The Exponential Family

STA 211: The Mathematics of Regression

Yue Jiang

March 28, 2023

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

An important disclaimer

This is not a mathematical statistics class. There are semester-long (and multiple semester-long) courses on these topics, and so what we cover in just two lectures scarcely touches on even the basics.

However, familiarity with some of these concepts are needed to more fully grasp generalized linear models, especially since the definition of a GLM directly depends on distributions in the exponential family. As such, we will be presenting a very abridged treatment of some of the fundamentals needed to proceed.

The exponential family

The **exponential family** of probability distributions are those that can be expressed in a specific form. Suppose X is a random variable with a distribution that depends on (a) parameter(s) $\boldsymbol{\theta}$. A random variable is said to belong to the exponential family if it can be expressed as:

$$f(x|\boldsymbol{\theta}) = h(x) \exp\left(\eta(\boldsymbol{\theta})^T T(x) - \psi(\boldsymbol{\theta})\right),$$

The exponential family

$$f(x|\boldsymbol{\theta}) = h(x)\exp\left(\eta(\boldsymbol{\theta})^T T(x) - \psi(\boldsymbol{\theta})\right),$$

Note in the exponent the $\eta(\boldsymbol{\theta})^T T(x)$ term, which represents the summation $\sum_{i=1}^k \eta_i(\boldsymbol{\theta}) T_i(x)$.

In this expression, each $\eta_i(\boldsymbol{\theta})$ and $\psi(\boldsymbol{\theta})$ are real-valued functions of the parameter(s) $\boldsymbol{\theta}$, and each $T_i(x)$ and $h(x)$ are real-valued functions of the data.

If we have just a single parameter θ in the expression above, then we have a member of a **one-parameter exponential family** distribution, expressible as

$$f(x|\theta) = h(x)\exp\left(\eta(\theta)T(x) - \psi(\theta)\right).$$

The exponential family

For simplicity's sake, for now let's consider one-parameter exponential family distributions:

$$f(x|\theta) = h(x) \exp\left(\eta(\theta) T(x) - \psi(\theta)\right).$$

The binomial distribution

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$.

- What is the probability mass function $f(x|p)$ corresponding to X? (there's a technicality here that you probably haven't had to bother with before)
- Is X a member of the one-parameter exponential family?
- If yes, identify the components $\eta(p)$, $\psi(p)$, $T(x)$, and $h(x)$. If not, explain why not

The binomial distribution

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is given by:

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} 1_{(x \in \{0,1,\cdots,n\})} \\ &= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n 1_{(x \in \{0,1,\cdots,n\})} \\ &= \binom{n}{x} \exp\left\{ x \log\left(\frac{p}{1-p}\right) + n \log(1-p) \right\} 1_{(x \in \{0,1,\cdots,n\})} \end{aligned}$$

- Is X a member of the one-parameter exponential family?
- If yes, identify the components $\eta(p)$, $\psi(p)$, $T(x)$, and $h(x)$. If not, explain why not

The binomial distribution

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is given by:

$$f(x|p) = \binom{n}{x} \exp\left\{ x \log\left(\frac{p}{1-p}\right) + n \log(1-p) \right\} 1_{(x \in \{0,1,\cdots,n\})}$$

Yes, this is a member of the one-parameter exponential family, with

$$\begin{aligned}
\eta(p) &= \log\left(\frac{p}{1-p}\right) \\
T(x) &= x \\
\psi(p) &= -n \log(1-p) \\
h(x) &= \binom{n}{x} 1_{(x \in \{0,1,\cdots,n\})}
\end{aligned}$$

The binomial distribution

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is given by:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} 1_{(x \in \{0, 1, \cdots, n\})}$$

The **sufficient statistic** is $T(x) = x$. A sufficient statistic is a statistic that provides "all the information about θ" that the entire sample could have provided (more on this when we see an example).

- What does this mean intuitively, specifically in the context of this binomial distribution?

More on sufficiency

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is given by:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} 1_{(x \in \{0,1,\cdots,n\})}$$

Keep in mind that sufficient statistics must be functions of the data only, *not* the parameter itself (even though the distribution of X might depend on the parameter).

Intuitively, sufficiency suggests that since $T(x)$ contains all the information about θ, all we need for inference *regarding* θ is given by $T(x)$ (and thus we don't actually need all of the X themselves - just the $T(x)$).

More on sufficiency

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is given by:

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} 1_{(x \in \{0,1,\cdots,n\})}$$

$T(x) = x$ is sufficient for p; all we need to know from the data in order to make inference on p is the number of successes themselves, x, not which specific observations were successes or failures.

The normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Then the parameter is $\boldsymbol{\theta} = (\mu, \sigma)$

- What is the dimension of $\boldsymbol{\theta}$?
- What is the probability density function $f(x|\boldsymbol{\theta})$? Again, watch out for the **support** of X
- Is X a member of the exponential family? If yes, identify the components $\eta_i(\boldsymbol{\theta})$, $\psi(\boldsymbol{\theta})$, $T_i(x)$, and $h(x)$. If not, explain why not

The normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Then the parameter is $\boldsymbol{\theta} = (\mu, \sigma)$.

$$\begin{aligned} f(x|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} 1_{x \in \mathbb{R}} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sigma)\right\} 1_{x \in \mathbb{R}} \end{aligned}$$

- Is X a member of the exponential family? If yes, identify the components $\eta_i(\boldsymbol{\theta})$, $\psi(\boldsymbol{\theta})$, $T_i(x)$, and $h(x)$. If not, explain why not

The normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Then the parameter is $\boldsymbol{\theta} = (\mu, \sigma)$.

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma} - \frac{\mu^2}{2\sigma^2} - \log(\sigma^2)\right\} 1_{x \in \mathbb{R}}$$

Yes, this is a member of the two-parameter exponential family, with

$$\eta_1(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}; \eta_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}$$
$$T_1(x) = x^2; T_2(x) = x$$
$$\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \log(\sigma)$$
$$h(x) = \frac{1}{\sqrt{2\pi}} 1_{x \in \mathbb{R}}$$

i.i.d. sampling

Suppose we have n i.i.d. samples from the same distribution that is in the exponential family. Then the joint density of these sample is:

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} h(x_i) \exp\left\{\eta(\boldsymbol{\theta})^T T(x_i) - \psi(\boldsymbol{\theta})\right\} \\ &= \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left\{\eta(\boldsymbol{\theta})^T \sum_{i=1}^{n} T(x_i) - n\psi(\boldsymbol{\theta})\right\} \end{aligned}$$

- What important fact do you notice above?

The normal distribution

Suppose $X \sim N(\mu, \sigma^2)$. Then the parameter is $\boldsymbol{\theta} = (\mu, \sigma)$.

$$
\begin{aligned}
\eta_1(\boldsymbol{\theta}) &= -\frac{1}{2\sigma^2}; \eta_2(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2} \\
T_1(x) &= x^2; T_2(x) = x \\
\psi(\boldsymbol{\theta}) &= \frac{\mu^2}{2\sigma^2} + \log(\sigma^2) \\
h(x) &= \frac{1}{\sqrt{2\pi}} 1_{x \in \mathbb{R}}
\end{aligned}
$$

We see that for an i.i.d. sample from a normal distribution with mean μ and variance σ^2, $\left(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i\right)$ are sufficient for the parameters (μ, σ^2).

- What does this mean in plain English?

Canonical form

$$f(x|\boldsymbol{\theta}) = h(x)\exp\left(\eta(\boldsymbol{\theta})^T T(x) - \psi(\boldsymbol{\theta})\right),$$

Notice that η and ψ are both functions of $\boldsymbol{\theta}$.

For an invertible function $\eta()$, suppose we define the variable $\boldsymbol{\eta} = \eta(\boldsymbol{\theta})$ such that $\boldsymbol{\theta} = \eta^{-1}(\boldsymbol{\eta})$ (sorry for using η as both the function and the variable).

Canonical form

We can thus re-write the exponential family in its **canonical form** using the $\boldsymbol{\eta}$ (the **canonical parameters**):

$$f(x|\boldsymbol{\theta}) = h(x)\exp\left(\eta(\boldsymbol{\theta})^T T(x) - \psi(\boldsymbol{\theta})\right)$$
$$f(x|\boldsymbol{\eta}) = h(x)\exp\left(\boldsymbol{\eta}^T T(x) - \psi(\eta^{-1}(\boldsymbol{\theta}))\right)$$

Notice that here, the canonical parameter(s) are directly multiplied with the sufficient statistic(s), and the $\psi()$ function is composed with $\eta^{-1}()$ as it acts on the (untransformed) parameters $\boldsymbol{\theta}$.

The binomial distribution (again)

Suppose $X \sim Bin(n, p)$ where n is assumed known and we have a single parameter $0 < p < 1$. Then the probability $P(X = x)$ is in the exponential family with:

$$f(x|p) = \binom{n}{x} \exp\left\{ x \log\left(\frac{p}{1-p}\right) + n \log(1-p) \right\} 1_{(x \in \{0, 1, \cdots, n\})}$$

Taking $\eta = \log\left(\frac{p}{1-p}\right)$, then we have $1 - p = \frac{1}{1+e^{\eta}}$, and so in canonical form, the binomial distribution is expressed as

$$f(x|p) = \binom{n}{x} \exp\left\{ \eta x - n \log(1 + e^{\eta}) \right\} 1_{(x \in \{0, 1, \cdots, n\})}$$

