Machine Learning I

Bjoern Andres, Shengxian Zhao

Machine Learning for Computer Vision
TU Dresden

Winter Term 2022/2023

Contents. This part of the course introduces the concept of labeled data and the supervised learning problem.

Example: A medical test with $n \in \mathbb{N}$ design parameters $\theta \in \Theta = \mathbb{R}^n$ measures $m \in \mathbb{N}$ quantities and indicates by $y \in Y = \{0, 1\}$ whether a measurement $x \in X = \mathbb{R}^m$ is considered to be healthy $(y = 0)$ or pathological $(y = 1)$.

$$X \xrightarrow{g_\theta} Y$$

Example: A medical test with $n \in \mathbb{N}$ design parameters $\theta \in \Theta = \mathbb{R}^n$ measures $m \in \mathbb{N}$ quantities and indicates by $y \in Y = \{0, 1\}$ whether a measurement $x \in X = \mathbb{R}^m$ is considered to be healthy ($y = 0$) or pathological ($y = 1$).

$$X \xrightarrow{g_\theta} Y$$

Informally, **supervised learning** is the problem of finding, in a family $g : \Theta \to Y^X$ of functions, one function $g_\theta : X \to Y$ that minimizes a weighted sum of two objectives:

- g_θ deviates little from a finite set $\{(x_s, y_s)\}_{s \in S}$ of input-output-pairs, called **labeled data**
- g_θ has low complexity, as quantified by a function $R : \Theta \to \mathbb{R}_0^+$, called a **regularizer**

Remarks:

- The family g defines a parameterization of functions from inputs X to outputs Y.
- g can be chosen so as to constrain the set of functions from X to Y in the first place.
- For instance, Θ can be a set of forms, g the functions defined by these forms, and R the length of these forms.

We concentrate exclusively on the special case where Y is finite.

We concentrate exclusively on the special case where Y is finite.

To begin with, we even concentrate on the case where $Y = \{0, 1\}$. Hence, we consider a family $g\colon \Theta \to \{0, 1\}^X$.

We concentrate exclusively on the special case where Y is finite.

To begin with, we even concentrate on the case where $Y = \{0, 1\}$. Hence, we consider a family $g \colon \Theta \to \{0, 1\}^X$.

We allow ourselves to take a detour by not optimizing over a family $g : \Theta \to \{0, 1\}^X$ directly but instead optimizing over a family $f : \Theta \to \mathbb{R}^X$ and defining g wrt. f via a function $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a **loss function**, such that

$$\forall \theta \in \Theta \ \forall x \in X \colon \quad g_\theta(x) \in \operatorname*{argmin}_{\hat{y} \in \{0,1\}} \ L(f_\theta(x), \hat{y}) \ . \tag{1}$$

We concentrate exclusively on the special case where Y is finite.

To begin with, we even concentrate on the case where $Y = \{0,1\}$. Hence, we consider a family $g\colon \Theta \to \{0,1\}^X$.

We allow ourselves to take a detour by not optimizing over a family $g : \Theta \to \{0,1\}^X$ directly but instead optimizing over a family $f : \Theta \to \mathbb{R}^X$ and defining g wrt. f via a function $L : \mathbb{R} \times \{0,1\} \to \mathbb{R}_0^+$, called a **loss function**, such that

$$\forall \theta \in \Theta \ \forall x \in X\colon \quad g_\theta(x) \in \operatorname*{argmin}_{\hat{y} \in \{0,1\}} \ L(f_\theta(x), \hat{y}) \ . \tag{1}$$

Example: 0/1-loss

$$\forall r \in \mathbb{R} \ \forall \hat{y} \in \{0,1\}\colon \quad L(r, \hat{y}) = \begin{cases} 0 & r = \hat{y} \\ 1 & \text{otherwise} \end{cases} \ . \tag{2}$$

We concentrate exclusively on the special case where Y is finite.

To begin with, we even concentrate on the case where $Y = \{0, 1\}$. Hence, we consider a family $g \colon \Theta \to \{0, 1\}^X$.

