Hometask 1

Khasianov Rasul

1. A single-parameter exponential family is a set of probability distributions whose probability density function can be expressed in the form:

$$f_X(x|\theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\theta))$$

Bernouilli distribution:

$$\begin{aligned} f(k;\theta) &= \theta^k \cdot (1-\theta)^{1-k} \\ \log f(k;\theta) &= k\log\theta + (1-k)\log(1-\theta) = \\ &= k\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta) \\ \Rightarrow f(k;\theta) &= \exp\left(k\log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right) \end{aligned}$$

Denote:

$$\begin{aligned} \eta(\theta) &= \log\frac{\theta}{1-\theta} & T(k) &= k \\ A(\theta) &= -\log(1-\theta) & h(k) &= 1 \end{aligned}$$

Thus, Bernouilli distribution for $\theta \in (0,1)$ form an exponential family.

2. Uniform distribution $U([0,\theta])$ does not belong to an exponential family and the sufficient statistic is $T(x) = \max(X_1,\dots,X_n) = X_{(n)}$.

$$f(x|\theta) = \frac{1}{\theta}I(x_i \in [0,\theta])$$

   - A statistic $t = T(x)$ is **sufficient** for underlying parameter θ precisely if the conditional probability distribution of the data X, given the statistic

$t = T(x)$, doesn't depend on the parameter θ, i.e. $P(x|T(x) = t, \theta) = P(x|T(X) = t)$.

Factorization criterion: if the probability density function is $L(x;\theta)$, then T is sufficient for θ if and only if nonnegative functions g and h can be found such that:

$$L(x;\theta) = g(T(x);\theta)h(x)$$

Likelihood function is

$$L(x;\theta) = \theta^{-1} I(x_1 \in [0,\theta]) \dots \theta^{-1} I(x_n \in [0,\theta]) = \\ = \theta^{-n} I(x_{(n)} \leqslant \theta) \cdot I(x_{(1)} \geqslant 0)$$

Thus, $X_{(n)}$ is a sufficient statistic, because we can denote functions $g(t;\theta) = \frac{1}{\theta^n} I(s \leqslant \theta)$ and $h(x) = I(x_{(1)} \geqslant 0)$

- The support of the distribution cannot depend on θ. That's why a uniform distribution is not belong to an exponential family.

3. Kullback-Leibler divergence is:

$$D_{KL}(P||Q) = E_P\left[\log\frac{P}{Q}\right] = \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx$$

Normal distribition is:

$$f(x|\mu_i, \Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)}{2}\right)$$

Substitute the values:

$$D_{KL}(P||Q) = E_P\left[\log\exp\left(-\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2} + \frac{(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)}{2}\right)\right] = \\ = \frac{1}{2}E_P\left[-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right] =$$

Note:

$$Tr(X^TAX) = X^TAX = Tr(AXX^T) \\ E(X^TAX) = E(Tr(AXX^T)) = TrEAXX^T = Tr(AE(XX^T))$$

That's why:

$$= \frac{1}{2}E_P\left[-Tr((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)) + Tr((x-\mu_2)^T\Sigma^{-1}(x-\mu_2))\right] =$$
$$= \frac{1}{2}E_P\left[-Tr(\Sigma^{-1}(x-\mu_1)(x-\mu_1)^T) + Tr(\Sigma^{-1}(x-\mu_2)(x-\mu_2)^T)\right] =$$
$$= \frac{1}{2}E_P\left[-Tr(\Sigma^{-1}\Sigma) + Tr(\Sigma^{-1}(xx^T - 2x\mu_2^T + \mu_2\mu_2^T)\right] =$$
$$= -\frac{n}{2} + \frac{1}{2}Tr(\Sigma^{-1}(\Sigma + \mu_1\mu_1^T - 2\mu_1\mu_2^T + \mu_2\mu_2^T)) =$$
$$= -\frac{n}{2} + \frac{n}{2} + \frac{1}{2}(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1) =$$
$$= \frac{1}{2}(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)$$

4. Let $X_1, \ldots, X_n$ – i.i.d random variables, that are uniformly $[\theta_1, \theta_2]$ distributed. Besides that $X_{(r)}$ is a random variable, that equals to an order statistic. Let $x_{(1)}, x_{(n)} \in [\theta_1, \theta_2]$. Thus, the conditional probability:

