

Chapter ML:III

III. Decision Trees

- ❑ Decision Trees Basics
- ❑ Impurity Functions
- ❑ Decision Tree Algorithms
- ❑ Decision Tree Pruning

Impurity Functions

Splitting

Let t be a leaf node of an incomplete decision tree, and let $D(t)$ be the subset of the example set D that is represented by t . [\[Illustration\]](#)

Possible criteria for a splitting of $X(t)$:

1. Size of $D(t)$.
2. Purity of $D(t)$.
3. Ockham's Razor.

Impurity Functions

Splitting

Let t be a leaf node of an incomplete decision tree, and let $D(t)$ be the subset of the example set D that is represented by t . [\[Illustration\]](#)

Possible criteria for a splitting of $X(t)$:

1. Size of $D(t)$.

$D(t)$ will not be partitioned further if the number of examples, $|D(t)|$, is below a certain threshold.

2. Purity of $D(t)$.

$D(t)$ will not be partitioned further if all examples in D are members of the same class.

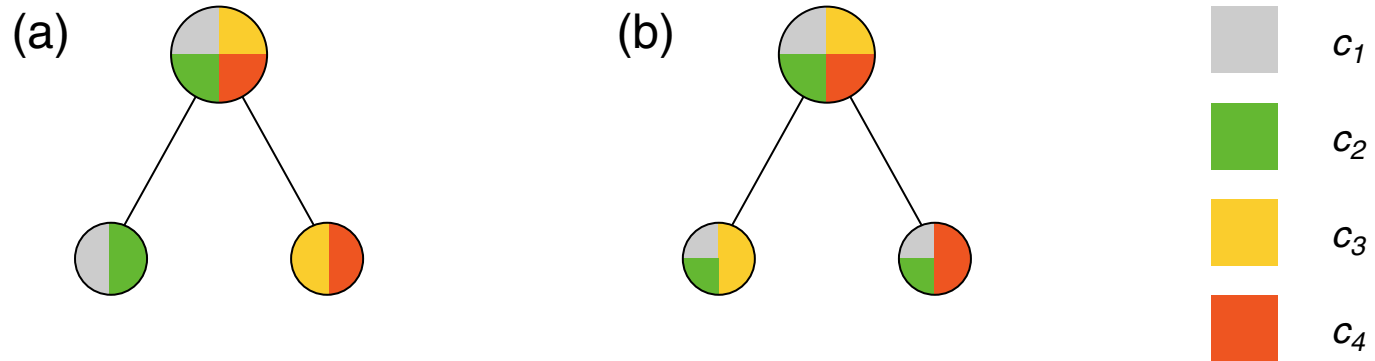
3. Ockham's Razor.

$D(t)$ will not be partitioned further if the resulting decision tree is not improved significantly by the splitting.

Impurity Functions

Splitting (continued)

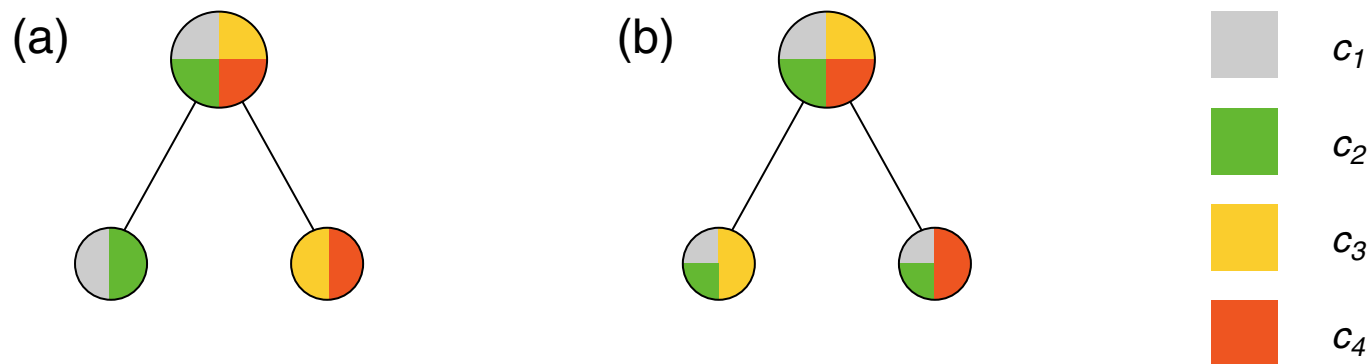
Let D be a set of examples over a feature space X and a set of classes $C = \{c_1, c_2, c_3, c_4\}$. Distribution of D for two possible splittings of X :



