implies that f(n) attains its minimum at a boundary point of the sequence (Lemma 4.2).
- Lemma 4.1 does not apply to the first or last uniform sequences, as the model breaks down when x = 0 or x = 1. We handle these special cases by showing that $f(n)$ is strictly decreasing over the first uniform sequence (Lemma 4.5) and strictly increasing over the last (Lemma 4.6), so that over any uniform sequence, f(n) attains its minimum at a boundary point of the sequence.
- Finally, Lemma 4.7 shows that if over any uniform sequence f(n) attains its minimum at a boundary point of the sequence, then $f(n)$ attains its global minimum at a boundary point (note however that the global minimum is not necessarily unique).

□

We first show that f(x) is strictly convex n over uniform sequences:

**Lemma 4.1** ***If*** *the impurity measure* I *is strictly convex* $\cap$ *on*

$$\Omega = \{(p_1, \ldots, p_m) : 0 \leq p_i \leq 1 \textit{ for all } i, \textit{ and } \sum_{i=1}^{m} p_i = 1\}$$

*then for all points* x,, $x_b$ *that delimit uniform sequences, the splitting rule* $f(x)$ *is strictly convex* n *on* [x,, $x_b$], *where* 0 < x, < $x_b$ < *1.*

**Proof:** Let $x, y \in [x_,, x_b]$, $x \neq y$, and $A \in (0, I)$. To show that $f(x)$ is strictly convex ∩, we must show that

$$f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda)f(y)$$

or, expanding $f$ in terms of $I$, that

$$\begin{aligned} &(Ax + (1-\lambda)y)I(\vec{\ell}(\lambda x + (1-\lambda)y)) + (1 - (\lambda x + (1-\lambda)y))\, I(\vec{r}(\lambda x + (1-\lambda)y)) \\ > {}& \lambda\left(xI(\vec{\ell}(x)) + (1-x)I(\vec{r}(x))\right) + (1-\lambda)\left(yI(\vec{\ell}(y)) + (1-y)I(\vec{r}(y))\right) \end{aligned}$$

or, after rearranging terms, that

$$\begin{aligned} &(\lambda x + (1-\lambda)y)I(\vec{\ell}(\lambda x + (1-\lambda)y)) + (1 - (\lambda x + (1-\lambda)y))\, I(\vec{r}(\lambda x + (1-\lambda)y)) \\ > {}& \left(\lambda xI(\vec{\ell}(x)) + (1-\lambda)yI(\vec{\ell}(y))\right) + (\lambda(1-x)I(\vec{r}(x)) + (1-\lambda)(1-y)I(\vec{r}(y))) \end{aligned}$$

This is true if

$$(Ax + (1-\lambda)y)I(\vec{\ell}(\lambda x + (1-\lambda)y)) > \lambda xI(\vec{\ell}(x)) + (1-\lambda)yI(\vec{\ell}(y)) \tag{11}$$

and

$$(1 - (\lambda x + (1-\lambda)y))\, I(\vec{r}(\lambda x + (1-\lambda)y)) > \lambda(1-x)I(\vec{r}(x)) + (1-\lambda)(1-y)I(\vec{r}(y)) \tag{12}$$

These are verified in Claims *1* and *2*.

**Claim 1** *(11)* holds.
Let

$$q = \frac{Ax}{\lambda x + (1-\lambda)y} \tag{13}$$

**Claim 1a** $0 < q < 1$.
$0 < x, y, A < 1$ implies $0 < Ax + (1-\lambda)y < 1$ and the claim follows.

**Claim 1b** $\vec{\ell}(x) \neq \vec{\ell}(y)$.
Assume that $\vec{\ell}(x) = \vec{\ell}(y)$. There exists an i such that $L_i \neq 0$. If i $\neq$ k, then

$$\frac{L_i}{Nx} = \frac{L_i}{Ny}$$

which yields $x = y$, a contradiction. If $i =$ k, then

$$1 - \frac{x_{min}}{x} = 1 - \frac{x_{min}}{Y}$$

which again yields $x = y$. Thus in either case we get a contradiction, so we must have $\vec{\ell}(x) \neq \vec{\ell}(y)$.

Since $I$ is strictly convex ∩, $0 < q < 1$, and $\vec{\ell}(x) \neq \vec{\ell}(y)$, we have

$$I(q\vec{\ell}(x) + (1-q)\vec{\ell}(y)) > qI(\vec{\ell}(x)) + (1-q)I(\vec{\ell}(y)) \tag{14}$$

Since

$$\begin{aligned} 1 - q &= 1 - \frac{\lambda x}{\lambda x + (1-\lambda)y} \\ &= \frac{(1-\lambda)y}{\lambda x + (1-\lambda)y} \end{aligned} \tag{15}$$

the argument of the impurity measure on the left of **(14)** becomes

$$\begin{aligned} q\vec{\ell}(x) + (1-q)\vec{\ell}(y) &= \left(\frac{\lambda x}{\lambda x + (1-\lambda)y}\right)\vec{\ell}(x) + \left(\frac{(1-\lambda)y}{\lambda x + (1-\lambda)y}\right)\vec{\ell}(y) \\ &= \frac{\lambda x\vec{\ell}(x) + (1-\lambda)y\vec{\ell}(y)}{\lambda x + (1-\lambda)y} \end{aligned} \tag{16}$$

We now show that

$$\frac{\lambda x\vec{\ell}(x) + (1-\lambda)y\vec{\ell}(y)}{\lambda x + (1-\lambda)y} = \vec{\ell}(\lambda x + (1-\lambda)y) \tag{17}$$

For $i \neq$ k

$$\begin{aligned} \frac{\lambda x \ell_i(x) + (1-\lambda)y\ell_i(y)}{\lambda x + (1-\lambda)y} &= \frac{\lambda x\frac{L_i}{Nx} + (1-\lambda)y\frac{L_i}{Ny}}{\lambda x + (1-\lambda)y} \\ &= \frac{\lambda\frac{L_i}{N} + (1-\lambda)\frac{L_i}{N}}{\lambda x + (1-\lambda)y} \\ &= \frac{\frac{L_i}{N}}{\lambda x + (1-\lambda)y} \\ &= \frac{L_i}{N(\lambda x + (1-\lambda)y)} \\ &= \ell_i(\lambda x + (1-\lambda)y) \end{aligned}$$

and for $i = k$

$$\begin{aligned} \frac{\lambda x \ell_k(x) + (1-\lambda)y\ell_k(y)}{\lambda x + (1-\lambda)y} &= \frac{\lambda x(1 - \frac{x_{min}}{x}) + (1-\lambda)y(1 - \frac{x_{min}}{y})}{\lambda x + (1-\lambda)y} \\ &= \frac{\lambda x - \lambda x_{min} + (1-\lambda)y - (1-\lambda)x_{min}}{\lambda x + (1-\lambda)y} \\ &= \frac{\lambda x + (1-\lambda)y - \lambda x_{min} - (1-\lambda)x_{min}}{\lambda x + (1-\lambda)y} \\ &= \frac{\lambda x + (1-\lambda)y - \lambda x_{min} - x_{min} + \lambda x_{min}}{\lambda x + (1-\lambda)y} \\ &= \frac{\lambda x + (1-\lambda)y - x_{min}}{\lambda x + (1-\lambda)y} \\ &= 1 - \frac{x_{min}}{\lambda x + (1-\lambda)y} \\ &= \ell_k(\lambda x + (1-\lambda)y) \end{aligned}$$

so that *(17)* holds. *(16)* and (17) imply

$$\begin{aligned} q\vec{\ell}(x) + (1-q)\vec{\ell}(y) &= \frac{\lambda x\vec{\ell}(x) + (1-\lambda)y\vec{\ell}(y)}{\lambda x + (1-\lambda)y} \\ &= \vec{\ell}(\lambda x + (1-\lambda)y) \end{aligned} \tag{18}$$

Substituting (13), (15), and (18) into (14) gives

$$I(\vec{\ell}(\lambda x + (1-\lambda)y)) > \frac{\lambda x}{\lambda x + (1-\lambda)y} I(\vec{\ell}(x)) + \frac{(1-\lambda)y}{\lambda x + (1-\lambda)y} I(\vec{\ell}(y))$$

or, multiplying thru by $\lambda x + (1-\lambda)y \neq 0$

$$(\lambda x + (1-\lambda)y) I(\vec{\ell}(\lambda x + (1-\lambda)y)) > \lambda x I(\vec{\ell}(x)) + (1-\lambda) y I(\vec{\ell}(y))$$

as claimed.

**Claim 2** (12) holds.
Let

$$q = \frac{\lambda(1-x)}{1-(\lambda x + (1-\lambda)y)} \tag{19}$$

**Claim 2a** $0 < q < 1$.
$0 < x, y, \lambda < 1$ implies $0 < \lambda x + (1-\lambda)y < 1$, which in turn implies that $0 < 1-(\lambda x+(1-\lambda)y) < 1$ and hence that $q > 0$. We show that $q < 1$ by contradiction. Assume that $q \geq 1$. Then

$$\begin{aligned} q \geq 1 \quad &\Rightarrow \quad \frac{\lambda(1-x)}{1-(\lambda x + (1-\lambda)y)} \geq 1 \\ &\Rightarrow \quad \lambda(1-x) \geq 1 - (\lambda x + (1-\lambda)y) \\ &\Rightarrow \quad \lambda - \lambda x \geq 1 - \lambda x - (1-\lambda)y) \\ &\Rightarrow \quad (1-\lambda)y \geq 1 - \lambda \\ &\Rightarrow \quad y \geq 1 \end{aligned}$$

which is a contradiction. The claim follows.

**Claim 2b** $\vec{r}(x) \neq \vec{r}(y)$.
Assume that $\vec{r}(x) = \vec{r}(y)$. There exists an $i$ such that $R_i \neq 0$. If $i \neq k$, then

$$\frac{R_i}{N(1-x)} = \frac{R_i}{N(1-y)}$$

which yields $x = y$, a contradiction. If $i = k$, then

$$1 + \frac{x_{max} - 1}{1-x} = 1 + \frac{x_{max} - 1}{1-y}$$

which again yields $x = y$. Thus in either case we get a contradiction, so we must have $\vec{r}(x) \neq \vec{r}(y)$.

Since **I** is strictly convex $\cap$, $0 < q < 1$, and $\vec{r}(x) \neq \vec{r}(y)$, we have

$$I(q\vec{r}(x) + (1-q)\vec{r}(y)) > qI(\vec{r}(x)) + (1-q)I(\vec{r}(y)) \tag{20}$$

Since

$$1 - q \quad = \quad 1 - \frac{\lambda(1-x)}{1-(\lambda x + (1-\lambda)y)}$$

$$\begin{aligned}
& - \frac{1-\ (Ax+(1-\lambda)y)-\lambda(1-x)}{1-(Ax+\$1-\lambda)y)} \\
& - \frac{1-Ax-(1-\lambda)y-A+Ax}{1-(Ax+(1-A)y)} \\
& - \frac{(1-A)-(1-\lambda)y}{1-(Ax+(1-\lambda)y)} \\
& - \frac{(1-\lambda)(1-y)}{1-(Ax+(1-\lambda)y)} && (21)
\end{aligned}$$

the argument of the impurity measure on the left of *(20)* becomes

$$\begin{aligned}
q\vec{r}(x)+(1-q)\vec{r}(y) &= \left(\frac{\lambda(1-x)}{1-(\lambda x+(1-\lambda)y)}\right)\vec{r}(x)+\left(\frac{(1-\lambda)(1-y)}{1-(\lambda x+(1-\lambda)y)}\right)\vec{r}(y) \\
&= \frac{\lambda(1-x)\vec{r}(x)+(1-\lambda)(1-y)\vec{r}(y)}{1-(\lambda x+(1-\lambda)y)} && (22)
\end{aligned}$$

We now show that

$$\frac{\lambda(1-x)\vec{r}(x)+(1-\lambda)(1-y)\vec{r}(y)}{1-(Ax+(1-\lambda)y)} = \vec{r}(\lambda x + (1-\lambda)y) \qquad (23)$$

