Machine Learning II

Bjoern Andres

Machine Learning for Computer Vision
TU Dresden

https://mlcv.cs.tu-dresden.de/courses/26-summer/ml2/

Summer Term 2026

Motivation. Even the most general learning and inference problem w.r.t. constrained data $(S, X, x, \mathcal{Y})$ we have considered is too restrictive for certain applications:

- Features $x_s \in X$ are defined for single elements $s \in S$ only.
- Dependencies between decisions $y_s, y_{s'} \in \{0, 1\}$ for distinct $s, s' \in S$ are only due to hard constraints definded by the feasible set $\mathcal{Y} \subset \{0, 1\}^S$.

Example: Pixel classification: Given a digital image, we need to decide for every pixel $s \in S$, by the contents of the image around that pixel, whether the pixel is of interest $(y_s = 1)$ or not of interest $(y_s = 0)$.

Typically, decisions at neighboring pixels $s, s' \in S$ are more likely to be equal $(y_s = y_{s'})$ than unequal $(y_s \neq y_{s'})$, and we wish to learn how this increased probability depends on the contents of the image.

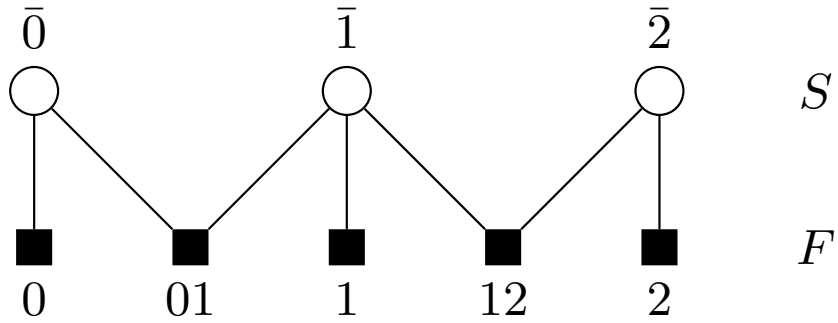
The mathematical abstractions of learning we have considered so far are insufficient to express these dependencies.

In order to lift this restriction, we will define the **supervised structured learning** problem and the **structured inference** problem in which

- features are associated with subsets of S
- decisions can be tied by probabilistic dependencies.

More specifically, we will

- introduce a family $H : \Theta \to \mathbb{R}^{X \times Y}$ of functions that quantify by $H_\theta(x, y)$ how incompatible features $x \in X$ are with a combination of decisions $y \in \{0, 1\}^S$
- define supervised structured learning as a problem of finding one function from this family
- define structured inference as the problem of finding a combination of decisions $y \in \{0, 1\}^S$ that minimizes $H_\theta(x, \cdot)$.



Definition. A triple (S, F, E) is called a **factor graph** with **variable nodes** S and **factor nodes** F iff $S \cap F = \varnothing$ and $(S \cup F, E)$ is a bipartite graph such that $\forall e \in E \, \exists s \in S \, \exists f \in F \colon e = \{s, f\}$.

- For any factor node $f \in F$, we denote by $S_f = \{s \in S \mid \{s, f\} \in E\}$ the set of those variable nodes that are neighbors of f.
- For any variable node $s \in S$, we denote by $F_s = \{f \in F \mid \{s, f\} \in E\}$ the set of those factor nodes that are neighbors of s.

Definition. Unlabeled structured data is a tuple $T = (S, F, E, \{X_f\}_{f \in F}, x)$ such that the following conditions hold:

- (S, F, E) is a factor graph.
- Every set X_f is non-empty, called the **feature space** of f.
- $x \in \prod_{f \in F} X_f$, where the Cartesian product $\prod_{f \in F} X_f$ is called the **feature space** of T.

Labeled structured data is a tuple $(S, F, E, \{X_f\}_{f \in F}, x, y)$ such that $(S, F, E, \{X_f\}_{f \in F}, x)$ is unlabeled structured data, and $y \in \{0, 1\}^S$.

Definition 1. An instance of **supervised structured learning** is a tuple $(D, \Theta, H, L, R, \lambda)$ with

- labeled structured data $D = (S, F, E, \{X_f\}_{f \in F}, x, y)$,
- $\Theta \neq \varnothing$, called the **feasible set**,
- $H : \Theta \to \mathbb{R}^{X \times Y}$ with $X = \prod_{f \in F} X_f$ and $Y = \{0, 1\}^S$, called an **energy function**,
- $L : \mathbb{R}^Y \times Y \to \mathbb{R}_0^+$, called a **loss function**,
- $R : \Theta \to \mathbb{R}_0^+$, called a **regularizer**, and
- $\lambda \in \mathbb{R}_0^+$, called a **regularization parameter**.