Impurity Functions

Splitting (continued)

Let D be a set of examples over a feature space X and a set of classes $C = \{c_1, c_2, c_3, c_4\}$. Distribution of D for two possible splittings of X :



- ❑ Splitting (a) minimizes the *impurity* of the subsets of D in the leaf nodes and should be preferred over splitting (b). This argument presumes that the misclassification costs are independent of the classes.
- ❑ The impurity is a function defined on $\mathcal{P}(D)$, the set of all subsets of an example set D .

Impurity Functions

Definition 4 (Impurity Function ι)

Let $k \in \mathbb{N}$. An impurity function $\iota : [0; 1]^k \rightarrow \mathbb{R}$ is a partial function defined on the standard $k-1$ -simplex, denoted Δ^{k-1} , for which the following properties hold:

- (a) ι becomes minimum at points $(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1)$.
- (b) ι is symmetric with regard to its arguments, p_1, \dots, p_k .
- (c) ι becomes maximum at point $(1/k, \dots, 1/k)$.

Impurity Functions

Definition 5 (Impurity of an Example Set $\iota(D)$)

Let D be a set of examples, let $C = \{c_1, \dots, c_k\}$ be set of classes, and let $c : X \rightarrow C$ be the ideal classifier for X . Moreover, let $\iota : [0; 1]^k \rightarrow \mathbb{R}$ an impurity function. Then, the impurity of D , denoted as $\iota(D)$, is defined as follows:

$$\iota(D) = \iota \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|}, \dots, \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_k\}|}{|D|} \right)$$

Impurity Functions

Definition 5 (Impurity of an Example Set $\iota(D)$)

Let D be a set of examples, let $C = \{c_1, \dots, c_k\}$ be set of classes, and let $c : X \rightarrow C$ be the ideal classifier for X . Moreover, let $\iota : [0; 1]^k \rightarrow \mathbb{R}$ an impurity function. Then, the impurity of D , denoted as $\iota(D)$, is defined as follows:

$$\iota(D) = \iota \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|}, \dots, \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_k\}|}{|D|} \right)$$

Definition 6 (Impurity Reduction $\Delta\iota$)

Let D_1, \dots, D_s be a partitioning of an example set D , which is induced by a splitting of a feature space X . Then, the resulting impurity reduction, denoted as $\Delta\iota(D, \{D_1, \dots, D_s\})$, is defined as follows:

$$\Delta\iota(D, \{D_1, \dots, D_s\}) = \iota(D) - \sum_{j=1}^s \frac{|D_j|}{|D|} \cdot \iota(D_j)$$

Remarks:

- ❑ The standard $k-1$ -simplex comprises all k -tuples with non-negative elements that sum to 1:
$$\Delta^{k-1} = \left\{ (p_1, \dots, p_k) \in \mathbf{R}^k : \sum_{i=1}^k p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i \right\}$$
- ❑ Observe the different domains of the impurity function ι in the Definitions 4 and 5, namely, $[0; 1]^k$ and D . The domains correspond to each other: the set of examples, D , defines via its class portions an element from $[0; 1]^k$ and vice versa.
- ❑ The properties in the definition of the impurity function ι suggest to minimize the external path length of T with respect to D in order to minimize the overall impurity characteristics of T .
- ❑ Within the *DT-construct* algorithm usually a greedy strategy (local optimization) is employed to minimize the overall impurity characteristics of a decision tree T .

