Chapter 6

Supervised structured learning

6.1 Intuition

Even in the most general learning and inference problem w.r.t. constrained data $(S, X, x, \mathcal{Y})$ we have considered so far, attributes $x_s \in X$ are defined for single elements $s \in S$ only, and solutions are such that decisions $y_s, y_{s'} \in \{0,1\}$ for distinct $s, s' \in S$ are independent unless they are tied by constraints of a feasible set $\mathcal{Y} \subset \{0,1\}^S$.

This mathematical abstraction of learning is too restrictive for certain applications. For example, consider a task where we are given a digital image and 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 may wish to learn how this increased probability depends on the contents of the image. None of the mathematical abstractions of learning we have considered so far is sufficient to express this dependency.

In order to lift this restriction, we will now define a supervised learning problem as well as an inference problem in which attributes are associated with subsets of S, and in which decisions can be tied by probabilistic dependencies. Therefor, we will introduce a family $H : \Theta \to \mathbb{R}^{X \times Y}$ of functions that quantify by $H_\theta(x, y)$ how incompatible attributes $x \in X$ are with a combination of decisions $y \in \{0,1\}^S$. We will define supervised structured learning as a problem of finding one function from this family. We will define structured inference as the problem of finding a combination of decisions $y \in \{0,1\}^S$ that minimizes $H_\theta(x, \cdot)$.

6.2 Definition

Definition 6 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 $S_f = \{s \in S \,|\, \{s, f\} \in E\}$ the set of those variable nodes that are neighbors of f.

Definition 7 A tuple $T = (S, F, E, \{X_f\}_{f \in F}, x)$ is called *unlabeled structured data* iff (S, F, E) is a factor graph, every set X_f is non-empty, called the *attribute space* of f, and $x \in \prod_{f \in F} X_f$, where the Cartesian product $\prod_{f \in F} X_f$ is called the *attribute space* of T. A tuple $(S, F, E, \{X_f\}_{f \in F}, x, y)$ is called *labeled structured data* iff $(S, F, E, \{X_f\}_{f \in F}, x)$ is unlabeled structured data, and $y \in \{0,1\}^S$.

Definition 8 For any labeled structured data $T = (S, F, E, \{X_f\}_{f \in F}, x, y)$, the attribute space $X = \prod_{f \in F} X_f$, the set $Y = \{0,1\}^S$, any $\Theta \neq \varnothing$ and family of functions $H : \Theta \to \mathbb{R}^{X \times Y}$, any $R : \Theta \to \mathbb{R}_0^+$, called a *regularizer*, any $L : \mathbb{R}^Y \times Y \to \mathbb{R}_0^+$, called a *loss function*, and any $\lambda \in \mathbb{R}_0^+$, called a *regularization parameter*, the instance of the *supervised structured learning problem*

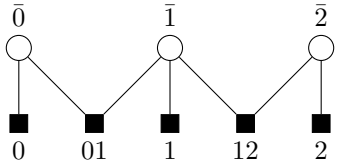


Figure 6.1: The factor graph with $S = \{\bar{0}, \bar{1}, \bar{2}\}$ and $F = \{0, 1, 2, 01, 12\}$ depicted above makes explicit that a function $H : \{0,1\}^S \to \mathbb{R}$ factorizes according to $H(y) = h_0(y_{\bar{0}}) + h_1(y_{\bar{1}}) + h_2(y_{\bar{2}}) + h_{01}(y_{\bar{0}}, y_{\bar{1}}) + h_{12}(y_{\bar{1}}, y_{\bar{2}})$.

w.r.t. T, Θ, H, R, L and λ is defined as

$$\inf_{\theta \in \Theta} \quad \lambda R(\theta) + L(H_\theta(x, \cdot), y) \tag{6.1}$$

Intuitively, H_θ is a function that quantifies by $H_\theta(x, y)$ how incompatible attributes $x \in X$ are with a combination of decisions $y \in \{0,1\}^S$. Consequently, $H_\theta(x, \cdot)$ is a functional that assigns an incompatibility to every combination of decisions.

Definition 9 For any unlabeled structured data $T = (S, F, E, \{X_f\}_{f \in F}, x)$ and any $\hat{H} \colon X \times \{0,1\}^S \to \mathbb{R}$, the instance of the *inference problem* w.r.t. T and $\hat{H}$ is defined as

$$\min_{y \in \{0,1\}^S} \quad \hat{H}(x, y) \tag{6.2}$$

6.3 Conditional graphical models

6.3.1 Data

Throughout Section 6.3, we consider labeled data $(S, F, E, \{X_f\}_{f \in F}, x, y)$ and an attribute space $X = \prod_{f \in F} X_f$ such that, for every $f \in F$, there is an $n_f \in \mathbb{N}$ and $X_f = \mathbb{R}^{n(f)}$.

6.3.2 Family of functions

Definition 10 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

$$\forall y \in \{0,1\}^S \colon \quad H(y) = \sum_{f \in F} h_f(y_{S_f}) \ . \tag{6.3}$$

An example is shown in Fig. 6.1.

