PGM Class Notes

Sanmit Narvekar
sanmit@cs.utexas.edu

October 22, 2012

1 Entropy

Suppose we have some random variables $X = (x_1, \ldots, x_p)$ and functions ϕ_α such that:

$$\mathbb{E}[\phi_\alpha(X)] = \mu_\alpha$$

where $\alpha \in I$, where I is the index set. Two questions we can ask are (1) what are the distributions $p(x)$ that satisfy these constraints, and (2) which distribution among these maximizes the entropy?

Recall that entropy (lack of information) is defined as:

$$\begin{aligned} H(p) &= -\int p(x) \log p(x) dx && (1) \\ &= -\sum_x p(x) \log p(x) && (2) \end{aligned}$$

Thus, we want to find

$$\max_{p \in \mathcal{P}} -\sum_x p(x) \log p(x) \tag{3}$$

such that

$$\sum_x p(x) \phi_x(x) = \mu_x$$

$$\sum_x p(x) = -1$$

$$p(x) \geq 0$$

Since we have constraints, we optimize by taking the Lagrangian

$$\max_{p\in\mathcal{P}} H(p) = -\sum_x p(x)\log p(x) + \sum_\alpha \lambda_\alpha[\sum_x p(x)\phi_\alpha(x) - \mu_\alpha] + \bar{\lambda}(\sum_x p(x) - 1) \tag{4}$$

$$= -\log p(x) - 1 + \sum_\alpha \lambda_\alpha \phi_\alpha(x) + \bar{\lambda} \tag{5}$$

$$= 0 \tag{6}$$

Thus

$$p(x) = \exp(\bar{\lambda} - 1)\exp(\sum_\alpha \lambda_\alpha \phi_\alpha(x)) \tag{7}$$

$$\propto \exp(\sum_\alpha \lambda_\alpha \phi_\alpha(x)) \tag{8}$$

2 Notation and Definitions

We now introduce some notation and definitions for the models we will discuss in the following sections.

We will denote random variables $X = (X_1, X_2, \ldots, X_p)$ where $X_s \in \mathcal{X}$ (some set of values). Additionally, we will define sufficient statistics $\phi_\alpha : X^p \mapsto \mathbb{R}$, where $\alpha \in I$ and $|I| = d$. $\theta_\alpha \in \mathbb{R}$ is associated with each ϕ_α. Then, we can write the exponential family as

$$p(x) = \exp\left\{\sum_\alpha \theta_\alpha \phi_\alpha(x) - A(\theta)\right\} \tag{9}$$

Defining

$$< \theta, \phi(x) >= \sum_\alpha \theta_\alpha \phi_\alpha(x) \tag{10}$$

gives us

$$A(\theta) = \log \int \exp\{< \theta, \phi(x) > dx\} \tag{11}$$

$$\text{OR} \quad \log \sum_x \exp < \theta, \phi(x) > \tag{12}$$

The domain of this function is

$$\Omega = \{\theta \in \mathbb{R}^d; \quad A(\theta) < \infty\}$$

Minimal Exponential Families

An exponential family is minimal if and only if $\nexists \alpha \in \mathbb{R}^d$ such that $< \theta, \phi(x) > \quad = b$ (b is a constant). That is, there doesn't exist a linear combination of sufficient statistics.

Proof: (I missed some parts here...)

Suppose $\exists a \in \mathbb{R}^d$ such that $< a, \phi(x) >= b$

$$\theta \in \mathbb{R}^d$$

$$\begin{aligned} p_{\theta+a}(x) &\propto \exp\{< \theta + a, \phi(x)\} && (13) \\ &\propto \ldots && (14) \end{aligned}$$

3 Ising Model

This comes up frequently when modeling spin in electrons. Consider again r.v.'s $X = (X_1, \ldots, X_p)$ where $X_s \in \{-1, 1\}$. So, for example, we could be interested in the joint distribution of spin in all molecules.

Consider a grid of atoms (?), where each node is represented by a r.v. representing the spin. Our sufficient statistics are

$$\{X_s, \quad s \in \mathcal{V}\} = \{\theta_s\}$$

$$\{X_s X_t, \quad (s,t) \in \mathcal{E}\} = \{\theta_{st}\}$$

We can write this as an exponential family

$$p(x) = \exp\left\{\sum_{s \in \mathcal{V}} \theta_s X_s + \sum_{(s,t) \in \mathcal{E}} \theta_{st} X_s X_t - A(\theta)\right\} \tag{15}$$

and the domain is

$$\Omega = \mathbb{R}^d$$

$$d = p + |\mathcal{E}|$$

You can verify this is minimal. We can also make higher order generalizations by considering $X_s X_t X_k$, etc.

