Machine Learning I

Bjoern Andres

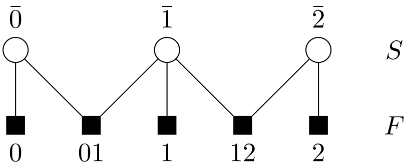
Machine Learning for Computer Vision
TU Dresden

Contents. This part of the course is about supervised structured learning of conditional graphical models.

Definition. 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

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

Example: 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. A tuple $(S, F, E, \{X_f\}_{f \in F}, \Theta, \{h_f\}_{f \in F})$ is called a **conditional graphical model** with attribute space $X := \prod_{f \in F} X_f$ and parameter space Θ iff the following conditions hold:

- (S, F, E) is a factor graph
- $\Theta \neq \emptyset$
- For every $f \in F$:
  - X_f is non-empty, called the **attribute space** of f
  - $h_f : \Theta \to \mathbb{R}^{X_f \times \{0,1\}^{S_f}}$, called a **factor**.

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

$$\forall \theta \in \Theta \ \forall x \in X \ \forall y \in \{0,1\}^S : \quad H_\theta(x, y) = \sum_{f \in F} h_{f\theta}(x_f, y_{S_f}) \qquad (2)$$

is called the family of **energy functions** of the conditional graphical model.

Family of Functions

- 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 $\theta \in \Theta$:

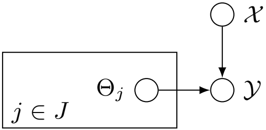
$$h_{f\theta}(x_f, y_{S_f}) = \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle \tag{3}$$

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

$$\xi(x, y) = \sum_{f \in F} \varphi_f(x_f, y_{S_f}) \tag{4}$$

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

$$\begin{aligned} H_\theta(x, y) &= \sum_{f \in F} h_{f\theta}(x_f, y_{S_f}) \\ &= \sum_{f \in F} \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle \\ &= \left\langle \theta, \sum_{f \in F} \varphi_f(x_f, y_{S_f}) \right\rangle \\ &= \langle \theta, \xi(x, y) \rangle \end{aligned} \tag{5}$$



Probabilistic Model

- 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 $\theta_j \in \mathbb{R}$

Factorization

- We assume:

$$P(\mathcal{X}, \mathcal{Y}, \Theta) = P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X}) \prod_{j \in J} P(\Theta_j) \tag{6}$$

- Thus:

$$\begin{aligned} P(\Theta \mid \mathcal{X}, \mathcal{Y}) &= \frac{P(\mathcal{X}, \mathcal{Y}, \Theta)}{P(\mathcal{X}, \mathcal{Y})} \\ &= \frac{P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X})\, \prod_{j \in J} P(\Theta_j)}{P(\mathcal{X}, \mathcal{Y})} \\ &\propto P(\mathcal{Y} \mid \mathcal{X}, \Theta) \prod_{j \in J} P(\Theta_j) \end{aligned} \tag{7}$$

Distributions

Definition. 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

$$Z(x,\theta) = \sum_{y \in \{0,1\}^S} e^{-H_\theta(x,y)} \tag{8}$$

$$p(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \tag{9}$$

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

$$p_{\mathcal{Y}|\mathcal{X},\Theta}(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \tag{10}$$

$$\forall j \in J\colon \qquad p_{\Theta_j}(\theta_j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_j^2/2\sigma^2} \quad . \tag{11}$$

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

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

is identical to the supervised structured learning problem w.r.t. L, R and λ such that

$$L(H_\theta(x,\cdot), y) = H_\theta(x,y) + \ln Z(x,\theta) \tag{12}$$

$$= H_\theta(x,y) + \ln \sum_{y' \in \{0,1\}^S} e^{-H_\theta(x,y')} \tag{13}$$

$$= \langle \theta, \xi(x,y) \rangle + \ln \sum_{y' \in \{0,1\}^S} e^{-\langle \theta, \xi(x,y') \rangle} \tag{14}$$

$$R(\theta) = \|\theta\|_2^2 \tag{15}$$

$$\lambda = \frac{1}{2\sigma^2} \tag{16}$$

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

$$\frac{\partial}{\partial \theta_j} \ln Z = \frac{1}{Z(x,\theta)} \sum_{y' \in \{0,1\}^S} (-\xi_j(x,y')) e^{-\langle \theta, \xi(x,y') \rangle} \tag{17}$$

$$= \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}}(-\xi_j(x,y')) \tag{18}$$

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

$$= \mathsf{COV}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}}(\xi_j(x,y'), \xi_k(x,y')) \tag{19}$$

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

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

$$\underset{y \in \{0,1\}^S}{\operatorname{argmax}} \quad p_{\mathcal{Y}|\mathcal{X},\Theta}(x, y, \theta) \tag{20}$$

is identical to the structured inference problem with $\hat{H}(x, y) = H_\theta(x, y)$.

Summary. Supervised structured learning of conditional graphical models whose factors are linear functions is a convex optimization problem.