The **objective function** is the $\varphi \colon \Theta \to \mathbb{R}_0^+$ such that

$$\forall \theta \in \Theta \colon \quad \varphi(\theta) = \lambda R(\theta) + L(H_\theta(x, \cdot), y) \ . \tag{1}$$

A **solution** is any $\hat{\theta} \in \Theta$ such that

$$\varphi(\hat{\theta}) = \inf \{\varphi(\theta) \mid \theta \in \Theta\} \ . \tag{2}$$

Definition 2. An instance of **structured inference** is a tuple $(D, \hat{H}, L)$ with

- unlabeled structured data $D = (S, F, E, \{X_f\}_{f \in F}, x)$,
- $\hat{H} \colon X \times Y \to \mathbb{R}$ with $Y = \{0,1\}^S$, called an **energy function**, and
- $L : \mathbb{R}^Y \times Y \to \mathbb{R}_0^+$, called a **loss function**.

The **objective function** is the $\varphi \colon Y \to \mathbb{R}_0^+$ such that

$$\forall y \in Y \colon \quad \varphi(y) = L(\hat{H}(x, \cdot), y) \ . \tag{3}$$

A **solution** is any $\hat{y} \in Y$ such that

$$\varphi(\hat{y}) = \min \{\varphi(y) \mid y \in Y\} \ . \tag{4}$$

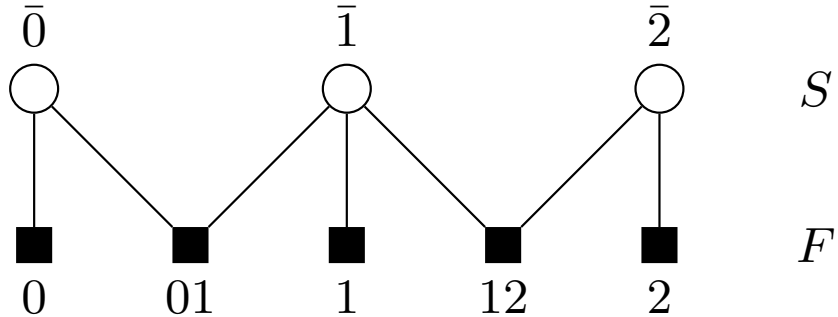
Summary.

- **Structured data** consists of a factor graph (S, F, E) and features $x_f \in X_f$ for every factor $f \in F$.
- **Supervised structured learning** is an optimization problem whose feasible solutions θ define functions $H_\theta : X \times Y \to \mathbb{R}$ whose values $H_\theta(x, y)$ quantify an incompatibility of features $x \in X$ and combinations of decisions $y \in \{0, 1\}^S$.
- **Structured inference** is an optimization problem whose solutions $\hat{y} \in \{0, 1\}^S$ have minimum incompatibility $\hat{H}(x, \hat{y})$ with features x.

Definition 3. For any factor graph $G = (S, F, E)$, a function $H : \{0,1\}^S \to \mathbb{R}$ is said to **factorize** w.r.t. G iff, for every $f \in F$, there exists a function a function $h_f : \{0,1\}^{S_f} \to \mathbb{R}$, called a **factor** of H, such that

$$\forall y \in \{0,1\}^S: \quad H(y) = \sum_{f \in F} h_f(y_{S_f}) \ . \tag{5}$$

Example 1. A function $H : \{0,1\}^S \to \mathbb{R}$ factorizes w.r.t. the factor graph



iff there exist suitable functions $h_0, h_{01}, h_1, h_{12}, h_2$ such that, for any $y \in \{0,1\}^S$: $H(y) = h_0(y_{\bar{0}}) + h_1(y_{\bar{1}}) + h_2(y_{\bar{2}}) + h_{01}(y_{\bar{0}}, y_{\bar{1}}) + h_{12}(y_{\bar{1}}, y_{\bar{2}})$.

Definition 4. A **conditional graphical model** is a tuple $M = (S, F, E, \{X_f\}_{f \in F}, \Theta, \{h_f\}_{f \in F})$ such that the following conditions hold:

- (S, F, E) is a factor graph
- $\Theta \neq \varnothing$, called the **parameter space** of M,
- for every $f \in F$: $X_f \neq \varnothing$, called the **feature space** of f in M,
- for every $f \in F$: $h_f : \Theta \to \mathbb{R}^{X_f \times \{0,1\}^{S_f}}$, called a **factor** of M.

The set $X := \prod_{f \in F} X_f$ is called the **feature space** of M.

The family $H : \Theta \to \mathbb{R}^{X \times \{0,1\}^S}$ such that

$$\forall \theta \in \Theta \; \forall x \in X \; \forall y \in \{0,1\}^S : \quad H_\theta(x, y) = \sum_{f \in F} h_{f\theta}(x_f, y_{S_f}) \tag{6}$$

is called the family of **energy functions** of M.

