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 introduces the problem of **classifying** data w.r.t. any given finite number of classes.
- This problem is introduced as an unsupervised learning problem w.r.t. constrained data whose feasible labelings are characteristic functions of **maps**.

A map $\varphi\colon A \to B$ is a binary relation $\varphi \subseteq A \times B$ with the properties

$$\forall a \in A\ \exists b \in B: \quad (a,b) \in \varphi \tag{1}$$

$$\forall a \in A\ \forall b, b' \in B: \quad (a,b) \in \varphi \ \wedge \ (a,b') \in \varphi \ \Rightarrow \ b = b' \ . \tag{2}$$

They are characterized by those functions $y : A \times B \to \{0,1\}$ that satisfy

$$\forall a \in A\colon \quad \sum_{b \in B} y_{ab} = 1 \ . \tag{3}$$

We reduce the problem of learning and inferring maps to the problem of learning and inferring decisions, by defining **constrained data** $(S, X, x, \mathcal{Y})$ with

$$S = A \times B \tag{4}$$

$$\mathcal{Y} = \left\{ y \in \{0,1\}^S \;\middle|\; \forall a \in A\colon \sum_{b \in B} y_{ab} = 1 \right\} \ . \tag{5}$$

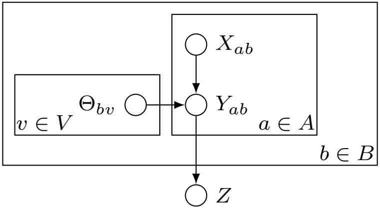
More specifically, we consider

- a finite, non-empty set V, called a set of **features**
- the **feature space** $X = B \times \mathbb{R}^V$ such that, for any $(a,b) \in A \times B$, the class label b is the first feature of (a,b), i.e.:

$$\forall a \in A \; \forall b \in B \; \exists \hat{x} \in \mathbb{R}^V\colon \quad x_{ab} = (b, \hat{x}) \tag{6}$$

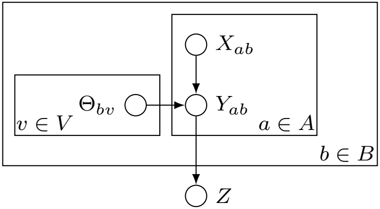
We consider **linear functions** with a separate set of coefficients for every class label. Specifically, we consider $\Theta = \mathbb{R}^{B \times V}$ and $f : \Theta \to \mathbb{R}^X$ such that

$$\forall \theta \in \Theta \; \forall b \in B \; \forall \hat{x} \in \mathbb{R}^V\colon \quad f_\theta((b, \hat{x})) = \sum_{v \in V} \theta_{bv} \, \hat{x}_v = \langle \theta_{b\cdot}, \hat{x} \rangle \ . \tag{7}$$



Probabilistic model:

- For any $(a, b) \in A \times B$, let X_{ab} be a random variable whose value is a vector $x_{ab} \in B \times \mathbb{R}^V$, the **feature vector** of (a, b).
- For any $(a, b) \in A \times B$, let Y_{ab} be a random variable whose value is a binary number $y_{ab} \in \{0, 1\}$, called the **decision** of classifying a as b
- For any $b \in B$ and any $v \in V$, let Θ_{bv} be a random variable whose value is a real number $\theta_{bv} \in \mathbb{R}$, a **parameter** of the function we seek to learn
- Let Z be a random variable whose value is a subset $\mathcal{Z} \subseteq \{0, 1\}^{A \times B}$ called the set of **feasible decisions**. For multiple label classification, we are interested in $\mathcal{Z} = \mathcal{Y}$, the set of the characteristic functions of all maps from A to B.



Probabilistic model: We assume the factorization

$$P(X, Y, Z, \Theta) = P(Z \mid Y) \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta) \prod_{(b,v) \in B \times V} P(\Theta_{bv}) \prod_{(a,b) \in A \times B} P(X_{ab})$$

- Supervised learning:

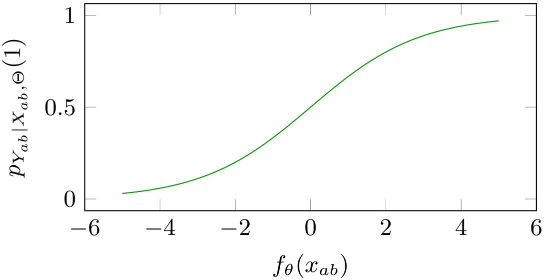
$$
\begin{aligned}
P(\Theta \mid X, Y, Z) &= \frac{P(X, Y, Z, \Theta)}{P(X, Y, Z)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid X, Y)\, P(X, Y)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid Y)\, P(X, Y)} \\
&= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)} \\
&\propto P(Y \mid X, \Theta)\, P(\Theta) \\
&= \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta) \prod_{(b,v) \in B \times V} P(\Theta_{bv})
\end{aligned}
$$

- Inference:

$$
\begin{aligned}
P(Y \mid X, Z, \theta) &= \frac{P(X, Y, Z, \Theta)}{P(X, Z, \Theta)} \\
&= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Z, \Theta)} \\
&\propto P(Z \mid Y)\, P(Y \mid X, \Theta) \\
&= P(Z \mid Y) \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta)
\end{aligned}
$$

