Statistical Machine Learning Distributions

Lecture Notes 2: Scale Mixture Distribution

Professor: Zhihua Zhang

2.1 Distribution Function

The CDF of a discrete random variable X can be expressed as the sum of its probability mass function(pmf) $f_X(x)$ as follows:

$$F(x) = \sum_{x_i \le x} f_X(x_i)$$

The CDF of a continuous random variable X can be expressed as the integral of its probability density function(pdf) $f_X(x)$ as follows:

$$F(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

and

$$F'(x) = f_X(x)$$

Lemma 2.1 *Let F be the CDF for a random variable X, then we have*

(1) $\Pr(X = x) = F(x) - F(x^-)$

(2) $\Pr(x < X \le y) = F(y) - F(x)$

(3) $\Pr(X > x) = 1 - F(x)$

(4) If X is continuous, then

$$F(b) - F(a) = \Pr(a < X < b) = \Pr(a \le X < b) = \Pr(a < X \le b) = \Pr(a \le X \le b)$$

Definition 2.3 *Suppose X is a random variable with CDF $F(x)$. The inverse CDF is defined by:*

$$F^{-1}(q) = inf\{x : F(x) > q\}$$

for $q \in [0, 1]$. It's is also called ***quantile function****.*

Definition 2.4 *The* **mode** *of a discrete probability distribution is the value at which its pmf takes its maximum value. The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.*

Remarks:

(1) The pmf is always less than or equal to 1, but the pdf can be greater than 1. For example, the uniform distribution on $[0, 1/5]$, the pdf is $f(x) = 5$. The pdf also can be infinite, e.g., $f(x) = \frac{2}{3}x^{-\frac{1}{3}}$.

(2) $\sum f(x) = 1$ or $\int f(x) = 1$ sometimes is written as $\int \mathrm{d}F(x) = 1$ or $\int F(\mathrm{d}x) = 1$.

(3) We call X and Y are equal in distribution iff $F_X(x) = F_Y(x)$ for any x. Notice that it is **not** the same as $X = Y$. For example, $\Pr(X = 1) = \Pr(X = -1) = \frac{1}{2}$. Let $Y = -X$, then X and Y are equal in distribution but $X \neq Y$.

2.2 Discrete Distribution Examples

2.2.1 Uniform Discrete Distribution

Random variable $X \in \{x_1, x_2, ..., x_n\}$ has a uniform discrete distribution pmf f if

$$f(x) = \begin{cases} \frac{1}{n} & x = x_i, i = 1, 2, ..., n \\ 0 & \text{otherwise} \end{cases}$$

2.2.2 Point Mass Distribution

Random variable X has a point mass distribution pmf f if

$$f(x) = \begin{cases} 1 & x = a \\ 0 & \text{otherwise} \end{cases}$$

2.2.3 Bernoulli Distribution

Random variable $X \in \{0, 1\}$ has a Bernoulli distribution pmf f if

$$f(x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases}$$

where $p \in [0, 1]$. It can also be written as $f(x) = p^x(1-p)^{1-x}$. In binary classification problem, Bernoulli distribution is always used to model the category $y = f(x)$. If $y > 0.5$, it's in class 1, else in class 2.

Example 2.5 *In* ***logistic regression****,*

$$\Pr(y = 0) = [1 + exp(-a^T x)]^{-1}$$

$$\Pr(y = 1) = [1 + exp(-a^T x)]^{-1} exp(-a^T x)$$

So the likelihood function is:

$$\mathcal{L} = \prod_{i=1}^{n} [1 + exp(-a^T x_i)]^{-y_i} [1 + exp(a^T x_i)]^{(y_i - 1)}$$

where $y_i \in \{0, 1\}$. Take the logarithm of the likelihood:

$$log(\mathcal{L}) = \sum_{i=1}^{n} -y_i log(1 + exp(-a^T x_i)) + (y_i - 1) log(1 + exp(a^T x_i))$$

2.2.4 Poisson Distribution

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if for $k = 0, 1, 2, ...$, the probability mass function of X is given by:

$$f(x; \lambda) = \Pr(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \geq 0$$

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = e^{-x} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-x} e^x = 1$$

Remark: If $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

2.2.5 Binomial Distribution

A discrete random variable X is said to have a binomial distribution with parameter n and p, we write $X \sim \text{Binomial}s(n, p)$. The probability mass function is given by:

$$f(x; n, p) = \Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

for $x = 0, 1, 2, ..., n$, where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient. It can be interpreted that the probability of exact k successes after n trials.

Remark: If $X_1 \sim \text{Binomial}(n_1, p)$, $X_2 \sim \text{Binomial}(n_2, p)$, then $X_1+X_2 \sim \text{Binomial}(n_1+n_2, p)$.

By the way, we introduce something about gamma function and a generalization form of $\binom{n}{k}$.

