Statistical Machine Learning Distributions

Lecture Notes 7: Exponential Family

Professor: Zhihua Zhang

Definition 7.1 (Statistic)**.** *Given random variables(vectors) $X_1, \cdots, X_n$ with respect to sets of possible values $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n$, respectively. A random vector $\boldsymbol{t}_n : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}^{k(n)}$ is called a $k(n)$ dimensional statistic.*

Example 7.1. $t_n(X_1, \cdots, X_n) = (X_1, \cdots, X_n)$.

Example 7.1 is the simplest statistic, usually we want to achieve data reduction by statistic, i.e., $k(n) < n$. Sometimes $k(n)$ are independent with n.

Example 7.2.
$\boldsymbol{t}_n = \frac{1}{n}(X_1 + \cdots + X_n), k(n) = 1$
$\boldsymbol{t}_n = [n, (X_1 + \cdots + X_n), (X_1^2 + \cdots + X_n^2)], k(n) = 3$, *the zero order, first and second moment.*
$\boldsymbol{t}_n = [n, \text{median}(X_1, \cdots, X_n)]$, *the median.*
$\boldsymbol{t}_n = \max\{X_1, \cdots, X_n\} - \min\{X_1, \cdots, X_n\}$, *the range.*

Definition 7.2 (Sufficient Statistic)**.** *The sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n$ is a sufficient statistic for $X_1, X_2, \cdots, X_n$ if for $n \geq 1$, the joint density for $X_1, X_2, \cdots, X_n$ given θ has the form*

$$p(x_1, x_2, \cdots, x_n|\theta) = h_n(\boldsymbol{t}_n, \theta) g(x_1, x_2, \cdots, x_n)$$

for some function $h_n \geq 0$, $g > 0$.

Theorem 7.1. *The sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n$ is sufficient for infinitely exchangeable $X_1, X_2, \cdots$ if and only if for any $n \geq 1$, the density $p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n)$ is independent of θ.*

Proof. For any $\boldsymbol{t}_n = t_n(X_1, X_2, \cdots, X_n)$,

$$p(x_1, x_2, \cdots, x_n|\theta) = p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n) p(\boldsymbol{t}_n, \theta)$$

If $p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n)$ is independent of θ, then $p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n)$ is g, $p(\boldsymbol{t}_n, \theta)$ is h_n. So $\boldsymbol{t}_n$ is a sufficient statistic.

If $\boldsymbol{t}_n$ is sufficient, then $p(x_1, x_2, \cdots, x_n|\theta) = h_n(\boldsymbol{t}_n, \theta) g(x_1, x_2, \cdots, x_n)$,$h_n \geq 0$, $g > 0$. Taking integral on both sides, we have

$$\int_{\{t_n(x_1, \cdots, x_n) = \boldsymbol{t}_n\}} p(x_1, \cdots, x_n|\theta) dx_1 \cdots dx_n = \int_{\{t_n(x_1, \cdots, x_n) = \boldsymbol{t}_n\}} h_n(\boldsymbol{t}_n, \theta) g(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

Note that $h_n(\boldsymbol{t}_n, \theta)$ in the right side is unrelated to the integral, $\int g(\boldsymbol{x}) d\boldsymbol{x}$ can be seemed as a function of $\boldsymbol{t}_n$, denoted by $G(\boldsymbol{t}_n)$ and $\int p(\boldsymbol{x}|\theta) d\boldsymbol{x}$ can be seemed as $p(\boldsymbol{t}_n|\theta)$. Hence, we have

$$\begin{aligned} & p(\boldsymbol{t}_n|\theta) = h_n(\boldsymbol{t}_n, \theta) G(\boldsymbol{t}_n) \\ \Longrightarrow \quad & h_n(\boldsymbol{t}_n, \theta) = \frac{p(\boldsymbol{t}_n|\theta)}{G(\boldsymbol{t}_n)} \end{aligned}$$