For $i \neq$ k

$$\begin{aligned}
\frac{\lambda(1-x)r_i(x)+(1-\lambda)(1-y)r_i(y)}{1-(\lambda x+(1-\lambda)y)} &= \frac{\lambda(1-x)\frac{R_i}{N(1-x)}+(1-\lambda)(1-y)\frac{R_i}{N(1-y)}}{1-(\lambda x+(1-\lambda)y)} \\
&- \frac{\lambda\frac{R_i}{N}+(1-\lambda)\frac{R_i}{N}}{1-(Ax+(1-\lambda)y)} \\
&- \frac{\frac{R_i}{N}}{1-(Ax+(1-\lambda)y)} \\
&= \frac{R_i}{N(I-(\lambda x+(1-\lambda)y))} \\
&= r_i(\lambda x+(1-\lambda)y)
\end{aligned}$$

and for i = k

$$\begin{aligned}
\frac{\lambda(1-x)r_k(x)+(1-\lambda)(1-y)r_k(y)}{1-(Ax+(1-\lambda)y)} &- \frac{\lambda(1-x)(1+\frac{x_{max}-1}{1-x})+(1-\lambda)(1-y)(1+\frac{x_{max}-1}{1-x})}{1-(Ax+(1-\lambda)y)} \\
&- \frac{\lambda(1-x)+\lambda(x_{max}-1)+(1-\lambda)(1-y)+(1-\lambda)(x_{max}-1)}{1-(Ax+(1-\lambda)y)} \\
&- \frac{\lambda(1-x)+(1-\lambda)(1-y)+(x_{max}-1)}{1-(Ax+(1-\lambda)y)} \\
&= \frac{\lambda(1-x)}{1-(Ax+(1-\lambda)y)} - \frac{(1-\lambda)(1-y)}{1-(Ax+(1-\lambda)y)} + \frac{x_{max}-1}{1-(Ax+(1-\lambda)y)} \\
&= q+(1-q)+\frac{x_{max}-1}{1-(\lambda x+(1-\lambda)y)}
\end{aligned}$$

$$\begin{aligned} &= 1 + \frac{x_{max} - 1}{1 - (Ax + (1 - \lambda)y)} \\ &= r_k(\lambda x + (1 - \lambda)y) \end{aligned}$$

so that (23)holds. (22)and (23)imply

$$\begin{aligned} q\vec{r}(x) + (1-q)\vec{r}(y) &= \frac{\lambda(1-x)\vec{r}(x) + (1-\lambda)(1-y)\vec{r}(y)}{1 - (\lambda x + (1-\lambda)y)} \\ &= \vec{r}(\lambda x + (1-\lambda)y) \qquad (24) \end{aligned}$$

Substituting (19), (21), and (24)into (20)gives

$$I(\vec{r}(\lambda x + (1-\lambda)y)) > \frac{\lambda(1-x)}{1-(\lambda x + (1-\lambda)y)} I(\vec{r}(x)) + \frac{(1-\lambda)(1-y)}{1-(\lambda x + (1-\lambda)y)} I(\vec{r}(y))$$

or, multiplying thru by $1 - (Ax + (1 - \lambda)y) \neq 0$

$$(1 - (Ax + (1-\lambda)y))I(\vec{r}(\lambda x + (1-\lambda)y)) > \lambda(1-x)I(\vec{r}(x)) + (1-\lambda)(1-y)I(\vec{r}(y))$$

as claimed

Thus we have shown that $f(x)$ is strictly convex ∩ on $[x_a, x_b]$, as required. □

We next show that convexity of $f(x)$over a uniform sequence implies that $f(n)$attains its minimum at a boundary point of the sequence:

**Lemma** 4.2 *If $f(x)$ is strictly convex ∩ on $[x_a, x_b]$, then $f(n)$ attains its minimum at a boundary point, that is, at either $n_a$ or $n_b$.*

**Proof:** Let $\mathcal{L}_{ab}$ denote the line segment connecting the points $(x_a, f(x_a))$ and $(x_b, f(x_b))$. We have three cases to consider:

**Case 1** $f(x_a) < f(x_b)$
In this case, $\mathcal{L}_{ab}$ has positive slope (see Figure 5). Since $f(x)$ is strictly convex ∩, it lies above this line for all $x \in (x_a, x_b)$; consequently

$$f(x_a) < f(x) \text{ for all } x \in (x_a, x_b)$$

which implies, since $f(n)$is just $f(x)$ sampled at the points $\frac{n}{N}$, that

$$f(n_a) < f(n) \text{ for alln} \in \{n_a + 1, \ldots n_b\}$$

**Case 2** $f(x_a) > f(x_b)$
In this case, $\mathcal{L}_{ab}$ has negative slope (see Figure 6). Since $f(x)$ is strictly convex ∩, it lies above this line for all $x \in (x_a, x_b)$; consequently

$$f(x_b) < f(x) \text{ for all } x \in (x_a, x_b)$$

which implies, since $f(n)$is just $f(x)$sampled at the points $\frac{n}{N}$, that

$$f(n_b) < f(n) \text{ for all } n \in \{n,, \ldots n_b - 1\}$$

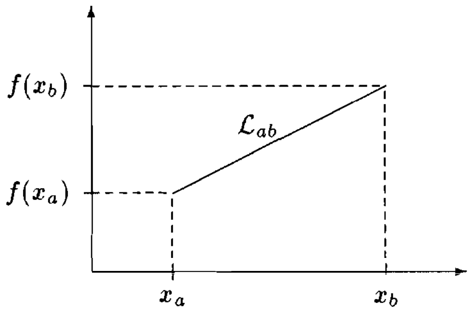


Figure 5: Case *1:* $f(x_a) \blacktriangleleft f(x_b)$.

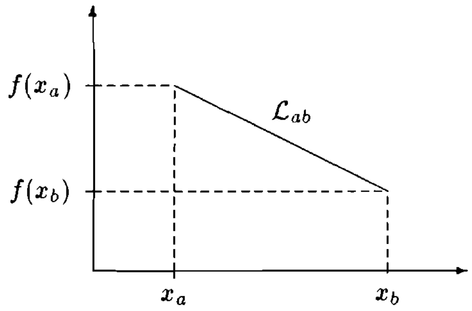


Figure 6: Case 2: $f(x_a) > f(x_b)$

<u>**Case 3**</u> $f(x_a) = f(x_b)$
In this case, $\mathcal{L}_{ab}$ has zero slope (see Figure 7). Since $f(x)$ is strictly convex $\cap$, it lies above $\mathcal{L}_{ab}$ for all $x \in (x_a, x_b)$, i.e.

$$f(x_a) = f(x_b) < f(x) \text{ for all } x \blacksquare (x_a, x_b)$$

which implies, since $f(n)$ is just $f(x)$ sampled at the points $\frac{n}{N}$, that

$$f(n_a) = f(n_b) < f(n) \text{ for all } n \text{ E } \{n_a + 1, \ldots n_b - 1\}$$

In all cases, $f(n)$ attains its minimum at one of the boundary points, either $n_a$ or $n_b$. □
The next two lemmas will be used to show that $f(n)$ is strictly decreasing over the first uniform sequence and strictly increasing over the last.

***Lemma 4.3*** *For* $n = 1, \ldots N$ − I

$$\vec{r}(0) = \vec{\ell}(N) = \frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n)$$

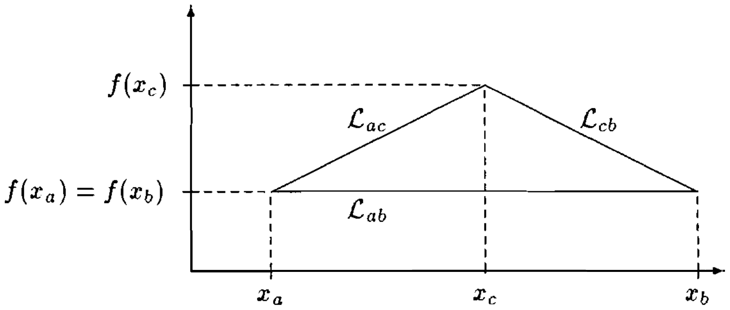


Figure 7: Case *3:* $f(x_a) = f(x_b)$.

**Proof:** Note that $\vec{r}(0) = \vec{\ell}(N)$ is the purity vector of the entire sequence of data at the node to be split. For any $n$, we have for each class $i$

$$\begin{aligned} r_i(0) &= \frac{L_i(n) + R_i(n)}{N} \\ &= \frac{L(n)}{N}\frac{L_i(n)}{L(n)} + \frac{R(n)}{N}\frac{R_i(n)}{R(n)} \\ &= \frac{n}{N}\ell_i(n) + \frac{N-n}{N}r_i(n) \\ &= \frac{n}{N}\ell_i(n) + \left(1 - \frac{n}{N}\right) r_i(n) \end{aligned}$$

which proves the result. □

The next lemma, due to Breiman et al. (Breiman et al., 1984), shows that splitting a node can never increase the average impurity:

**Lemma** *4.4 For $n$ = 1,...$N$ − 1, if $\vec{\ell}(n) \neq \vec{r}(n)$ then*

$$f(0) = f(N) > f(n)$$

**Proof:** By Lemma *4.3,* we have for each n

$$\vec{r}(0) = \vec{\ell}(N) = \frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n)$$

Thus for $\vec{\ell}(n) \neq \vec{r}(n)$

$$\begin{aligned} f(0) &= I(\vec{r}(0)) \\ &= I\left(\frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n)\right) \\ &> \frac{n}{N}I\left(\vec{\ell}(n)\right) + \left(1 - \frac{n}{N}\right)I\left(\vec{r}(n)\right) \\ &= f(n) \end{aligned}$$

since **I** is strictly convex ∩. Likewise

$$\begin{aligned} f(N) &= I(\vec{\ell}(N)) \\ &= I(\vec{r}(0)) \\ &= f(0) \\ &> f(n) \end{aligned}$$

as required. □

We next show that f(n) is strictly decreasing over the first uniform sequence:

**Lemma** *4.5 Let $n_b$ be the right boundary point of the first uniform sequence. Then for* n = *1,...$n_b$*

$$\Delta f(n) \equiv f(n) - f(n-1) < 0 \tag{25}$$

**Proof:** Assuming that the first uniform sequence consists of samples from class k, $\vec{\ell}(1) = \vec{e}_k$, where $\vec{e}_k$ is a vector of length $m$ with a 1 in the $k^{th}$ position and zeros elsewhere. Furthermore $\vec{r}(1) \neq \vec{\ell}(1)$ (otherwise the node to be split would consist purely of class k samples), so by Lemma 4.4

$$f(1) - f(0) < 0$$

Thus (25) holds for n = 1. For n ≥ 2, Lemma 4.3 implies that

$$\frac{n-1}{N}\vec{\ell}(n-1) + \left(1 - \frac{n-1}{N}\right)\vec{r}(n-1) = \frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n) \tag{26}$$

Since n ≤ $n_b$, n is either within the first uniform sequence or is a boundary point of that sequence. It follows that for 2 ≤ n ≤ $n_b$, $\vec{\ell}(n) = \vec{\ell}(n-1) = \vec{e}_k$, so (26) becomes

$$\frac{n-1}{N}\vec{\ell}(n) + \left(1 - \frac{n-1}{N}\right)\vec{r}(n-1) = \frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n)$$

$$\frac{N-n+1}{N}\vec{r}(n-1) = \frac{1}{N}\vec{\ell}(n) + \frac{N-n}{N}\vec{r}(n)$$

$$\vec{r}(n-1) = \frac{1}{N-n+1}\vec{\ell}(n) + \frac{N-n}{N-n+1}\vec{r}(n)$$

Since $\vec{r}(n) \neq \vec{\ell}(n)$ (otherwise the node to be split would consist purely of class k samples), it follows from the convexity of **I** that

$$\begin{aligned} I(\vec{r}(n-1)) &= I\left(\frac{1}{N-n+1}\vec{\ell}(n) + \frac{N-n}{N-n+1}\vec{r}(n)\right) \\ &> \frac{1}{N-n+1}I(\vec{\ell}(n)) + \frac{N-n}{N-n+1}I(\vec{r}(n)) \end{aligned}$$

$$\frac{N-n+1}{N}I(\vec{r}(n-1)) > \frac{1}{N}I(\vec{\ell}(n)) + \frac{N-n}{N}I(\vec{r}(n))$$

$$\begin{aligned}
&= \frac{n-(n-1)}{N} I(\vec{\ell}(n)) + \frac{N-n}{N} I(\vec{r}(n)) \\
&= \frac{n}{N} I(\vec{\ell}(n)) - \frac{n-1}{N} I(\vec{\ell}(n)) + \frac{N-n}{N} I(\vec{r}(n)) \\
&= \frac{n}{N} I(\vec{\ell}(n)) - \frac{n-1}{N} I(\vec{\ell}(n-1)) + \frac{N-n}{N} I(\vec{r}(n))
\end{aligned}$$

$$\frac{n-1}{N} I(\vec{\ell}(n-1)) + \left(1 - \frac{n-1}{N}\right) I(\vec{r}(n-1)) > \frac{n}{N} I(\vec{\ell}(n)) + \left(1 - \frac{n}{N}\right) I(\vec{r}(n))$$

and thus f(n – 1) > $f(n)$, as required. □

We next show that f(n) is strictly increasing over the last uniform sequence:

**Lemma** *4.6 Let $n_a$ be the left boundary point of the last uniform sequence. Then* for $n = n_a + 1, \ldots$ N

$$\Delta f(n) \equiv f(n) - f(n-1) > 0 \tag{27}$$

**Proof:** Assuming the last uniform sequence consists of samples from class k, F(N – 1) = $\vec{e}_k$, where $\vec{e}_k$ is a vector of length $m$ with a 1 in the k$^{\text{th}}$ position and zeros elsewhere. Furthermore $\vec{r}(N-1) \neq \vec{\ell}(N-1)$ (otherwise the node to be split would consist purely of class k samples), so by Lemma 4.4

$$f(N) - f(N-1) > 0$$

Thus (27) holds for n = N. For n ≤ N – 1, Lemma **4.3** implies that

$$\frac{n}{N}\vec{\ell}(n) + \left(1 - \frac{n}{N}\right)\vec{r}(n) = \frac{n-1}{N}\vec{\ell}(n-1) + (1 - \frac{n-1}{N}\Big)\,\vec{r}(n-1) \tag{28}$$