Canonical form

$$f(x|p) = \binom{n}{x} \exp\left\{\eta x - n\log(1+e^{\eta})\right\} 1_{(x\in\{0,1,\cdots,n\})}$$

In canonical form, we have

$$\begin{aligned}
\eta &= \frac{p}{1-p} \\
T(x) &= x \\
A(\eta) &= n\log(1+e^{\eta}) \\
h(x) &= \binom{n}{x} 1_{(x\in\{0,1,\cdots,n\})}
\end{aligned}$$

The log-partition function

This function composition $\psi(\eta^{-1}(\boldsymbol{\theta})$ is known as the log-partition function (let's give it a new name, $A(\boldsymbol{\eta})$, in terms of the canonical parameters).

We can use this function to easily calculate the mean and variance of distributions in the exponential family by *differentiating* the log-partition (often a lot easier than performing messy integration):

$$\frac{\partial A}{\partial \eta_i} = E(T_i(x))$$

$$\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = Cov(T_i(x), T_j(x))$$

The first and second derivatives of $A(\boldsymbol{\eta})$ are the mean and variances of the sufficient statistic, respectively.

The log-partition function

Remember, in canonical form the binomial distribution had log-partition $A(\eta) = n\log(1+e^{\eta})$, where $\eta = \log\left(\frac{p}{1-p}\right)$, and sufficient statistic $T(x)$.

$$\begin{aligned}
E(T(x)) = E(x) &= \frac{\partial A}{\partial \eta} = n\frac{e^{\eta}}{1+e^{\eta}} \\
&= np \\
Var(T(x)) = Var(x) &= \frac{\partial^2 A}{\partial \eta^2} = n\frac{e^{\eta}}{(1+e^{\eta})^2} \\
&= np(1-p)
\end{aligned}$$

The mean and variance of the binomial distribution.

Maximum likelihood estimation

Suppose we have n i.i.d. samples from the same distribution that is in the exponential family (let's say canonical form). Then the joint density of these sample is:

$$f(\mathbf{x}|\boldsymbol{\eta}) = \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left\{\boldsymbol{\eta}^T \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\eta})\right\}$$

- What important fact do you notice above?

Maximum likelihood estimation

The product of exponential family distributions is also in the exponential family (with sufficient statistic $\sum_{i=1}^{n} T(x_i)$)

So, for instance, the sufficient statistic for the joint distribution of n i.i.d. binomial random variables is $\sum_{i=1}^{n} x_i$. All we need for inference on the parameter p from the n observations is the sum of the x_is.

Maximum likelihood estimation

$$
\begin{aligned}
f(\mathbf{x}|\boldsymbol{\eta}) &= \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left\{\boldsymbol{\eta}^T \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\eta})\right\} \\
\log \mathcal{L}(\boldsymbol{\eta}|\mathbf{x}) &= \log\left(\prod_{i=1}^{n} h(x_i)\right) + \boldsymbol{\eta}^T \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\eta}) \\
\nabla_{\boldsymbol{\eta}} \log \mathcal{L} &= \sum_{i=1}^{n} T(x_i) - n\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}).
\end{aligned}
$$

Setting the last line equal to zero suggests the MLE for the mean parameter of an exponential family distribution is simply a method of moments estimator:

$$E(T(x)) = \nabla_{\boldsymbol{\eta}} A(\widehat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$$

Random closing comments

We just saw a nice property of the exponential family - that the expectation of sufficient statistics for the model are the empirical average of the sufficient statistics, and furthermore that this is the MLE.

This is just one of the many nice properties of exponential family distributions. A random aside for when you see these again - exponential family distributions have all sorts of nice properties that we don't have time to cover (e.g., existence of conjugate priors, connection to estimation theory, guaranteed log-concave likelihoods, etc.). You'll cover these in later classes!

Homework

1. Consider a uniform distribution on $(0, \theta)$ (that is, $f(x) = 1/\theta$ if $x \in (0, \theta)$). Is this a member of the exponential family? If so, identify the components in canonical form and use the log-partition function to calculate the MLE for θ from an i.i.d. sample. If not, explain why not.
2. Consider a normal distribution $N(\mu, \mu)$ for $\mu > 0$ (that is, where the variance equals the mean). Is this a member of the exponential family? If so, identify the components in the canonical form and use the log-partition function to calculate the MLE for μ from an i.i.d. sample. If not, explain why not.
3. Consider a distribution $f(x|\lambda) = \frac{\lambda}{x^{1+\lambda}}$ for $\lambda > 0$ and $x > 1$. Is this a member of the exponential family? If so, identify the components in the canonical form and use the log-partition function to calculate the MLE for λ from an i.i.d. sample. If not, explain why not.