Impurity Functions

Impurity Functions Based on the Misclassification Rate

Definition for two classes [impurity function] :

$$\iota_{\text{misclass}}(p_1, p_2) = 1 - \max\{p_1, p_2\} = \begin{cases} p_1 & \text{if } 0 \leq p_1 \leq 0.5 \\ 1 - p_1 & \text{otherwise} \end{cases}$$

Impurity Functions

Impurity Functions Based on the Misclassification Rate

Definition for two classes [impurity function] :

$$\iota_{\text{misclass}}(p_1, p_2) = 1 - \max\{p_1, p_2\} = \begin{cases} p_1 & \text{if } 0 \leq p_1 \leq 0.5 \\ 1 - p_1 & \text{otherwise} \end{cases}$$

$$\iota_{\text{misclass}}(D) = 1 - \max \left\{ \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|}, \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \right\}$$

Impurity Functions

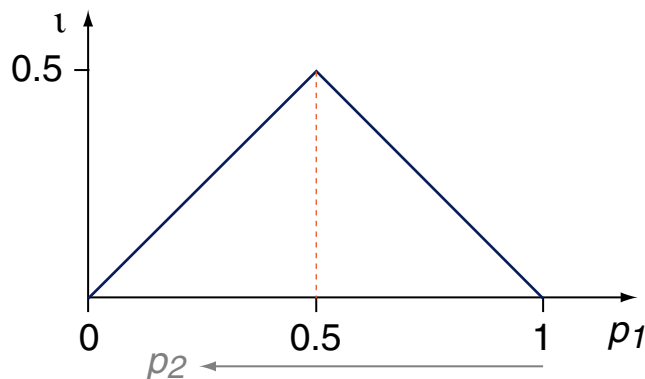
Impurity Functions Based on the Misclassification Rate

Definition for two classes [impurity function] :

$$\iota_{\text{misclass}}(p_1, p_2) = 1 - \max\{p_1, p_2\} = \begin{cases} p_1 & \text{if } 0 \leq p_1 \leq 0.5 \\ 1 - p_1 & \text{otherwise} \end{cases}$$

$$\iota_{\text{misclass}}(D) = 1 - \max \left\{ \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|}, \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \right\}$$

Graph of the function $\iota_{\text{misclass}}(p_1, 1 - p_1)$:



[Graph: Entropy, Gini]

Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Definition for k classes:

$$\iota_{\text{misclass}}(p_1, \dots, p_k) = 1 - \max_{i=1, \dots, k} p_i$$

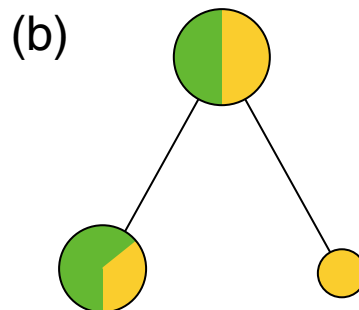
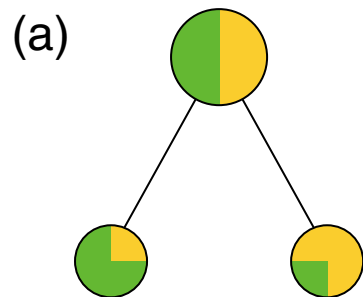
$$\iota_{\text{misclass}}(D) = 1 - \max_{c \in C} \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c\}|}{|D|}$$

Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Problems:

- ❑ $\Delta \ell_{\text{misclass}} = 0$ may hold for all possible splittings.
- ❑ The impurity function that is induced by the misclassification rate underestimates pure nodes, as illustrated in splitting (b):