Since n ≥ $n_a$ + 1, $n$ is either within the last uniform sequence or is a boundary point of that sequence. It follows that for $n_a + 1 \leq$ n ≤ N – 1, $\vec{r}(n) = \vec{r}(n-1) = \vec{e}_k$, so (28) becomes

$$\frac{n}{N}\vec{\ell}(n) + \frac{N - n}{N}F(n-1) = \frac{n-1}{N}\vec{\ell}(n-1) + \frac{N-n+1}{N}\vec{r}(n-1)$$

$$\frac{n}{N}\vec{\ell}(n) = \frac{n-1}{N}\vec{\ell}(n-1) + \frac{1}{N}\vec{r}(n-1)$$

or

$$\vec{\ell}(n) = \frac{n-1}{n}\vec{t}(n-1) + \frac{1}{n}\vec{r}(n-1)$$

Since $\vec{r}(n-1) \neq \vec{\ell}(n-1)$ (otherwise the node to be split would consist purely of class $k$ samples), it follows from the convexity of **I** that

$$\begin{aligned}
I(\vec{\ell}(n)) &= I\left(\frac{n-1}{n}\vec{\ell}(n-1) + \frac{1}{n}\vec{r}(n-1)\right) \\
&> \frac{n-1}{n} I(\vec{\ell}(n-1)) + \frac{1}{n} I(\vec{r}(n-1))
\end{aligned}$$

$$\begin{aligned}
\frac{n}{N}I(\vec{\ell}(n)) &= \frac{n-1}{N}I(\vec{\ell}(n-1)) + \frac{1}{N}I(\vec{r}(n-1)) \\
&= \frac{n-1}{N}I(\vec{\ell}(n-1)) + \frac{(N-n+1)-(N-n)}{N}I(\vec{r}(n-1)) \\
&= \frac{n-1}{N}I(\vec{\ell}(n-1)) + \frac{N-n+1}{N}I(\vec{r}(n-1)) - \frac{N-n}{N}I(\vec{r}(n-1)) \\
&= \frac{n-1}{N}I(\vec{\ell}(n-1)) + \left(1-\frac{n-1}{N}\right)I(\vec{r}(n-1)) - \left(1-\frac{n}{N}\right)I(\vec{r}(n))
\end{aligned}$$

$$\frac{n}{N}I(\vec{\ell}(n)) + \left(1-\frac{n}{N}\right)I(\vec{r}(n)) > \frac{n-1}{N}I(\vec{\ell}(n-1)) + \left(1-\frac{n-1}{N}\right)I(\vec{r}(n-1))$$

and thus $f(n) > f(n-1)$, as required. □

The next lemma connects the global behavior of f(n) to its local behavior over uniform sequences:

**Lemma 4.7** *The set of global minimizers off(n) includes only boundary points.*

**Proof:** Assume not. Then a global minimum occurs at an interior point $n_0$ of a uniform sequence. This uniform sequence cannot be the first or last, as we have shown that $f(n)$ is strictly decreasing over the first (Lemma 4.5) and strictly increasing over the last (Lemma 4.6). For any other uniform sequence, Lemma 4.2 shows that $f(n)$ is smaller at one of the boundary points $n_b$ of this sequence than at any interior point, so

$$f(n_0) > f(n_b)$$

and $n_0$ is not a global minimum, which is a contradiction. The lemma follows.

This completes the proof of Theorem 4.1.

## 4.3 Convex ∩ Impurity Measures

The main result of this section is

**Theorem 4.2** *If the impurity measure $I$ is convex ∩ on*

$$\Omega = \{(p_1, \ldots, p_m) : 0 \leq p_i \leq 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m} p_i = 1\}$$

*then the set of global minima of the splitting rule f(n) includes at least one boundary point, but may also include all interior and boundary points of one or more uniform sequences (see Figure 3).*

**Proof:** The proof consists of the following steps:

- We first show that $f(x)$ is convex ∩ over uniform sequences (Lemma 4.8).
- We then show that the convexity of $f(x)$ over a uniform sequence implies that $f(n)$ either attains its minimum at a boundary point of the sequence or is constant over that sequence. ('Lemma 4.9).

- Lemma 4.8 does not apply to the first or last uniform sequences, as the model breaks down when x = 0 or x = 1. We handle these special cases by showing that $f(n)$ is non-increasing over the first uniform sequence (Lemma 4.11) and non-decreasing over the last (Lemma 4.12), so that over any uniform sequence, f (n) either attains its minimum at a boundary point of the sequence or is constant over that sequence.
- Finally, Lemma 4.13 shows that if over any uniform sequence $f$(n) attains its minimum at a boundary point of the sequence or is constant over that sequence, then the set of global minima of f (n) includes at least one boundary point, but may also include all interior and boundary points of one or more uniform sequences (hence the global minimum is not necessarily unique).

□

We first show that f (x) is convex $\cap$ over uniform sequences:

**Lemma 4.8** If *the impurity measure I is convex* $\cap$ *on*

$$\Omega = \{(p_1, \ldots, p_m) : \; 0 \leq p_i \leq 1 \textit{ for all } i, \textit{ and } \sum_{i=1}^{m} p_i = 1)$$

*then* ***for*** *all points* x,, $x_b$ *that delimit uniform sequences, the splitting rule* f (x) *is convex* $\cap$ *on* $[x_a, x_b]$, *where* $0 < x_a < x_b < 1$.

**Proof:** Replace ">" with "≥" in the proof of Lemma 4.1. □
We next show that convexity of f (x) over a uniform sequence implies that f (n) either attains its minimum at a boundary point of the sequence or is constant over that sequence:

**Lemma 4.9** Iff$(x)$ *is convex* $\cap$ *on* [x,, $x_b$], *then* $f$(n) *either attains its minimum at a boundary point, that is, at either* n, *or* $n_b$, *or is constant for* n ∈ *{n,,* n, + 1, . . . , $n_b$}.

**Proof:** Let $\mathcal{L}_{ab}$ denote the line segment connecting the points $(x_a, \text{f}(x_a))$ and $(x_b, f(x_b))$. We have three cases to consider:

<u>Case 1</u> f $(x_a)$ < f $(x_b)$
In this case, $\mathcal{L}_{ab}$ has positive slope (see Figure 5). Since $f(x)$ is convex $\cap$, it lies on or above this line for all $x \in$ (x,, $x_b$); consequently

$$f(x_a) < f(x) \text{ for all } x \in (x_a, x_b)$$

which implies, since f (n) is just f (x) sampled at the points $\frac{n}{N}$, that

$$\text{f (n,)} < \text{f (n) for all n} \in \{n_a + 1, \ldots n_b\}$$

Case $f(x_a) >$ f $(x_b)$
In this case, $\mathcal{L}_{ab}$ has negative slope (see Figure 6). Since f (x) is convex $\cap$, it lies on or above this line for all x ∈ $(x_a, x_b)$; consequently

$$\text{f } (x_b) < \text{f } (x) \text{ for all } x \in (x_a, x_b)$$

which implies, since $f$(n) is just $f$(x) sampled at the points $\frac{n}{N}$, that

$$\text{f } (n_b) < \text{f (n) for all n E } \{n_a, \ldots n_b - 1\}$$

Case 3 $f(x_a) = f(x_b)$
In this case, $\mathcal{L}_{ab}$ has zero slope (see Figure 7). Let $x_c$ be any point between $x_a$ and $x_b$ for which $f(x_c) > f(x_a)$. If there is no such point, then $f(x)$ is constant on $[x_a, x_b]$, and consequently $f(n)$, which is just $f(x)$ sampled at the points $\frac{n}{N}$, is also constant for n $\in \{n,, \ldots n_b\}$.

If there is such a point, let $\mathcal{L}_{ac}$ denote the line connecting $(x_a, f(x_a))$ and $(x_c, f(x_c))$, and $\mathcal{L}_{cb}$ denote the line connecting $(x_c, f(x_c))$ and $(x_b, f(x_b))$. Since $f(x_c) > f(x_a) = f(x_b)$ $\mathcal{L}_{ac}$ has positive slope, and $\mathcal{L}_{cb}$ has negative slope. Since $f(x)$ is convex $\cap$, it lies on or above $\mathcal{L}_{ac}$ for all $x \in (x_a, x_c)$ and on or above $\mathcal{L}_{cb}$ for all $x \in (x_c, x_b)$; consequently

$$f(x_a) < f(x) \text{ for all } x \in (x_a, x_c)$$

and

$$f(x_b) < f(x) \text{ for all } x \in (x_c, x_b)$$

Thus

$$f(x_a) = f(x_b) < f(x) \text{ for all } x \in (x_a, x_b)$$

which implies, since $f(n)$ is just $f(x)$ sampled at the points $\frac{n}{N}$, that

$$f(n_a) = f(n_b) < f(n) \text{ for all } n \in \{n_a + 1, \ldots n_b - 1\}$$

In all cases, either

- $f(n)$ attains its minimum at one of the endpoints, either $n_a$ or $n_b$, or
- $f(n)$ is constant for $n \in \{n_a, \ldots n_b\}$.

□

The next lemma, due to Breiman et al. (Breiman et al., 1984), shows that splitting a node can never increase the average impurity:

**Lemma 4.10** *For $n = 1, \ldots N - 1$,*

$$f(0) = f(N) \geq f(n)$$

**Proof:** Replace ">" with "≥" in the proof of Lemma *4.4.*
We next show that f$(n)$ is non-increasing over the first uniform sequence:

**Lemma 4.11** *Let $n_b$ be the right boundary point of the first uniform sequence. Then for $n = 1, \ldots n_b$*

$$\Delta f(n) \equiv f(n) - f(n-1) \leq 0$$

**Proof:** Use Lemma *4.10* to conclude that $f(1) \leq f(0)$, then replace ">" with "≥" in the remainder of the proof of Lemma *4.5.* □
We next show that $f(n)$ is non-decreasing over the last uniform sequence:

**Lemma 4.12** *Let $n_a$ be the left boundary point of the last uniform sequence. Then for $n = n_a + 1, \ldots N$*

$$\Delta f(n) \equiv f(n) - f(n-1) \geq 0$$

**Proof:** Use Lemma *4.10* to conclude that $f(N) \geq f(N-1)$, then replace ">" with "≥" in the remainder of the proof of Lemma *4.6.* □
The next lemma connects the global behavior of $f(n)$ to its local behavior over uniform sequences:

**Lemma 4.13** *The set of global minimizers of the splitting rule $f(n)$ includes at least one boundary point, but may also include all interior and boundary points of one or more uniform sequences.*

**Proof:** Assume that the set of global minimizers includes only interior points. Let $n_0$ be one such point. We have three cases to consider:

> Case 1 If $n_0$ is within the first uniform sequence, then by Lemma *4.11* $f(n)$ must be constant over this sequence; it follows that all interior points and the right boundary point of this sequence belong to the set of global minimizers.
>
> Case 2 If $n_0$ is within the last uniform sequence, then by Lemma *4.12* $f(n)$ must be constant over this sequence; it follows that all interior points and the left boundary point of this sequence belong to the set of global minimizers.
>
> Case 3 If $n_0$ is within a uniform sequence other than the first or last, then by Lemma *4.9* $f(n)$ must be constant over this sequence; it follows that all interior points and both boundary points of this sequence belong to the set of global minimizers.

Since $n_0$ was an arbitrary element of the set of global minimizers, it follows that this set contains at least one boundary point, and may contain the interior and boundary points of one or more uniform sequences. □
This completes the proof of Theorem *4.2*.

## 4.4 A Calculus Proof of the Minima-Free Property

Above we showed that the minima-free property follows from the convexity of the splitting rule $f(x)$, which in turn follows from the convexity of the impurity measure **I** via Lemma *4.1*. These results were obtained using primarily geometric arguments. In this section we show that similar results can be obtained using calculus; in particular, we have

**Theorem 4.3** *If the Hessian $\nabla^2 I$ of the impurity measure $I(p_1, \ldots, p)$ is negative definite on*

$$\Omega = \left\{ (p_1, \ldots, p,)\text{:} \quad 0 \le p_i \le 1 \textit{ for all } i, \textit{ and } \sum_{i=1}^{m} p_i = 1 \right\}$$

*then the splitting rule $f(x)$ is strictly convex $\cap$*

**Theorem 4.4** *If the Hessian $\nabla^2 I$ of the impurity measure $I(p_1, \ldots, p_m)$ is negative semidefinite on*

$$\left\{ (p_1, \ldots, p_m) : \; 0 \le p_i \le 1 \textit{ for all } i, \textit{ and } \sum_{i=1}^{m} p_i = 1 \right\}$$

*then the splitting rule $f(x)$ is convex $\cap$.*

In order to apply the above theorems, we cannot just test the matrix of second derivatives defined by

$$\nabla^2 I(p_1, \ldots p_m) = \begin{bmatrix} \frac{\partial^2 I}{\partial p_1 \partial p_1} & \frac{\partial^2 I}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_1 \partial p_m} \\ \frac{\partial^2 I}{\partial p_2 \partial p_1} & \frac{\partial^2 I}{\partial p_2 \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_2 \partial p_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 I}{\partial p_m \partial p_1} & \frac{\partial^2 I}{\partial p_m \partial p_2} & \cdots & \frac{\partial^2 I}{\partial p_m \partial p_m} \end{bmatrix}$$

for negative definiteness or semidefiniteness, as this fails to take account of the fact. that $p_1, \ldots, p_m$ are related by

$$\sum_{i=1}^{m} p_i = 1$$

Instead we must first substitute the above constraint for, say, $p_m$ to get a function of the $m-1$ independent variables $p_1, \ldots, p_{m-1}$ which we designate as $I$:

$$\tilde{I}(p_1, \ldots, p_{m-1}) \equiv I\left(p_1, \ldots, p_{m-1}, 1 - \sum_{i=1}^{m-1} p_i\right)$$

It follows by construction that

- $\nabla^2 \tilde{I}$ is negative definite on

  $$\tilde{\Omega} = \left\{(p_1, \ldots p_{m-1}) : 0 \leq p_i \leq 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m-1} p_i \leq 1\right\}$$

  if and only if $\nabla^2 I$ is negative definite on R.