We allow ourselves to take a detour by not optimizing over a family $g : \Theta \to \{0, 1\}^X$ directly but instead optimizing over a family $f : \Theta \to \mathbb{R}^X$ and defining g wrt. f via a function $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a **loss function**, such that

$$\forall \theta \in \Theta \; \forall x \in X \colon \quad g_\theta(x) \in \operatorname*{argmin}_{\hat{y} \in \{0,1\}} \; L(f_\theta(x), \hat{y}) \ . \tag{1}$$

Example: 0/1-loss

$$\forall r \in \mathbb{R} \; \forall \hat{y} \in \{0, 1\} \colon \quad L(r, \hat{y}) = \begin{cases} 0 & r = \hat{y} \\ 1 & \text{otherwise} \end{cases} \ . \tag{2}$$

Next, we define the supervised learning problem rigorously.

Definition. For any finite, non-empty set S, called a set of **samples**, any $X \neq \emptyset$, called an **attribute space** and any $x : S \to X$, the tuple (S, X, x) is called **unlabeled data**.

Definition. For any finite, non-empty set S, called a set of **samples**, any $X \neq \emptyset$, called an **attribute space** and any $x : S \to X$, the tuple (S, X, x) is called **unlabeled data**.

For any $y : S \to \{0, 1\}$, given in addition and called a **labeling**, the tuple (S, X, x, y) is called **labeled data**.

Definition. For any labeled data $T = (S, X, x, y)$, any $\Theta \neq \emptyset$ and $f : \Theta \to \mathbb{R}^X$, any $R : \Theta \to \mathbb{R}_0^+$, called a **regularizer**, any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a **loss function**, and any $\lambda \in \mathbb{R}_0^+$:

- The instance of the **supervised learning problem** wrt. T, Θ, f, R, L and λ has the form

$$\inf_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{3}$$

Definition. For any labeled data $T = (S, X, x, y)$, any $\Theta \neq \emptyset$ and $f : \Theta \to \mathbb{R}^X$, any $R : \Theta \to \mathbb{R}_0^+$, called a **regularizer**, any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a **loss function**, and any $\lambda \in \mathbb{R}_0^+$:

- The instance of the **supervised learning problem** wrt. T, Θ, f, R, L and λ has the form

$$\inf_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{3}$$

- The instance of the **separation problem** wrt. T, Θ, f and R has the form

$$\begin{aligned} \inf_{\theta \in \Theta} \quad & R(\theta) && \quad (4) \\ \text{subject to} \quad & \forall s \in S : \quad f_\theta(x_s) = y_s && \quad (5) \end{aligned}$$

Definition. For any labeled data $T = (S, X, x, y)$, any $\Theta \neq \emptyset$ and $f : \Theta \to \mathbb{R}^X$, any $R : \Theta \to \mathbb{R}_0^+$, called a **regularizer**, any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a **loss function**, and any $\lambda \in \mathbb{R}_0^+$:

- The instance of the **supervised learning problem** wrt. T, Θ, f, R, L and λ has the form

$$\inf_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{3}$$

- The instance of the **separation problem** wrt. T, Θ, f and R has the form

$$\begin{aligned} \inf_{\theta \in \Theta} \quad & R(\theta) && (4) \\ \text{subject to} \quad & \forall s \in S : \quad f_\theta(x_s) = y_s && (5) \end{aligned}$$

- The instance of the **bounded separability problem** wrt. T, Θ, f, R and $m \in \mathbb{N}$ is to decide whether there exists a $\theta \in \Theta$ such that

$$\begin{aligned} & R(\theta) \leq m && (6) \\ \forall s \in S : \quad & f_\theta(x_s) = y_s && (7) \end{aligned}$$

Definition. For any unlabeled data $T = (S, X, x)$, any $\hat{f} : X \to \mathbb{R}$ and any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, the instance of the **inference problem** wrt. $T, \hat{f}$ and L is defined as

$$\min_{y' \in \{0,1\}^S} \sum_{s \in S} L(\hat{f}(x_s), y'_s) \tag{8}$$

Lemma. The solutions to the inference problem are the $y : S \to \{0, 1\}$ such that

$$\forall s \in S\colon \quad y_s \in \underset{\hat{y} \in \{0,1\}}{\operatorname{argmin}} \ L(\hat{f}(x_s), \hat{y}) \ . \tag{9}$$

Moreover, if $\hat{f}(X) \subseteq \{0, 1\}$ and L is the 01-loss, then

$$\forall s \in S\colon \quad y'_s = \hat{f}(x_s) \ . \tag{10}$$

Summary. Supervised learning is an optimization problem. It consists in finding, in a family of functions, one function that minimizes a weighted sum of two objectives:

1. The function deviates little from given labeled data, as quantified by a loss function
2. The function has low complexity, as quantified by a regularizer.