Lecture Notes 3: Scale Mixture Distribution

Professor: Zhihua Zhang

3.1 Scale Mixture Distribution

We will show several distributions can be seen as the scale mixture of distribution, which is defined as follows,

$$X \sim F(\theta)$$
$$\theta \sim G(\lambda)$$

, So, $T(x) = \int_\theta F(\theta)G(\lambda)d\theta$ can be seen as a scale mixture of F, where the scale has distribution G.

3.1.1 Student's t-distribution

The Student's t-distribution is a scale of Gaussian distribution, where the scale has a Gamma distribution. Let $X \sim N(\mu, \frac{\sigma^2}{r})$, $r \sim Gamma(\frac{\nu}{2}, \frac{\nu}{2})$, then the integral will be:

$$\begin{aligned}
&\int_0^\infty \frac{r^{-1/2}}{\sqrt{2\pi}\sigma} e^{-\frac{r(x-\mu)^2}{2\sigma^2}} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} r^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}r} dr \\
=&\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})\sigma\sqrt{2\pi}} \int_0^\infty r^{\frac{\nu}{2}-\frac{3}{2}} e^{-\frac{r}{2}(\frac{(x-\mu)^2}{\sigma^2}+\frac{\nu}{2})} dr \\
=&\frac{\nu^{\frac{\nu}{2}}\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\pi}\Gamma(\frac{\nu}{2})} \left[\frac{(x-\mu)^2}{\sigma^2} + \frac{\nu}{2}\right]^{\frac{\nu+1}{2}}
\end{aligned}$$

Note that during the integral, we use a math trick. Since we know $\int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = 1$ from Gamma distribution, so we can get $\int_0^\infty x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha)}{\beta^\alpha}$. This trick will be often used in the follows.

3.1.2 Laplace Distribution

The Laplace distribution is a scale of Gaussian distribution, where the scale has a exponential distribution. Let $X \sim N(\mu, r)$, $r \sim Exponential(\frac{1}{2\sigma^2})$, then we can get the mixture distribution:

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi r}} e^{-\frac{(x-\mu)^2}{2r}} \frac{1}{2\sigma^2} e^{-\frac{r}{2\sigma^2}} dr \\
=&\frac{1}{2\sigma^2\sqrt{2\pi}} \int_0^\infty r^{\frac{1}{2}-1} e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{r}+\frac{r}{\sigma^2}\right)} dr \\
=&\frac{1}{2\sigma} e^{\frac{|x-\mu|}{\sigma}}
\end{aligned}$$

3.1.3 Negative Binomial Distribution

Negative Binomial Distribution is a scale of Poisson distribution, where the scale has a Gamma distribution. Let $K \sim Poisson(\lambda)$, $\lambda \sim Gamma(r, \frac{1-p}{p})$, then we can get the mixture distribution:

$$\begin{aligned}
&\int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \frac{\lambda^{r-1} e^{-\frac{1-p}{p}\lambda}}{\Gamma(r)(\frac{p}{1-p})^r} d\lambda \\
=&\frac{1}{k!\Gamma(r)(\frac{p}{1-p})^r} \int_0^\infty \lambda^{k+r-1} e^{-\frac{\lambda}{p}} d\lambda \\
=&\binom{k+r-1}{k} p^k (1-p)^r
\end{aligned}$$

Homework 1. $\sum_{k=0}^{\infty} Gamma(x|k,\beta) Poisson(k|\lambda)$.

3.2 Statistical Inference (I)

3.2.1 Jeffrey Prior

In order to show Jeffrey prior, we first introduce **Fisher information**. In mathematical statistics, the Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends.

Assume we have a model for random variable X, for example $\mathbb{P}(X|\theta)$. $\mathbb{P}(X|\theta)$ can be seen as a joint function of x and θ. Let $f(x,\theta) = \mathbb{P}(X|\theta)$. Then Fisher information of X about θ is given by:

$$\begin{aligned}
I(\theta) =&\mathbb{E}[(\frac{\partial \log f(x,\theta)}{\partial \theta})^2] \\
=&\int \left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2 f(x,\theta) d\theta
\end{aligned}$$

Lemma 3.1 *Under certain condition,*

$$I(\theta) = -\mathbb{E}[\frac{\partial^2 \log f}{\partial \theta^2}]$$

Proof.

$$\begin{aligned}
\frac{\partial^2 \log f}{\partial \theta^2} =&\frac{\partial}{\partial \theta}\left(\frac{f'}{f}\right) \\
=&\frac{f''}{f} - \left(\frac{f'}{f}\right)^2 \\
=&\frac{f''}{f} - \left(\frac{\partial \log f}{\partial \theta}\right)^2
\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}[\frac{\partial^2 \log f}{\partial \theta^2}] &= \int \frac{\partial^2 \log f}{\partial \theta^2} f dx \\ &= \int \frac{\partial^2 f}{\partial \theta^2} dx - I(\theta) \\ &= \frac{\partial^2}{\partial \theta^2} \int f dx - I(\theta) \\ &= -I(\theta)\end{aligned}$$

□

Now let we go to see Jeffrey prior. When we do MAP(maximum a posteriori), we usually meet $\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)BP(\theta)$. Usually $\mathbb{P}(X|\theta)$ is easy to get, but $\mathbb{P}(\theta)$ (prior) needs our hypothesis. How to choose hypothesis? If we set a prior with hyper-parameter, the training process will be difficult. Jeffrey prior tells us how to choose hypothesis:

$$\mathbb{P}(\theta) \propto \sqrt{I(\theta)}$$

Remark: Jeffrey prior has property called **invariant under reparameterization**, which means if we replace θ with φ, and there is a one to one rejection between θ and φ. Then we can get:

$$\begin{aligned}\mathbb{P}(\varphi) &= \mathbb{P}(\theta) \left| \frac{\partial \theta}{\partial \varphi} \right| \\ &\propto \sqrt{I(\theta) \left(\frac{\partial \theta}{\partial \varphi} \right)^2} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{\partial \log f}{\partial \theta} \right)^2 \right] \left(\frac{\partial \theta}{\partial \varphi} \right)^2} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{\partial \log f}{\partial \varphi} \right)^2 \right]}\end{aligned}$$

Assume $X \sim N(\mu, \sigma^2)$.
Case 1:

3.2.2 Problem: $X = \theta + \epsilon$