- $\nabla^2 \tilde{I}$ is negative semidefinite on $\tilde{\Omega}$ if and only if $\nabla^2 I$ is negative semidefinite on $\Omega$.

Thus the problem of determining whether $V^2 1$ is negative definite on $\Omega$ reduces to the problem of determining whether

$$\nabla^2 \tilde{I}(p_1, \ldots p_{m-1}) = \begin{bmatrix} \frac{\partial^2 \tilde{I}}{\partial p_1 \partial p_1} & \frac{\partial^2 \tilde{I}}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 \tilde{I}}{\partial p_1 \partial p_{m-1}} \\ \frac{\partial^2 \tilde{I}}{\partial p_2 \partial p_1} & \frac{\partial^2 \tilde{I}}{\partial p_2 \partial p_2} & \cdots & \frac{\partial^2 \tilde{I}}{\partial p_2 \partial p_{m-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \tilde{I}}{\partial p_{m-1} \partial p_1} & \frac{\partial^2 \tilde{I}}{\partial p_{m-1} \partial p_2} & \cdots & \frac{\partial^2 \tilde{I}}{\partial p_{m-1} \partial p_{m-1}} \end{bmatrix}$$

is negative definite on $\tilde{\Omega}$.

Proofs of these theorems are given more for completeness than because they add anything to the presentation, and can be skipped without any loss of continuity. In fact, Theorems 4.3 and 4.4 are much weaker than their geometric counterparts, as:

- they require the impurity measure to have continuous first and second partial derivatives, and
- they follow directly from Lemmas 4.1 and 4.8, in view of Facts **3** and 4.

We prove Theorems 4.3 and 4.4 below.

**Proof of Theorem 4.3:** We show that $f''(x) < 0$ on (x,, $x_b$); Fact **3** then implies that $f(x)$ is strictly convex $\cap$. In order to differentiate $f(x)$, we must know how to differentiate a function taking a vector argument. To review how this is done, we will compute the first two derivatives of the simpler function

$$w(x) = I(\vec{r}(x)) = I(r_1(x), r_2(x), \ldots r_m(x))$$

which are

$$
\begin{aligned}
w'(x) &= \frac{d}{dx} I(r_1(x), r_2(x), \ \ldots \ r_m(x)) \\
&= \sum_{i=1}^{m} \frac{\partial I(r_1(x), r_2(x), \ \ldots \ r_m(x))}{\partial r_i} \frac{\partial r_i}{\partial x} \quad \text{by the chain rule} \\
&= \sum_{i=1}^{m} \frac{\partial I(\vec{r}(x))}{\partial r_i} \frac{\partial r_i}{\partial x} \\
&= \left[ \frac{\partial I(\vec{r}(x))}{\partial r_1} \ \frac{\partial I(\vec{r}(x))}{\partial r_2} \ \cdots \frac{\partial I(\vec{r}(x))}{\partial r_m} \right] \cdot \left[ \frac{\partial r_1}{\partial x} \ \frac{\partial r_2}{\partial x} \ \cdots \frac{\partial r_m}{\partial x} \right]^T \\
&= \nabla I(\vec{r}(x)) \, . \, \overset{a}{\mathbf{,}} [r_1(x) r_2(x) \ \ldots r_m(x)]^T \\
&= \nabla I(\vec{r}(x)) \cdot \frac{\partial}{\partial x} \vec{r}(x) \\
&= \nabla I(\vec{r}(x)) \cdot \vec{r}\,'(x) \\
w''(x) &= \frac{d}{dx} (\nabla I(\vec{r}(x)) \cdot \vec{r}\,'(x)) \\
&= \frac{d}{dx} \left( \sum_{i=1}^{m} \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i'(x) \right) \\
&= \sum_{i=1}^{m} \frac{d}{dx} \left( \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i'(x) \right) \\
&= \sum_{i=1}^{m} \left( \frac{d}{dx} \left( \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i'(x) \right) + \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i''(x) \right) \\
&= \sum_{i=1}^{m} \left( \sum_{j=1}^{m} \frac{\partial^2 I(\vec{r}(x))}{\partial r_j \partial r_i} r_j'(x) r_i'(x) + \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i''(x) \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial^2 I(\vec{r}(x))}{\partial r_j \partial r_i} r_j'(x) r_i'(x) + \sum_{i=1}^{m} \frac{\partial I(\vec{r}(x))}{\partial r_i} r_i''(x) \\
&= \vec{r}\,'(x)^T \nabla^2 I(\vec{r}(x)) \vec{r}\,'(x) + \nabla I(\vec{r}(x)) \cdot \vec{r}\,''(x)
\end{aligned}
$$

where $\nabla^2 I(\vec{r}(x))$ is the Hesssian matrix of the impurity measure $I$ evaluated at the point $\vec{r}(x)$. This completes the review. We now compute the first, and second derivatives of

$$f(x) = xI(\vec{\ell}(x)) + (1-x)I(\vec{r}(x))$$

which are

$$
\begin{aligned}
f'(x) &= I(\vec{\ell}(x)) + x\nabla I(\vec{\ell}(x)) \, . \, \vec{\ell}\,'(x) - I(\vec{r}(x)) + (1-x)\nabla I(\vec{r}(x)) \, . \, \vec{r}\,'(x) \\
f''(x) &= \nabla I(\vec{\ell}(x)) \, . \, \vec{\ell}\,'(x) + \nabla I(\vec{\ell}(x)) \cdot \vec{\ell}\,'(x) + x\vec{\ell}\,'(x)^T \nabla^2 I(\vec{\ell}(x)) \vec{\ell}\,'(x) + x\nabla I(\vec{\ell}(x)) \, . \, \vec{\ell}\,''(x) \\
&\quad - \nabla I(\vec{r}(x)) \, . \, \vec{r}\,'(x) - \nabla I(\vec{r}(x)) \, . \, \vec{r}\,'(x) + (1-x)\vec{r}\,'(x)^T \nabla^2 I(\vec{r}(x)) \vec{r}\,'(x) + (1-x)\nabla I(\vec{r}(x)) \, . \, \vec{r}\,''(x)
\end{aligned}
$$

$$\begin{aligned} &= \nabla I(\vec{\ell}(x)) \cdot \left(2\vec{\ell}'(x) + x\vec{\ell}''(x)\right) + x\vec{\ell}'(x)^T \nabla^2 I(\vec{\ell}(x))\vec{\ell}'(x) \\ &\quad + \nabla I(\vec{r}(x)) \cdot (-2\vec{r}'(x) + (1-x)\vec{r}''(x)) + (1-x)\vec{r}'(x)^T \nabla^2 I(\vec{r}(x))\vec{r}'(x) \end{aligned} \tag{29}$$

We now show that two terms in the above expression are zero by showing that

$$2\vec{\ell}'(x) + x\vec{\ell}''(x) = \vec{0} \tag{30}$$

and

$$-2\vec{r}'(x) + (1-x)\vec{r}''(x) = \vec{0} \tag{31}$$

Claim 1 *(30)* holds.
For $i \neq$ k

$$\begin{aligned} \ell_i &= \frac{L_i}{Nx} \\ \ell_i' &= -\frac{L_i}{Nx^2} \\ \ell_i'' &= 2\frac{L_i}{Nx^3} \\ \Rightarrow \qquad 2\ell_i' + x\ell_i'' &= 2\left(-\frac{L_i}{Nx^2}\right) + x\left(2\frac{L_i}{Nx^3}\right) \\ &= 0 \end{aligned}$$

and for i = k

$$\begin{aligned} \ell_k &= 1 - \frac{x_{min}}{x} \\ \ell_k' &= \frac{Xmin}{x^2} \\ \ell_k'' &= -2\frac{Xmin}{x^3} \\ \Rightarrow \qquad 2\ell_k' + x\ell_k'' &= 2\left(\frac{x_{min}}{x^2}\right) + x\left(-2\frac{x_{min}}{x^3}\right) \\ &= 0 \end{aligned}$$

Thus each component of $2\vec{\ell}'(x) + x\vec{\ell}''(x)$ is zero, as claimed.

Claim 2 *(31)* holds.
For $i \neq$ k

$$\begin{aligned} r_i &= \frac{R_i}{N(1-x)} \\ r_i' &= \frac{R_i}{N(1-x)^2} \\ r_i'' &= 2\frac{R_i}{N(1-x)^3} \\ \Rightarrow \qquad -2r_i' + (1-x)r_i'' &= -2\left(\frac{R_i}{N(1-x)^2}\right) + (1-x)\left(2\frac{R_i}{N(1-x)^3}\right) \\ &= 0 \end{aligned}$$

and for i = k

$$\begin{aligned} r_k &= \frac{x_{max}-1}{1-x}+1 \\ r'_k &= \frac{x_{max}-1}{(1-x)^2} \\ r''_k &= 2\frac{x_{max}-1}{(1-x)^3} \\ \Rightarrow \qquad -2r'_k+(1-x)r''_k &= -2\left(\frac{x_{max}-1}{(1-x)^2}\right)+(1-x)\left(2\frac{x_{max}-1}{(1-x)^3}\right) \\ &= 0 \end{aligned}$$

Thus each component of $-2\vec{r}'(x)$ + *(1 –* $x)\vec{r}''(x)$ is zero, as claimed.

In view of *(30)* and (31), two terms drop out of (29), leaving

$$f''(x) = x\vec{\ell}'(x)^T\nabla^2 I(\vec{\ell}(x))\vec{\ell}'(x) + (1-x)\vec{r}'(x)^T\nabla^2 I(\vec{r}(x))\vec{r}'(x)$$

The following observations are sufficient to ensure that $f''(x) < 0$ on $(x_a, x_b)$:

- $x > 0$ and $1 - x > 0$.
- At least one of $\vec{\ell}'(x)$, $\vec{r}'(x)$ is not equal to the zero vector. For if $\vec{\ell}'(x) = \vec{r}'(x) =$ **6,** then

$$\begin{aligned} \vec{\ell}'(x)=\vec{0} \quad &\Rightarrow \quad \ell'_i(x) = -\frac{L_i}{Nx^2} = 0 \quad \text{forifk} \\ &\Rightarrow \quad L_i = 0 \quad \text{forifk} \\ &\Rightarrow \quad \text{all samples left of } x \text{ are class } k \\ \vec{r}'(x)=\vec{0} \quad &\Rightarrow \quad r'_i(x) = \frac{R_i}{N(l-x)^2} = 0 \quad \text{forifk} \\ &\Rightarrow \quad R_i = 0 \quad \text{forifk} \\ &\Rightarrow \quad \text{all samples right of } x \text{ are class k} \end{aligned}$$

  so that all samples in the sequence are class k, and we would not be splitting it in the first place.
- The Hessian $V^2 1$ is negative definite on $\Omega$, which implies that

$$y \;\; \nabla^2 I(\vec{p})\vec{y} < 0 \qquad \text{for all } \vec{y} \neq 0 \text{ and } \vec{p} \in \Omega$$

Since $f''(x) < 0$, *f(x)* is strictly convex n by Fact *3*. □
**Proof of Theorem 4.4:** Exactly as in the proof of Theorem *4.3,* we obtain

$$f''(x) = x\vec{\ell}'(x)^T\nabla^2 I(\vec{\ell}(x))\vec{\ell}'(x) + (1-x)\vec{r}'(x)^T\nabla^2 I(\vec{r}(x))\vec{r}'(x)$$

but the negative semidefiniteness of the impurity measure allows us to conclude only that $f''(x) \leq$ *0,* and therefore that f*(x)* is convex ***n*** by Fact 4. □

## 4.5 Rational Splitting Rules and the Minima-Free Property

Some splitting rules, notably the information gain ratio (Quinlan, 1990), do not fit into the framework developed thus far; we must therefore develop additional tools to analyze these cases. In particular, we consider splitting rules of the form

$$F(x) = \frac{g(x)}{h(x)}$$

and we seek conditions on $g(x)$ and $h(x)$ which ensure that $F(x)$ attains its minimum at a boundary point. The main result of this section is:

**Theorem 4.5** *Consider the splitting rule*

$$F(x) = \frac{g(x)}{h(x)}$$

*If $g(x)$ and $h(x)$ are twice differentiable and satisfy the following:*

$g''(x) \leq 0$ *over uniform sequences.*

- $g(x) < 0$ *for* $x \in (0,1)$.
- $h''(x) < 0$ *for* $x \in (0,1)$.
- $h(x) > 0$ *for* $x \in (0,1)$.
- *For* $G(x) = g'h - h'g$,

$$G(0) = \lim_{x \to 0^+} G(x) = \lim_{x \to 0^+} g'h - h'g \leq 0$$

*and*

$$G(1) = \lim_{x \to 1^-} G(x) = \lim_{x \to 1^-} g'h - h'g \geq 0$$

*then F(x) attains its minimum at a boundary point.*

**Proof:** Fix a uniform sequence extending from $x,$ to $x_b$. We must show that $F(x)$ attains its minimum at either $x,$ or $x_b$. We have immediately that for $x \in (x_a, x_b)$:

- $g''(x) \leq 0$.
- $g(x) < 0$.
- $h''(x) < 0$.
- $h(x) > 0$.

from which it follows that $g''h - h''g < 0$ for all $x \in (x_a, x_b)$ and thus by Lemma 4.14 that $F(x)$ attains its minimum at either $x,$ or $x_b$. Since this holds for any uniform sequence, $F(x)$ must attain its minimum over $[0,1]$ at a boundary point.

