The exponential family and generalized linear models

Yimeng Zhang

2017/03/29

mostly based on Chapter 8 of "*An introduction to probabilistic graphical models*" by Michael I. Jordan, later referred to as "the book"

Contents

- Exponential family
    - Examples
    - Properties
    - ML estimation
- Generalized linear models
    - response function

Exponential family

$$p(x \mid \eta) = h(x) \exp\left\{\eta^T T(x) - A(\eta)\right\}$$

1. transform x to $T(x)$, "sufficient statistics". $T(x)$ is a fixed function, without any parameter to fit.
2. dot product $T(x)$ with η, then take exp, thus *exponential*
3. $A(\eta)$ is just normalization. See Eq. (8.2) of the book,.
4. $h(x)$ weights different x. mostly used in Poisson, trivial in other distributions.

$$A(\eta) = \log \int h(x) \exp\{\eta^T T(x)\} dx \tag{8.2}$$

where we see that $A(\eta)$ can be viewed as the logarithm of a normalization factor. The set of η for which this integral is finite is referred to as the *natural parameter space*.

It is also common to write the exponential family distribution in the following way:

$$p(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp\{\eta^T T(x)\}, \tag{8.3}$$

which is equivalent to if we let $A(\eta) = \log Z(\eta)$. Although we focus on Eq. (8.1) throughout this chapter, we will also make use of Eq. (8.72) in later chapters.

Exp family - examples

Bernoulli distribution, pp. 4 of the book.

$$\begin{aligned} p(x \mid \pi) &= \pi^x (1-\pi)^{1-x} \\ &= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}. \end{aligned}$$

and throughout the chapter, is to take the exponen
on. Thus we see that the Bernoulli distribution is

$$\begin{aligned} \eta &= \frac{\pi}{1-\pi} \\ T(x) &= x \\ A(\eta) &= -\log(1-\pi) = \log(1+e^{\eta}) \\ h(x) &= 1. \end{aligned}$$

e relationship between η and π is invertible. Solving

$$\pi = \frac{1}{1+e^{-\eta}},$$

Exp family - examples

Poisson distribution, pp. 5 of the book.

ion of a Poisson random variable is given as follows:

$$p(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

ve obtain:

$$p(x \mid \lambda) = \frac{1}{x!} \exp\{x \log \lambda - \lambda\}.$$

ion is an exponential family distribution, with:

$$\begin{aligned} \eta &= \log \lambda \\ T(x) &= x \\ A(\eta) &= \lambda = e^\eta \\ h(x) &= \frac{1}{x!}. \end{aligned}$$

invert the relationship between η and λ:

$$\lambda = e^\eta.$$

Exp family - examples

Gaussian distribution, pp. 5 of the book.

The (univariate) Gaussian distribution can be written as follows:

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{8.18}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \ln\sigma\right\}. \tag{8.19}$$

This is in the exponential family form, with:

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \tag{8.20}$$

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{8.21}$$

$$A(\eta) = \frac{\mu^2}{2\sigma^2} + \ln\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2) \tag{8.22}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}. \tag{8.23}$$

Note in particular that the univariate Gaussian distribution is a two-parameter distribution and that its sufficient statistic is a vector.

I omit tricky examples. See `misc/exponential_families.ipynb` on https://github.com/leelabcnbc/machine-learning-notes for a more

Exp family - examples

My take away:

1. a distribution can be parameterized (specified) in unusual ways, not just mean and variance.
2. there is isomorphism between different parameterizations, it's hard to say which is more "correct" than the other. This ambiguity is still true, even if we parameterize in the form of exponential family. Say $T(x) = x$ or $T(x) = 2x$.
3. In some forms (say exponential), many distributions can be considered as connected ones, sharing many good properties

Exp family - properties

Looks that A is just some normalization factor, but it actually has everything you need about the distribution.

- $P(x \mid \eta^*)$ is scalar distribution, parameterized by a scalar η.
- the first derivative of $A(\eta)$ at η^* gives you the mean of $T(x)$, second derivative gives variance of $T(x)$.

Example: pp. 7 of the book.

Exp family - properties

Recall that in the case of the Bernoulli distribution we have $A(\eta) = \log(1 + e^{\eta})$. Taking a first derivative yields:

$$\begin{aligned}
\frac{dA}{d\eta} &= \frac{e^{\eta}}{1 + e^{\eta}} && (8.32)\\
&= \frac{1}{1 + e^{-\eta}} && (8.33)\\
&= \mu, && (8.34)
\end{aligned}$$

which is the mean of a Bernoulli variable.

Taking a second derivative yields:

$$\begin{aligned}
\frac{d^2 a}{d\eta^2} &= \frac{d\mu}{d\eta} && (8.35)\\
&= \mu(1 - \mu), && (8.36)
\end{aligned}$$

This makes ML estimation of exponential family super easy.

Exp family - ML estimation

Check pp. 12 of the book

Consider an IID data set, $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. From Eq. (8.58) we obtain the following log likelihood:

$$l(\eta \,|\, \mathcal{D}) = \log\left(\prod_{n=1}^{N} h(x_n)\right) + \eta^T \left(\sum_{n=1}^{N} T(x_n)\right) - NA(\eta). \tag{8.59}$$

Taking the gradient with respect to η yields:

$$\nabla_\eta l = \sum_{n=1}^{N} T(x_n) - N\nabla_\eta A(\eta), \tag{8.60}$$

and setting to zero gives:

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{n=1}^{N} T(x_n). \tag{8.61}$$

Finally, defining $\mu \triangleq E[T(x)]$, and recalling Eq. (8.40), we obtain:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} T(x_n) \tag{8.62}$$

Last equation is the essence of ML estimation. Compute mean of $T(x)$ from data, and find η with matching $E[T(x)]$.

Exp family - ML estimation

- In pp. 7 of the book, we know that, given η, we can compute mean of $T(x)$.
- Given mean of $T(x)$, we can also solve for the (unique) η, under certain conditions.
- That's all about ML estimation.

Generalized linear models

- we use GLM to model relationship between some input x and output y.
- y is an exponential family distribution whose parameter is conditioned on x: $\eta(x)$.
- That's all.

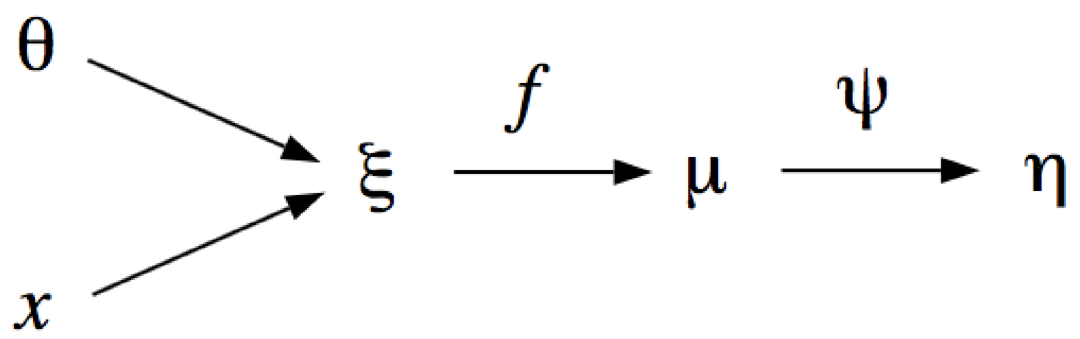

Figure 8.3: A diagram summarizing the relationships between the variables in a GLIM model.