Chapter ML:VI

VI. 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. Impurity reduction of $D(t)$.

Impurity Functions

Splitting (continued)

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 $D(t)$.
   $D(t)$ is not split if $|D(t)|$ is below a threshold.

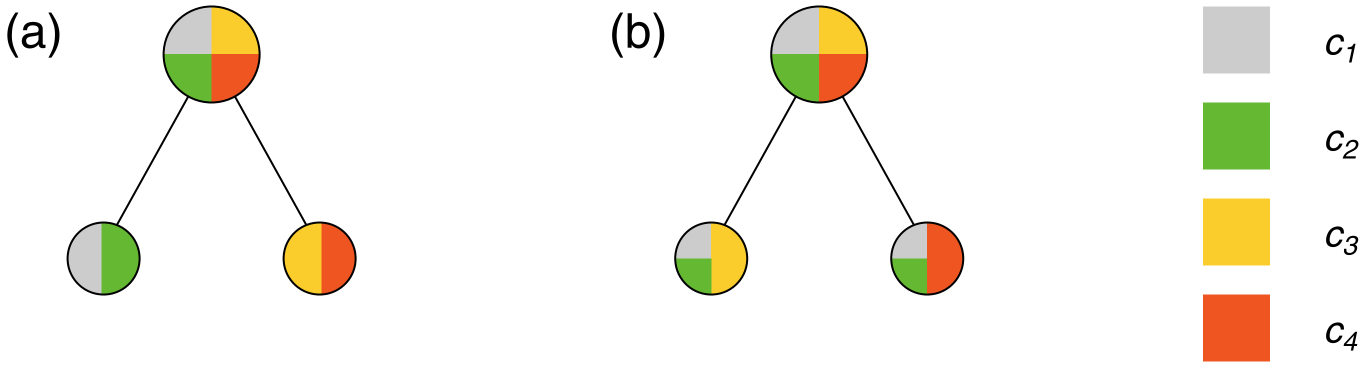
2. Purity of $D(t)$.
   $D(t)$ is not split if all examples in $D(t)$ are members of the same class.

3. Impurity reduction of $D(t)$.
   $D(t)$ is not split if its impurity reduction, $\Delta\iota$, is below a threshold.

Impurity Functions

Splitting (continued)

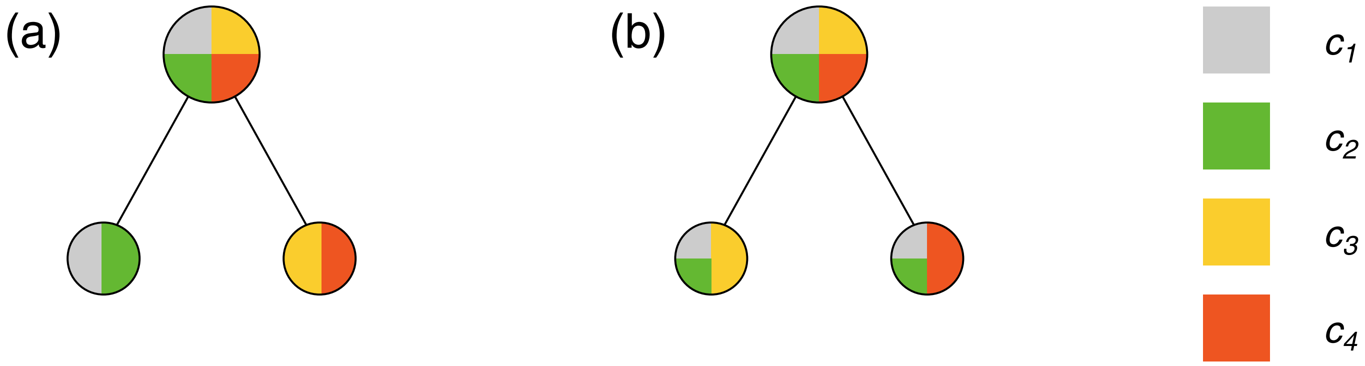
Let X be a multiset of feature vectors, $D \subseteq X$ a multiset of examples, and $C = \{c_1, c_2, c_3, c_4\}$ a set of classes. Distribution of D for two splittings of X :



