Machine Learning I

Bjoern Andres

Machine Learning for Computer Vision
TU Dresden

https://mlcv.cs.tu-dresden.de/courses/26-winter/ml1/

Winter Term 2026/2027

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

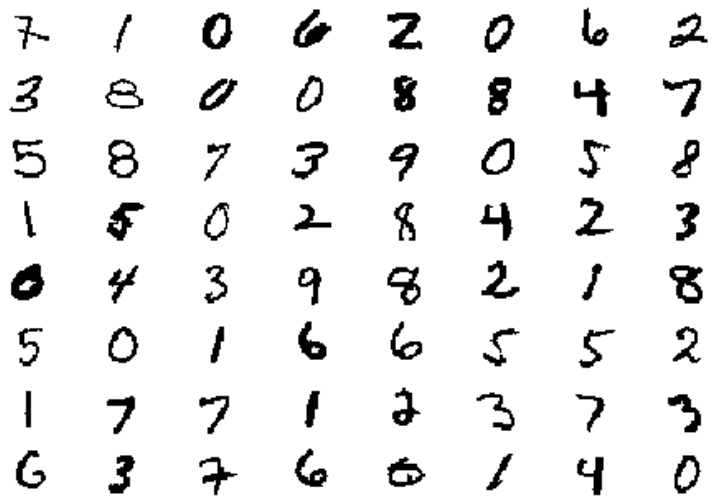
- We introduce the problem as a specialization of supervised learning by defining labeled data, a family of functions, a regularizer and a loss function.
- We prove that the problem is NP-hard, by reduction of the set cover problem

Binary Decision Trees

We consider labeled data $T = (S, X, x, y)$ such that $X = \{0, 1\}^V$, i.e., we consider **binary inputs**. Hence:

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

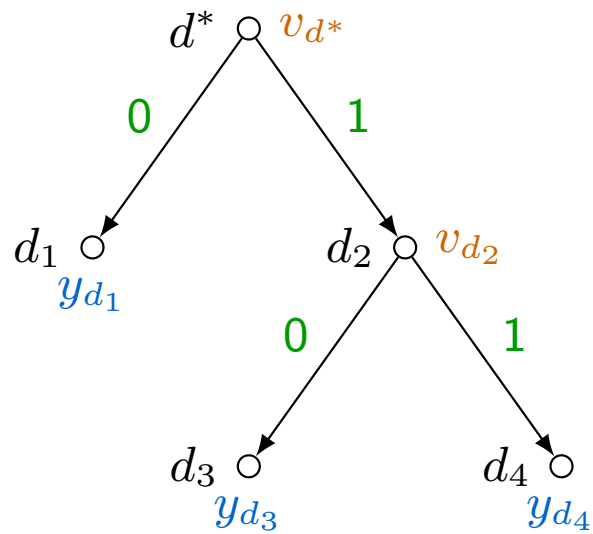
Example.



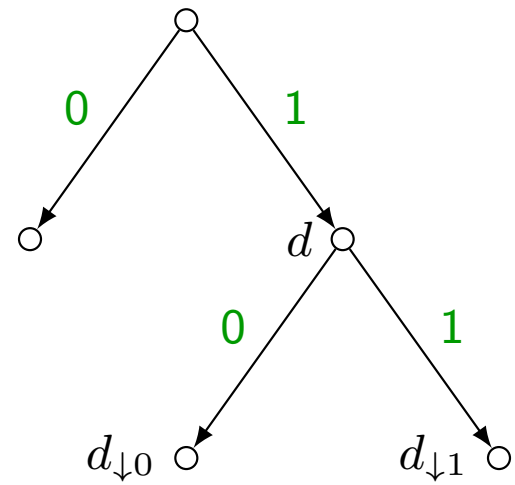
LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. doi:10.1109/5.726791

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 $d_{\downarrow j} \in D \cup D'$ be 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 if and only if

- $(D_2 \cup D_2', E')$ 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 (see exercises).

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

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

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 the separability problem for BDTs is NP-hard, by reducing the set cover problem, using a construction by Haussler (1988).

