

Lecture 6: Exponential Families

Lecturer: Sasha Rush

Scribes: Meena Jagadeesan, Yufeng Ling, Tomoka Kan, Wenting Cai

6.1 Introduction

(Wainwright and Jordan (textbook) presents a more detailed coverage of the material in this lecture.)

This lecture, we will unify all of the fundamentals presented so far:

$p(\theta)$	$p(x)$	$p(y x) / p(x, y)$
Beta, Dir	Discrete	Classification
MVN, IW	MVN	Linear Regression
	Exponential Families	Generalized Linear Models
	Undirected Graphic Models	Conditional UGM
	Variational Inference	

We will focus on coming up with a general form for Discrete and MVN through exponential families. We will also come up with a general form for classification and linear regression through generalized linear models.

6.2 Definition of Exponential Family

The definition is

$$\begin{aligned}
 p(x | \theta(\mu)) &= \frac{1}{Z(\theta)} h(x) \exp\{\theta^T \phi(x)\} \\
 &= h(x) \exp\{\theta^T \phi(x) - A(\theta)\}
 \end{aligned}$$

where

μ	mean parameters
$\theta(\mu)$	natural / canonical / exponential parameters
$Z(\theta)A(\theta)$	also written as $Z(\theta(\mu))$ or $Z(\mu)$, the partition function and log partition
$\phi(x)$	sufficient statistics of x , potential functions, “features”
$h(x)$	scaling term, in most cases, we have $h(x) = 1$

Note that there is “minimal form” and “overcomplete form”.

6.3 Examples of Exponential Families

6.3.1 Bernoulli/Categorical

First, we consider the Bernoulli as an exponential family. Like last lecture, we rewrite the distribution as an exp of log.

$$\begin{aligned}
 \text{Ber}(x|\mu) &= \mu^x (1 - \mu)^{(1-x)} \\
 &= \exp x \log \mu + (1 - x) \log(1 - \mu) \\
 &= \underbrace{\exp \log \left(\frac{\mu}{1 - \mu} \right)}_{\theta} \underbrace{x}_{\phi(x)} + \underbrace{\log(1 - \mu)}_{-A(\mu)}
 \end{aligned}$$

For the **minimal form**, we have

$$\begin{aligned}
h(x) &= 1 \\
\phi_1(x) &= x \\
\theta_1(\mu) &= \log \frac{\mu}{1-\mu} \text{ ("log odds")} \\
\mu &= \sigma(\theta) \\
A(\mu) &= -\log(1-\mu) \\
A(\theta) &= -\log(1-\sigma(\theta)) = \theta + \log(1+e^{-\theta})
\end{aligned}$$

For the **overcomplete form**, we have

$$\begin{aligned}
\phi(x) &= \begin{bmatrix} x \\ 1-x \end{bmatrix} \\
\theta &= \begin{bmatrix} \log \mu \\ \log(1-\mu) \end{bmatrix}
\end{aligned}$$

For the Categorical/Multinoulli distribution, we have

$$\theta = \begin{bmatrix} \log \mu_1 \\ \vdots \\ \log \mu_n \end{bmatrix}$$

where $\sum_c \mu_c = 1$.

Side note: Writing out in overcomplete form usually comes with some restraints.

6.3.2 Univariate Gaussians

$$\begin{aligned}
\mathcal{N}(x \mid \mu, \sigma^2) &= (2\pi\sigma^2)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\
&= \underbrace{(2\pi\sigma^2)^{-1/2}}_{A(\mu, \sigma^2)} \exp\left\{-\underbrace{\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x}_{\theta^T \phi(x)} - \underbrace{\frac{1}{2\sigma^2}\mu^2}_{A(\mu, \theta^2)}\right\}
\end{aligned}$$

$$\begin{aligned}
\phi(x) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} \\
\theta &= \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \\
A(\mu, \sigma^2) &= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mu^2 \\
\mu &= -\frac{\theta_1}{2\theta_2} \\
\sigma^2 &= -\frac{1}{2\theta_2} \\
A(\theta) &= -\frac{1}{2} \log(-2\theta_2) - \frac{\theta_1^2}{4\theta_2}
\end{aligned}$$

6.3.3 Bad distributions

Two simple distributions that do not fit this form are the uniform distribution $\text{Uniform}(0,1)$ (check this as an exercise), and the Student-T distribution.

6.4 Properties of Exponential Families

Most inference problems involve a mapping between natural parameters and mean parameters, so this is a natural framework.