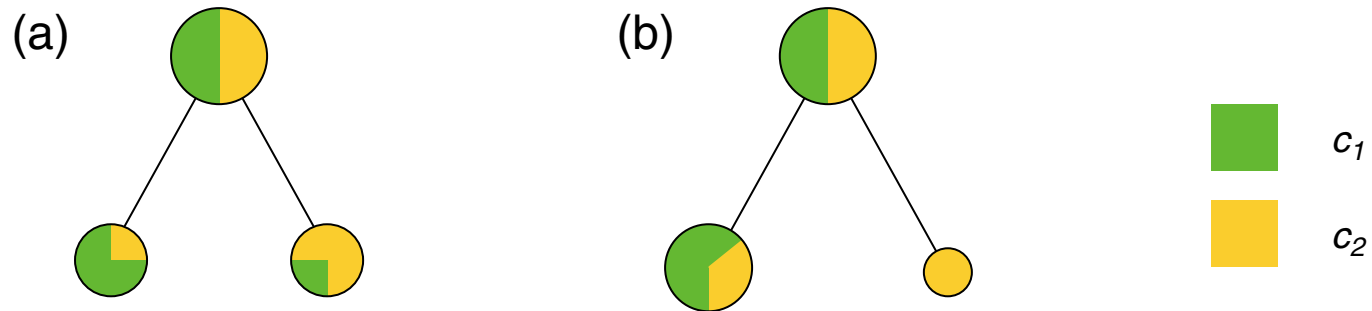


Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Problems:

- $\Delta \iota_{\text{misclass}} = 0$ may hold for all possible splittings.
- The impurity function that is induced by the misclassification rate underestimates pure nodes, as illustrated in splitting (b):



$$\Delta \iota_{\text{misclass}} = \iota_{\text{misclass}}(D) - \left(\frac{|D_1|}{|D|} \cdot \iota_{\text{misclass}}(D_1) + \frac{|D_2|}{|D|} \cdot \iota_{\text{misclass}}(D_2) \right)$$

left splitting: $\Delta \iota_{\text{misclass}} = \frac{1}{2} - \left(\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} \right) = \frac{1}{4}$

right splitting: $\Delta \iota_{\text{misclass}} = \frac{1}{2} - \left(\frac{3}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot 0 \right) = \frac{1}{4}$

Impurity Functions

Definition 7 (Strict Impurity Function)

Let $\iota : [0; 1]^k \rightarrow \mathbb{R}$ be an impurity function and let $\mathbf{p}, \mathbf{p}' \in \Delta^{k-1}$. Then ι is called strict, if it is strictly concave:

$$(c) \rightarrow (c') \quad \iota(\lambda \mathbf{p} + (1 - \lambda) \mathbf{p}') > \lambda \iota(\mathbf{p}) + (1 - \lambda) \iota(\mathbf{p}'), \quad 0 < \lambda < 1, \mathbf{p} \neq \mathbf{p}'$$

Impurity Functions

Definition 7 (Strict Impurity Function)

Let $\iota : [0; 1]^k \rightarrow \mathbb{R}$ be an impurity function and let $\mathbf{p}, \mathbf{p}' \in \Delta^{k-1}$. Then ι is called strict, if it is strictly concave:

$$(c) \rightarrow (c') \quad \iota(\lambda \mathbf{p} + (1 - \lambda) \mathbf{p}') > \lambda \iota(\mathbf{p}) + (1 - \lambda) \iota(\mathbf{p}'), \quad 0 < \lambda < 1, \mathbf{p} \neq \mathbf{p}'$$

Lemma 8

Let ι be a *strict* impurity function and let D_1, \dots, D_s be a partitioning of an example set D , which is induced by a splitting of a feature space X . Then the following inequality holds:

$$\underline{\Delta} \iota(D, \{D_1, \dots, D_s\}) \geq 0$$

The equality is given iff for all $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, s\}$ holds:

$$\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|} = \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D_j : c(\mathbf{x}) = c_i\}|}{|D_j|}$$

Remarks:

- ❑ Equality means that the partitioning of D resembles exactly the class distribution of D .
- ❑ Strict concavity entails Property (c) of the [impurity function](#) definition.
- ❑ For two classes, strict concavity means $\iota(p_1, 1 - p_1) > 0$, where $0 < p_1 < 1$.
- ❑ If ι is a twice differentiable function, strict concavity is equivalent with a negative definite Hessian of ι .
- ❑ With properly chosen coefficients, polynomials of second degree fulfill the properties (a) and (b) of the [impurity function](#) definition as well as strict concavity. See impurity functions based on the [Gini index](#) in this regard.
- ❑ The impurity function that is induced by the misclassification rate is concave, but it is not strictly concave.
- ❑ The proof of Lemma 8 exploits the strict concavity property of ι .