Impurity Functions

Splitting (continued)

Let X be a multiset of feature vectors, $D \subseteq X$ a multiset of examples, and $C = \{c_1, c_2, c_3, c_4\}$ a set of classes. Distribution of D for two 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 $k \in \mathbf{N}$. An impurity function $\iota : [0;1]^k \to \mathbf{R}$ is a function defined on the standard $k{-}1$-simplex, denoted Δ^{k-1}, for which the following properties hold:

(a) ι becomes minimum at points $(1,0,\ldots,0), (0,1,\ldots,0), \ldots, (0,\ldots,0,1)$.

(b) ι is symmetric with regard to its arguments, $p_1,\ldots,p_k$.

(c) ι becomes maximum at point $(1/k,\ldots,1/k)$.

Impurity Functions

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

Let X be a multiset of feature vectors, $C = \{c_1, \ldots, c_k\}$ a set of classes and $D \subseteq X \times C$ a multiset of examples. Moreover, let $\iota : [0;1]^k \to \mathbf{R}$ be an impurity function. Then, the impurity of D, denoted as $\iota(D)$, is defined as follows:

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

Impurity Functions

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

Let X be a multiset of feature vectors, $C = \{c_1, \ldots, c_k\}$ a set of classes and $D \subseteq X \times C$ a multiset of examples. Moreover, let $\iota : [0;1]^k \rightarrow \mathbf{R}$ be an impurity function. Then, the impurity of D, denoted as $\iota(D)$, is defined as follows:

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

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

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

$$\Delta\iota\big(D,\ \{D_1, \ldots, D_m\}\big) \ = \ \iota(D) - \sum_{l=1}^{m} \frac{|D_l|}{|D|} \cdot \iota(D_l)$$

Remarks:

- Δ^{k-1} denotes the standard $k-1$-simplex, which contains all k-tuples with non-negative elements that sum to 1:

$$\Delta^{k-1} = \Big\{(p_1, \ldots, p_k) \in \mathbf{R}^k : \textstyle\sum_{i=1}^{k} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\Big\}$$

- Observe the different domains of the impurity function ι in the definitions for ι and $\iota(D)$, namely, $[0;1]^k$ and D. The domains correspond to each other: the set of examples, D, defines via its class ratios 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] :

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

Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Definition for two classes [impurity function] :

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

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

Impurity Functions

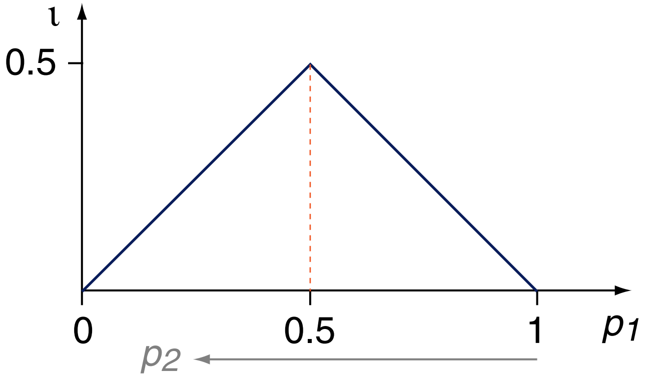
Impurity Functions Based on the Misclassification Rate (continued)

Definition for two classes [impurity function] :

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

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

Graph of the function $\iota_{misclass}(p_1, 1 - p_1)$, i.e., for two classes:



[Graphs: misclassification, entropy, Gini]

Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Definition for k classes:

$$\iota_{\textit{misclass}}(p_1, \ldots, p_k) = 1 - \max_{i=1,\ldots,k} \; p_i$$

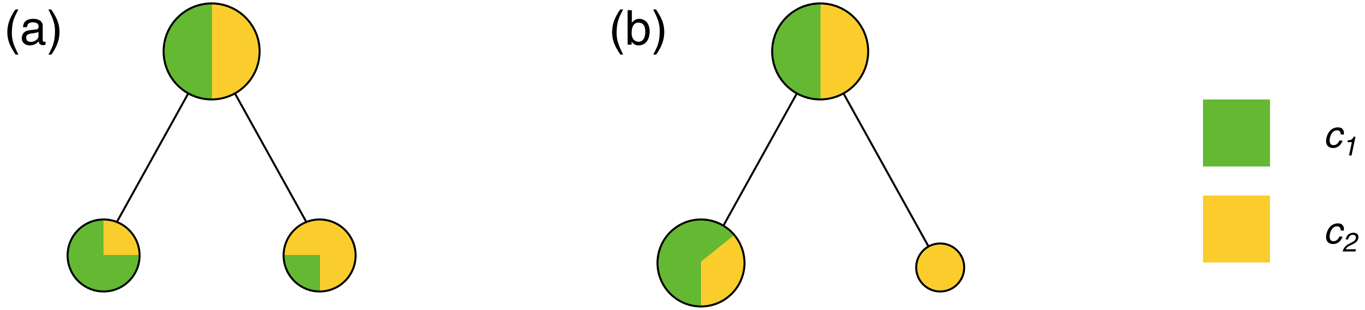
$$\iota_{\textit{misclass}}(D) = 1 - \max_{c \in C} \; \frac{|\{(\mathbf{x}, c) \in D\}|}{|D|}$$

Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Problems:

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

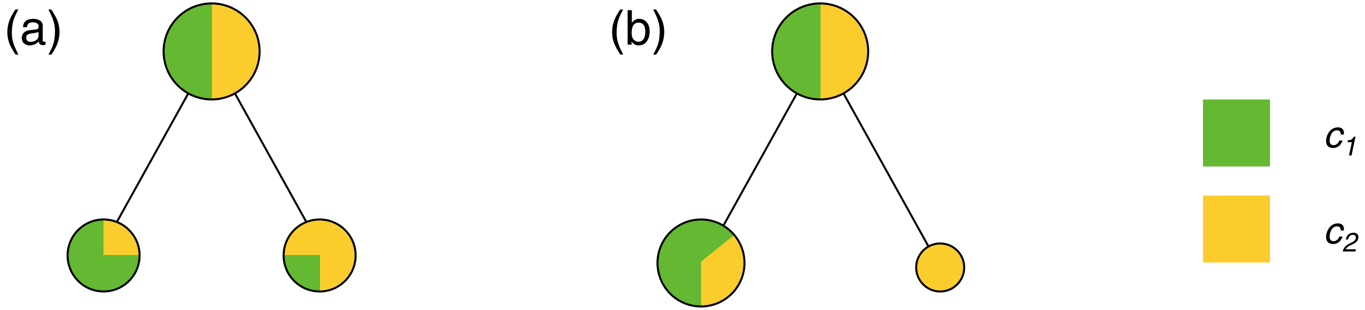


Impurity Functions

Impurity Functions Based on the Misclassification Rate (continued)

Problems:

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



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

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

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

Impurity Functions

Definition 7 (Strict Impurity Function)

Let $\iota : [0;1]^k \to \mathbf{R}$ 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) → (c') $\quad \iota\big(\lambda\mathbf{p} + (1-\lambda)\mathbf{p}' \big) \; > \; \lambda\, \iota(\mathbf{p}) \; + \; (1-\lambda)\, \iota(\mathbf{p}'), \quad 0 < \lambda < 1,\; \mathbf{p} \neq \mathbf{p}'$

Impurity Functions

Definition 7 (Strict Impurity Function)

