Machine Learning I

Jannik Irmai, David Stein, Bjoern Andres

Machine Learning for Computer Vision
TU Dresden

https://mlcv.cs.tu-dresden.de/courses/24-winter/ml1/

Winter Term 2024/2025

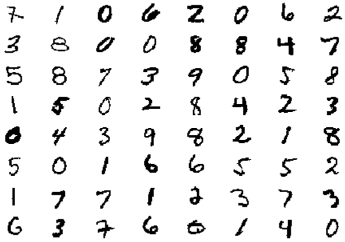
Contents. This part of the course is about the supervised learning of binary decision trees.

- We introduce the problem as a specialization of supervised learning by defining labeled data, a family of functions, a regularizer and a loss function.
- We prove that the problem is NP-hard, by relating it to the exact cover by 3-sets problem.

We consider labeled data with **binary features**. More specifically, we consider some finite, non-empty set V, called the set of features, and labeled data $T = (S, X, x, y)$ such that $X = \{0,1\}^V$. Hence:

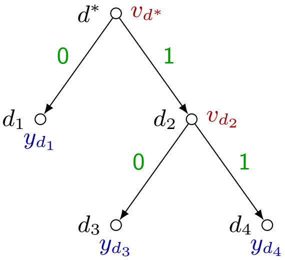
$$x\colon S \to \{0,1\}^V$$
$$y\colon S \to \{0,1\}$$

Example.

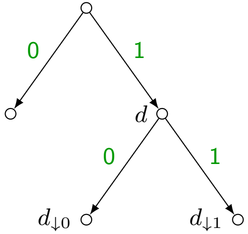


Definition. A tuple $(V, Y, D, D', d^*, E, \delta, v, y)$ is called a V-variate Y-valued **binary decision tree** (BDT) if and only if the following conditions hold:

1. $V \neq \emptyset$ is finite (called the set of **variables**)
2. $Y \neq \emptyset$ is finite (called the set of **values**)
3. $(D \cup D', E)$ is a finite, non-empty directed binary tree with root d^*
4. every $d \in D'$ is a leaf
5. $\delta \colon E \to \{0, 1\}$
6. every $d \in D$ has precisely two out-edges, $e = (d, d'), e' = (d, d'')$, such that $\delta(e) = 0$ and $\delta(e') = 1$
7. $v \colon D \to V$
8. $y \colon D' \to Y$



Definition. For any BDT $(V, Y, D, D', d^*, E, \delta, v, y)$, any $d \in D$ and any $j \in \{0, 1\}$, let $d_{\downarrow j} \in D \cup D'$ the unique node such that $e = (d, d_{\downarrow j}) \in E$ and $\delta(e) = j$.



Definition. For any BDT $\theta = (V, Y, D, D', d^*, E, \delta, v, y)$ and any $d \in D \cup D'$, the tuple $\theta[d] = (V, Y, D_2, D'_2, d, E', \delta', v', y')$ is called the **binary decision subtree** of θ rooted at d iff

- $(D_2 \cup D'_2, E')$ is the subtree of $(D \cup D', E)$ rooted at d
- δ', v' and y' are the restrictions of δ, v and y to the subsets D_2, D'_2 and E'

Lemma. For any BDT $\theta = (V, Y, D, D', d^*, E, \delta, v, y)$ and any $d \in D \cup D'$, the binary decision subtree $\theta[d]$ is itself a V-variate Y-valued BDT.

Definition. For any BDT $\theta = (V, Y, D, D', d^*, E, \delta, v, y)$, the function defined by θ is the $f_\theta : \{0,1\}^V \to Y$ such that $\forall x \in \{0,1\}^V$:

$$f_\theta(x) = \begin{cases} y(d^*) & \text{if } D = \emptyset \\ f_{\theta[d^*_{\downarrow 0}]}(x) & \text{if } D \neq \emptyset \wedge x_{v(d^*)} = 0 \\ f_{\theta[d^*_{\downarrow 1}]}(x) & \text{if } D \neq \emptyset \wedge x_{v(d^*)} = 1 \end{cases}$$
$$= \begin{cases} y(d^*) & \text{if } D = \emptyset \\ (1 - x_{v(d^*)}) f_{\theta[d^*_{\downarrow 0}]}(x) + x_{v(d^*)} f_{\theta[d^*_{\downarrow 1}]}(x) & \text{otherwise} \end{cases}$$

Remark. The set Θ of V-variate $Y = \{0,1\}$-valued BDTs can be identified with a subset of V-variate disjunctive normal forms.

