Exponential family

Sep 2022

Murphy chap9

Sufficient statistic

Let $X_1, ..., X_n$ be a random sample from a distribution parametrized by θ, T be a statistic. If for all possible value t of T,

$$P(X_1, ..., X_n \mid T = t, \theta) = P(X_1, ..., X_n \mid T = t)$$

then we say T is a sufficient statistic for the parameter θ

Sufficient statistic

$T = r(X_1, ..., X_n)$ is a sufficient statistic for θ if and only if the joint density or mass function $f_n(x|\theta)$ of $X_1, ..., X_n$ can be factored as follows for all values $x = (x_1, ..., x_n)$ and all admissible θ:

$$f_n(x|\theta) = u(x)v(r(x), \theta) \qquad (*)$$

here u, v are nonnegative functions, $u(x)$ may depend on x, but does not depend on θ

Proof for discrete distribution

$$A(t) = \{x : r(x) = t\}$$

(==>) Suppose T is sufficient. Then, for every given value t of T, every point $\boldsymbol{x} \in A(t)$, and every value of $\theta \in \Omega$, the conditional probability $\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta)$ will not depend on θ and will therefore have the form

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = u(\boldsymbol{x}).$$

If we let $v(t, \theta) = \Pr(T = t|\theta)$, it follows that

$$\begin{aligned} f_n(\boldsymbol{x}|\theta) &= \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta)\Pr(T = t|\theta) \\ &= u(\boldsymbol{x})v(t, \theta). \end{aligned}$$

Sufficient statistic

$T = r(X_1, ..., X_n)$ is a sufficient statistic for θ if and only if the joint density or mass function $f_n(x|\theta)$ of $X_1, ..., X_n$ can be factored as follows for all values $x = (x_1, ..., x_n)$ and all admissible θ:

$$f_n(x|\theta) = u(x)v(r(x), \theta) \qquad (*)$$

here u, v are nonnegative functions, $u(x)$ may depend on x, but does not depend on θ

Proof for discrete distribution

$$A(t) = \{x : r(x) = t\}$$

(<==) Suppose (*) For every point $\boldsymbol{x} \in A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{\Pr(\boldsymbol{X} = \boldsymbol{x}|\theta)}{\Pr(T = t|\theta)} = \frac{f_n(\boldsymbol{x}|\theta)}{\sum_{y \in A(t)} f_n(\boldsymbol{y}|\theta)}$$

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = \frac{u(\boldsymbol{x})}{\sum_{y \in A(t)} u(\boldsymbol{y})}.$$ <— use (*), does not depend on theta

for every point $\boldsymbol{x}$ that does not belong to $A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) = 0.$$ <— does not depend on theta

Sufficient statistic

$T = r(X_1, ..., X_n)$ is a sufficient statistic for θ if and only if the joint density or mass function $f_n(x|\theta)$ of $X_1, ..., X_n$ can be factored as follows for all values $x = (x_1, ..., x_n)$ and all admissible θ:

$$f_n(x|\theta) = u(x)v(r(x), \theta) \qquad (*)$$

here u, v are nonnegative functions, $u(x)$ may depend on x, but does not depend on θ

It is sufficient to verify (*) for x such that the density or mass > 0

T is sufficient if and only if the posterior of theta depends on the data only through T

Multiple statistics:

$$f_n(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v[r_1(\boldsymbol{x}), \ldots, r_k(\boldsymbol{x}), \theta].$$

Sufficient statistic - example

Poisson distribution with mean theta

Let $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. $\qquad T = r(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$

For every set of nonnegative integers $x_1, \ldots, x_n$, the joint p.f. $f_n(\boldsymbol{x}|\theta)$ of $X_1, \ldots, X_n$ is as follows:

$$f_n(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) e^{-n\theta}\theta^{r(\boldsymbol{x})}.$$

$$u(\boldsymbol{x}) = \prod_{i=1}^{n}(1/x_i!) \text{ and } v(t, \theta) = e^{-n\theta}\theta^{t}$$

Sufficient statistic - example

Normal distribution with unknown mean mu and known variance sigma

Let $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. $\quad T = r(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$

For $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$, the joint p.d.f. of $\boldsymbol{X}$ is as follows:

$$f_n(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

This equation can be rewritten in the form

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right) \exp\left(\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right).$$

Let $u(\boldsymbol{x})$ be the constant factor and the first exponential factor

$$v(t, \mu) = \exp(\mu t/\sigma^2 - n\mu^2/\sigma^2)$$

Sufficient statistic - example

Normal distribution with known mean mu and known variance sigma

Let $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. $T_1 = \sum_{i=1}^{n} X_i$ and $T_2 = \sum_{i=1}^{n} X_i^2$