Let $\iota : [0;1]^k \rightarrow \mathbf{R}$ 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) → (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 $D_1, \ldots, D_m$ be a splitting of an example set D, which is induced by a splitting of X. Then the following inequality holds:

$$\Delta\iota(D, \{D_1, \ldots, D_m\}) \geq 0$$

The equality is given iff for all $i \in \{1, \ldots, k\}$ and $l \in \{1, \ldots, m\}$ holds:

$$\frac{|\{(\mathbf{x}, c_i) \in D\}|}{|D|} = \frac{|\{(\mathbf{x}, c_i) \in D_l\}|}{|D_{|}}$$

Remarks:

- Equality means that the splitting 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) $A_1, \ldots, A_k$. 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 all outcomes of an experiment, 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) $A_1, \ldots, A_k$, and let $\mathcal{B}$ be another experiment with the exclusive outcomes (events) $B_1, \ldots, B_m$. Then the conditional entropy of the conditional experiment $\mathcal{A} \mid \mathcal{B}$, i.e., "the entropy of $\mathcal{A}$ if the outcome of $\mathcal{B}$ is known", is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{l=1}^{m} P(B_l) \cdot H(\mathcal{A} \mid B_l),$$

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

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) $A_1, \ldots, A_k$, and let $\mathcal{B}$ be another experiment with the exclusive outcomes (events) $B_1, \ldots, B_m$. Then the conditional entropy of the conditional experiment $\mathcal{A} \mid \mathcal{B}$, i.e., "the entropy of $\mathcal{A}$ if the outcome of $\mathcal{B}$ is known", is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{l=1}^{m} P(B_l) \cdot H(\mathcal{A} \mid B_l),$$

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

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) $A_1, \ldots, A_k$, and let $\mathcal{B}$ be another experiment with the exclusive outcomes (events) $B_1, \ldots, B_m$. Then the conditional entropy of the conditional experiment $\mathcal{A} \mid \mathcal{B}$, i.e., "the entropy of $\mathcal{A}$ if the outcome of $\mathcal{B}$ is known", is defined as follows:

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{l=1}^{m} P(B_l) \cdot H(\mathcal{A} \mid B_l),$$

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

The information gain owed to experiment $\mathcal{B}$ is defined as follows:

$$H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A}) - \sum_{l=1}^{m} P(B_l) \cdot H(\mathcal{A} \mid B_l)$$

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 question, whose answer will inform us about which of the events in $A \in \mathcal{A}$ occurred, is the following:
  - "Does $\mathbf{x}$ belong to class c ?" or
  - "$C{=}c$?" (random variable C has realization c)

  Likewise, experiment $\mathcal{B}$ corresponds to evaluating feature j of feature vector $\mathbf{x}$. A question, whose answer will inform us about which of the events in $B \in \mathcal{B}$ occurred, is the following:
  - "Does $\mathbf{x}$ have value a for feature j ?" or
  - "$X_j{=}a$?" (random variable X_j has realization a)
- Rationale: Typically, the events "$\mathbf{x}$ belongs to class c" and "$\mathbf{x}$ has value a for feature j" are statistically dependent. Hence, the entropy of the event "$\mathbf{x}$ belongs to class c" will become smaller if we learn about the value of feature j 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*:

$$P(A_i) = P(A_i \mid B_l), \quad i \in \{1, \ldots, k\},\ l \in \{1, \ldots, m\},$$

  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:

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

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for two classes [impurity function] :

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

Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for two classes [impurity function] :

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

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

Impurity Functions

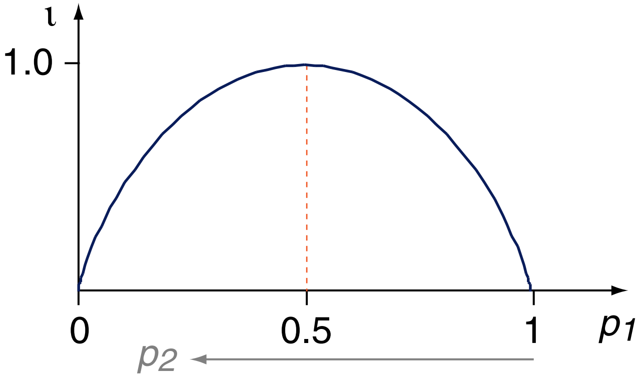
Impurity Functions Based on Entropy (continued)