Note however that this result does not rule out the possibility of $F(x)$ attaining it:; minimum at $x = 0$ or $x = 1$ (or equivalently, of $F(n)$ attaining its minimum at $n = 0$ or $n =$ N). We want to rule out this possibility. because if $F(n)$ attains its minimum at n $= 0$, say, then since we are only evaluating $F(n)$ at the points $n = 1, 2, \ldots, N - 1$, it may be that the minimum of $F(n)$ over this set occurs at $n = 1$ , which may not be a boundary point. To exclude this possibility, we note that

- $F(x)$ cannot attain its minimum at $x = 0$ because by Lemma *4.15*, $F(x)$ is strictly decreasing over the first uniform sequence.
- $F(x)$ cannot attain its minimum at $x = 1$ because by Lemma *4.16*, $F(x)$ is strictly increasing over the last uniform sequence.

Thus we have shown that $F(x)$ attains its minimum over [0,1] at a boundary point, and furthermore that this point is not $x = 0$ or $x = 1$. □
We prove Lemmas *4.14*, *4.15* and *4.16* below:

**Lemma 4.14** *Consider o function of the form*

$$F(x) = \frac{g(x)}{h(x)}$$

*If* $h \neq 0$ *and* $g''h - gh'' < 0$ *over on interval* $(x_a, x_b)$, *then* $F$ *attains its minimum on* $[x_a, x_b]$ *at either* $x_a$ *or* $x_b$.

**Proof:** We show that $g''h - gh'' < 0$ for all $x \in (x_a, x_b)$ implies that $F''(x^*) < 0$ for any critical point $x^*$ in the interval $(x_a, x_b)$, that is, for any point $x^*$ satisfying $F'(x^*) = 0$. This effectively forces all extrema of $F(x)$ in this interval to he maxima. The first and second derivatives of $F$ are:

$$\begin{aligned} F' &= \frac{g'h - gh'}{h^2} \\ F'' &= \frac{(g''h + g'h' - g'h' - gh'')\,h^2 - (g'h - gh')\,2hh'}{h^4} \\ &= \frac{(g''h - gh'')h^2 - (g'h - gh')2hh'}{h^4} \end{aligned}$$

Now, for any critical point $x^*, F'(x^*) = 0$, which implies that $g'h - gh' = 0$. Hence at the critical point,

$$\begin{aligned} F''(x^*) &= \frac{(g''h - gh'')\,h^2 - (g'h - gh')\,2hh'}{h^4} \\ &- \frac{(g''h - gh'')\,h^2}{h^4} \quad \text{since } g'h - gh' = 0 &&(32) \\ &= \frac{g''(x^*)h(x^*) - g(x^*)h''(x^*)}{h^2} &&(33) \\ &< 0 \quad \text{since } x^* \in (x_a, x_b) &&(34) \end{aligned}$$

Thus $F''(x^*) < 0$ for any critical point $x^* \in (x_a, x_b)$, which implies that $F(x)$ has no minima in this interval. It follows that $F$ either increases, decreases, or increases and then decreases on $[x_a, x_b]$; in all cases $F$ attains its minimum over $[x_a, x_b]$ at either $x_a$ or $x_b$. □

**Lemma 4.15** *Assume the first uniform sequence lies between* 0 *and* $x_b$. *Let* $F(x)$, $G(x)$ *be as above. If* $h \neq 0$ *and* $g''h - gh'' < 0$ *for* $x \in (0, x_b)$, *and if*

$$G(0) = \lim_{x \to 0+} G(x) = \lim_{x \to 0+} g'h - h'g \leq 0$$

*then* $F(x)$ *is strictly decreasing over the first uniform sequence*

**Proof:** For any $x \in (0, x_b)$,

$$\begin{aligned} G'(x) &= g''h - g'h' + g'h' - gh'' \\ &= g''h - gh'' \\ &< 0 \end{aligned}$$

by assumption. Since $G^1(x) < 0$ for $x \in (0, x_b)$ and $G(0) \leq 0$, it follows that $G(x) < 0$ for $x \in (0, x_b)$. Thus for $x \in (0, x_b)$

$$\begin{aligned} F^1(x) &= \frac{g'h - gh'}{h^2} \\ &= \frac{G(x)}{h^2} \\ &< 0 \end{aligned}$$

so that $F(x)$ is strictly decreasing over the first uniform sequence. □

**Lemma** *4.16 Assume the last uniform sequence lies between $x_a$ and 1. Let $F(x)$, $G(x)$ be as above. If $h \neq 0$ and $g''h - gh'' < 0$ for $x \in (x_a, 1)$, and if*

$$G(l) = \lim_{x \to 1^-} G(x) = \lim_{x \to 1^-} g'h - h'g \geq 0$$

*then $F(x)$ is strictly increasing over the last uniform sequence.*

**Proof:** For any $x \in (x_a, 1)$,

$$\begin{aligned} G'(x) &= g''h - g'h' + g'h' - gh'' \\ &= g''h - gh'' \\ &< 0 \end{aligned}$$

by assumption. Since $G'(x) < 0$ for $x \in (x_a, 1)$ and $G(l) \geq 0$, it follows that $G(x) > 0$ for $x \in (x_a, 1)$. Thus for $x \in (x,, 1)$

$$\begin{aligned} F'(x) &= \frac{g'h - gh'}{h^2} \\ &= \frac{G(x)}{h^2} \\ &> 0 \end{aligned}$$

so that $F(x)$ is strictly increasing over the last uniform sequence. □

This completes the proof of Theorem 4.5.

## 4.6 Repeated Values

In this section, we show that the above results, which were obtained assuming no repeated values, carry over to the case in which repeated values are allowed. We distinguish two types of repeated value:

**Definition 7** A <u>*uniform repeated value*</u> (URV) *is an attribute value to which more than one instance is mapped, and all such instances belong to the same class.*

**Definition 8** A <u>*non-uniform repeated value*</u> *(NRV) is an attribute value to which more than one instance is mapped, and at least two such instances belong to different classes.*

The definition of a boundary point given in Section 3 is not general enuf to handle repeated values, so we use the definition given in (Fayyad & Irani, 1992b):

**Definition 9** A *point c is a* <u>*boundary point*</u> *if there exists instances $s_1$, $s_2$ such that:*

- *$s_1$ and $s_2$ belong to different classes,*
- *c lies between the projected valves of $s_1$ and $s_2$, i.e. $A(s_1) < c < A(s_2)$, and*
- *No sample s maps to a valve between $A(s_1)$ and $A(s_2)$, that is, for all instances s*

$$A(s) \notin (A(s_1), A(s_2))$$

With this definition of a boundary point, all of the results obtained up to this point hold even when repeated values are allowed:

**Theorem 4.6** *Let f be a splitting rule which, when applied to sequences with no repeated values, has the following properties:*

1. *Over any uniform sequence, it attains its minimum at a boundary point of that sequence.*
2. *It is strictly decreasing over the first uniform sequence.*
3. *It is strictly increasing over the last uniform sequence.*

*Then f has the same properties over sequences with repeated values.*

**Proof:** Fix a sequence S of training data, possibly containing URVs and NRVs; then we must show that the splitting rule $f$ has properties 1–3 over S. These properties relate to the behavior of $f$ over uniform sequences, so we pick a particular uniform sequence $U$ from S, possibly containing URVs, and investigate the behavior of $f$ over U.

Let $S'$ be the sequence produced by "unstacking" the URVs in S, meaning that the instances mapped to a given URV are remapped to distinct values in the interval $(\text{URV} - \epsilon, \text{URV} + \epsilon)$, where $\epsilon$ is chosen small enuf that no other instances map to this interval. Then $S'$ consists of uniform sequences, possibly separated by zero or more NRVs. One such uniform sequence, call it $U'$, corresponds to $U$ with all URVs unstacked.

Let $S''$ be the sequence produced by unstacking the NRVs in $S'$; clearly this does not affect $U'$, which contains no NRVs (because it was obtained from a uniform sequence in S). Now, S" is a sequence with no repeated values, and $U'$ is a uniform sequence from S", so by the assumptions of the theorem we can conclude that:

- f attains its minimum at a boundary point of $U'$
- If $U'$ is the first uniform sequence, then $f$ is strictly decreasing over $U'$
- If $U'$ is the last uniform sequence, then $f$ is strictly increasing over $U'$

The splitting rules f we consider have the property that the behavior of f over a. uniform sequence is unaffected by how the samples are arranged outside that sequence, so long as the number of samples of each class to the left and right of that sequence stays the same. This implies that the behavior of f over $U'$ when $U'$ is regarded as a subsequence of $S'$ is exactly the same as its behavior when $U'$ is regarded as a subsequence of $S''$. It follows that the above properties also apply to the behavior of $f$ over $U'$ when $U'$ is regarded as a subsequence of $S'$.

Next, we plot how $f$ varies over $U'$ and also how $f$ varies over $U$ (these plots are called *splitting curves).* The relation between these curves is that the splitting curve of $f$ over $U$ is just the splitting curve of $f$ over $U'$ with some sections taken out (those corresponding to unstacked URVs). We observe that:

- In view of the fact that $U$ and $U'$ have the same boundary points, that $f$ attains its minimum over $U'$ at a boundary point implies that f will attain its minimum over $U$ at a boundary point (because throwing away certain sections over which $f$ does not attain its minimum does not affect the location of the minimum).
- The fact that $f$ is strictly decreasing on $U'$ when $U'$ is the first uniform sequence implies that $f$ will be strictly decreasing on $U$ when $U$ is the first uniform sequence (because throwing away certain sections of a curve does not affect its monotonicity properties).
- The fact that $f$ is strictly increasing on $U'$ when $U'$ is the last uniform sequence implies that $f$ will be strictly increasing on $U$ when $U$ is the last uniform sequence (because throwing away certain sections of a curve does not affect its monotonicity properties).

The theorem follows. □

**Corollary 4.7** *For the case in which repeated values are allowed, the set of global minima of a splitting rule f satisfying the conditions of Theorem 4.6 includes only boundary points.*

**Proof:** Assume not. Then a global minimum occurs at an interior point $n_0$ of a uniform sequence. This uniform sequence cannot be the first or last, as Theorem 4.6 shows that $f$ is strictly decreasing over the first uniform sequence and strictly increasing over the last. For any other uniform sequence, Theorem 4.6 shows that $f$ is smaller at one of the boundary points $n_b$ of this sequence than at any interior point, so

$$f(n_0) > f(n_b)$$

and $n_0$ is not a global minimum, which is a contradiction. The result follows. □

**Corollary 4.8** *For the case in which repeated values are allowed, splitting rules of the form*

$$f(x;\mathbf{I}) = xI(\vec{\ell}(x)) + (1-x)I(\vec{r}(x))$$

*where* **I** *is strictly convex* $\cap$ *cut only at boundary points.*

**Proof:** This follows from Corollary 4.7, since splitting rules of this form satisfy the conditions of Theorem 4.6 by Lemmas 4.2, 4.5 and 4.6. □

**Corollary 4.9** *For the case in which repeated values are allowed, splitting rules of the form*

$$F(x) = \frac{g(x)}{h(x)}$$

*where g and h satisfy the conditions of Theorem 4.5 cut only at boundary points.*

**Proof:** This follows from Corollary 4.7, since splitting rules of this form satisfy the conditions of Theorem 4.6 by Lemmas 4.14, 4.15 and 4.16. □

Similar arguments can be used to show that all of the results obtained previously assuming no repeated values carry over to the case in which repeated values are allowed.

# 5 Applications

We now apply the theory developed above to the Entropy, Gini, and Inaccuracy impurity measures, as well as to the purity gain and purity gain ratio, which respectively generalize the information gain and information gain ratio to impurity measures other than Entropy.