So,

$$\begin{aligned} & p(x_1, x_2, \cdots, x_n|\theta) = \frac{p(\boldsymbol{t}_n|\theta)}{G(\boldsymbol{t}_n)} g(x_1, x_2, \cdots, x_n) \\ \Longrightarrow \quad & p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n) = \frac{p(x_1, x_2, \cdots, x_n|\theta)}{p(\boldsymbol{t}_n|\theta)} = \frac{g(x_1, x_2, \cdots, x_n)}{G(\boldsymbol{t}_n)} \end{aligned}$$

Thus we can see $p(x_1, x_2, \cdots, x_n|\theta, \boldsymbol{t}_n)$ is independent with θ. □

Example 7.3 (Bernoulli Distribution)**.** *For a Bernoulli sequence $X_1, \cdots, X_n$,*

$$\begin{aligned} p(x_1, \cdots, x_n) &= \int_0^1 p(x_1, \cdots, x_n|\theta) dF(\theta) \\ &= \int_0^1 \prod_{i=1}^n B(x|\theta) dF(\theta) \\ &= \int_0^1 \theta^{S_n}(1-\theta)^{n-S_n} dF(\theta) \end{aligned}$$

where $S_n = x_1 + \cdots + x_n$. So,

$$p(x_1, \cdots, x_n|\theta) = \theta^{S_n}(1-\theta)^{n-S_n}$$

Let $\boldsymbol{t}_n = [n, S_n]$, $p(x_1, \cdots, x_n|\theta)$ can be factorized into $h_n = \theta^{S_n}(1-\theta)^{n-S_n}$ and $g = 1$. So $\boldsymbol{t}_n$ is the sufficient statistic of Bernoulli distribution.

Example 7.4 (Normal Distribution)**.**

$$\begin{aligned} p(x_1, \cdots, x_n|\mu, \lambda) &= \prod_{i=1}^n (\frac{\lambda}{2\pi})^{\frac{1}{2}} \exp(-\frac{\lambda}{2}(x_i - \mu)^2) \\ &= (\frac{\lambda}{2\pi})^{\frac{n}{2}} \exp(-\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2) \\ &= (\frac{\lambda}{2\pi})^{\frac{n}{2}} \exp(-\frac{\lambda}{2}[n(\bar{x} - \mu) + nS_n^2) \end{aligned}$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i, S_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. So the sufficient statistic of normal distribution can be $[n, \bar{X}_n, S_n^2]$. Note that the sufficient statistic is not unique, for example, $[n, \bar{X}_n, \frac{1}{n}\sum_{i=1}^n X_i^2]$ is also sufficient statistic of normal distribution.

7.1 Exponential Family

Definition 7.3 (one-parameter exponential family)**.** *A p.d.f or p.m.f $p(x|\theta)$, labelled by $\theta \in \Theta \subseteq \mathbb{R}$ is said to belong to one-parameter exponential family if it is of the form*

$$p(x|\theta) = f(x)g(\theta)\exp\left(c \cdot \phi(\theta)h(x)\right)$$

where $g^{-1}(\theta) = \int f(x)\exp\left(c \cdot \phi(\theta)h(x)\right)dx < \infty$ is a regularization factor. Denoted by $E_f(f, g, h, \phi, c, \theta)$.

Definition 7.4. *The family is called regular if $\mathcal{X}, (X \in \mathcal{X})$ does not dependent on θ, otherwise is called non-regular.*

Proposition 7.1 (Sufficient statistic for E_f)**.** *If $X_1, \cdots, X_n \in \mathcal{X}$ is an exchangeable sequence such that given regular $E_f(X|f, g, h, \phi, c, \theta)$,*

$$p(x_1, \cdots, x_n) = \int_\theta \prod_{i=1}^{n} E_f(x_i|f, g, h, \phi, c) dF(\theta)$$

for some $dF(\theta)$. Then $\boldsymbol{t}_n = t_n(X_1, \cdots, X_n) = [n, h(X_1) + \cdots + h(X_n)]$ is sufficient statistic.