Definition for two classes [impurity function] :

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

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

Graph of the function $\iota_{entropy}(p_1, 1 - p_1)$, i.e., for two classes:

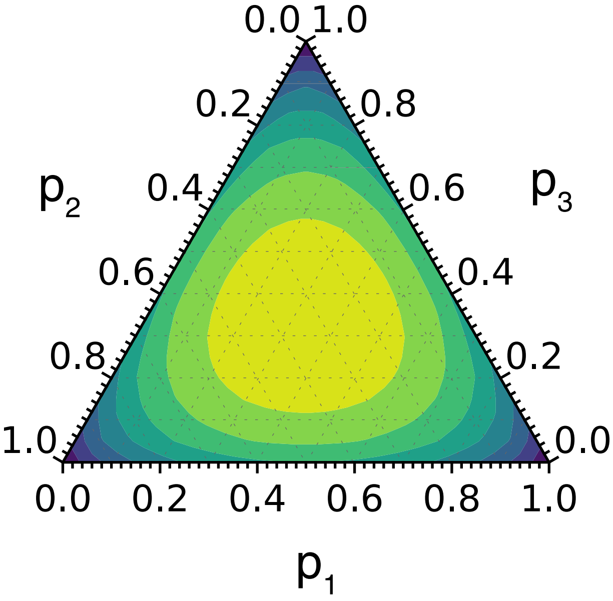


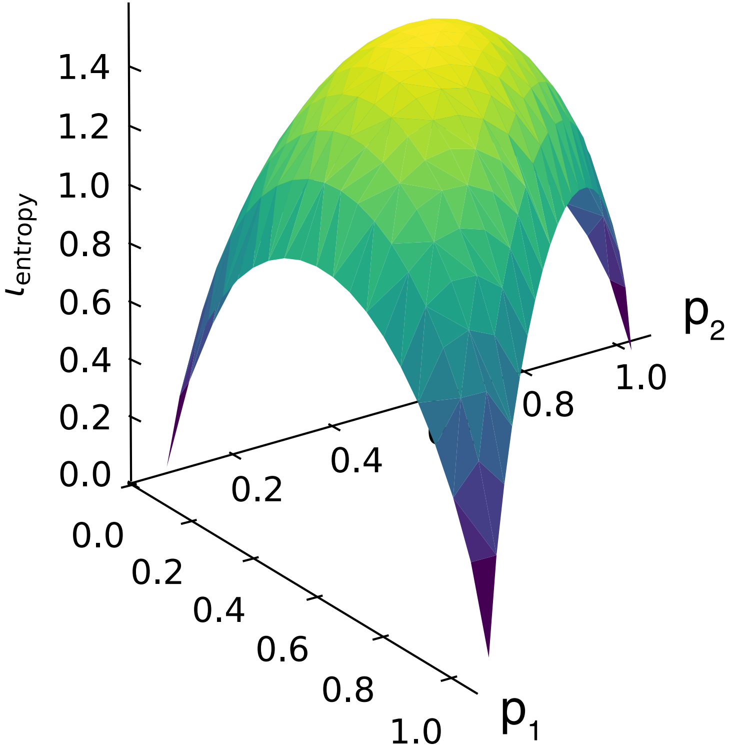
[graphs: misclassification, entropy, Gini]

Impurity Functions

Impurity Functions Based on Entropy (continued)

Graph of the function $\iota_{\mathit{entropy}}(p_1, p_2, 1-p_1-p_2)$, i.e., for three classes:





Impurity Functions

Impurity Functions Based on Entropy (continued)

Definition for k classes:

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

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

Impurity Functions

Impurity Functions Based on Entropy (continued)

