1 Approximation

For every natural number $k \geq 1$, we can define a CI E_k as follows. Let E_k be the CI with exactly 2^k examples $\{e_1, \ldots, e_{2^k}\}$ on k binary features $\{f_1, \ldots, f_k\}$: there is exactly one example for every of the 2^k feature assignments. An example $e \in E_k$ is a positive example if $|\{f \in feat(E_k) \mid f(e) = 1\}|$ is even and negative otherwise. The set S_k denotes $\{f_1, \ldots, f_k\}$.

1.1 For Size

Let D_k be the set of all the examples $e \in E_k$ such that $f_i(e) = 1$ for every $i \in [k-2]$ and denote by $\overline{D_k}$ the set $E_k \setminus D_k$. Now we are ready to define a new feature f^* as follows: $f^*(e) = 1$ if either e is a positive example or $e \in D_k$ and $f^*(e) = 0$ otherwise. See Figure 1 for a visual representation of E_3 and its decomposition in D_3 and $\overline{D_3}$. For simplicity, the set S_k^* denotes $\{f_1, \ldots, f_k, f^*\}$.

f_1	f_2	f_3	f^*		
0	0	0	1	+	
0	0	1	0	−	$\overline{D_3}$
0	1	0	0	−	
0	1	1	1	+	
1	0	0	1	−	
1	0	1	1	+	
1	1	0	1	+	D_3
1	1	1	1	−	

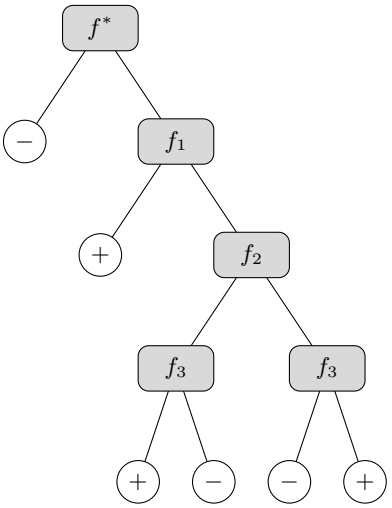

Figure 1 The CI E_3 and the DT T_3.

▶ **Lemma 1.** *For every integer $k \geq 1$, the set of features S_k is the only minimal support set in S_k^* for E_k.*

Proof. First we show that S_k is a support set: let $e^- \in E_k^-$ and $e^+ \in E_k^+$, by construction there is one feature $f \in S_k$ where $f(e^+) \neq f(e^-)$. Now it is time to show that, for any $i \in [k]$, the set $S_k^i = \{f_1, \ldots, \overline{f_i}, \ldots, f_k, f^*\} = S_k^* \setminus \{f_i\}$ is not a support set for E_k. For $i \in [k-2]$, let e_i^- and e_i^+ be the negative and the positive examples such that $f(e_i^-) = f(e_i^+) = 1$ if k is odd (= 0 if k is even) for every $f \in S_k^i$: e_i^- and e_i^+ can not be distinguished by a feature in S_k^i and so S_k^i is not a support set (note there is one such pair E_k). For $i \in \{k-1, k\}$, let e_i^- and e_i^+ be the negative and the positive examples in D_k such that $f(e_i^-) = f(e_i^+)$ for every $f \in S_k^i$: e_i^- and e_i^+ can not be distinguished by a feature in S_k^i and so S_k^i is not a support set (note there are two of such pairs in E_k). ◀

▶ **Lemma 2.** *For every integer $k \geq 1$, a reduced DT T with features in S_k is a DT for E_k if and only if T is a complete DT of height $k+1$. In particular such DT has $2^{k+2} - 1$ nodes ($2^{k+1} - 1$ of those are inner nodes).*

Proof. In this proof we assume that a leaf is either positive or negative depending on the parity of the number of right arcs present in the unique path from the root to that leaf. We start with the forward direction: let T be a reduced DT that is not a complete DT of height $k+1$. Let P be a path of T from the root to a leaf ℓ of length at most k: at most $k-1$ features appear in P and so there exists a feature $f_i \in S_k$ that does not appear in P. Since by Lemma 1 S_k^i is not a support set for E_k, there exit a negative example e^- and a positive example e^+ that can not be distinguished by S_k^i, this means that $\{e^-, e^+\} \subseteq E_T(\ell)$ and so T is not a DT for E_k.

