CS761 Spring 2013 Advanced Machine Learning

Exponential Families and Graphical Models

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

1 The Maximum Entropy Principle

Consider the *maximum entropy* problem: Given iid items $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim p^*$, let

$$\hat{\mu}_j \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_j(\mathbf{x}_i), \tag{1}$$

where $\phi_j : \mathcal{X} \mapsto \mathbb{R}$ is some function for $j = 1 \ldots d$. The value $\hat{\mu}_j$ is the empirical expectation of ϕ_j.

Can we recover p^* from $\hat{\mu}_1 \ldots \hat{\mu}_d$? Consider density p absolutely continuous w.r.t. a base measure ν, which usually is the counting measure for probability mass function p, or the Lebesgue measure on $\mathbb{R}$ for continuous p. We say p is consistent with the data, if

$$\mathbb{E}_p[\phi_j(\mathbf{x})] \equiv \int_{\mathcal{X}} \phi_j(x) p(x) \nu(dx) = \hat{\mu}_j \text{ for } j \in \{1 \ldots d\}. \tag{2}$$

In general (when d is small), there will be many distributions that are consistent with the data. The maximum entropy principle is to pick the distribution

$$\hat{p} = \text{argmax}_p \qquad \int_{\mathcal{X}} -p(x) \log p(x) \nu(dx) \tag{3}$$

$$\text{s.t.} \qquad \mathbb{E}_p[\phi_j(\mathbf{x})] = \hat{\mu}_j \text{ for all } j = 1 \ldots d. \tag{4}$$

On one hand, there is no guarantee that $\hat{p} = p^*$. On the other hand, $\hat{p}$ has a very interesting form:

$$\hat{p}(\mathbf{x}) \propto \exp\left(\sum_{j=1}^{d} \theta_j \phi_j(\mathbf{x})\right) \tag{5}$$

with parameters $\theta_j \in \mathbb{R}$.

2 Exponential Families

The solution above is in the so called exponential families. Formally, let $\phi = (\phi_1, \ldots, \phi_d)^\top$ be d *sufficient statistics*, where $\phi_i : \mathcal{X} \mapsto \mathbb{R}$. Note X here in general is a high dimensional object itself, corresponding to all the nodes in a Graphical model. Let $\theta = (\theta_1, \ldots, \theta_d)^\top$ be an associated *canonical parameters*. The *exponential family* is a family of probability densities:

$$p_\theta(\mathbf{x}) = \exp\left(\theta^\top \phi(\mathbf{x}) - A(\theta)\right) \tag{6}$$

★ *The key is the linear interaction (inner product) between parameters θ and sufficient statistics ϕ.*

A is the log partition function

$$A(\theta) = \log \int \exp\left(\theta^\top \phi(\mathbf{x})\right) \nu(d\mathbf{x}). \tag{7}$$

★ $A = \log Z$.

Define

$$\Omega = \{\theta \in \mathbb{R}^d \mid A(\theta) < \infty\} \tag{8}$$

i.e., those parameters for which the density is normalizable. A *regular* exponential family is where Ω is an open set. A *minimal* exponential family is where the ϕ's are linearly independent, namely there does not exist a nonzero $\alpha \in \mathbb{R}^d$ such that $\alpha^\top \phi(\mathbf{x}) =$ constant for all $\mathbf{x}$. If an exponential family is not minimal, it is called *overcomplete*. Both minimal and overcomplete representations are useful.

Example 1 (Bernoulli) *Let $p(x) = \beta^x(1-\beta)^{1-x}$ for $x \in \{0,1\}$ and $\beta \in (0,1)$. Although it does not look like an exponential family, one can equivalently express the density as*

$$\begin{aligned} p(x) &= \exp\left(x \log \beta + (1-x)\log(1-\beta)\right) &(9)\\ &= \exp\left(x\theta - \log(1+\exp(\theta))\right), &(10)\end{aligned}$$

where $\theta = \log \frac{\beta}{1-\beta}$. Note (9) is in exponential family form with $\phi_1(x) = x, \phi_2(x) = 1-x$, $\theta_1 = \log\beta, \theta_2 = \log(1-\beta)$, and $A(\theta) = 0$. Further note that $\alpha_1 = \alpha_2 = 1$ makes $\alpha^\top\phi(x) = 1$ for all x, thus (9) is an overcomplete representation. In contrast, (10) is a minimal exponential family with $\phi(x) = x, \theta = \log\frac{\beta}{1-\beta}, A(\theta) = \log(1+\exp(\theta))$.

Many standard distributions (e.g., Gaussian, exponential, Poisson, Beta) are in the exponential family, see standard textbooks for details. Not all familiar distributions are in the exponential family. For example, the Laplace distribution with unknown mean cannot be written as an exponential family (thanks Vincent Tan for supplying a proof).