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

$$\Delta\iota_{\mathit{entropy}} = \underbrace{\iota_{\mathit{entropy}}(D)}_{H(\mathcal{A})} \quad - \quad \underbrace{\sum_{l=1}^{m} \frac{|D_l|}{|D|} \cdot \iota_{\mathit{entropy}}(D_l)}_{H(\mathcal{A}\,|\,\mathcal{B})}$$

Impurity Functions

Impurity Functions Based on Entropy (continued)

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

$$\Delta\iota_{\mathit{entropy}} = \underbrace{\iota_{\mathit{entropy}}(D)}_{H(\mathcal{A})} \quad - \quad \underbrace{\sum_{l=1}^{m} \frac{|D_l|}{|D|} \cdot \iota_{\mathit{entropy}}(D_l)}_{H(\mathcal{A} \mid \mathcal{B})}$$

Mapping:

- $A_i,\ i = 1, \ldots, k$, denotes the event that $\mathbf{x} \in X(t)$ belongs to class c_i.
  The experiment $\mathcal{A}$ corresponds to the classification $c : X(t) \to C$.

- $B_l,\ l = 1, \ldots, m$, denotes the event that $\mathbf{x} \in X(t)$ has value a_l for feature j.
  The experiment $\mathcal{B}$ corresponds to evaluating feature A and entails the following splitting:

  $$X(t) = X(t_1) \cup \ldots \cup X(t_m) = \{\mathbf{x} \in X(t) : \mathbf{x}|_j = a_1\} \ \cup \ldots \cup \ \{\mathbf{x} \in X(t) : \mathbf{x}|_j = a_m\}$$

- $\iota_{\mathit{entropy}}(D) \ = \ \iota_{\mathit{entropy}}(P(A_1), \ldots, P(A_k)) \ = \ -\sum_{i=1}^{k} P(A_i) \cdot \log_2(P(A_i)) \ = \ H(\mathcal{A})$

- $\frac{|D_l|}{|D|} \cdot \iota_{\mathit{entropy}}(D_l) \ = \ P(B_l) \cdot \iota_{\mathit{entropy}}(P(A_1 \mid B_l), \ldots, P(A_k \mid B_l)) \ = \ H(\mathcal{A} \mid B_l),\ l = 1, \ldots, m$

- $P(A_i),\ P(B_l),\ P(A_i \mid B_l)$ 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 (continued)

Definition for two classes [impurity function] :

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

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

Impurity Functions

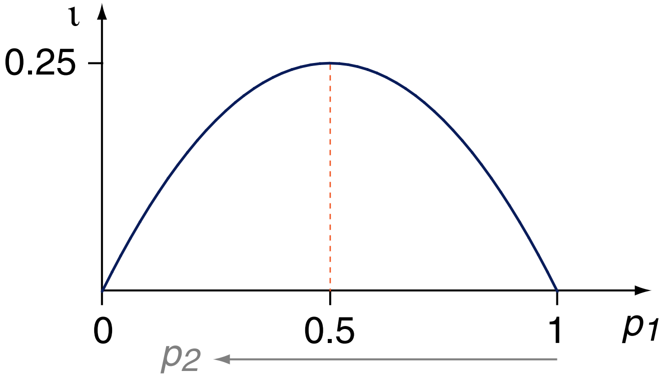
Impurity Functions Based on the Gini Index (continued)

Definition for two classes [impurity function] :

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

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

Graph of the function $\iota_{Gini}(p_1, 1 - p_1)$, i.e., for two classes:



[Graphs: misclassification, entropy, Gini]

Impurity Functions

Impurity Functions Based on the Gini Index (continued)

Definition for k classes:

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

$$\iota_{\mathit{Gini}}(D) = \left(\sum_{i=1}^{k} \frac{|\{(\mathbf{x}, c_i) \in D\}|}{|D|} \right)^2 - \sum_{i=1}^{k} \left(\frac{|\{(\mathbf{x}, c_i) \in D\}|}{|D|} \right)^2$$

$$= 1 - \sum_{i=1}^{k} \left(\frac{|\{(\mathbf{x}, c_i) \in D\}|}{|D|} \right)^2$$