Impurity Functions

Impurity Functions Based on Entropy

Definition 9 (Entropy)

Let A denote an event and let $P(A)$ denote the occurrence probability of A . Then the entropy (self-information, information content) of A is defined as $-\log_2(P(A))$.

Let \mathcal{A} be an experiment with the exclusive outcomes (events) A_1, \dots, A_k . Then the mean information content of \mathcal{A} , denoted as $H(\mathcal{A})$, is called Shannon entropy or entropy of experiment \mathcal{A} and is defined as follows:

$$H(\mathcal{A}) = - \sum_{i=1}^k P(A_i) \cdot \log_2(P(A_i))$$

Remarks:

- ❑ The smaller the occurrence probability of an event, the larger is its entropy. An event that is certain has zero entropy.
- ❑ The Shannon entropy combines the entropies of an experiment's outcomes, using the outcome probabilities as weights.
- ❑ In the entropy definition we stipulate the identity $0 \cdot \log_2(0) = 0$.
- ❑ Related. Entropy encoding methods such as Huffman coding. [\[Wikipedia\]](#)
- ❑ Related. The perplexity of a discrete probability distribution p is defined as $2^{H(p)}$. [\[Wikipedia\]](#)

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition 10 (Conditional Entropy, Information Gain)

Let \mathcal{A} be an experiment with the exclusive outcomes (events) A_1, \dots, A_k , and let \mathcal{B} be another experiment with the outcomes B_1, \dots, B_s . Then the conditional entropy of the combined experiment $(\mathcal{A} \mid \mathcal{B})$ is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{j=1}^s P(B_j) \cdot H(\mathcal{A} \mid B_j),$$

where $H(\mathcal{A} \mid B_j) = - \sum_{i=1}^k P(A_i \mid B_j) \cdot \log_2(P(A_i \mid B_j))$

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition 10 (Conditional Entropy, Information Gain)

Let \mathcal{A} be an experiment with the exclusive outcomes (events) A_1, \dots, A_k , and let \mathcal{B} be another experiment with the outcomes B_1, \dots, B_s . Then the conditional entropy of the combined experiment $(\mathcal{A} \mid \mathcal{B})$ is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{j=1}^s P(B_j) \cdot H(\mathcal{A} \mid B_j),$$

where $H(\mathcal{A} \mid B_j) = - \sum_{i=1}^k P(A_i \mid B_j) \cdot \log_2(P(A_i \mid B_j))$

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition 10 (Conditional Entropy, Information Gain)

Let \mathcal{A} be an experiment with the exclusive outcomes (events) A_1, \dots, A_k , and let \mathcal{B} be another experiment with the outcomes B_1, \dots, B_s . Then the conditional entropy of the combined experiment $(\mathcal{A} \mid \mathcal{B})$ is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{j=1}^s P(B_j) \cdot H(\mathcal{A} \mid B_j),$$

where $H(\mathcal{A} \mid B_j) = - \sum_{i=1}^k P(A_i \mid B_j) \cdot \log_2(P(A_i \mid B_j))$

The information **gain** due to experiment \mathcal{B} is defined as follows:

$$H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A}) - \sum_{j=1}^s P(B_j) \cdot H(\mathcal{A} \mid B_j)$$

Remarks [Bayes for classification] :

- ❑ Information gain is defined as reduction in entropy.
- ❑ In the context of decision trees, experiment \mathcal{A} corresponds to classifying feature vector \mathbf{x} with regard to the target concept. A possible question, whose answer will inform us about which event $A_i \in \mathcal{A}$ occurred, is the following: “Does \mathbf{x} belong to class c_i ?”
Likewise, experiment \mathcal{B} corresponds to evaluating feature B of feature vector \mathbf{x} . A possible question, whose answer will inform us about which event $B_j \in \mathcal{B}$ occurred, is the following: “Does \mathbf{x} have value b_j for feature B ?”
- ❑ Rationale: Typically, the events “target concept class” and “feature value” are statistically dependent. Hence, the entropy of the event “ \mathbf{x} belongs to class c_i ” will become smaller if we learn about the value of some feature of \mathbf{x} (recall that the class of \mathbf{x} is unknown).
We experience an information gain with regard to the outcome of experiment \mathcal{A} , which is rooted in our information about the outcome of experiment \mathcal{B} . Under no circumstances the information gain will be negative; the information gain is zero if the involved events are *conditionally independent*:

$$P(A_i) = P(A_i \mid B_j), \quad i \in \{1, \dots, k\}, \quad j \in \{1, \dots, s\},$$

which leads to a split as specified as the special case in Lemma 8.

Remarks (continued) :

- ❑ Since $H(\mathcal{A})$ is constant, the feature that provides the maximum information gain (= the maximally informative feature) is given by the minimization of $H(\mathcal{A} \mid \mathcal{B})$.
- ❑ The expanded form of $H(\mathcal{A} \mid \mathcal{B})$ reads as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = - \sum_{j=1}^s P(B_j) \cdot \sum_{i=1}^k P(A_i \mid B_j) \cdot \log_2(P(A_i \mid B_j))$$

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for two classes [impurity function] :

$$\iota_{entropy}(p_1, p_2) = -(p_1 \cdot \log_2(p_1) + p_2 \cdot \log_2(p_2))$$

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for two classes [impurity function] :

$$\iota_{entropy}(p_1, p_2) = -(p_1 \cdot \log_2(p_1) + p_2 \cdot \log_2(p_2))$$

$$\iota_{entropy}(D) = - \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} \cdot \log_2 \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} + \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \cdot \log_2 \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \right)$$

Impurity Functions

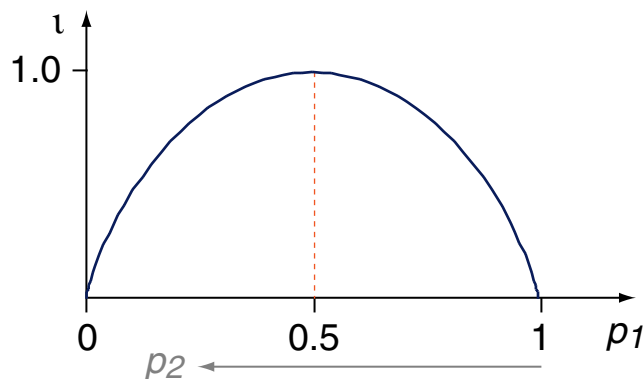
Impurity Functions Based on Entropy (continued)

Definition for two classes [\[impurity function\]](#):

$$\iota_{entropy}(p_1, p_2) = -(p_1 \cdot \log_2(p_1) + p_2 \cdot \log_2(p_2))$$

$$\iota_{entropy}(D) = - \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} \cdot \log_2 \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} + \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \cdot \log_2 \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|} \right)$$

Graph of the function $\iota_{entropy}(p_1, 1 - p_1)$:

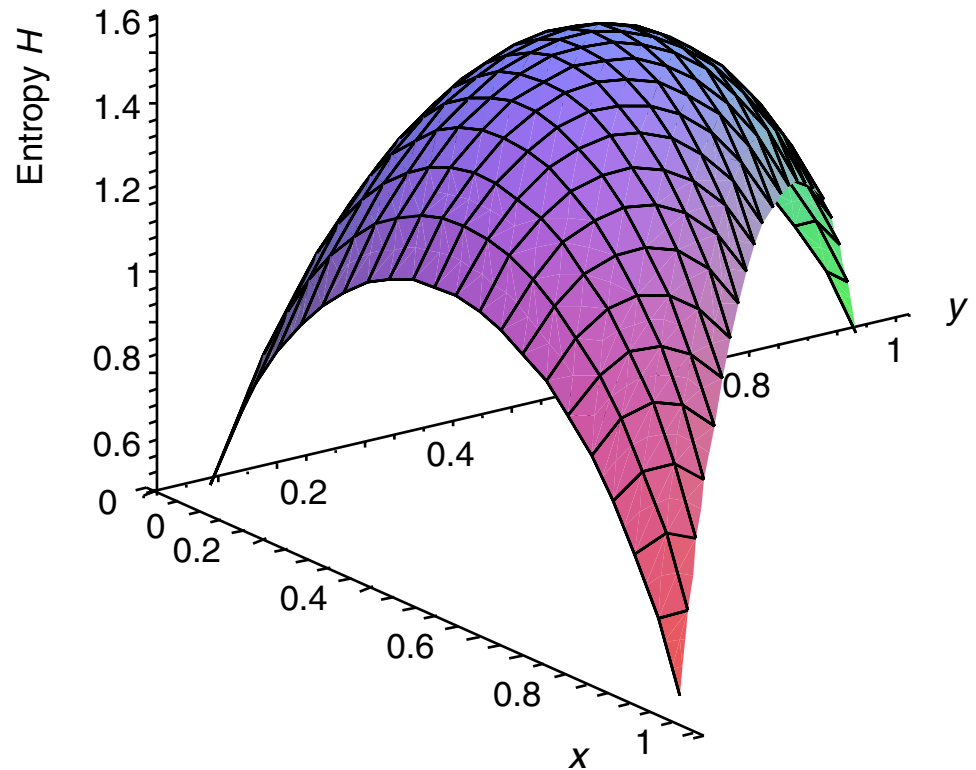
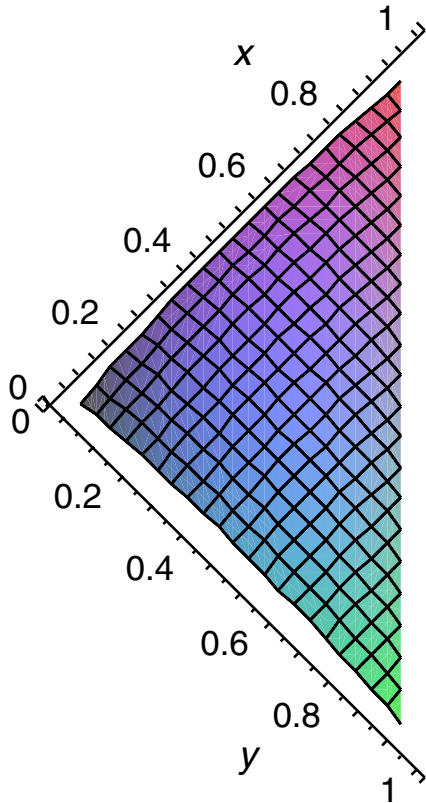


[Graph: [Misclassification, Gini](#)]

Impurity Functions

Impurity Functions Based on Entropy (continued)

Graph of the function $\iota_{\text{entropy}}(p_1, p_2, 1 - p_1 - p_2)$:



Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for k classes:

$$\iota_{entropy}(p_1, \dots, p_k) = - \sum_{i=1}^k p_i \cdot \log_2(p_i)$$

$$\iota_{entropy}(D) = - \sum_{i=1}^k \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|} \cdot \log_2 \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|}$$

Impurity Functions

Impurity Functions Based on Entropy (continued)

$\Delta \iota_{entropy}$ corresponds to the information gain $H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B})$:

$$\underbrace{\Delta \iota_{entropy} = \iota_{entropy}(D)}_{H(\mathcal{A})} - \underbrace{\sum_{j=1}^s \frac{|D_j|}{|D|} \cdot \iota_{entropy}(D_j)}_{H(\mathcal{A} \mid \mathcal{B})}$$

Impurity Functions

Impurity Functions Based on Entropy (continued)

$\Delta \iota_{\text{entropy}}$ corresponds to the information gain $H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B})$:

$$\Delta \iota_{\text{entropy}} = \underbrace{\iota_{\text{entropy}}(D)}_{H(\mathcal{A})} - \underbrace{\sum_{j=1}^s \frac{|D_j|}{|D|} \cdot \iota_{\text{entropy}}(D_j)}_{H(\mathcal{A} \mid \mathcal{B})}$$