Example 7.5 (Bernoulli Distribution)**.**

$$\begin{aligned} p(x|\theta) &= \theta^x (1-\theta)^{1-x}, \qquad x \in \{0,1\}, \theta \in [0,1] \\ &= (1-\theta) \left(\frac{\theta}{1-\theta} \right)^x \\ &= (1-\theta) \exp(x \ln \frac{\theta}{1-\theta}) \end{aligned}$$

So, $f(x) = 1, g(\theta) = 1-\theta, c = 1, h(x) = x, \phi(\theta) = \ln \frac{\theta}{1-\theta}$.

Example 7.6 (Poisson Distribution)**.**

$$\begin{aligned} p(x|\theta) &= \frac{\theta^x \cdot e^{-\theta}}{x!} \\ &= \frac{1}{x!} \exp(-\theta) \exp(x \ln x) \end{aligned}$$

So, $f(x) = \frac{1}{x!}, g(\theta) = e^{-\theta}, c = 1, h(x) = x, \phi(\theta) = \ln x$.

Example 7.7 (Normal Distribution with Unknown Variance)**.**

$$\begin{aligned} p(x|\theta) &= N(x|0, \sigma^2) \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp(-\frac{x^2}{2\sigma^2}) \\ &= \left(\frac{1}{2\pi} \right)^{\frac{1}{2}} \theta^{-\frac{1}{2}} \exp(-\frac{x^2}{2\theta}) \end{aligned}$$

So, $f(x) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}}, g(\theta) = \theta^{-\frac{1}{2}}, c = -\frac{1}{2}, h(x) = x^2, \phi(\theta) = \theta^{-1}$.

Example 7.8 (Uniform Distribution, non-regular)**.**

$$p(x|\theta) = U(x|[0,\theta]) = \frac{1}{\theta}$$

So, $f(x) = 1, g(\theta) = \theta^{-1}, c = 1, h(x) = 0, \phi(\theta) = 0$. Since $\mathcal{X}$ is $[0,\theta]$, related to θ, so it is non regular.

$$
\begin{aligned}
f_X(x_1, \ldots, x_n) &= \frac{1}{\theta}\mathbf{1}_{\{0 \le x_1 \le \theta\}} \cdots \frac{1}{\theta}\mathbf{1}_{\{0 \le x_n \le \theta\}} \\
&= \frac{1}{\theta^n}\mathbf{1}_{\{0 \le \min\{x_i\}\}}\mathbf{1}_{\{\max\{x_i\} \le \theta\}}
\end{aligned}
$$

where $\mathbf{1}\{...\}$ *is the indicator function. So the sufficient statistic* $\boldsymbol{t}_n = [n, \max\{x_i\}]$.

Definition 7.5 (k-parameters exponential family)**.** *A p.d.f or p.m.f* $p(x|\theta)$, $x \in \mathcal{X}$, *which is labelled by* $\theta \in \Theta \subseteq \mathbb{R}$ *is said to belong to k-parameters exponential family if it is of the form*

$$
p(x|\theta) = f(x)g(\theta)\exp\left(\sum_{j=1}^{k} c_j \cdot \phi_j(\theta)h_j(x)\right)
$$

Denoted by $E_{f_k}(x|f, g, h, \phi, c, \theta)$.

Proposition 7.2 (Sufficient statistic for E_{f_k})**.** *If* $X_1, \cdots, X_n \in \mathcal{X}$ *is an exchangeable sequence such that given regular* $E_{f_k}(X|f, g, h, \phi, c, \theta)$,

$$
p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} E_{f_k}(x_i|f, g, h, \phi, c, \theta)
$$

Then $\boldsymbol{t}_n = t_n(X_1, \cdots, X_n) = [n, \sum_{i=1}^{n} h_1(X_i), \cdots, \sum_{i=1}^{n} h_k(X_i)]$ *is sufficient statistic of* $X_1, \cdots, X_n$.