Definition. For any set S and any $\Sigma \subseteq 2^S$, Σ is called a **cover** of S if and only if $\cup\Sigma = S$.

Definition. An instance of the decision problem **set cover** is a triple (S, Σ, m) with S being any finite set, $\emptyset \notin \Sigma \subseteq 2^S$ and $m \in \mathbb{N}$. A solution is any $\Sigma' \subseteq \Sigma$ such that Σ' is a cover of S and $|\Sigma'| \leq m$.

Theorem (Karp 1972). The set cover problem is NP-complete.

Definition (Haussler 1988). For any instance (S', Σ, m) of set cover, the **Haussler data** induced by (S', Σ, m) is the labeled data (S, X, x, y) such that $S = \{1\} \cup S'$, $X = \{0,1\}^{\Sigma}$, $x_1 = (1, \ldots, 1) =: 1^{\Sigma}$,

$$\forall s \in S' \; \forall \sigma \in \Sigma\colon \quad x_s(\sigma) = \begin{cases} 0 & \text{if } s \in \sigma \\ 1 & \text{otherwise} \end{cases} , \tag{7}$$

$y_1 = 1$ and $\forall s \in S'\colon y_s = 0$.

Lemma (Haussler 1988). For any instance (S', Σ, m) of set cover, for the Haussler data (S, X, x, y) induced by (S', Σ, m), and for any $\Sigma' \subseteq \Sigma$:

$$\bigcup_{\sigma \in \Sigma'} \sigma = S' \quad \Leftrightarrow \quad \forall s \in S'\colon \prod_{\sigma \in \Sigma'} x_s(\sigma) = 0 \ .$$

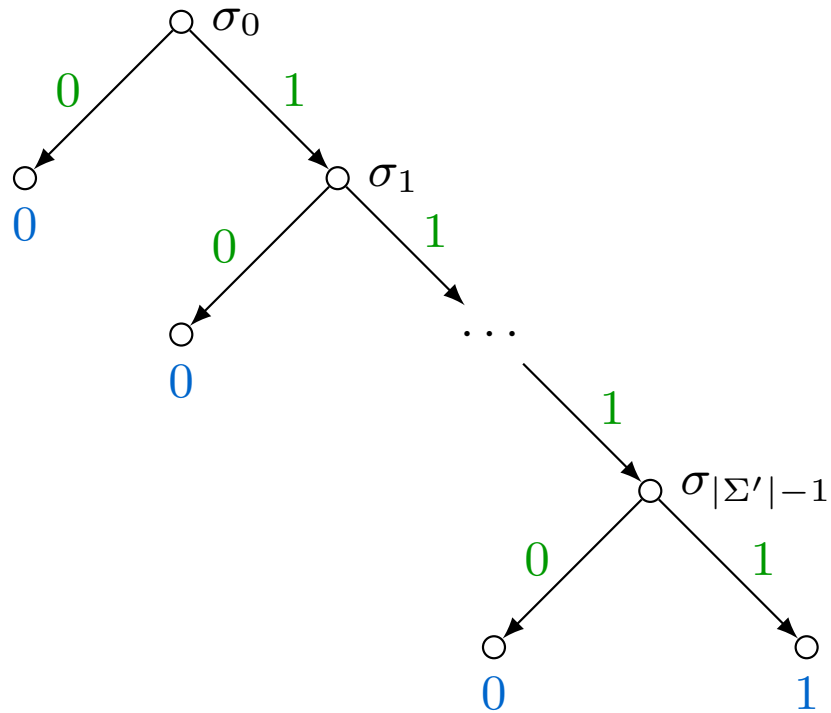
Proof.

$$\begin{aligned} & \bigcup_{\sigma \in \Sigma'} \sigma = S' \\ \Leftrightarrow \quad & \forall s \in S' \; \exists \sigma \in \Sigma'\colon s \in \sigma \\ \Leftrightarrow \quad & \forall s \in S' \; \exists \sigma \in \Sigma'\colon x_s(\sigma) = 0 \\ \Leftrightarrow \quad & \forall s \in S'\colon \prod_{\sigma \in \Sigma'} x_s(\sigma) = 0 \ . \end{aligned}$$

Lemma. set cover $\leq_p$ separability for BDTs.