For $-\infty < x_i < \infty$ $(i = 1, \ldots, n)$, the joint p.d.f. of $\boldsymbol{X}$ is as follows:

$$f_n(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

This equation can be rewritten in the form

$$f_n(\boldsymbol{x}|\mu) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right) \exp\left(\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right).$$

Sufficient statistic - example

Normal distribution with known mean mu and known variance sigma

Another pair of sufficient statistics !!

$$T_1' = \overline{X}_n \quad \text{and} \quad T_2' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Then

$$T_1' = \frac{1}{n}T_1 \quad \text{and} \quad T_2' = \frac{1}{n}T_2 - \frac{1}{n^2}T_1^2.$$

Also, equivalently,

$$T_1 = nT_1' \quad \text{and} \quad T_2 = n(T_2' + T_1'^2).$$

1-1 correspondence between T1,T2 and T1',T2'

Recall that mu and sigma determines a normal distribution

Exponential family

A pdf or pmf $p(\mathbf{x}|\boldsymbol{\theta})$, for $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}^m$ and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, is said to be in the **exponential family** if it is of the form

$$
\begin{aligned}
p(\mathbf{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] && (9.1) \\
&= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] && (9.2)
\end{aligned}
$$

where

$$
\begin{aligned}
Z(\boldsymbol{\theta}) &= \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} && (9.3) \\
A(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta}) && (9.4)
\end{aligned}
$$

Here $\boldsymbol{\theta}$ are called the **natural parameters** or **canonical parameters**, $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^d$ is called a vector of **sufficient statistics**, $Z(\boldsymbol{\theta})$ is called the **partition function**, $A(\boldsymbol{\theta})$ is called the **log partition function** or **cumulant function**, and $h(\mathbf{x})$ is the a scaling constant, often 1. If $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, we say it is a **natural exponential family**.

Definition

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] \tag{9.1}$$

$$= h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})] \tag{9.2}$$

Equation 9.2 can be generalized by writing

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp[\eta(\boldsymbol{\theta})^T \boldsymbol{\phi}(\mathbf{x}) - A(\eta(\boldsymbol{\theta}))] \tag{9.5}$$

where η is a function that maps the parameters $\boldsymbol{\theta}$ to the canonical parameters $\boldsymbol{\eta} = \eta(\boldsymbol{\theta})$.

Examples

Bernoulli

The Bernoulli for $x \in \{0, 1\}$ can be written in exponential family form as follows:

$$\mathrm{Ber}(x|\mu) = \mu^x(1-\mu)^{1-x} = \exp[x\log(\mu) + (1-x)\log(1-\mu)] = \exp[\boldsymbol{\phi}(x)^T\boldsymbol{\theta}] \tag{9.6}$$

where $\boldsymbol{\phi}(x) = [\mathbb{I}(x=0), \mathbb{I}(x=1)]$ and $\boldsymbol{\theta} = [\log(\mu), \log(1-\mu)]$.

Another (minimal) statistics, it is a function of theta in (9.6)

$$\mathrm{Ber}(x|\mu) = (1-\mu)\exp\left[x\log\left(\frac{\mu}{1-\mu}\right)\right] \tag{9.8}$$

we have $\phi(x) = x$, $\theta = \log\left(\frac{\mu}{1-\mu}\right)$

Examples

Univariate Gaussian

The univariate Gaussian can be written in exponential family form as follows:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2] \quad (9.20)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp[-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}\mu^2] \quad (9.21)$$

$$= \frac{1}{Z(\boldsymbol{\theta})}\exp(\boldsymbol{\theta}^T\boldsymbol{\phi}(x)) \quad (9.22)$$

where

$$\boldsymbol{\theta} = \begin{pmatrix} \mu/\sigma^2 \\ \frac{-1}{2\sigma^2} \end{pmatrix} \quad (9.23)$$

$$\boldsymbol{\phi}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad (9.24)$$

$$Z(\mu,\sigma^2) = \sqrt{2\pi}\sigma\exp[\frac{\mu^2}{2\sigma^2}] \quad (9.25)$$

$$A(\boldsymbol{\theta}) = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi) \quad (9.26)$$

Log partition function A(theta)