Example 7.9 (Normal Distribution with Unknown Mean and Variance)**.** *Let* $\theta = [\mu, \lambda]$,

$$
\begin{aligned}
p(x|\theta) &= N(x|\mu, \lambda) \\
&= (\frac{\lambda}{2\pi})^{\frac{1}{2}}\exp(-\frac{\lambda}{2}(x-\mu)^2)) \\
&= (\frac{1}{2\pi})^{\frac{1}{2}}\lambda^{\frac{1}{2}}\exp(-\frac{\lambda}{2}\mu^2)\exp(\lambda\mu x - \frac{1}{2}\lambda x^2)
\end{aligned}
$$

So, $g(\theta) = \lambda^{\frac{1}{2}}\exp(-\frac{\lambda}{2}\mu^2)$, $c_1 = 1, c_2 = -\frac{1}{2}$, $\phi_1(\theta) = \lambda\mu, \phi_2(\theta) = \lambda$, $h_1(x) = x, h_2(x) = x^2$.

7.2 Canonical(Natural) Exponential Family

The p.d.f of exponential family can be rewritten into another form:

$$
p(\boldsymbol{y}|\varphi) = cef(\boldsymbol{y}|a, b, \varphi) = a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi))
$$

where $y = (y_1, \cdots, y_k), \phi = (\varphi_1, \cdots, \varphi_k)$. Comparing to the previous form, we can see $y_i = h_i(x), \varphi_i = c_i\phi(\theta)$.

Proposition 7.3 (moments of cef)**.** *For* y *in definition of cef, we have*

$$
E[\boldsymbol{y}|\varphi] = \int \boldsymbol{y}a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi))d\boldsymbol{y}
$$

Since $\int a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi))d\boldsymbol{y} = 1$, *taking derivation on both sides.*

$$
\begin{aligned}
& \int a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi))(\boldsymbol{y} - \triangledown_{\varphi} b(\varphi))d\boldsymbol{y} = 0 \\
\Longrightarrow \quad & \int a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi))\boldsymbol{y}d\boldsymbol{y} = \int a(\boldsymbol{y})\exp(\boldsymbol{y}^T\varphi - b(\varphi)) \triangledown_{\varphi} b(\varphi))d\boldsymbol{y} \\
\Longrightarrow \quad & E[\boldsymbol{y}] = \triangledown_{\varphi} b(\varphi))
\end{aligned}
$$

Example 7.10 (Possion Distribution)**.**

$$
e^{-\lambda}\frac{\lambda^x}{x!} = \frac{1}{x!}\exp(x\log\lambda - \lambda) = \frac{1}{x!}\exp(x\theta - e^{\theta}), \lambda = e^{\theta}
$$

So, $E[\boldsymbol{y}] = \triangledown_{\varphi} b(\varphi)) = \lambda$

Theorem 7.2. *If* $X = (X_1, \cdots, X_n)$ *is random variable from a regular exponential family distribution such that*

$$
p(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i)[g(\theta)]^n \exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\sum_{i=1}^{n} h_j(x_i)\right).
$$

Then the conjugate family for θ *has the form*

$$
p(\boldsymbol{x}|\tau) = [K(\tau)]^{-1}[g(\theta)]^{\tau_0}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right)
$$

where $k(\tau) = \int_{\theta}[g(\theta)]^{\tau}\exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j\right)d\theta < \infty$

Example 7.11 (Bernoulli Likelihood)**.**

$$
\begin{aligned}
p(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{(1-\boldsymbol{x}_i)} \\
&= (1-\theta)^n\exp\left((\log\frac{\theta}{1-\theta})\sum_{i=1}^{n} x_i\right)
\end{aligned}
$$

So,

$$
\begin{aligned}
p(\theta|\tau) &\propto (1-\theta)^{\tau_0}\exp\left(\log\frac{\theta}{1-\theta}\tau_1\right) \\
&\propto (1-\theta)^{\tau_0}\left(\frac{\theta}{1-\theta}\right)^{\tau_1} \\
&\propto \theta^{\tau_1}(1-\theta)^{\tau_0-\tau_1}
\end{aligned}
$$

Hence, the conjugate prior of Bernoulli distribution is beta distribution.