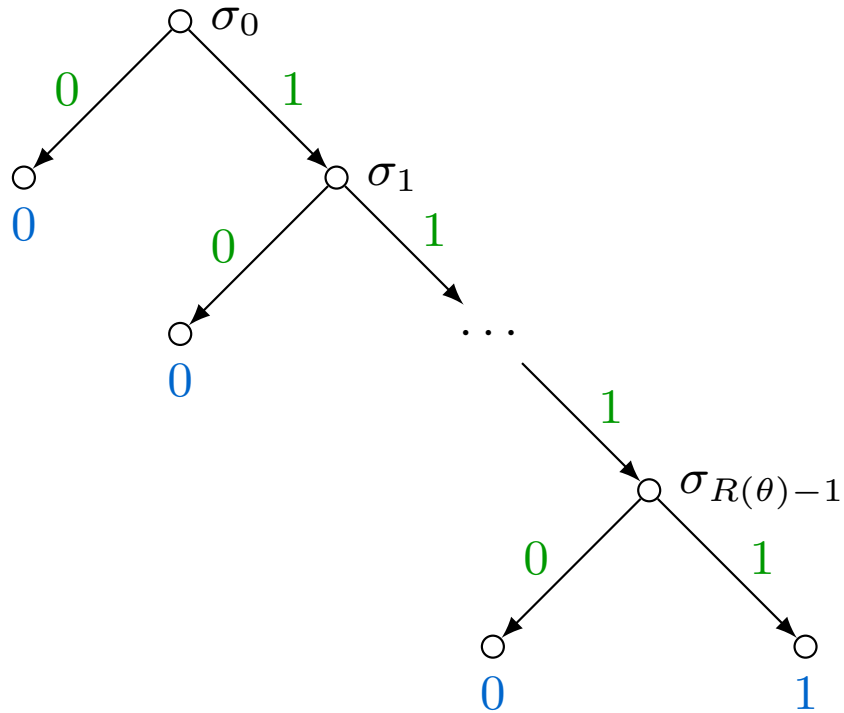
Proof. Let (S', Σ, m) be any instance of set cover. Let (S, X, x, y) be the Haussler data induced by (S', Σ, m). We will show: There exists a cover $\Sigma' \subseteq \Sigma$ of S' with $|\Sigma'| \leq m$ if and only if there exists a Σ-variate, $\{0,1\}$-valued BDT θ such that $R(\theta) \leq m$ and $\forall s \in S\colon f_\theta(x_s) = y_s$.

($\Rightarrow$) Let $\Sigma' \subseteq \Sigma$ be a cover of S' and $|\Sigma'| \leq m$. Choose any bijection $\sigma\colon |\Sigma'| \to \Sigma'$. Consider the BDT θ depicted below.



This BDT θ decides the labeled data correctly. Moreover, $R(\theta) = |\Sigma'| \leq m$.

($\Leftarrow$) Let $\theta = (V, Y, D, D', d^*, E, \delta, \sigma, y')$ be a BDT that solves the instance of separability for BDTs w.r.t. the Haussler data (S, X, x, y). W.l.o.g., assume for any $d \in D$, that $d_{\downarrow 0}$ is a leaf, and that $y'(d_{\downarrow 0}) = 0$. Thus, θ is of the form depicted below.



Since f_θ decides the Haussler data correctly, it is necessary that

$$\forall s \in S' \; \exists j \in R(\theta) : \quad x_s(\sigma_j) = 0 \; . \tag{8}$$

Let $\Sigma' := \{\sigma_j \mid j \in R(\theta)\}$. Now:

$$\forall s \in S' \; \exists \sigma \in \Sigma' : \quad x_s(\sigma) = 0 \ . \tag{9}$$

Analogous to the proof of Haussler's lemma, it follows that Σ' is a cover of S'. Moreover, $|\Sigma'| = R(\theta) \leq m$. □

Corollary. Separability for BDTs is NP-hard.

Corollary. Supervised learning of BDTs is NP-hard.

Summary:

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