Definition. For any BDT $\theta = (V, Y, D, D', d^*, E, \delta, v, y)$, the **depth** of θ is the $R(\theta) \in \mathbb{N}$ such that

$$R(\theta) = \begin{cases} 0 & \text{if } D = \emptyset \\ 1 + \max\{R(\theta[d^*_{\downarrow 0}]), R(\theta[d^*_{\downarrow 1}])\} & \text{otherwise} \end{cases} . \quad (1)$$

Definition. For any labeled data $T = (S, X, x, y)$ with $X = \{0,1\}^V$, the set Θ of all V-variate $\{0,1\}$-valued BDTs, the family $f : \Theta \to \{0,1\}^X$ of functions defined by these BDTs, the depth R of BDTs as a regularizer, the 0-1-loss L and any $\lambda \in \mathbb{R}_0^+$:

- The instance of the **supervised learning** problem of BDTs has the form

$$\min_{\theta \in \Theta} \quad \lambda R(\theta) + \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{2}$$

- The **separation problem** of BDTs has the form

$$\inf_{\theta \in \Theta} \quad R(\theta) \tag{3}$$

$$\text{subject to} \quad \forall s \in S: \quad f_\theta(x_s) = y_s \tag{4}$$

- For any $m \in \mathbb{N}$, the **separability problem** of BDTs is to decide whether there exists a BDT $\theta \in \Theta$ such that

$$R(\theta) \leq m \tag{5}$$

$$\forall s \in S: \quad f_\theta(x_s) = y_s \ . \tag{6}$$

Remark. separability $\leq_p$ separation $\leq_p$ supervised learning[1].

[1] $\leq_p$: Karp reduction.

Next, we show that *separability* is NP-hard by reducing the **exact cover by 3-sets** problem, using a construction by Haussler (1988).

Definition. Let S be any set and $\Sigma \subseteq 2^S$. Σ is called a **cover** of S if and only if

$$\bigcup_{\sigma \in \Sigma} \sigma = S \ . \tag{7}$$

A cover of S is called **exact** if and only if

$$\forall \{\sigma, \sigma'\} \in \binom{\Sigma}{2} : \quad \sigma \cap \sigma' = \emptyset \ . \tag{8}$$

Definition. Let S be any set and $\Sigma \subseteq 2^S$. Deciding whether there exists a $\Sigma' \subseteq \Sigma$ such that Σ' is an exact cover of S is called the instance of the **exact cover problem** w.r.t. S and Σ. If, in addition, $|S|$ is an integer multiple of 3 and any $U \in \Sigma$ is such that $|U| = 3$, the instance of the exact cover problem wrt. S and Σ is also called the instance of the **exact cover by 3-sets problem** wrt. S and Σ.

Proof. For any instance (S', Σ) of the exact cover by 3-sets problem and the $n \in \mathbb{N}$ such that $|S'| = 3n$, we construct the instance of the separability problem of BDTs such that

- $V = \Sigma$
- $S = S' \mathbin{\dot\cup} \{0\}$
- $x : S \to \{0,1\}^\Sigma$ such that $x_0 = 0$ and

$$\forall s \in S' \ \forall \sigma \in \Sigma : \quad x_s(\sigma) = \begin{cases} 1 & \text{if } s \in \sigma \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

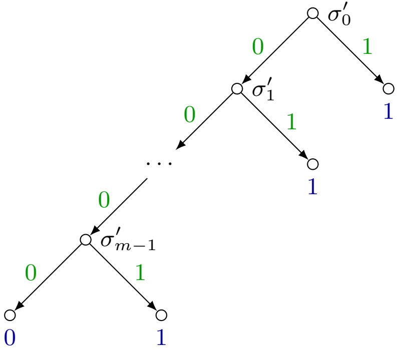
- $y : S \to \{0,1\}$ such that $y_0 = 0$ and $\forall s \in S' : y_s = 1$.
- $m = n$