Here are three properties of exponential families:

Property 1 Derivatives of $A(\theta)$ provide us the cumulants of the distribution $\mathbb{E}(\phi(x))$, $\text{var}(\phi(x))$:

Proof. For univariate, first order:

$$\begin{aligned}
 \frac{dA}{d\theta} &= \frac{d}{d\theta} (\log Z(\theta)) \\
 &= \frac{d}{d\theta} \log \left(\underbrace{\int \exp\{\theta\phi\} h(x) dx}_{\text{needed to integrate to 1}} \right) \\
 &= \frac{\int \phi \exp\{\theta\phi\} h(x) dx}{\int \exp(\theta\phi) h(z) dx} \\
 &= \frac{\int \phi \exp\{\theta\phi\} h(x) dx}{\exp(A(\theta))} \\
 &= \int \phi(x) \underbrace{\exp(\theta\phi(x) - A(\theta)) h(x)}_{p(x)} dx \\
 &= \int \phi(x) p(x) dx \\
 &= \mathbb{E}(\phi(x))
 \end{aligned}$$

The same property holds for multivariates (refer to textbook for proof). □

Bernoulli:

$$\begin{aligned}
 A(\theta) &= \theta + \log(1 + e^{-\theta}) \\
 \frac{dA}{d\theta} &= 1 - \frac{e^{-\theta}}{1 + e^{-\theta}} = \frac{1}{\underbrace{1 + e^{-\theta}}_{\text{sigmoid}}} = \sigma(\theta) = \mu
 \end{aligned}$$

Univariate Normal Left as exercise.

Property 2 MLE has a nice form (through “moment matching”)

Proof.

$$\begin{aligned}
 \underset{\theta}{\operatorname{argmax}} \log p(\text{data} \mid \theta) &= \underset{\theta}{\operatorname{argmax}} \left(\sum_d \theta^T \phi(x_d) \right) - NA(\theta) \\
 &= \underset{\theta}{\operatorname{argmax}} \theta^T \underbrace{\left(\sum_d \phi(x_d) \right)}_{\text{sum of sufficient statistics}} - \underbrace{NA(\theta)}_{\text{amount of points}}
 \end{aligned}$$

We take a derivative to obtain:

$$\begin{aligned}\frac{d(\cdot)}{d\theta} &= \sum_d \phi(x_d) - N \frac{dA(\theta)}{d\theta} \\ &= \sum_d \phi(x_d) - N \mathbb{E}(\phi(x)) \\ &= 0\end{aligned}$$

$$E(\phi(x)) = \underbrace{\frac{\sum \phi(x_d)}{N}}_{\text{set mean parameter to sample means that gives us MLE}}$$

□

Property 3 Exponential families have conjugate priors.

Proof. We first introduce some notations.

$$\begin{aligned}\eta &\text{ - parameters} \\ \bar{s} &= \sum_d \phi(x_d) / N \\ p(\text{data} \mid \eta) &\propto \exp[(N\bar{s})\eta - NA(\eta)] \\ p(\eta \mid N_0, \bar{s}_0) &\propto \exp[(N_0, \bar{s}_0)\eta - N_0 \underbrace{A(\eta)}_{\text{not log partition, which has to be a function strictly of parameters}}] \\ p(\eta \mid \text{data}) &\propto \exp((N\bar{s} + N_0\bar{s}_0)^T \eta - (N_0 + N)A(\eta))\end{aligned}$$

The above two distributions have the same sufficient statistics – so we have a conjugate prior. It also tells us that it is not a coincidence that we kept obtaining pseudo counts. (More references will be put up to describe this).

□

6.5 Definition of Generalized Linear Models

While exponential families generalize $p(x)$, GLMs generalize $p(y|x)$.

$$p(y|x, w) = h(y) \exp\{\theta(\underbrace{\mu(x)}_{\text{predict mean}})^T \phi(y) - A(\theta)\}$$

where $\mu(x) = \underbrace{g^{-1}}_{\text{squashing const}} (w^T x + b)$ where g is an appropriate linear transformation.

This can be summarized through the following sequence of transformations:

$$x \xrightarrow{g^{-1}(w^T x + b)} \mu \rightarrow \theta \rightarrow p(y \mid x).$$

6.6 Examples of Generalized Linear Models

We present three examples:

Example 1 Exponential family - Normal distribution with $\sigma^2 = 1$ and g^{-1} is the identity function. This gives us the linear regression

$$\mu = w^T x + b \quad \mathbb{R} \rightarrow \mathbb{R}.$$

Example 2 Exponential family - Bernoulli distribution and g^{-1} is the sigmoid function $\sigma : \mathbb{R} \rightarrow (0, 1)$. Now, $\mu = \sigma(w^T x + b)$ and $\theta = \log \left(\frac{\mu}{1-\mu} \right)$. This is how we define logistic regression. This gives us

$$p(y | x) = \sigma(w^T x + b)^y (1 - \sigma(w^T x + b))^{1-y}$$

Example 3 Exponential family - Categorical distribution with g^{-1} as the softmax function.

$$\begin{aligned} \mu_c &= \text{softmax}(w_c^T x + b_c)_c \\ \theta_c &= \log \mu_c \end{aligned}$$