4 Overcomplete Exponential Family

Consider again the r.v.'s $X = (X_1, \dots, X_p)$, this time drawn from a finite discrete set $X_s \in \mathcal{X} = \{0, 1, \dots, r-1\}$.

Our sufficient statistics are the indicator functions

$$I_{s,j}(x) = \begin{cases} 1 & \text{if } X_s = j, \\ 0 & \text{otherwise} \end{cases} \equiv \{\theta_{s,j}\}_{s \in \mathcal{V}, j \in \mathcal{X}} \tag{16}$$

$$I_{st,jk}(x) = \begin{cases} 1 & \text{if } X_s = j, X_t = k \\ 0 & \text{otherwise} \end{cases} \equiv \{\theta_{st,jk}\}_{(s,t) \in \mathcal{E}; j,k \in \mathcal{X}} \tag{17}$$

Here, the number of parameters $d = pr + |\mathcal{E}|r^2$, and the domain is $\Omega = \mathbb{R}^d$. Note that this is not minimal.

4.1 Subclasses

Comes up in metric labeling. Consider you have a metric ρ over $\mathcal{X}$, where

$$\rho(j,j) = 0$$
$$\rho(j,k) \geq 0$$
$$\rho(j,l) \leq \rho(j,k) + \rho(k,l)$$

We can write the metric weights (?)

$$\theta_{st,jk} = -\rho(j,k)$$

(I missed some parts over here)

This model doesn't like neighboring nodes to take values that are far apart (it does this by giving those assignments lower probabilities). This is used in computer vision, e.g. for segmentation.

5 Gaussian

Consider again the r.v.'s $X = (X_1, \dots, X_p)$, drawn from $\mathbb{R}^p$. The sufficient statistics are:

$$\{X_s, X_s^2; s \in \mathcal{V}\} \equiv \theta_s = \Theta_{ss}$$
$$\{X_s X_t; (s,t) \in \mathcal{E}\} \equiv \Theta_{st}$$

where Θ is the p x p negative inverse of the covariance matrix:

$$\Theta = -\Sigma^{-1}$$

$$\theta = -\Sigma^{-1}\mu$$

If an entry (i, j) in Θ is nonzero, there is an edge between node i and j. We can write the exponential family as

$$p(x) = \exp\left\{\sum_s \theta_s X_s + \sum_{(s,t)\in\mathcal{E}} \Theta_{st} X_s X_t - A(\theta, \Theta)\right\} \tag{18}$$

This is minimal, and the domain is

$$\Omega = \{(\theta, \Theta) \quad ; \theta \in \mathbb{R}^p, \Theta \leq 0\}$$

6 Mixture Models

Consider the r.v.'s $X = (X_1, \ldots, X_p)$, drawn from a finite discrete set $X_s \in \mathcal{X} = \{0, 1, \ldots, r-1\}$. X_s is unobserved, and Y_s is observed.

$$\begin{aligned} p(X_s) &\propto \exp\left\{\sum_j \theta_{s,j} \mathbb{I}_j(X_s)\right\} && (19) \\ &\equiv p(X_s = j) && (20) \\ &\propto \exp(\theta_{sj}) && (21) \end{aligned}$$

(I missed some more parts over here. Also can't tell if the $\mathbb{I}$ is supposed to be a $\prod$)

$$Y_s | X_s = j \sim \mathcal{N}(\ldots)$$
$$p(Y_s | X_s = j) \propto \exp \gamma_{sj} y_s + \gamma'_{sj} y_s^2$$
$$p(Y_s | X_s) \propto \exp\left\{\sum_j \mathbb{I}_j(X_s)(\gamma_{sj} y_s + \gamma'_{sj} y_s^2)\right\}$$

$$\begin{aligned} p(X_s, Y_s) &= p(X_s) p(Y_s | X_s) && (22) \\ &\propto \exp\left\{\sum_j \mathbb{I}_j(X_s)\theta_{sj} + \sum_j \mathbb{I}_j(X_s)\left[\gamma_{sj} y_s + \gamma'_{sj} y_s^2\right]\right\} && (23) \end{aligned}$$

7 Latent Dirichlet Allocation

Used to model things like documents. For example, the probability of a particular word given a topic $p(W = j|Z = k) = \gamma_{jk}$. Random variable U is used to determine the topic Z.

$$p(W|Z) \propto \exp\left\{\sum_{jk} \mathbb{I}_j(w)\mathbb{I}_k(z)\theta_{jk}\right\} \tag{24}$$

$$p(Z = k|U) = U_k \tag{25}$$

$$p(Z|U) = \exp\left\{\sum_{k} \mathbb{I}_k(z)\log U_k\right\} \tag{26}$$

$$p(U) \propto \exp\left\{\sum_{k} \alpha_k \log U_k\right\} \tag{27}$$