- We consider a conditional graphical model $(S, F, E, \{X_f\}_{f\in F}, \Theta, \{h_f\}_{f\in F})$ and its family H of energy functions.
- We assume that Θ is a finite-dimensional, real vector space, i.e., there exists a finite, non-empty set J and $\Theta = \mathbb{R}^J$.
- We assume that every function h_f is linear in Θ, i.e., for every $f \in F$, there exists a $\varphi_f : X_f \times \{0,1\}^{S_f} \to \mathbb{R}^J$ such that for any $x_f \in X_f$, any $y_{S_f} \in \{0,1\}^{S_f}$ and any $\theta \in \Theta$:

$$h_{f\theta}(x_f, y_{S_f}) = \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle \tag{7}$$

- For convenience, we define $\xi : X \times \{0,1\}^S \to \mathbb{R}^J$ such that for any $x \in X$ and any $y \in \{0,1\}^S$:

$$\xi(x,y) = \sum_{f\in F} \varphi_f(x_f, y_{S_f}) \tag{8}$$

Thus, we obtain for any $\theta \in \Theta$, any $x \in X$ and any $y \in Y$:

$$\begin{aligned} H_\theta(x,y) &= \sum_{f\in F} h_{f\theta}(x_f, y_{S_f}) = \sum_{f\in F} \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle = \left\langle \theta, \sum_{f\in F} \varphi_f(x_f, y_{S_f}) \right\rangle \\ &= \langle \theta, \xi(x,y) \rangle \end{aligned} \tag{9}$$

- Let $\mathcal{X}$ be a random variable whose value is an element $x \in X$ of the attribute space.
- Let $\mathcal{Y}$ be a random variable whose value is a combination of decisions $y \in \{0,1\}^S$
- For any $j \in J$, let Θ_j a random variable whose value is a parameter $\theta_j \in \mathbb{R}$
- We assume:

$$P(\mathcal{X}, \mathcal{Y}, \Theta) = P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X}) \prod_{j \in J} P(\Theta_j) \tag{10}$$

Thus:

$$\begin{aligned} P(\Theta \mid \mathcal{X}, \mathcal{Y}) &= \frac{P(\mathcal{X}, \mathcal{Y}, \Theta)}{P(\mathcal{X}, \mathcal{Y})} \\ &= \frac{P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X}) \prod_{j \in J} P(\Theta_j)}{P(\mathcal{X}, \mathcal{Y})} \\ &\propto P(\mathcal{Y} \mid \mathcal{X}, \Theta) \prod_{j \in J} P(\Theta_j) \end{aligned} \tag{11}$$

Definition 5. For any conditional graphical model, the **partition function** $Z\colon X \times \Theta \to \mathbb{R}$ and **Gibbs distribution** $p\colon X \times \{0,1\}^S \times \Theta \to [0,1]$ are defined by the forms

$$Z(x,\theta) = \sum_{y \in \{0,1\}^S} e^{-H_\theta(x,y)} \tag{12}$$

$$p(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \tag{13}$$

We consider a $\sigma \in \mathbb{R}^+$ and

$$p_{\mathcal{Y}|\mathcal{X},\Theta}(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \tag{14}$$

$$\forall j \in J\colon \qquad p_{\Theta_j}(\theta_j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_j^2/2\sigma^2} \quad . \tag{15}$$

Lemma 1. Estimating maximally probable parameters θ, given attributes x and decisions y, i.e.,

$$\operatorname*{argmax}_{\theta \in \mathbb{R}^J} \quad p_{\Theta|\mathcal{X},\mathcal{Y}}(\theta, x, y)$$

is an instance of supervised structured learning with L, R and λ such that

$$L(H_\theta(x, \cdot), y) = H_\theta(x, y) + \ln Z(x, \theta) \tag{16}$$

$$= H_\theta(x, y) + \ln \sum_{y' \in \{0,1\}^S} e^{-H_\theta(x, y')} \tag{17}$$

$$= \langle \theta, \xi(x, y) \rangle + \ln \sum_{y' \in \{0,1\}^S} e^{-\langle \theta, \xi(x, y') \rangle} \tag{18}$$

$$R(\theta) = \|\theta\|_2^2 \tag{19}$$

$$\lambda = \frac{1}{2\sigma^2} \tag{20}$$

Lemma 2. The first and second partial derivatives of the logarithm of the partition function have the forms

$$\frac{\partial}{\partial \theta_j} \ln Z = \frac{1}{Z(x,\theta)} \sum_{y' \in \{0,1\}^S} (-\xi_j(x,y')) e^{-\langle \theta, \xi(x,y') \rangle} \tag{21}$$

$$= \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}} (-\xi_j(x,y')) \tag{22}$$

$$\frac{\partial^2}{\partial \theta_j \, \partial \theta_k} \ln Z = \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}} (\xi_j(x,y') \xi_k(x,y'))$$

$$- \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}} (\xi_j(x,y')) \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}} (\xi_k(x,y'))$$

$$= \mathsf{COV}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}} (\xi_j(x,y'), \xi_k(x,y')) \tag{23}$$

Lemma 3. Supervised structured learning of a conditional graphical model is a convex optimization problem.

Lemma 4. Estimating maximally probable decisions y, given attributes x and parameters θ, i.e.

$$\underset{y \in \{0,1\}^S}{\operatorname{argmax}} \quad p_{\mathcal{Y}|\mathcal{X},\Theta}(x, y, \theta) \tag{24}$$

is an instance of structured inference with $\hat{H}(x, y) = H_\theta(x, y)$. Moreover, any solution $\hat{y}$ is such that

$$\hat{H}(x, \hat{y}) = \min \{\hat{H}(x, y) \mid y \in Y\} \ . \tag{25}$$