In order to prove the backward direction, we assume that T is a reduced and complete DT of height $k+1$ with features in S_k. Let P be a path of T from the root to a leaf ℓ of length $k+1$. Since T is reduced, every feature of S_k appears exactly once in P. Since by Lemma 1 S_k is a support set, there is only one example e_ℓ that ends ℓ, that is $e_\ell \in E_T(\ell)$. From this proof, it follows that every reduced DT T with features in $\S_k$ for E_k has $2^{k+2}-1$ nodes ($2^{k+1}-1$ of those are inner nodes). ◀

A more general description of T_k can be found commented

For every integer $k \geq 1$, let us describe a DT T_k as follows. The root r of T_k has feature f^*. The left child of r is a negative leaf and the right child v_1 has feature f_1. For every $i \in [k-2]$, the left child of v_i is a positive leaf and the right child v_{i+1} has feature f_{i+1}. Finally v_k and v'_k are respectively the left and right child of v_{k-1}, both having feature f_k. The children of v_k and v'_k are leaves that are either positive or negative depending on the parity of the number of right arcs present in the unique path from the root to that leaf. In particular note that T_k has $2k+5$ nodes ($k+2$ of those are inner nodes). See Figure 1 for a visual representation of T_3.

▶ **Lemma 3.** *For every integer $k \geq 1$, T_k is a DT for E_k.*

Proof. By construction, r and its feature f^* send every negative example to its left child c_ℓ, which is a negative leaf, except for the two negative examples in D_k, that is, if $\{e_1^-, e_2^-\} = E_k^- \cap D_k$ then $E_{T_k}(c_\ell) = E_k^- \setminus \{e_1^-, e_2^-\}$ and $E_{T_k}(v_1) = E_k^+ \cup \{e_1^-, e_2^-\}$.

Let e be an example in D_k; by construction, for every $i \in [k-2]$ if $e \in E_{T_k}(v_i)$ then $e \in E_{T_k}(v_{i+1})$ and by induction we obtain that $e \in E_{T_k}(v_{k-1})$. Let e be an example in $\overline{D_k}$ and $j \in [k-2]$ be the minimum integer such that $f_j(e) = 0$. This means that $e \notin E_{T_k}(v_{j+1})$ and e is classified by the left child of the node v_j. We have just proved that $D_k = E_{T_K}(v_{k-1})$ and that T_k classifies $\overline{D_k}$. Now it is straightforward to show that the subtree of T_k rooted at v_{k-1} classifies D_k. ◀

Let E be a CI and S be a support set for E. We denote with $dts(E,S)$ the minimum size of a DT for E that uses exactly all the features in S. By Lemma 2, we have that $dts(E_k, S_k) = 2^{k+2}-1$. Moreover, since by Lemma 3 T_k is a DT for E_k, we have that $dts(E_k, S_k^*) \leq |T_k| = 2k+5$. In conclusion we have that

$$\frac{dts(E_k, S_k)}{dts(E_k, S_k^*)} \geq \frac{2^{k+2}+1}{2k+5} \approx \frac{2^{k+1}}{k}.$$

1.2 For Height

For every integer $k \geq 1$, let us describe a DT T^k as follows. The tree T^k has v_1 as root. For every $i \in [k]$, the node v_i has v_{2i} and v_{2i+1} as left child and right child respectively. Moreover, if $i \leq k$, the node v_i has feature f_i and is a leaf otherwise (negative if i is even and positive if i is odd). Note that T^k has height $log(k)+1$.

Let F_k be the set of all the examples in E_k that are classified by T^k and denote by $\overline{F_k} = E_k \setminus F_k$. Now we are ready to define a new feature f' follows: $f'(e) = 0$ if $e \in F_k$ and $f'(e) = 1$ otherwise. We also introduce another new feature f'' as follows: $f''(e) = 0$ if either $e \in F_k$ or e is a negative example and $f''(e) = 1$ otherwise. For simplicity the set S'_k denotes $\{f_1, \ldots, f_k, f', f''\}$.

