PGM Scribe Note: Exponential Family 3

Xiaolong Li

October 24, 2012

1 Mean Parameter

Consider an exponential family parameterized by θ:

$$p(x, \theta) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}.$$

Define *mean parameter* as the expectation of the sufficient statistics, $\mu_\alpha = \mathbb{E}_X[\phi_\alpha(x)]$, where $\alpha \in I$ and $|I| = d$ is the index set. More over, we can define the mean parameter space as the set of all realizable mean parameters

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_{\mathbb{P}}[\phi(x)] \text{ for some distribution } \mathbb{P} \text{ over } x\}.$$

Proposition 1. *$\mathcal{M}$ is a convex set.*

Proof. Let $\mu_1 = \mathbb{E}_{\mathbb{P}_1}[\phi(x)]$, $\mu_2 = \mathbb{E}_{\mathbb{P}_2}[\phi(x)]$. Since for any $0 \leq \lambda \leq 1$, the convex combination $\lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2$ of two probability measure $\mathbb{P}_1$ and $\mathbb{P}_2$ is still a valid probability measure, and by linearility of measure

$$\lambda\mu_1 + (1-\lambda)\mu_2 = \lambda\mathbb{E}_{\mathbb{P}_1}[\phi(x)] + (1-\lambda)\mathbb{E}_{\mathbb{P}_2}[\phi(x)] = \mathbb{E}_{\lambda\mathbb{P}_1 + (1-\lambda)\mathbb{P}_2}[\phi(x)]$$

This proves the claim. □

Proposition 2. *Suppose $X = (X_1, \ldots, X_p)$, $X_i \in \mathcal{X}$ where $\mathcal{X}$ is a finite set, then $\mathcal{M}$ is a polytope.*

Proof. $\mu \in \mathcal{M}$ if and only if $\mu = \sum_{x \in \mathcal{X}^p} p(x)\phi(x)$, where $p(x)$ is distribution over finite set $\mathcal{X}^p$. Thus, by definition, $\mathcal{M}$ is the convex hull of finite set $\{\phi(x_1), \ldots, \phi(x_n)\}$, $n = |\mathcal{X}|^p$. Since the convex hull of finite many points is always a polytope, we have proved the claim. □

As a remark, there are two ways to characterize a polytope. One way, called V-representation, is to characterize it as the convex hull of finite many points; another way, called H-representation, is to characterize it as the intersection of finite many closed halfspaces. Note that for H-representation, a polytope might not be bounded. Therefore, if we want to make it equivalent to V-representation, an additional boundedness condition is necessary.

Consider $|\mathcal{X}| = 2$, then $\mathcal{M}$ has up to 2^p vertices and it is forbidding to use V-representation. However, it might require far less number of halfsplaces if we use H-representation. We know that if $\mathcal{M}$ can be represented efficiently, the inference will also be efficient.

(Potential project topic: In general, r half spaces could create exponential vertices. But what about the special case for tree? Is it efficiently V-representable?)

2 Examples

2.1 Multi-variate Gaussian

The sufficient statistics is X and XX^T, then

$$\mathcal{M} = \{(\mu, \Sigma) \mid \mu = \mathbb{E}_{\mathbb{P}}[X], \Sigma = \mathbb{E}_{\mathbb{P}}[XX^T] \text{ for some distribution } \mathbb{P} \text{ over } x\}.$$

(μ, Σ) is a valid parameter for some Gaussian distribution if and only if $\Sigma \succeq \mu\mu^T$, which characterizes the shape of $\mathcal{M}$. [1]

2.2 Ising Model

Suppose $X = (X_1, X_2, \ldots, X_p)$, $X_s \in \{0, 1\}$.[2] The sufficient statistics is X_s and $X_s X_t$ for $(s, t) \in E$, so the mean parameter is

$$\mu_s = \mathbb{E}[X_s] = \mathbb{P}(X_s = 1) \in [0, 1]$$

and

$$\mu_{st} = \mathbb{E}[X_s, X_t] = \mathbb{P}(X_s = 1, X_t = 1) \in [0, 1].$$

There are many relationships we can find for the mean parameter, for exmaple

$$\mu_{st} = \mathbb{P}(X_s = 1, X_t = 1) \leq \max\{\mathbb{P}(X_s = 1), \mathbb{P}(X_t = 1)\} = \max\{\mu_s, \mu_t\}$$

[1] We write $A \succeq B$ if $A - B$ is positive semi-definite.