Example 7.12 (Possion Likelihood)**.**

$$
\begin{aligned}
p(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{x_i}\exp(-\theta)}{x_i!} \\
&= \prod_{i=1}^{n} (x_i!)^{-1}\exp(-n\theta)\exp(\log\theta\sum_{i=1}^{n} x_i)
\end{aligned}
$$

So,

$$
\begin{aligned}
p(\theta|\tau) &\propto \exp(-\tau_0\theta)\exp(\tau_1\log\theta) \\
&\propto \theta^{\tau_1}\exp(-\tau_0\theta)
\end{aligned}
$$

Hence, the conjugate prior of Possion distribution is gamma distribution.

Example 7.13 (Normal Likelihood)**.** *Let* $\theta = (\mu, \lambda)$

$$
\begin{aligned}
p(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right) \\
&= (2\pi)^{-\frac{n}{2}}\left[\lambda^{\frac{1}{2}}\exp(-\frac{\lambda}{2}\mu^2)\right]^n \exp(\mu\lambda\sum_{i=1}^{n} x_i - \frac{\lambda}{2}\sum_{i=1}^{n} x_i^2)
\end{aligned}
$$

So,

$$
\begin{aligned}
p(\theta|\tau) &\propto \left[\lambda^{\frac{1}{2}}\exp(-\frac{\lambda}{2}\mu^2)\right]^{\tau_0}\exp(\mu\lambda\tau_1 - \frac{\lambda}{2}\tau_2) \\
&\propto \lambda^{\frac{\tau_0}{2}}\exp(-\frac{\tau_0\lambda}{2}\mu^2)\exp(\mu\lambda\tau_1 - \frac{\lambda}{2}\tau_2) \\
&\propto \lambda^{\frac{\tau_0}{2}}\exp\left(-\frac{\tau_0\lambda}{2}(\mu - \frac{\tau_1}{\tau_0})^2\right)\exp(\frac{\lambda\tau_1^2}{2\tau_0})\exp(-\frac{\lambda\tau_2}{2}) \\
&\propto (\tau_0\lambda)^{\frac{1}{2}}\exp\left(-\frac{\tau_0\lambda}{2}(\mu - \frac{\tau_1}{\tau_0})^2\right)\exp(\frac{\lambda\tau_1^2}{2\tau_0})\exp(-\frac{\lambda\tau_2}{2})\lambda^{\frac{\tau_0}{2}}(\tau_0\lambda)^{-\frac{1}{2}} \\
&\propto (\tau_0\lambda)^{\frac{1}{2}}\exp\left(-\frac{\tau_0\lambda}{2}(\mu - \frac{\tau_1}{\tau_0})^2\right)\exp(-\frac{\lambda}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0}))\tau_0^{-\frac{1}{2}}\lambda^{\frac{\tau_0+1}{2}-1}
\end{aligned}
$$

Note that $(\tau_0\lambda)^{\frac{1}{2}}\exp\left(-\frac{\tau_0\lambda}{2}(\mu - \frac{\tau_1}{\tau_0})^2\right)$ *can be seemed as a normal prior,* $p(\mu|\lambda,\tau)$ *and* $\exp(-\frac{\lambda}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0}))\tau_0^{-\frac{1}{2}}\lambda^{\frac{\tau_0+1}{2}-1}$ *can be seemed as a gamma prior,* $p(\lambda|\tau)$*.*

Theorem 7.3 (posterior)**.** $p(\theta|x,\tau) = p(\theta|\tau + \boldsymbol{t}_n(x))$*, where* $\tau + \boldsymbol{t}_n(x) = (\tau_0 + n, \tau_1 + \sum_{i=1}^{n} h_1(x_i), \cdots, \tau_k + \sum_{i=1}^{n} h_k(x_i))$*.*