## 5.1 Entropy

It is well-known that the Entropy impurity measure

$$\text{Ent}\,(p_1, \ldots p_m) = \sum_{i=1}^{m} -p_i \log(p_i) \tag{35}$$

is strictly convex ∩ (see, for example (Cover & Thomas, 1991)); it follows that the splitting rule f(n : Ent) inherits all of the properties discussed in Section 4, including the minima-free property. We thus obtain, via a slightly different approach (Theorem 4.1), Fayyad and Irani's result (Fayyad & Irani, 1992b) that Entropy always cuts at boundary points. For completeness, we include a proof that the Entropy impurity measure is strictly convex ∩:

**Theorem 5.1** *The Entropy impurity measure (35) is strictly convex ∩, that is, for $\vec{p}, \vec{q} \in \Omega$, $\lambda \in (0,1)$*

$$Ent(\lambda \vec{p} + (1-\lambda)\vec{q}) > \lambda Ent(\vec{p}) + (1-\lambda) Ent(\vec{q})$$

**Proof:** Recall that $p$ and $q$ are purity vectors, i.e.

$$\vec{p} = (p_1, \ldots p_m)$$

Fix $\vec{p}$, $\vec{q}$, and $\lambda$, and let

$$w(y) = -y \log(y)$$

Claim 1 *For each $i$, $w(\lambda p_i + (1-\lambda)q_i) \geq \lambda w(p_i) + (1-\lambda)w(q_i)$, with equality if and only if $p_i = q_i$.* If $p_i = q_i$, it is clear that equality holds. Thus assume $p_i \neq q_i$. We have three cases to consider:

<u>**Case 1**</u> $pi > 0,\ q_i = 0$
We have

$$\begin{aligned} w(\lambda p_i + (1-\lambda)q_i) &= w(\lambda p_i) \\ &= -\lambda p_i \log(\lambda p_i) \\ &= -\lambda p_i \log(\lambda) - \lambda p_i \log(p_i) \\ &> -\lambda p_i \log(p_i) \\ &= \lambda w(p_i) \\ &= \lambda w(p_i) + (1-\lambda) w(q_i) \end{aligned}$$

and the claim holds.

<u>**Case2**</u> $p_i = 0,\ q_i > 0$
We have

$$\begin{aligned} w(\lambda p_i + (1-\lambda)q_i) &= w((1-\lambda)q_i) \\ &= -(1-\lambda)q_i \log((1-\lambda)q_i) \\ &= -(1-\lambda)q_i \log((1-\lambda)) - (1-\lambda)q_i \log(q_i) \\ &> -(1-\lambda)q_i \log(q_i) \\ &= (1-\lambda) w(q_i) \\ &= \lambda w(p_i) + (1-\lambda) w(q_i) \end{aligned}$$

and the claim holds.

<u>**Case 3**</u> $p_i > 0,\ q_i > 0$
The first and second derivatives of $w(y) = -y \log(y)$ are:

$$\begin{aligned} w'(y) &= -\log(y) - 1 \\ w''(y) &= -\frac{1}{y} \\ &< 0 \qquad \text{for } y > 0 \end{aligned}$$

Thus $w(y)$ is strictly convex n for $y > 0$ *by* Fact **3,** and since $p_i \neq q_i$ and $\lambda \in (0,1)$, it follows that

$$w(\lambda p_i + (1-\lambda)q_i) > \lambda w(p_i) + (1-\lambda) w(q_i)$$

as required. This completes the proof of Claim *1*

Thus for each i, we have

$$w(\lambda p_i + (1-\lambda)q_i) \geq \lambda w(p_i) + (1-\lambda) w(q_i)$$

and, since $\vec{p} \neq \vec{q}$, for at least one i we have

$$w(\lambda p_i + (1-\lambda)q_i) > \lambda w(p_i) + (1-\lambda) w(q_i).$$

It follows by summing over $i$ that

$$\sum_{i=1}^{m} w(\lambda p_i + (1-\lambda)q_i) > \sum_{i=1}^{m} \lambda w(p_i) + \sum_{i=1}^{m} (1-\lambda) w(q_i)$$

or

$$\sum_{i=1}^{m} w(\lambda p_i + (1-\lambda)q_i) > \lambda \sum_{i=1}^{m} w(p_i) + (1-\lambda)\sum_{i=1}^{m} w(q_i)$$

or equivalently

$$\mathrm{Ent}(\lambda\vec{p} + (1-\lambda)\vec{q}) > \lambda\mathrm{Ent}(\vec{p}) + (1-\lambda)\mathrm{Ent}(\vec{q})$$

as required. □

## *5.2* Gini

The Gini impurity measure

$$I(p_1,\ldots,p_m) = 1 - \sum p_i^2$$

has been shown to be strictly convex n (Breiman et al., *1984)*. Here we extend this result:

**Theorem 5.2** *For $a > 1$, the impurity measure*

$$I(p_1,\ldots,p_m) = 1 - \sum p_i^\alpha$$

*is strictly convex $\cap$.*

**Proof:** Let $\lambda \in (0,1)$, and

$$\vec{p},\ \vec{q} \in \Omega = \{(p_1,\ldots,p_m) :\ 0 \le p_i \le 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m} p_i = 1\}$$

where $\vec{p} \neq \vec{q}$. Then we must show that

$$I(\lambda\vec{p} + (1-\lambda)\vec{q}) > \lambda I(\vec{p}) + (1-\lambda)I(\vec{q}) \tag{36}$$

but before doing so we prove a few claims.

<u>**Claim 1**</u> *$Xy^a + (1-\lambda) - (Xy + (1-\lambda))^\alpha \geq 0$, with equality if and only if $y = 1$, where $y > 0$,* $0 < \lambda < 1$, *and* $\alpha > 1$.

Let

$$w(y) = \lambda y^\alpha + (1-\lambda) - (Xy + (1-\lambda))^\alpha$$

Then

$$\begin{aligned} w'(y) &= \lambda\alpha y^{\alpha-1} - a(\lambda y + (1-\lambda))^{\alpha-1}\lambda \\ &= \lambda\alpha y^{\alpha-1}\underbrace{\left(1 - \left(\lambda + \frac{1-\lambda}{y}\right)^{\alpha-1}\right)}_{w_1(y)} \\ &= \lambda\alpha y^{\alpha-1} w_1(y) \end{aligned}$$

where $w_1(y) = 1 - \left(A + \frac{1-\lambda}{y}\right)^{\alpha-1}$. Furthermore,

$$\begin{aligned} w_1'(y) &= -(\alpha-1)\left(\lambda + \frac{1-\lambda}{y}\right)^{\alpha-2}\left(-\frac{1-\lambda}{y^2}\right) \\ &= \frac{(a-1)(1-A)}{y^2}\left(\lambda + \frac{1-\lambda}{y}\right)^{\alpha-2} \\ &> 0 \qquad\qquad \text{since } a > 1, 0 < A < 1, y > 0 \end{aligned}$$

so that $w_1(y)$ is strictly increasing for $y > 0$. Since

$$\begin{aligned} w_1(1) &= 1 - \left(\lambda + \frac{1-\lambda}{1}\right)^{\alpha-1} \\ &= 0 \end{aligned}$$

and $w_1(y)$ is strictly increasing, it follows that

$$w_1(y) \begin{cases} < 0 & \text{for } 0 < y < 1 \\ = 0 & \text{for } y = 1 \\ > 0 & \text{for } y > 1 \end{cases}$$

and thus that

$$w'(y) = \lambda\alpha y^{\alpha-1} w_1(y) \begin{cases} < 0 & \text{for } 0 < y < 1 \\ = 0 & \text{for } y = 1 \\ > 0 & \text{for } y > 1 \end{cases}$$

so that $w(y)$ is strictly decreasing for $y < 1$ and strictly increasing for $y > 1$. Since

$$\begin{aligned} w(1) &= \lambda(1)^\alpha + (1-\lambda) - (\lambda(1) + (1-\lambda))^\alpha \\ &= 0 \end{aligned}$$

this implies that

$$w(y) \begin{cases} > 0 & \text{for } 0 < y < 1 \\ = 0 & \text{for } y = 1 \\ > 0 & \text{for } y > 1 \end{cases}$$

which proves the claim.

**Claim 2** *For each $i$, $\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha \geq 0$, with equality if and only if $p_i = q_i$, when: $0 < A < 1$ and $a > 1$.*
We have four cases:

**Case 1** $p_i = 0, q_i = 0$.
In this case

$$\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha = 0$$

so that equality holds for $p_i = q_i = 0$

**Case 2** $p_i = 0, q_i > 0$
In this case

$$\begin{aligned}\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha &= (1-\lambda)q_i^\alpha - ((1-\lambda)q_i)^\alpha \\ &= (1-\lambda)q_i^\alpha\left(1-(1-\lambda)^{\alpha-1}\right) \\ &> 0 \qquad \text{since } 0<\lambda<1, \alpha>1\end{aligned}$$

so that the inequality is strict for $q_i \neq p_i = 0$.
**Case3** $p_i > 0, q_i = 0$.
In this case

$$\begin{aligned}\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha &= \lambda p_i^\alpha - (\lambda p_i)^\alpha \\ &= \lambda p_i^\alpha\left(1-\lambda^{\alpha-1}\right) \\ &> 0 \qquad \text{since } 0<\lambda<1, \alpha>1\end{aligned}$$

so that the inequality is strict for $p_i \neq q_i = 0$.
**Case 4** $p_i > 0, q_i > 0$
Let

$$y = \frac{p_i}{q_i} > 0$$

Then

$$\begin{aligned}\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha &= q_i^\alpha\left(\lambda\left(\frac{p_i}{q_i}\right)^\alpha + (1-\lambda) - \left(\lambda\left(\frac{p_i}{q_i}\right) + (1-\lambda)\right)^\alpha\right) \\ &= q_i^\alpha\left(Xy* + (1-\lambda) - (Xy + (1-\lambda))^\alpha\right) \\ &\begin{cases} >0 & \text{for } y \neq 1 \text{ (i.e. } p_i \neq q_i) \text{ by Claim } 1 \\ =0 & \text{for } y = 1 \text{ (i.e. } p_i = q_i) \text{ by Claim } 1\end{cases}\end{aligned}$$

so that if $p_i = q_i$ equality holds, and if $p_i \neq q_i$ the inequality is strict.

This proves the claim.

**Claim 3** $\sum_{i=1}^m (\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha) > 0$, *where* $0 < \lambda < 1$ *and* $a > 1$.
Each term in the sum is $\geq 0$, by Claim 2. If all terms were zero, then we would have $p_i = q_i$ for all $i$, or $\vec{p} = \vec{q}$, in violation of the assumption that $\vec{p} \neq \vec{q}$. Thus the sum is strictly $> 0$, which proves the claim.

We now have the necessary tools to prove *(36)*:

$$\begin{aligned}I(\lambda\vec{p} + (1-\lambda)\vec{q}) - \lambda I(\vec{p}) - (1-\lambda)I(\vec{q}) &= 1 - \sum_{i=1}^m (\lambda p_i + (1-\lambda)q_i)^\alpha - \lambda\left(1 - \sum_{i=1}^m p_i^\alpha\right) - (1-\lambda)\left(1 - \sum_{i=1}^m p_i^\alpha\right) \\ &= \sum_{i=1}^m (\lambda p_i^\alpha + (1-\lambda)q_i^\alpha - (\lambda p_i + (1-\lambda)q_i)^\alpha) \\ &> 0 \qquad\qquad \text{by Claim } 3\end{aligned}$$

It follows that **I** is strictly convex ∩. □

## 5.3 Inaccuracy

Inaccuracy is in some sense the natural impurity measure to use, since what we really want to do is to maximize the accuracy of the classification over the test data. However, this ignores the lact that decision tree construction is a multi-step process, and it is not clear that choosing the split to minimize an Inaccuracy-based splitting rule at each step will yield a tree that performs well on the test data. In fact, experimental evidence suggests that the Inaccuracy impurity measure yields larger trees (Brodley, 1995) and lower predictive accuracy (Pazzani, Merz, Murphy, Ali, Hume, & Brunk, 1994) than the Entropy impurity measure. Such problems led to the abandonment of Inaccuracy in favor of measures such as Entropy and Gini (Brodley, 1995; Lubinsky, 1995).

The problems with Inaccuracy can be traced to the fact that it is not strictly convex n, but only convex n, as shown below:

**Theorem 5.3** *The Inaccvracy impvrzty measure*

$$I(p_1, \ldots, p_m) = 1 - \max_i p_i$$

*is convex ∩, but not strictly so.*

**Proof:** For $\lambda \in [0,\mathrm{l}]$ and

$$\vec{p},\ \vec{q} \in \Omega = \{(p_1, \ldots, p_m) : 0 \leq p_i \leq 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m} p_i = 1\}$$

we must show that

$$I(\lambda\vec{p} + (1 - \lambda)\vec{q}) \geq \lambda I(\vec{p}) + (1 - \lambda)I(\vec{q})$$

Consider

$$\begin{aligned}
I(\lambda\vec{p} &+ (1-\lambda)\vec{q}) - (\lambda I(\vec{p}) + (1-\lambda)I(\vec{q})) \\
&= 1 - \max_i (\lambda p_i + (1-\lambda)q_i) - \left(\lambda(1 - \max_i p_i) + (1-\lambda)(1 - \max_j q_j)\right) \\
&= 1 - \lambda - (1-\lambda) + \lambda \max_i p_i + (1-\lambda)\max_j q_j - \max_i (\lambda p_i + (1-\lambda)q_i) \\
&= \lambda \max_i p_i + (1-\lambda)\max_j q_j - \max_i (\lambda p_i + (1-\lambda)q_i) \\
&\geq 0
\end{aligned}$$

The $\geq$ holds because one can vary the indices i in $p_i$ and $j$ in $q_j$ independently to achieve a potentially higher maximum value of

$$\mathrm{X}\max_{\imath} \mathrm{pi}+ (1-\lambda)\max_{r} q_j$$

than would be possible if they were required to be the same, as in

$$\max_i (\lambda p_i + (1-\lambda)q_i)$$

Thus

$$I(\lambda\vec{p} + (1 - \lambda)\vec{q}) \geq \lambda I(\vec{p}) + (1 - \lambda)I(\vec{q})$$

so that I is convex ∩, but not strictly so, since equality holds when the intersection of the set of indices that maximize $p_i$ with those that maximize $q_j$ is non-empty. □

That Inaccuracy is only convex ∩ admits, by Theorem 4.2, the possibility that $f(x;\text{Inacc})$ attains its minimum in a flat valley that is constant over a uniform sequence.[8] Indeed, it was recognized as long ago as 1984 that there exist non-trivial sequences for which the the Inaccuracy-based splitting rule is constant over the entire sequence (Breiman et al., 1984). Such sequences are not at all atypical; they have the property that for every cut point $c$, the same class is in the majority on either side of $c$ (an example is shown in Figure 8). To see why this causes problems, if an attribute $A$ is selected for which the splitting curve is constant, the algorithm will most likely split off the first or last instance along A, so that one of the newly created nodes contains only a single instance. If this occurs repeatedly, it could explain the large tree sizes and poor generalization performance of the Inaccuracy impurity measure.