$$P\left(X_{(r)} \leqslant x_{(r)} | X_{(1)} \geqslant x_{(1)}, X_{(n)} \leqslant x_{(n)}\right) =$$
$$= P(r \text{ elements in } [x_{(1)}, x_{(r)}] \mid \text{all n elements in } [x_{(1)}, x_{(n)}]) =$$
$$= \frac{P(r \text{ elements in } [x_{(1)}, x_{(r)}] \text{ and all n elements in } [x_{(1)}, x_{(n)}])}{P(\text{all n elements in } [x_{(1)}, x_{(n)}])} =$$
$$= \begin{cases} 0 & \text{if } x_{(r)} < x_{(1)} \\ \frac{\left(\frac{x_{(r)}-x_{(1)}}{\theta_2-\theta_1}\right)^r \cdot \left(\frac{x_{(n)}-x_{(1)}}{\theta_2-\theta_1}\right)^{n-r}}{\left(\frac{x_{(n)}-x_{(1)}}{\theta_2-\theta_1}\right)^n} = \left(\frac{x_{(r)}-x_{(1)}}{x_{(n)}-x_{(1)}}\right)^r, & \text{if } x_{(1)} \leqslant x_{(r)} \leqslant x_{(n)} \\ 1, & \text{if } x_{(r)} > x_{(n)} \end{cases}$$

Denote:

$$Y = \left(\frac{X_{(r)} - X_{(1)}}{X_{(n)} - X_{(1)}}\right)^r$$

Then:

$$P\left(Y \leqslant y | X_{(1)} = x_{(1)}, X_{(n)} = x_{(n)}\right) =$$
$$= P\left(X_{(r)} \leqslant x_{(1)} + y(x_{(n)} - x_{(1)}) | X_{(1)} = x_{(1)}, X_{(n)} = x_{(n)}\right) =$$
$$= \left(\frac{x_{(1)} + y(x_{(n)} - x_{(1)}) - x_{(1)}}{x_{(n)} - x_{(1)}}\right)^r = y^r$$

This probability doesn't depend on $x_{(1)}$ and $x_{(n)}$. Thus, Y doesn't depend on $(X_{(1)}, X_{(n)})$

5 (a) Fisher information:

$$I(\theta) = E\left[\left(\frac{\partial L(\theta, X)}{\partial\theta}\right)^2 \middle| \theta\right] = -E\left[\frac{\partial^2 L(\theta, X)}{\partial\theta^2} \middle| \theta\right]$$

$$\begin{aligned}
L(\theta, X) &= \ln f(x) = \ln \frac{1}{\pi(1 + (X-\theta)^2)} = \\
&= -\ln\pi - \ln(1 + (X-\theta)^2) \\
\frac{\partial L}{\partial\theta} &= \frac{2(X-\theta)}{1 + (X-\theta)^2} \\
\frac{\partial^2 L}{\partial\theta^2} &= \frac{-2(1 + (X-\theta)^2) + 4(X-\theta)^2}{(1 + (X-\theta)^2)^2} = \\
&= \frac{-2}{1 + (X-\theta)^2} + \frac{4(X-\theta)^2}{(1 + (X-\theta)^2)^2} = \\
I(\theta) &= -\int_{\mathbb{R}} f(x) \cdot \frac{\partial^2 L}{\partial\theta^2} dx = \frac{2}{\pi}\int_{\mathbb{R}} \left(\frac{1}{(1 + (x-\theta)^2)^2} - \frac{2(x-\theta)^2}{(1 + (x-\theta)^2)^3}\right) dx =
\end{aligned}$$

Thanks to:

$$\int_{\mathbb{R}} \frac{1}{(1+t^2)^2} dt = \frac{\pi}{2} \qquad\qquad \int_{\mathbb{R}} \frac{1}{(1+t^2)^3} dt = \frac{\pi}{8}$$

Thus:

$$I(\theta) = \frac{2}{\pi} \cdot \left(\frac{\pi}{2} - 2\frac{\pi}{8}\right) = \frac{1}{2}$$

(b) The likelihood is:

$$L = -n\ln\pi - \sum_i \ln(1 + (x_i - \mu)^2$$

The partial derivative of the log-likelihood function:

$$\frac{\partial L}{\partial\mu} = 2\sum_i \frac{y_i}{1 + y_i^2} = 0, \; y_i = x_i - \mu$$

Thus, the maximum likelihood estimation could be found by solving the nonlinear equation.

import numpy as np
from scipy import stats

```
from scipy.optimize import fsolve
X = stats.cauchy.rvs(loc=12, size=20000)
def f(y): return np.sum( (X - y) / (1+ (X - y) ** 2))
fsolve(f, 11)
```

Cramer–Rao bound:

$$D_\theta \hat{\theta}(x) \geqslant \frac{1}{nI(\theta)} = \frac{2}{n}$$

(c) Code:

```
data = []
for j in range(10000):
X = stats.cauchy.rvs(loc=10, size=10000)
def f(y):
t = X - y
return np.sum(t / (1+t ** 2))
data.append(fsolve(f, 11))
np.var(data)
```

These experiments gave me the same results with Cramer-Rao bound:

$$Result = 0.00019941428709298418 \qquad RC = \frac{2}{10000} = 0.0002$$

It means that the inequality is achieved almost exactly!