Example 2 (Ising Model) *Let $G = (V, E)$ be an undirected graph. Each node $s \in V$ is associated with a binary random variable $x_s \in \{0,1\}$. The Ising model is defined to be*

$$p_\theta(\mathbf{x}) = \exp\left(\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - A(\theta)\right). \tag{11}$$

The sufficient statistics are $d = |V|+|E|$ dimensional: $\phi(\mathbf{x}) = (\ldots x_s \ldots x_{st} \ldots)^\top$. This is a regular ($\Omega = \mathbb{R}^d$), minimal exponential family.

Example 3 (Potts Model) *Similar to Ising model but generalizing $x_s \in \{0,\ldots,r-1\}$. Let indicator functions $f_{sj}(\mathbf{x}) = 1$ if $x_s = j$ and 0 otherwise, and $f_{stjk}(\mathbf{x}) = 1$ if $x_s = j \wedge x_t = k$, and 0 otherwise. The Potts model is defined to be*

$$p_\theta(\mathbf{x}) = \exp\left(\sum_{sj}\theta_{sj}f_{sj}(\mathbf{x}) + \sum_{stjk}\theta_{stjk}f_{stjk}(\mathbf{x}) - A(\theta)\right). \tag{12}$$

Now $d = r|V| + r^2|E|$. This is regular but overcomplete, because $\sum_{j=0}^{r-1}\theta_{sj}(\mathbf{x}) = 1$ for any $s \in V$ and all $\mathbf{x}$. The Potts model is a special case where the parameters are tied: $\theta_{stkk} = \alpha$, and $\theta_{stjk} = \beta$ for $j \neq k$.

3 Mean Parametrization and Marginal Polytopes

Let p be any density (not necessarily in exponential family). Given sufficient statistics ϕ, the *mean parameters* $\mu = (\mu_1, \ldots, \mu_d)^\top$ is defined as

$$\mu_i = \mathbb{E}_p[\phi_i(\mathbf{x})] = \int \phi_i(\mathbf{x})p(\mathbf{x})\nu(dx). \tag{13}$$

Consider the set of mean parameters realized by *any* distribution (not necessarily exponential family):

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } \mathbb{E}_p[\phi(\mathbf{x})] = \mu\}. \tag{14}$$

Example 4 (The First Two Moments) *Let $\phi_1(x) = x, \phi_2(x) = x^2$. For any p (NB not necessarily Gaussian), the mean parameters $\mu = (\mu_1, \mu_2) = (\mathbb{E}(x), \mathbb{E}(x^2))^\top$. Since $\mathbb{V}(x) = \mathbb{E}(x^2) - \mathbb{E}^2(x) = \mu_2 - \mu_1^2 \geq 0$ for any p, we see that $\mathcal{M}$ is not $\mathbb{R}^2$ but rather the subset defined by $\mu_1 \in \mathbb{R}, \mu_2 \geq \mu_1^2$.*

★ *It is helpful to think about $p = uniform(-a, a)$.*

$\mathcal{M}$ is a convex subset of $\mathbb{R}^d$, since if $\mu^{(1)}, \mu^{(2)} \in \mathcal{M}$ there must exist $p^{(1)}, p^{(2)}$, and the convex combinations of $p^{(1)}, p^{(2)}$ realizes the convex combinations of $\mu^{(1)}, \mu^{(2)}$.

The *marginal polytope* is defined for any graphical model with multinomial random variables $x_s \in \{0, 1, \ldots, r-1\}$ (this can be generalizes so different x_s have different r_s). Consider the indicator functions defined in Example 3, which we call the standard overcomplete representation. The mean parameters has an intuitive interpretation:

$$\mu_{sj} = \mathbb{E}_p[f_{sj}(\mathbf{x})] = p(x_s = j) \text{ (node marginals)} \quad (15)$$

$$\mu_{stjk} = \mathbb{E}_p[f_{stjk}(\mathbf{x})] = p(x_s = j, x_t = k) \text{ (edge marginals)} \quad (16)$$

Furthermore, because

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \mu = \sum_{\mathbf{x}} \phi(\mathbf{x}) p(\mathbf{x}) \text{ for some } p\} \quad (17)$$

it is easy to see that $\mathcal{M} = conv\{\phi(\mathbf{x}), \forall \mathbf{x}\}$, i.e., the convex hull of point mass distributions that put mass on a single $\mathbf{x}$. This convex hall is called the marginal polytope.

Example 5 (The Marginal Polytope of a Tiny Ising Model) *Consider a tiny Ising model with two nodes $x_1, x_2 \in \{0, 1\}$ and an edge between them. The Ising model minimal representation is $\phi(x_1, x_2) = (x_1, x_2, x_1 x_2)^\top$. Note that there are only 4 different $\mathbf{x} = (x_1, x_2)$. Therefore, the marginal polytope is defined by the convex hull of*

$$\mathcal{M} = conv\{(0,0,0), (0,1,0), (1,0,0), (1,1,1)\}, \quad (18)$$

which is a polytope inside the unit cube. The three coordinates can be interpreted as $\mu_1 \equiv \mathbb{E}_p[x_1 = 1]$, $\mu_2 \equiv \mathbb{E}_p[x_2 = 1]$, $\mu_{12} \equiv \mathbb{E}_p[x_1 = x_2 = 1]$.