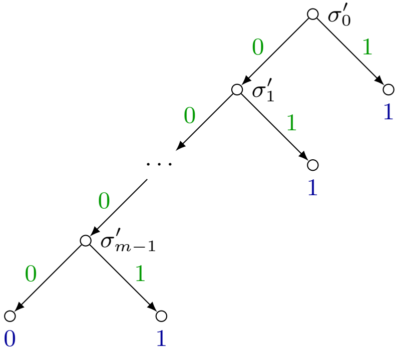
We show that the instance the exact cover problem has a solution iff the instance of the separability problem of BDTs has a solution.

($\Rightarrow$) Let $\Sigma' \subseteq \Sigma$ a solution to the instance of the exact cover problem.

Consider any order on Σ' and the bijection $\sigma' : n \to \Sigma'$ induced by this order.

We show that the BDT θ depicted below solves the instance of the separability problem of BDTs.





The BDT satisfies $R(\theta) = m$.

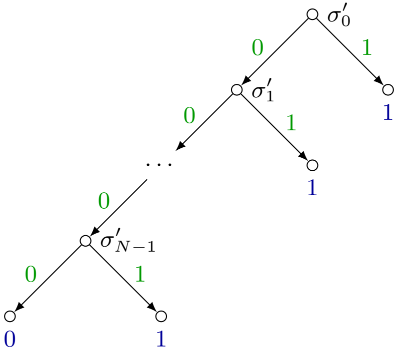
The BDT decides the labeled data correctly because

- $f_\theta(x_0) = 0 = y_0$
- At each of the m interior nodes, three additional elements of S' are mapped to 1. Thus, all $3m$ many elements $s \in S'$ are mapped to 1. That is $\forall s \in S' \colon f_\theta(x_s) = 1 = y_s$.

($\Leftarrow$) Let $\theta = (V, Y, D, D', d^*, E, \delta, \sigma, y')$ a BDT that solves the instance of the separability problem of BDTs.

W.l.o.g., we assume, for any interior node $d \in D$, that $d_{\downarrow 1}$ is a leaf and $y'(d_{\downarrow 1}) = 1$.

Hence, θ is of the form depicted below.



Therefore:

$$\forall x \in X\colon \quad f_\theta(x) = \begin{cases} 1 & \text{if } \exists j \in N\colon x(\sigma_j) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Thus,

$$\forall s \in S\colon \quad f_\theta(x_s) = \begin{cases} 1 & \text{if } \exists j \in N\colon s \in \sigma_j \\ 0 & \text{otherwise} \end{cases} \quad , \tag{11}$$

by definition of x in (9).

Consequently,

$$\bigcup_{j=0}^{N-1} \sigma_j = S' \ , \tag{12}$$

by definition of y such that $\forall s \in S'\colon y_s = 1$.

Moreover, $N = m$, because

$$3m = |S'| \overset{(12)}{=} \left| \bigcup_{j=0}^{N-1} \sigma_j \right| \leq \sum_{j=0}^{N-1} |\sigma_j| = \sum_{j=0}^{N-1} 3 = 3N \overset{(5)}{\leq} 3m \ .$$

Therefore:

$$\forall \{j, l\} \in \binom{[N]}{2} : \quad \sigma_k \cap \sigma_l = \emptyset \tag{13}$$

Thus,

$$\bigcup_{j=0}^{N-1} \sigma_j$$

is a solution to the instance of the exact cover by 3-sets problem defined by (S', Σ), by (12) and (13).

□

Summary:

- Supervised learning of BDTs is hard.
- More specifically, the NP-hard exact cover by 3-sets problem is reducible to the separability problem of BDTs, by construction of Haussler data.

Topics of upcoming exercises:

- A heuristic algorithm for the supervised learning of BDTs
- Supervised learning of disjunctive normal forms