[2] In last class, we said $X_s \in \{-1, 1\}$, but its all the same after linear transformation.

and many others. (But the relationship we mentioned in the class, which is $\sum_t \mu_{st} = \mu_s$, doesn't make sense to me. Any comments?)

However, it is not easy to enumerate all constaints that could fully characterize $\mathcal{M}$. In fact, the polytope

$$\mathcal{M} = \left\{ \{\mu_s\}_{s\in V} \{\mu_{st}\}_{(s,t)\in E} \mid \mu_s = \mathbb{E}_{\mathbb{P}}[X_s]\ \mu_{st} = \mathbb{E}_{\mathbb{P}}[X_s X_t] \text{ for some distribution } \mathbb{P} \right\}$$

is called a cut/correlation polytope and it has exponentially many halfspaces. It is an important topic in combinatorics and people don't fully understand it even now.

2.3 Over Complete Family

Suppose $X = (X_1, X_2, \ldots, X_p)$, $X_s \in \{0, 1, \ldots, r-1\}$. The sufficient statistics are $\{\mathbb{1}_{s,j}(x)\}_{s\in V}$ and $\{\mathbb{1}_{st,jk}(x)\}_{(s,t)\in E}$, so the mean parameter is

$$\mu_{s,j} = \mathbb{E}[\mathbb{1}_{s,j}(x)] = \mathbb{P}(X_s = j)$$

and

$$\mu_{st,jk} = \mathbb{E}[\mathbb{1}_{st,jk}(x)] = \mathbb{P}(X_s = j, X_t = k).$$

Since the model is over complete, we have marginalization constraints such as

$$\sum_k \mu_{st,jk} = \mu_{s,j} \text{ for all } s, t, j.$$

Due to this reason (maybe?), the corresponding polytope $\mathcal{M}$ is called the margingal polytope, which, as in the case of Ising model, could have exponential facets and is complicated in nature. The marginal polytope is sometimes denoted as $\mathcal{M}(G)$, where $G = (V, E)$ is the underlining pairwise constaint.

3 Mean Parameter in Inference

Consider an exponential family

$$p(x, \theta) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}.$$

where $X = (X_1, \ldots, X_p)$ and $\{\phi_\alpha(x)\}_{\alpha\in I}$ are given. Suppose we have n independent samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$. To find θ, the traditional way is to maximize the log-likelihood

$$\max_{\theta\in\Omega} \sum_{i=1}^{n} \log p_\theta(x^{(i)}).$$

It turns out that for exponential family, it has nice form (since log of exp is just linear)

$$\log p_\theta(x^{(i)}) = \langle \theta, \phi(x^{(i)}) \rangle - A(\theta),$$

$$\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(x^{(i)}) = \langle \theta, \frac{1}{n}\sum_{i=1}^{n} \phi(x^{(i)}) \rangle - A(\theta).$$

Let $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \phi(x^{(i)})$, then the maximization could be written as

$$\max_{\theta \in \Omega} \langle \theta, \hat{\mu} \rangle - A(\theta).$$

Notice that $\hat{\mu}$ is convex combination of $\phi(x^{(i)})$, so $\hat{\mu} \in \mathcal{M}$.

4 Properties of $A(\theta)$

(a) $A(\theta)$ is a convex function.

(b) $A(\theta)$ is strictly convex if the exponential family is minimal.

(c) $\nabla A(\theta) = \mathbb{E}[\phi(x)]$, $\nabla^2 A(\theta) = \mathrm{Cov}(\phi(x))$.

We have already proved (c) in previous class. (a) and (b) are important since it guarantees the convexity of log-likelihood. To prove (a) assuming it is twice-differentiable, just notice $\nabla^2 A(\theta) = \mathrm{Cov}[\phi(x)] \succeq 0$.

Proof of (b). If the exponential family is minimal, then $\langle a, \phi(x) \rangle$ is not constant for all a. If X is discrete, then this implies that $\mathrm{Var}(\langle a, \phi(x) \rangle) > 0$. If X is continuous and $\phi(x)$ is continuous, then it also implies that $\mathrm{Var}(\langle a, \phi(x) \rangle) > 0$. The last step is to notice that $a^T \nabla^2 A(\theta) a = \mathrm{Var}(\langle a, \phi(x) \rangle)$. □