Derive mean and variance of sufficient statistics from log partition function

1-parameter distribution

$$
\begin{aligned}
\frac{dA}{d\theta} &= \frac{d}{d\theta}\left(\log \int \exp(\theta\phi(x))h(x)dx\right) && (9.27)\\
&= \frac{\frac{d}{d\theta}\int \exp(\theta\phi(x))h(x)dx}{\int \exp(\theta\phi(x))h(x)dx} && (9.28)\\
&= \frac{\int \phi(x)\exp(\theta\phi(x))h(x)dx}{\exp(A(\theta))} && (9.29)\\
&= \int \phi(x)\exp(\theta\phi(x) - A(\theta))h(x)dx && (9.30)\\
&= \int \phi(x)p(x)dx = \mathbb{E}\left[\phi(x)\right] && (9.31)
\end{aligned}
$$

Log partition function A(theta)

Derive mean and variance of sufficient statistics from log partition function

1-parameter distribution

$$\frac{d^2 A}{d\theta^2} = \int \phi(x) \exp\left(\theta\phi(x) - A(\theta)\right) h(x)(\phi(x) - A'(\theta))dx \tag{9.32}$$

$$= \int \phi(x)p(x)(\phi(x) - A'(\theta))dx \tag{9.33}$$

$$= \int \phi^2(x)p(x)dx - A'(\theta)\int \phi(x)p(x)dx \tag{9.34}$$

$$= \mathbb{E}\left[\phi^2(X)\right] - \mathbb{E}\left[\phi(x)\right]^2 = \mathrm{var}\left[\phi(x)\right] \tag{9.35}$$

and hence

$$\nabla^2 A(\boldsymbol{\theta}) = \mathrm{cov}\left[\boldsymbol{\phi}(\mathbf{x})\right] \tag{9.37}$$

we see that $A(\boldsymbol{\theta})$ is a convex function

Log partition function A(theta)

$$\text{Ber}(x|\mu) = (1-\mu)\exp\left[x\log\left(\frac{\mu}{1-\mu}\right)\right] \tag{9.8}$$

$$\phi(x) = x,\, \theta = \log\left(\tfrac{\mu}{1-\mu}\right)$$

$$e^{-A(\theta)} = 1-\mu,\; 1+e^{\theta} = \frac{1}{1-\mu}$$

Example: the Bernoulli distribution

For example, consider the Bernoulli distribution. We have $A(\theta) = \log(1+e^{\theta})$, so the mean is given by

$$\frac{dA}{d\theta} = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} = \text{sigm}(\theta) = \mu \tag{9.38}$$

The variance is given by

$$\frac{d^2A}{d\theta^2} = \frac{d}{d\theta}(1+e^{-\theta})^{-1} = (1+e^{-\theta})^{-2}.e^{-\theta} \tag{9.39}$$

$$= \frac{e^{-\theta}}{1+e^{-\theta}}\frac{1}{1+e^{-\theta}} = \frac{1}{e^{\theta}+1}\frac{1}{1+e^{-\theta}} = (1-\mu)\mu \tag{9.40}$$

MLE for exponential family

The likelihood of an exponential family model has the form

$$p(\mathcal{D}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{N} h(\mathbf{x}_i)\right] g(\boldsymbol{\theta})^N \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T [\sum_{i=1}^{N} \boldsymbol{\phi}(\mathbf{x}_i)]\right) \tag{9.41}$$

We see that the sufficient statistics are N and

$$\boldsymbol{\phi}(\mathcal{D}) = [\sum_{i=1}^{N} \phi_1(\mathbf{x}_i), \ldots, \sum_{i=1}^{N} \phi_K(\mathbf{x}_i)] \tag{9.42}$$

$$A(\theta) = -log(\theta)$$

MLE for exponential family

N iid data points $\mathcal{D} = (x_1, \ldots, x_N)$, the log-likelihood is

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \boldsymbol{\theta}^T \phi(\mathcal{D}) - NA(\boldsymbol{\theta}) \tag{9.45}$$

Since $-A(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^T \phi(\mathcal{D})$ is linear in $\boldsymbol{\theta}$, we see that the log likelihood is concave, and hence has a unique global maximum. To derive this maximum, we use the fact that the derivative of the log partition function yields the expected value of the sufficient statistic vector (Section 9.2.3):

$$\nabla_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta}) = \phi(\mathcal{D}) - N\mathbb{E}\left[\boldsymbol{\phi}(\mathbf{X})\right] \tag{9.46}$$

Setting this gradient to zero, we see that at the MLE, the empirical average of the sufficient statistics must equal the model's theoretical expected sufficient statistics, i.e., $\hat{\boldsymbol{\theta}}$ must satisfy

$$\mathbb{E}\left[\boldsymbol{\phi}(\mathbf{X})\right] = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{x}_i) \tag{9.47}$$