It appears that Inaccuracy can be fixed, however. The key observation was made by Lubinsky (Lubinsky, 1995), who noted that the splitting rule

$$\text{Inacc.Gini} = f(n;\text{Inacc}) + \frac{1}{N} f(n;\text{Gini}) \tag{37}$$

breaks ties in Inaccuracy using Gini, by virtue of the following facts:

- $f(n;\text{Inacc})$ changes in increments of at most $\frac{1}{N}$, as it is the fraction of samples that are wrongly classified.
- $0 \leq f(n;\text{Gini}) \leq 1$.

Lubinsky reported that the Inacc.Gini splitting rule produced significantly smaller trees than Gini, with comparable error rates on all data sets but one (Lubinsky, 1995). The relevance of this to the present work is that the Inacc.Gini splitting rule (37) has the minima-free property:

**Theorem 5.4** Let I be a strictly convex ∩ impurity *measure*, and define the impurity *measure*

$$\text{Inacc.}I(\vec{p}) \equiv Inacc(\vec{p}) + \frac{1}{N}\frac{I(\vec{p})}{M}$$

where

$$M = 1 + \max_{\vec{p}\in\Omega} I(\vec{p})$$

Then the splitting rule $f(n;Inacc.I)$ only *cuts* at *boundary* points.

**Proof:** We show that the impurity measure Inacc.Iis strictly convex ∩. It then follows by Theorem 4.1 that f (n.;Inacc.I) only cuts at boundary points. For $\lambda \in (0,1)$ and

$$\vec{p},\ \vec{q} \in \Omega = \{(p_1,\ldots,p_m) : 0 \leq p_i \leq 1 \text{ for all } i, \text{ and } \sum_{i=1}^{m} p_i = 1\}$$

we must show that

$$\text{Inacc.I}(\lambda\vec{p} + (1-\lambda)\vec{q}) \geq \lambda\ \text{Inacc.I}(\vec{p}) + (1-\lambda)\ \text{Inacc.I}(\vec{q})$$

[8] Elomaa and Rousu (Elomaa & Rousu, 1996) have shown that $f(x;\text{Inacc})$ is "well-behaved", meaning that its value at one of the boundary points of a uniform sequence is at least as small as that at any interior point.

We have

$$\begin{aligned}
\text{Inacc.I}(\lambda\vec{p} + (1 - \text{A})\$ &= (\text{Inacc} + \frac{1}{NM}I)(\lambda\vec{p} + (1 - \text{A})\$ \\
&= \text{Inacc}(\lambda\vec{p} + (1 - \lambda)\vec{q}) + \frac{1}{NM}I(\lambda\vec{p} + (1 - \lambda)\vec{q}) \\
&> \lambda\ \text{Inacc}(\vec{p}) + (1 - \lambda)\ \text{Inacc}(\vec{q}) + \frac{1}{NM}(\lambda I(\vec{p}) + (1 - \lambda)I(\vec{q})) \\
&= \lambda(\text{Inacc} + \frac{1}{NM}I)(\vec{p}) + (1 - \text{A})(\text{Inacc} + \frac{1}{NM}I)(\vec{q}) \\
&= \text{A Inacc.I}(\vec{p}) + (1 - \text{A})\ \text{Inacc.I}(\vec{q})
\end{aligned}$$

as required

**Corollary 5.5** The *Inacc.Gini* splitting rule **(37)** has the minima-free property.

**Proof:** First note that

$$\begin{aligned}
\text{Inacc.Gini} &= f(n;\text{Inacc}) + \frac{1}{N}f(n;\text{Gini}) \\
&= \frac{n}{N}\text{Inacc}(\vec{\ell}(n)) + (1 - \frac{n}{N})\text{Inacc}(\vec{r}(n)) + \frac{1}{N}\left(\frac{n}{N}\text{Gini}(\vec{\ell}(n)) + (1 - \frac{n}{N})\text{Gini}(\vec{r}(n))\right) \\
&- \frac{n}{N}\left(\text{Inacc}(\vec{\ell}(n)) + \frac{1}{N}\text{Gini}(\vec{\ell}(n))\right) + (1 - \frac{n}{N})\left(\text{Inacc}(\vec{r}(n)) + \frac{1}{N}\text{Gini}(\vec{r}(n))\right) \\
&= \frac{n}{N}\left(\text{Inacc} + \frac{1}{N}\text{Gini}\right)(\vec{\ell}(n)) + (1 - \frac{n}{N})\left(\text{Inacc} + \frac{1}{N}\text{Gini}\right)(\vec{r}(n)) \\
&= f(n;\text{Inacc} + \frac{1}{N}\text{Gini})
\end{aligned}$$

Theorem 5.4 and the convexity of the Gini impurity measure (Theorem 5.2) then imply that Inacc.Gini has the minima-free property. □

That Inacc.Gini has the minima-free property may explain why it produces smaller trees than either Inaccuracy or Gini alone. Inacc.Gini generally splits where Inaccuracy does, and thus produces more balanced splits than Gini, which has a preference for splitting near the ends of the sequence (Breiman, 1996; Lubinsky, 1995); it follows that Inacc.Gini generally produces smaller trees than Gini. Moreover, the problem of Inaccuracy producing large trees as a result of being constant over the entire sequence does not occur for Inacc.Gini, which has the minima-free property; it follows that Inacc.Gini produces smaller trees than Inaccuracy.

## 5.4 Purity Gain and Purity Gain Ratio

We now generalize the information gain and information gain ratio to impurity measures other than entropy by introducing the purity gain, defined as

$$\text{purity gain} = f(0;\mathbf{I}) - f(x;I) \tag{38}$$

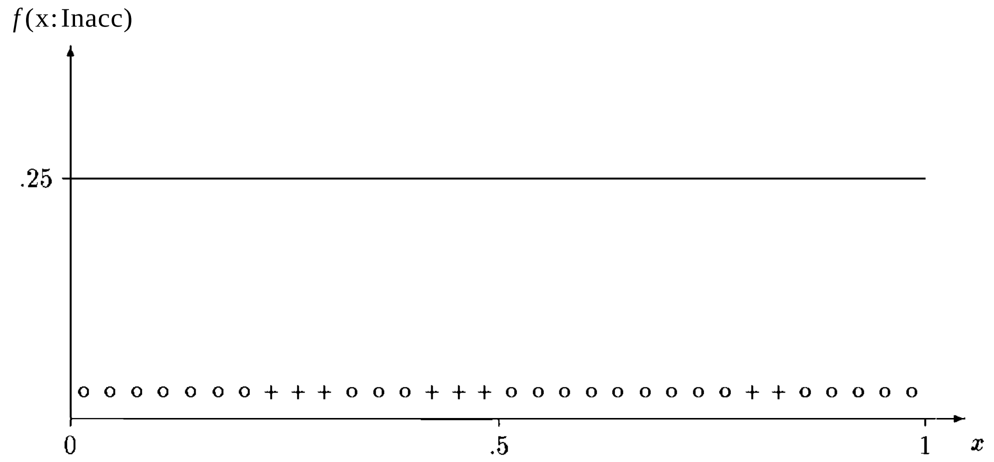


Figure 8: Sequence for which the splitting rule based on the Inaccuracy impurity measure is a constant function of the cut point.

and the purity gain ratio, defined as

$$\text{purity gain ratio} = \frac{\text{purity gain}}{-x\log(x) - (1-x)\log(1-x)} \tag{39}$$

In this section we derive sufficient conditions for the purity gain and purity gain ratio to cut only at boundary points, and show that the information gain and information gain ratio satisfy these conditions, thus showing that they cut only at boundary points. We begin with the easy case, purity gain:

**Theorem 5.6** *For any strictly convex $\cap$ impurity measure* I, *the purity gain (38) only cuts at boundary points.*

**Proof:** To discuss the purity gain, which is to be maximized, in the framework we have developed, which deals with splitting rules that are to be minimized, we define the negative purity gain as the negative of the purity gain, that is

$$\text{negative purity gain} = f(x;\mathbf{I}) - f(0;\mathbf{I})$$

Now, since the negative purity gain differs from the average impurity $f(x;I)$ by the constant $f(0;I)$, the negative purity gain inherits all of the properties of the average impurity $f(x;I)$, and in particular the minima-free property (Theorem 4.1). It follows that the negative purity gain achieves its minimum at a boundary point, or equivalently, that the purity gain achieves its maximum at a boundary point. □

**Corollary 5.7** *The information gain (2) cuts only at boundary points.*

**Proof:** This follows from Theorem 5.6 and the convexity of the Entropy impurity measure (Theorem 5.1). □

**Theorem 5.8** *Let* **I** *be a strictly convex* $\cap$ *impurity measure and let*

$$\begin{aligned} g(x) &= f(x; I) - f(0; I) \\ h(x) &= -x\log(x) - (1-x)\log(1-x) \\ G(x) &= g'h - h'g \end{aligned}$$

*Then if*

$$G(0) = \lim_{x\to 0+} G(x) = \lim_{x\to 0+} g'h - h'g \leq 0$$

*and*

$$G(1) = \lim_{x\to 1^-} G(x) = \lim_{x\to 1^-} g'h - h'g \geq 0$$

*the purity gain ratio* ***(39)*** *only cuts at boundary points.*

**Proof:** To discuss the purity gain ratio, which is to be maximized, in the frameworlc we have developed, which deals with splitting rules that are to be minimized, we define the negative purity gain ratio as the negative of the purity gain ratio, that is

$$\begin{aligned} \text{negative purity gain ratio} &= \frac{f(x; I) - f(0; I)}{-x\log(x) - (1-x)\log(1-x)} \\ &= \frac{g(x)}{h(x)} \end{aligned}$$

With the goal of applying Theorem 4.5, we compute the first two derivatives of h to check that $h'' < 0$:

$$\begin{aligned} h' &= \log(1-x) + 1 - \log(x) - 1 \\ &= \log(1-x) - \log(x) \\ h'' &= -\frac{1}{1-x} - \frac{1}{x} \\ &= -\frac{x+1-x}{x(1-x)} \\ &= -\frac{1}{x(1-x)} \end{aligned}$$

So $h''(x)$ is indeed $< 0$ for $x \in (0,1)$. Furthermore,

- $g''(x) \leq 0$ over uniform sequences, as
$$g''(x) = f''(x; \mathbf{I})$$
and $f''(x; \mathbf{I})$ is $\leq 0$ by Fact 4, in view of the convexity of $f(x; \mathbf{I})$ (Theorem 4.1).
- $g(x) < 0$ for $x \in (0,1)$ by Lemma 4.4.
- $h(x) > 0$ for $x \in (0,1)$.