Derivation:

- A_i , $i = 1, \dots, k$, denotes the event that $\mathbf{x} \in X(t)$ belongs to class c_i .
The experiment \mathcal{A} corresponds to the classification $c : X(t) \rightarrow C$.
- B_j , $j = 1, \dots, s$, denotes the event that $\mathbf{x} \in X(t)$ has value b_j for feature B .
The experiment \mathcal{B} corresponds to evaluating feature B and entails the following splitting:

$$X(t) = X(t_1) \cup \dots \cup X(t_s) = \{\mathbf{x} \in X(t) : \mathbf{x}|_B = b_1\} \cup \dots \cup \{\mathbf{x} \in X(t) : \mathbf{x}|_B = b_s\}$$

- $\iota_{\text{entropy}}(D) = \iota_{\text{entropy}}(P(A_1), \dots, P(A_k)) = -\sum_{i=1}^k P(A_i) \cdot \log_2(P(A_i)) = H(\mathcal{A})$
- $\frac{|D_j|}{|D|} \cdot \iota_{\text{entropy}}(D_j) = P(B_j) \cdot \iota_{\text{entropy}}(P(A_1 \mid B_j), \dots, P(A_k \mid B_j))$, $j = 1, \dots, s$
- $P(A_i), P(B_j), P(A_i \mid B_j)$ are estimated as relative frequencies based on D .

Impurity Functions

Impurity Functions Based on the Gini Index

Definition for two classes [\[impurity function\]](#) :

$$\iota_{Gini}(p_1, p_2) = 1 - (p_1^2 + p_2^2) = 2 \cdot p_1 \cdot p_2$$

Impurity Functions

Impurity Functions Based on the Gini Index

Definition for two classes [\[impurity function\]](#):

$$\iota_{Gini}(p_1, p_2) = 1 - (p_1^2 + p_2^2) = 2 \cdot p_1 \cdot p_2$$

$$\iota_{Gini}(D) = 2 \cdot \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} \cdot \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|}$$

Impurity Functions

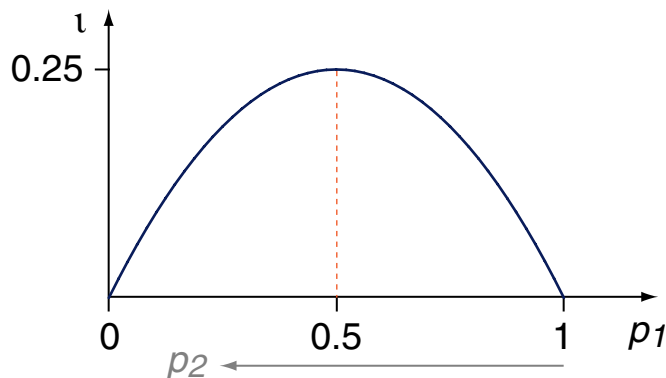
Impurity Functions Based on the Gini Index

Definition for two classes [\[impurity function\]](#):

$$\iota_{Gini}(p_1, p_2) = 1 - (p_1^2 + p_2^2) = 2 \cdot p_1 \cdot p_2$$

$$\iota_{Gini}(D) = 2 \cdot \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_1\}|}{|D|} \cdot \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_2\}|}{|D|}$$

Graph of the function $\iota_{Gini}(p_1, 1 - p_1)$:



[Graph: [Misclassification](#), [Entropy](#)]

Impurity Functions

Impurity Functions Based on the Gini Index (continued)

Definition for k classes:

$$\iota_{Gini}(p_1, \dots, p_k) = 1 - \sum_{i=1}^k (p_i)^2$$

$$\begin{aligned}\iota_{Gini}(D) &= \left(\sum_{i=1}^k \frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|} \right)^2 - \sum_{i=1}^k \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|} \right)^2 \\ &= 1 - \sum_{i=1}^k \left(\frac{|\{(\mathbf{x}, c(\mathbf{x})) \in D : c(\mathbf{x}) = c_i\}|}{|D|} \right)^2\end{aligned}$$