For every integer $k \geq 1$, let us describe a DT T^k_* as follows. The root of r has feature f' and its left branch is the DT T^k. The right child of r is a node u with feature f''. The left/right child of u is a negative/positive leaf. In particular note that T^k_* has height $log(k)+2$. See Figure 2 for a visual representation of E_3, its decomposition in F_3 and $\overline{F_3}$ and the DTs T^3 and T^3_*.

f_1	f_2	f_3	f'	f''		
0	0	1	0	0	$-$	
0	1	1	0	0	$+$	F_3
1	0	0	0	0	$-$	
1	0	1	0	0	$+$	
0	0	0	1	1	$+$	
0	1	0	1	0	$-$	
1	1	0	1	1	$+$	$\overline{F_3}$
1	1	1	1	0	$-$	

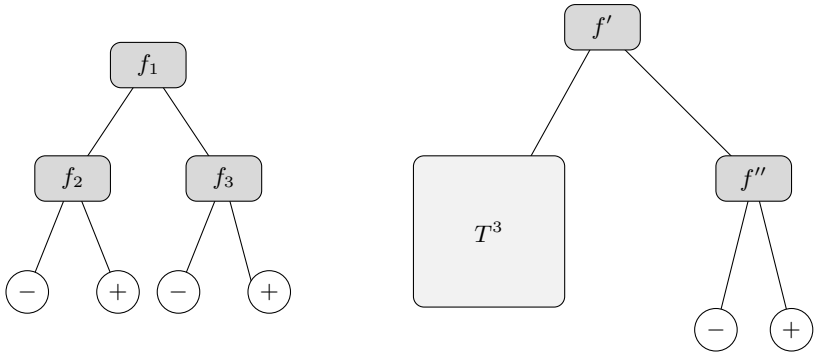

Figure 2 The CI E_3, the DT T^3 and the DT T^3_*

▶ **Lemma 4.** *For every integer $k \geq 1$, the set of features S_k is the only minimal support set in S'_k for E_k.*

Proof. By Lemma 1, the set S_k is a support set. Now it is time to show that, for every $i \in [k]$, the set $S^i_k = \{f_1, \ldots, \overline{f_i}, \ldots, f_k, f', f''\} = S'_k \setminus \{f_i\}$ is not a support set for E_k.

Let e_i be an example of F_k be such that, in T^k, e_i ends in a leaf ℓ_i which has v_i as an ancestor. Let P_i be the path from ℓ_i to v_i. Since T^k has height $log(k) < k$ there is at least a feature in S_k that does not appear in the path from r to ℓ_i: therefore such example e_i always exists.

Let Q_i be the path from the child of v_i that is not an ancestor of ℓ_i to a leaf ℓ'_i such that for every left arc uw of Q_i then $feat(u)(e_i) = 0$ and for every right arc uw of Q_i then $feat(u)(e_i) = 1$. Let e'_i be the example in E_k such that $f_j(e'_i) = f_j(e_i)$ for every $j \in [k] \setminus \{i\}$ and $f_i(e'_i) = 1 - f_i(e_i)$. Now it is crucial to note that $e'_i \in F_k$: indeed e'_i ends in the ℓ'_i (and its sign is different from the one of ℓ_i). Since e_i and e'_i are both examples of F_k, by construction e_i and e'_i can not be distinguished by either f' or f''.

This means that e_i and e'_i can not be distinguished by a feature in S^i_k and so S^i_k is not a support set. ◀

▶ **Lemma 5.** *For every integer $k \geq 1$, T^k_* is a DT for E_k.*

Proof. By construction, r and its feature f' send every example of F_k to its left child and every other example, that is $\overline{F_k}$, to the right child. By definition, the set F_k is classified by T^k and, by construction, the subtree of T^k_* rooted at the right child u of r classifies $\overline{F_k}$. Therefore, T^k_* classifies $F_k \cup \overline{F_k} = E_k$. ◀

Let E be a CI and S be a support set for E. We denote with $dth(E, S)$ the minimum height of a DT for E that uses exactly all the features in S. By Lemma 2, we have that

$dth(E_k, S_k) = k + 1$. Moreover, since by Lemma 5 T_*^k is a DT for E_k, we have that $dth(E_k, S'_k) \leq |T_*^k| = log(k) + 2$. In conclusion we have that

$$\frac{dth(E_k, S_k)}{dth(E_k, S'_k)} \geq \frac{k+1}{log(k)+2} \approx \frac{k}{log(k)}.$$