Thus we can apply Theorem 4.5 to conclude that the negative purity gain ratio

$$\mathrm{F}(x) = \frac{g(x)}{h(x)}$$

attains its minimum at a boundary point, or equivalently, that the purity gain ratio achieves its maximum at a boundary point. □

**Corollary 5.9** The ***information gain ratio (4) alway cuts at a boundary point.***

**Proof:** In view of Theorem 5.8 and the convexity of the Entropy impurity measure (Theorem 5.1), it suffices to show that $G(0) \leq 0$ and $G(1) \geq 0$, where

$$\begin{aligned} g(x) &= \mathrm{f}(x;\mathrm{Ent}) - \mathrm{f}(0;\mathrm{Ent}) \\ h(x) &= -x\log(x) - (1-x)\log(1-x) \\ G(x) &= g'h - h'g \end{aligned}$$

**Claim 1** $G(0) = 0$.
For a cut point $x$ inside the first uniform sequence, assumed to consist of class k samples, the fraction to the left of $x$ that belong to class $i$ is:

$$\ell_i(x) = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$$

and the fraction to the right that belong to class i is:

$$\begin{aligned} r_i(x) &= \begin{cases} \frac{N_i}{N(1-x)} & i \neq k \\ \frac{N_i - Nx}{N(1-x)} & i = k \end{cases} \\ &= \begin{cases} \frac{n_i}{1-x} & i \neq k \\ \frac{n_i - x}{1-x} & i = k \end{cases} \end{aligned}$$

where $N_i$ and $n_i$ are respectively the number and fraction of class i samples in the entire sequence. It follows that

$$\begin{aligned} g(x) &= \mathrm{f}(\mathrm{x};\mathrm{Ent}) - \mathrm{f}(0;\mathrm{Ent}) \\ &= x\mathrm{Ent}\left(\vec{\ell}(x)\right) + (1-x)\mathrm{Ent}\left(\vec{r}(x)\right) - \left(-\sum_{i=1}^{m} n_i \log n_i\right) \\ &= (1-x)\mathrm{Ent}\left(\vec{r}(x)\right) - \left(-\sum_{i=1}^{m} n_i \log n_i\right) \quad \text{since } \mathrm{Ent}\left(\vec{\ell}(x)\right) = 0 \\ &= (1-x)\left(-\frac{n_k - x}{1-x}\log\left(\frac{n_k - x}{1-x}\right) - \sum_{i\neq k}\frac{n_i}{1-x}\log\left(\frac{n_i}{1-x}\right)\right) - \left(-\sum_{i=1}^{m} n_i \log n_i\right) \\ &= -(n_k - x)\log\left(\frac{n_k - x}{1-x}\right) - \sum_{i\neq k} n_i \log\left(\frac{n_i}{1-x}\right) + \sum_{i=1}^{m} n_i \log n_i \end{aligned}$$

$$
\begin{aligned}
&= -(n_k - \mathrm{x})\log(n_k - x) + (n_k - \mathrm{x})\log(1 - \mathrm{x}) - \sum_{\mathrm{ifk}} n_i \log n_i + \sum_{i\neq k} n_i \log(1 - \mathrm{x}) + \sum_{\mathrm{i=1}}^{m} n_i \log n_i \\
&= -(n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (n_k - \mathrm{x})\log(1 - \mathrm{x}) + \sum_{\mathrm{ifk}} n_i \log(1 - \mathrm{x}) + n_k \log n_k \\
&= n_k \log n_k - (n_k - x)\log(n_k - x) + \left(n_k - x + \sum_{i\neq k} n_i\right)\log(1 - x) \\
&= n_k \log n_k - (n_k - x)\log(n_k - x) + (1 - x)\log(1 - x)
\end{aligned}
$$

and furthermore

$$
\begin{aligned}
g'(x) &= \log(n_k - \mathrm{x}) + 1 - \log(1 - x) - 1 \\
&= \log(n_k - \mathrm{x}) - \log(1 - \mathrm{x})
\end{aligned}
$$

Thus

$$
\begin{aligned}
G(x) &= g'h - \mathrm{gh'} \\
&= (\log(n_k - \mathrm{x}) - \log(1 - x))(-x\log(x) - (1 - \mathrm{x})\log(1 - \mathrm{x})) \\
&\quad - (n_k \log n_k - (n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (1 - \mathrm{x})\log(1 - x))(\log(1 - \mathrm{x}) - \log(x))
\end{aligned}
$$

$$
\begin{aligned}
G(0) &= \lim_{x\to 0^+} G(x) \\
&= \lim_{x\to 0^+} (n_k \log n_k - (n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (1 - \mathrm{x})\log(1 - \mathrm{x}))\log(x) \quad (= 0\,.\,(-\infty);\ \text{indeterminate}) \\
&= \lim_{x\to 0^+} \frac{\log(x)}{(n_k \log n_k - (n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (1 - \mathrm{x})\log(1 - x))^{-1}} \quad (= \infty/\infty;\ \text{indeterminate}) \\
&= \lim_{x\to 0^+} \frac{x^{-1}}{-(n_k \log n_k - (n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (1 - \mathrm{x})\log(1 - x))^{-2}(\log(n_k - \mathrm{x}) - \log(1 - \mathrm{x}))} \\
&\quad \text{by l'Hôpital's rule} \\
&= \lim_{x\to 0^+} \frac{-(n_k \log n_k - (n_k - x)\log(n_k - \mathrm{x}) + (1 - \mathrm{x})\log(1 - x))^2}{x(\log(n_k - \mathrm{x}) - \log(1 - x))} \quad (=0/0;\ \text{indeterminate}) \\
&= \lim_{x\to 0^+} \frac{-2(n_k \log n_k - (n_k - \mathrm{x})\log(n_k - \mathrm{x}) + (1 - x)\log(1 - x))(\log(n_k - x) - \log(1 - x))}{(\log(n_k - \mathrm{x}) - \log(1 - x)) + x\left(\frac{-1}{n_k - x} + \frac{1}{1 - x}\right)} \\
&\quad \text{by l'Hôpital's rule} \\
&= 0 \quad \text{since } 0 < n_k < 1
\end{aligned}
$$

as required.

**<u>Claim</u>** <u>2</u> G(1) = 0.
For a cut point x inside the last uniform sequence, assumed to consist of class k samples, the fraction to the left of x that belong to class *i* is:

$$
\ell_i(x) = \begin{cases} \frac{N_i}{Nx} & i \neq k \\ \frac{N_\imath - N(1-x)}{Nx} & i = k \end{cases}
$$

$$= \begin{cases} \frac{n_i}{x} & i \neq k \\ \frac{n_i-(1-x)}{x} & i = k \end{cases}$$

and the fraction to the right that belong to class $i$ is:

$$r_i(x) = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$$

It follows that

$$\begin{aligned}
g(x) &= \mathrm{f}(\mathrm{x};\mathrm{Ent}) - \mathrm{f}(0;\mathrm{Ent}) \\
&= x\mathrm{Ent}\left(\vec{\ell}(x)\right) + (1-x)\mathrm{Ent}\left(\vec{r}(x)\right) - \left(-\sum_{i=1}^{m} n_i \log n_i\right) \\
&= x\mathrm{Ent}\left(\vec{\ell}(x)\right) - \left(-\sum_{i=1}^{m} \mathrm{n\ o\ n}\ \right) \quad \text{since } \mathrm{Ent}\left(\vec{r}(x)\right) = 0 \\
&= x\left(-\frac{n_k-(1-x)}{x}\log\left(\frac{n_k-(1-x)}{x}\right) - \sum_{i\neq k}\frac{n_i}{x}\log\left(\frac{n_i}{x}\right)\right) - \left(-\sum_{i=1}^{m} n_i \log n_i\right) \\
&= -(n_k-(1-x))\log\left(\frac{n_k-(1-x)}{x}\right) - \sum_{\mathrm{ifk}} n_i \log\left(\frac{n_i}{x}\right) + \sum_{\mathrm{i=1}}^{m} n_i \log n_i \\
&= -(n_k-(1-x))\log(n_k-(1-x)) + (n_k-(1-x))\log x - \sum_{i\neq k} n_i \log n_i + \sum_{i\neq k} n_i \log x + \sum_{i=1}^{m} n_i \log n_i \\
&= n_k \log n_k - (n_k-(1-\mathrm{x}))\log(n_k-(1-\mathrm{x})) + (n_k-(1-\mathrm{x}))\log \mathrm{x} + \sum_{i\neq k} n_i \log \mathrm{x} \\
&= n_k \log n_k - (n_k-(1-x))\log(n_k-(1-\mathrm{x})) + \left(n_k-(1-\mathrm{x}) + \sum_{\mathrm{i}\neq k} n_i\right)\log x \\
&= n_k \log n_k - (n_k-(1-x))\log(n_k-(1-\mathrm{x})) + \mathrm{x}\log\mathrm{x}
\end{aligned}$$

and furthermore

$$\begin{aligned}
g'(x) &= \log \mathrm{x} + 1 - \log(n_k-(1-x)) - 1 \\
&= \log \mathrm{x} - \log(n_k-(1-\mathrm{x}))
\end{aligned}$$

Thus

$$\begin{aligned}
G(x) &= \mathrm{g'h} - gh' \\
&= (\log \mathrm{x} - \log(n_k-(1-x)))(-x\log(x) - (1-\mathrm{x})\log(1-x)) \\
&\quad - (n_k \log n_k - (n_k-(1-\mathrm{x}))\log(n_k-(1-\mathrm{x})) + \mathrm{x}\log x)(\log(1-\mathrm{x}) - \log(x))
\end{aligned}$$

$$\mathrm{G(1)} = \lim_{x\to 1^-} G(x)$$

$$\begin{aligned}
&= \lim_{x\to 1^-} -(n_k \log n_k - (n_k - (1-x))\log(n_k - (1-x)) + x\log x)\log(1-x) \\
&\qquad (= 0 \,.\, (-\infty);\ \text{indeterminate}) \\
&= \lim_{x\to 1^-} \frac{-\log(1-x)}{(n_k \log n_k - (n_k - (1-x))\log(n_k - (1-x)) + x\log x)^{-1}} \qquad (= \infty/\infty;\ \text{indeterminate}) \\
&= \lim_{x\to 1^-} \frac{(1-x)^{-1}}{-(n_k \log n_k - (n_k - (1-x))\log(n_k - (1-x)) + x\log x)^{-2}(\log x - \log(n_k - (1-x)))} \\
&\qquad \text{by l'Hôpital's rule} \\
&= \lim_{x\to 1^-} \frac{-(n_k \log n_k - (n_k - (1-x))\log(n_k - (1-x)) + x\log x)^2}{(1-x)(\log x - \log(n_k - (1-x)))} \qquad (=0/0;\ \text{indeterminate}) \\
&= \lim_{x\to 1^-} \frac{-2(n_k \log n_k - (n_k - (1-x))\log(n_k - (1-x)) + x\log x)(\log x - \log(n_k - (1-x)))}{-(\log x - \log(n_k - (1-x))) + (1-x)\left(\frac{1}{x} - \frac{1}{n_k-(1-x)}\right)} \\
&\qquad \text{by l'Hôpital's rule} \\
&= 0 \quad \text{since } 0 < n_k < 1
\end{aligned}$$

as required.

□

## 6 Conclusion

We have established that splitting rules $f(n;I)$ based on strictly convex $\cap$ impurity measures

- attain their minimum at a boundary point, and
- are strictly decreasing over the first uniform sequence and strictly increasing over the last.

We applied this theory to show that the Entropy and Gini impurity measures always cut at boundary points.

Splitting rules $f(n;I)$ based on impurity measures that are only convex $\cap$ were shown to have similar properties, and this result was used to show that the Inaccuracy impurity measure either cuts at a boudary point, or attains its minimum in a flat valley that is constant over a uniform sequence.

We also developed tools for verifying the minima-free property for rational splitting rules, and used these to show that the information gain ratio always cuts at boundary points.

## Reference

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification *and* Regression Trees (First edition). Wadsworth, Belmont, CA.

Breiman, L. (1996). Some properties of splitting criteria. Machine Learning, *24*, 41–47

Brodley, C. E. (1995). Automatic selection of split criterion during tree growing based on node location. In Proceedings of the $12^{th}$ Intl. Conf. *on* Machine Learning, pp. 73–80.

Buntine, W., & Niblett, T. ('1992). A further comparison of splitting rules for decision-tree induction. Machine Learning, 8, 75–85.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. Machine Learning, 3, 261–283.

Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory (First edition). Wiley Interscience.

Dietterich, T., Kearns, M., & Mansour, Y. (1996). Applying the weak learning framework to understand and improve C4.5. In Machine Learning: Proceedings of the 13th *Intl.* Conf., pp. 96–104.

Elomaa, T., & Rousu, J. (1996). On the well-behavedness of important attribute evaluation functions. Tech. rep. NC-TR-97-006, University of Helsinki, Department of Computer Science.

Fayyad, U. M. (1994). Branching on attribute values in decision tree generation. In National Conference on Artificial Intelligence, pp. 601–606.

Fayyad, U. M., & Irani, K. B. (1990). What should be minimized in a decision tree?. In *AAAI-90:* Proceedings of the 8'' *National* Conf. on Artificial Intelligence, pp. 749–754.

Fayyad, U. M., & Irani, K. B. (1992a). The attribute selection problem in decision tree generation. In *AAAI-92*, pp. 104–110.

Fayyad, U. M., & Irani, K. B. (1992b). On the handling of continuous-valued attributes in decision tree generation. Machine Learning, 8, 87–102.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the 13th Intl. Joint Conf. on Artificial Intelligence, Vol. 2, pp. 1022--1027.

Goodman, R. M., & Smyth, P. (1988). Decision tree design from a communication theory standpoint. IEEE Transactions on Information Theory, 34(5), 979–994.

Lewis, P. M. (1962). The characteristic selection problem in recognition systems. *IRE* Transactions on Information Theory, 8, 171–178.

López de Mántaras, R. (1991). A distance-based attribute selection measure for decision tree induction. Machine Learning, 6, 81–92.

Lubinsky, D. J. (1995). Increasing the performance and consistency of classification trees by using the accuracy criterion at the leaves. In Proceedings of the 12'' Intl. Conf. on Machine Learning, pp. 371–377.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, 58, 414–434.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In Proceedings of the 11th *Intl.* Conf. on Machine Learning, pp. 217–225. Morgan Kaufmann.

Peressini, A. L., Sullivan, F. E., & Uhl, Jr., J. J. (1988). The Mathematics of Nonlinear Programming (First edition). Springer-Verlag.

Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann.

Quinlan, J. R. (1990). Induction of decision trees. In Shavlik, J. W., & Dietterich, T. G. (Eds.), *Readings in Machine Learning*, pp. 57–69. Morgan Kaufmann.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, *4*, 77–90.

Quinlan, J. R., & Rivest, R. L. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, *80*, 227–248.

Sethi, I. K., & Sarvarayudu, G. P. R. (1982). Hierarchical classifier design using mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(4), 441–445.

Strang, G. (1988). *Linear Algebra and its Applications* (Third edition). Harcourt, Brace, Jovanovich.