4 The Log Partition Function A

For any regular exponential family, $A(\theta)$ is convex in θ. It is strictly convex:

$$A(\lambda\theta^1 + (1-\lambda)\theta^2) < \lambda A(\theta^1) + (1-\lambda)A(\theta^2), \ \forall \lambda \in (0,1), \theta^1 \neq \theta^2, \quad (19)$$

if the representation is minimal.

The gradient of A has a special property:

$$\frac{\partial A(\theta)}{\partial \theta_i} = \mathbb{E}_\theta[\phi_i(\mathbf{x})], \quad (20)$$

and therefore $\nabla A = \mu$ the mean parameters of p_θ. This can be viewed as a *forward mapping* from θ to μ.

★ *Note that this includes the inference problem (e.g., node marginals), and therefore computing the forward mapping would not be easy.*

Recall that $\mathcal{M}$ is defined by *all* distributions p, not limited to exponential family. One might therefore think that ∇A only covers a subset of $\mathcal{M}$. Remarkably, there are two interesting properties:

- $\nabla A : \Omega \mapsto \mathcal{M}$ is one-to-one iff the exponential representation is minimal.
- For minimal representation, ∇A is onto the interior $\mathcal{M}^0$ of $\mathcal{M}$. For each $\mu \in \mathcal{M}^0$, there exists some $\theta(\mu) \in \Omega$ such that $\mathbb{E}_{\theta(\mu)}[\phi(\mathbf{x})] = \mu$.

Because ∇A covers $\mathcal{M}^0$, and because the exponential family is a strict subset of all distributions, each μ is realized by other distributions not in the exponential family. Among all distributions realizing μ, $p_{\theta(\mu)}$ has the maximum entropy.

5 Conjugate Duality

The *conjugate dual function* A^* to A is defined as

$$A^*(\mu) = \sup_{\theta \in \Omega} \theta^\top \mu - A(\theta). \tag{21}$$

Such definition, where a quantity is expressed as the solution to an optimization problem, is called a *variational* definition.

The dual function is closely related to entropy. For any $\mu \in \mathcal{M}^0$, let $\theta(\mu)$ be the unique canonical parameter satisfying the *dual matching* condition:

$$\mathbb{E}_{\theta(\mu)}[\phi(\mathbf{x})] = \nabla A(\theta(\mu)) = \mu. \tag{22}$$

A^* takes the form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \mu \in \mathcal{M}^0 \\ \infty & \mu \notin \bar{\mathcal{M}}, \end{cases} \tag{23}$$

where $\bar{\mathcal{M}}$ is the closure of $\mathcal{M}$. For any boundary point $\mu \in \bar{\mathcal{M}} \setminus \mathcal{M}^0$ we have $A^*(\mu) = \lim_{n\to\infty} A^*(\mu^n)$ taken over any sequence $\{\mu^n\} \subset \mathcal{M}^0$ converging to μ.

The fact that $A^*(\mu) = \infty$ if $\mu \notin \bar{\mathcal{M}}$ means that optimizing A^* only needs to be carried over $\mathcal{M}$.

The dual of the dual gives back A:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \mu^\top \theta - A^*(\mu). \tag{24}$$

For all $\theta \in \Omega$, the supremum is attained uniquely at the $\mu \in \mathcal{M}^0$ by the moment matching conditions

$$\mu = \mathbb{E}_\theta[\phi(\mathbf{x})]. \tag{25}$$

★ *This is a very important observation. It is the foundation of variational inference. Do you want to compute the marginals? Note they are the mean parameters under standard overcomplete representation. Do you want to compute the mean parameters? Note by* (25) *they are the solution to optimization problem* (24)*. That is, even if you don't care about $A(\theta)$, solving $\sup_{\mu\in\mathcal{M}} \mu^\top\theta - A^*(\mu)$ will give you the mean parameter μ corresponding to θ, hence solving the inference problem associated with p_θ.*

Example 6 (Conjugate Duality for Bernoulli) *Recall the minimal exponential family for Bernoulli with $\phi(x) = x, A(\theta) = \log(1 + \exp(\theta)), \Omega = \mathbb{R}$. By definition*

$$A^*(\mu) = \sup_{\theta \in \mathbb{R}} \theta\mu - \log(1 + \exp(\theta)). \tag{26}$$

Taking derivatives and solve, for $\mu \in (0, 1)$ one arrives at

$$A^*(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu). \tag{27}$$

References

[1] Yasemin Altun and Alexander J. Smola. Unifying divergence minimization and statistical inference via convex duality. In Gbor Lugosi and Hans-Ulrich Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 139–153. Springer, 2006.

[2] Miroslav Dudk, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.

[3] Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.