Definition 11 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* $\prod_{f \in F} X_f = X$ and *parameter space* Θ iff (S, F, E) is a factor graph, $\Theta \neq \varnothing$ and, for every $f \in F$, $X_f \neq \varnothing$, called the *attribute space* of f, and $h_f : \Theta \to \mathbb{R}^{X_f \times \{0,1\}^{S_f}}$, called a *factor*.

The $H : \Theta \to \mathbb{R}^{X \times \{0,1\}^S}$ defined below is called the *energy function* of the conditional graphical model.

$$\forall \theta \in \Theta \ \forall x \in X \ \forall y \in \{0,1\}^S \colon \quad H_\theta(x, y) = \sum_{f \in F} h_{f\theta}(x_f, y_{S_f}) \tag{6.4}$$

Throughout Section 6.3, we consider such a conditional graphical model. We make two additional assumptions: Firstly, we assume that Θ is a finite-dimensional, real vector space, i.e., there exists a finite, non-empty set J and $\Theta = \mathbb{R}^J$. Secondly, 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 $\theta \in \Theta$:

$$h_{f\theta}(x_f, y_{S_f}) = \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle \tag{6.5}$$

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

$$\xi(x,y) = \sum_{f \in F} \varphi_f(x_f, y_{S_f}) \tag{6.6}$$

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

$$\begin{aligned} H_\theta(x,y) &= \sum_{f \in F} h_{f\theta}(x_f, y_{S_f}) \\ &= \sum_{f \in F} \langle \theta, \varphi_f(x_f, y_{S_f}) \rangle \\ &= \left\langle \theta, \sum_{f \in F} \varphi_f(x_f, y_{S_f}) \right\rangle \\ &= \langle \theta, \xi(x,y) \rangle \end{aligned} \tag{6.7}$$

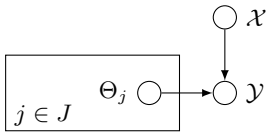
6.3.3 Probabilistic model

Random Variables

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

Conditional independence assumptions

We assume a probability distribution that factorizes according to the Bayesian net depicted below.



Factorization

- Firstly:

$$P(\mathcal{X}, \mathcal{Y}, \Theta) = P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X}) \prod_{j \in J} P(\Theta_j) \tag{6.8}$$

- Secondly:

$$\begin{aligned} P(\Theta \mid \mathcal{X}, \mathcal{Y}) &= \frac{P(\mathcal{X}, \mathcal{Y}, \Theta)}{P(\mathcal{X}, \mathcal{Y})} \\ &= \frac{P(\mathcal{Y} \mid \mathcal{X}, \Theta)\, P(\mathcal{X}) \prod_{j \in J} P(\Theta_j)}{P(\mathcal{X}, \mathcal{Y})} \\ &\propto P(\mathcal{Y} \mid \mathcal{X}, \Theta) \prod_{j \in J} P(\Theta_j) \end{aligned} \tag{6.9}$$

Forms

Definition 12 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

$$Z(x,\theta) = \sum_{y \in \{0,1\}^S} e^{-H_\theta(x,y)} \tag{6.10}$$

$$p(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \tag{6.11}$$

We consider in (6.9) the Gibbs distribution of our conditional graphical model, i.e.

$$p_{\mathcal{Y}|\mathcal{X},\Theta}(y,x,\theta) = \frac{1}{Z(x,\theta)} e^{-H_\theta(x,y)} \quad . \tag{6.12}$$

Moreover, we consider in (6.9) a $\sigma \in \mathbb{R}^+$ and, for every $j \in J$, the *normal distribution*

$$p_{\Theta_j}(\theta_j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_j^2/2\sigma^2} \quad . \tag{6.13}$$

6.3.4 Learning problem

Lemma 6 *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 identical to the supervised structured learning problem w.r.t. L, R and λ such that

$$L(H_\theta(x,\cdot),y) = H_\theta(x,y) + \ln Z(x,\theta) \tag{6.14}$$

$$= H_\theta(x,y) + \ln \sum_{y' \in \{0,1\}^S} e^{-H_\theta(x,y')} \tag{6.15}$$

$$= \langle \theta, \xi(x,y) \rangle + \ln \sum_{y' \in \{0,1\}^S} e^{-\langle \theta, \xi(x,y') \rangle} \tag{6.16}$$

$$R(\theta) = \|\theta\|_2^2 \tag{6.17}$$

$$\lambda = \frac{1}{2\sigma^2} \tag{6.18}$$

Exercise 2 *Prove Lemma 6.*

Lemma 7 *The first and second partial derivatives of the logarithm of the partition function have the forms*

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

$$= \mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}}(-\xi_j(x,y')) \tag{6.19}$$

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

$$= \mathrm{COV}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}}(\xi_j(x,y'), \xi_k(x,y')) \tag{6.20}$$

Exercise 3 *Prove Lemma 7.*

Lemma 8 *Supervised structured learning of a conditional graphical model is a convex optimization problem.*

Exercise 4 *Prove Lemma 8 using* (6.20).

6.3.5 Inference problem

Lemma 9 *Estimating maximally probable decisions y, given attributes x and parameters θ, i.e.*

$$\operatorname*{argmax}_{y \in \{0,1\}^S} \quad p_{\mathcal{Y}|\mathcal{X},\Theta}(x, y, \theta) \tag{6.21}$$

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

Exercise 5 *Prove Lemma 9.*