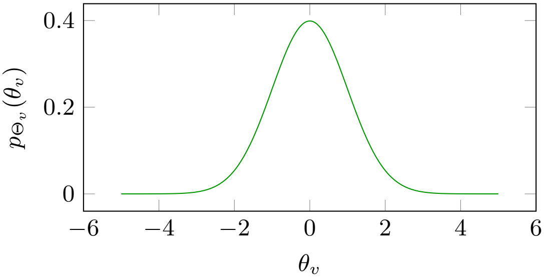
- **Sigmoid distribution**

$$\forall a \in A \ \forall b \in B: \qquad p_{Y_{ab}|X_{ab},\Theta}(1) = \frac{1}{1 + 2^{-f_\theta(x_{ab})}} \tag{8}$$



- **Normal distribution** with $\sigma \in \mathbb{R}^+$:

$$\forall b \in B \; \forall v \in V : \qquad p_{\Theta_{bv}}(\theta_{bv}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_{bv}^2/2\sigma^2} \tag{9}$$



- **Uniform distribution on a subset**

$$\forall \mathcal{Z} \subseteq \{0,1\}^{A\times B}\ \forall y \in \{0,1\}^{A\times B} \quad p_{Z|Y}(\mathcal{Z}, y) \propto \begin{cases} 1 & \text{if } y \in \mathcal{Z} \\ 0 & \text{otherwise} \end{cases}$$

Note that $p_{Z|Y}(\mathcal{Y}, y)$ is non-zero iff the relation $y^{-1}(1) \subseteq A \times B$ is a map.

Lemma. Estimating maximally probable parameters θ, given features x and decisions y, i.e., $\operatorname{argmax}_{\theta \in \mathbb{R}^{B \times V}} p_{\Theta|X,Y,Z}(\theta, x, y, \mathcal{Y})$ separates into $|B|$ independent l_2-regularized logistic regression problems, each with parameters in $\mathbb{R}^V$.

Proof. Analogous to the case of deciding, we now obtain:

$$\begin{aligned} &\operatorname*{argmax}_{\theta \in \mathbb{R}^{B \times V}} \quad p_{\Theta|X,Y,Z}(\theta, x, y, \mathcal{Y}) \\ = &\operatorname*{argmin}_{\theta \in \mathbb{R}^{B \times V}} \quad \sum_{(a,b) \in A \times B} \left(-y_{ab}\, f_\theta(x_{ab}) + \log\left(1 + 2^{f_\theta(x_{ab})}\right)\right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \ . \end{aligned}$$

Consider the unique $x' : A \times B \to \mathbb{R}^V$ with $\forall (a,b) \in A \times B \colon x_{ab} = (b, x'_{ab})$.

$$\begin{aligned} &\min_{\theta \in \mathbb{R}^{B \times V}} \sum_{(a,b) \in A \times B} \left(-y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log\left(1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle}\right)\right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \\ = &\min_{\theta \in \mathbb{R}^{B \times V}} \sum_{b \in B} \left(\sum_{a \in A} \left(-y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log\left(1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle}\right)\right) + \frac{\log e}{2\sigma^2} \|\theta_{b\cdot}\|_2^2 \right) \\ = &\sum_{b \in B} \min_{\theta_{b\cdot} \in \mathbb{R}^V} \left(\sum_{a \in A} \left(-y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log\left(1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle}\right)\right) + \frac{\log e}{2\sigma^2} \|\theta_{b\cdot}\|_2^2 \right) \ . \end{aligned}$$

Lemma. For any constrained data as defined above, any $\theta \in \mathbb{R}^{B \times V}$ and any $\hat{y} : A \times B \to \{0, 1\}$, $\hat{y}$ is a solution to the inference problem

$$\min_{y \in \mathcal{Y}} \quad \sum_{(a,b) \in A \times B} L(f_\theta(x_{ab}), y_{ab}) \tag{10}$$

iff there exists an $\varphi : A \to B$ such that

$$\forall a \in A\colon \quad \varphi(a) \in \max_{b \in B} \langle \theta_{b\cdot}, x'_{ab} \rangle \tag{11}$$

and

$$\forall (a,b) \in A \times B\colon \quad \hat{y}_{ab} = 1 \;\Leftrightarrow\; \varphi(a) = b \;\; . \tag{12}$$

Proof.

$$
\begin{aligned}
&\sum_{(a,b)\in A\times B} L(f_\theta(x_{ab}), y_{ab}) \\
&= \sum_{(a,b)\in A\times B} \left(L(f_\theta(x_{ab}), 1)\, y_{ab} + L(f_\theta(x_{ab}), 0)\,(1 - y_{ab})\right) \\
&= \sum_{(a,b)\in A\times B} \left(L(f_\theta(x_{ab}), 1) - L(f_\theta(x_{ab}), 0)\right) y_{ab} + \text{const.} \\
&= \sum_{(a,b)\in A\times B} \left(-f_\theta(x_{ab})\right) y_{ab} \\
&= \sum_{(a,b)\in A\times B} \left(-\langle \theta_{b\cdot}, x'_{ab}\rangle\right) y_{ab} && x_{ab} = (b, x'_{ab}) \\
&= \sum_{a\in A} \sum_{b\in B} \left(-\langle \theta_{b\cdot}, x'_{ab}\rangle\right) y_{ab}
\end{aligned}
$$

Summary.

- Classification can be cast as an unsupervised learning problem w.r.t. constrained data defined such that the feasible labelings are characteristic functions of maps.
- In the special case of supervised learning and the logistic loss function, this problem separates into as many independent independent logistic regression problems as there are classes. This is commonly called one-versus-rest learning.