The gamma function (represented by the capital Greek letter Γ) is an extension of the factorial function, with its argument shifted down by 1, to real and complex numbers. That is, if n is a positive integer:

$$\Gamma(n) = (n - 1)!$$

The gamma function is defined for all complex numbers except the negative integers and zero. For complex numbers with a positive real part, it is defined via a convergent improper integral:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \, \mathrm{d}t$$

.

As a generalization of factorial function, $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1) = 0! = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Also, we can define $\binom{r}{k}$ when r is a real number and k is a integer:

$$\binom{r}{k} = \begin{cases} \frac{r(r-1)...(r-k+1)}{k!} & k \geq 0 \\ 0 & k < 0 \end{cases}$$

where $\binom{r}{0} = 1$, $\binom{r}{1} = r$. Then we can get a new binomial theorem: $(1 + z)^r = \frac{f(0)}{0!} z^0 + \frac{f'(0)}{1!} z^1 + \cdots = \sum_k \frac{f^{(k)}(0)}{k!} z^k = \sum_k \binom{r}{k} z^k$, $|z| < 1$.

2.2.6 Negative Binomial Distribution

Suppose there is a sequence of independent Bernoulli trials, each trial having two potential outcomes called "success" and "failure". In each trial the probability of success is p and of failure is $1-p$. We are observing this sequence until a predefined number r of failures has occurred. Then the random number of successes we have seen, X, will have the negative binomial (or Pascal) distribution:

$$X \sim \mathrm{NB}(r, p).$$

The probability mass function of the negative binomial distribution is:

$$f(k; r, p) = \Pr(X = k) = \binom{k+r-1}{k} p^k (1-p)^r$$

for $k = 0, 1, 2, ...$

Note that

$$\begin{aligned} \binom{k+r-1}{k} &= \frac{(k+r-1)(k+r-2)\dots r}{k!} \\ &= \frac{(-1)^k(-r)(-r-1)\dots(-r-k+1)}{k!} \\ &= (-1)^k \binom{-r}{k} \end{aligned}$$

That's why it's called negative binomial distribution. Hence,

$$\sum \Pr(X = k) = (1-p)^r \sum (-1)^k \binom{-r}{k} p^k = (1-p)^r (1-p)^{-r} = 1$$

When $r = 1$, the negative binomial distribution is **geometric distribution**:$\Pr(X = k) = (1-p)^{k-1} p$.

Let $p = \frac{\lambda}{\lambda + r}$. If $r \to \infty$, then $p \to 0$. We can get Poisson distribution:

$$\begin{aligned} \lim_{r\to\infty} f(\lambda) &= \lim_{r\to\infty} \frac{(k+r-1)\dots r}{k!} \left(\frac{\lambda}{r+\lambda}\right)^k \left(\frac{r}{r+\lambda}\right)^r \\ &= \lim_{r\to\infty} \lambda^k \frac{(k+r-1)\dots r}{k!} \left(\frac{1}{r+\lambda}\right)^k \left(\frac{1}{\frac{\lambda}{r}+1}\right)^r \\ &= \lim_{r\to\infty} \frac{\lambda^k}{k!} \frac{(k+r-1)\dots r}{(\lambda+r)^k} \frac{1}{(1+\frac{\lambda}{r})^r} \\ &= \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

Bernoulli Distribution and Measure Let $\Omega = [0, 1]$, $P([a, b]) = b - a$, $0 \le a \le b \le 1$(Lebesgue measure). Fix $P \in (0, 1)$ and let

$$X(\omega) = \begin{cases} 1 & \omega \le p \\ 0 & \omega > p \end{cases}$$

Hence, $\Pr(X = 1) = \Pr(\omega \leq p) = \Pr([0, p]) = p$, $\Pr(X = 1) = \Pr(\omega > p) = \Pr((p, 1]) = 1 - p$.

Homework:

(1) If $\lim_{n\to\infty} a_n = a$, show that $\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$.

(2) If $nt > -1$, show that $(1 - t)^n \geq 1 - nt$.

(3) If $-x < n < m$, show that $(1 + \frac{x}{n})^n \leq (1 + \frac{x}{m})^m$.

2.3 Continuous Distribution Examples

2.3.1 Continuous Uniform Distribution

A continuous random variable X is said to have a uniform distribution in $[a, b]$, if the probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

2.3.2 Normal(Gaussian) Distribution

A continuous random variable X is said to have a Gaussian distribution with parameter μ and σ, if the probability density function of X is given by:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$. The cumulative distribution function of Gaussian random variable X with parameter $\mu = 0$ and $\sigma = 1$ ($X \sim \mathcal{N}(0, 1)$)is:

$$\Phi(z) = \Pr(X < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t$$

2.3.3 Dirac Distribution

The Dirac function, or δ function can be loosely thought of as a function on the real line which is zero everywhere except at the origin, where it is infinite,

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

and which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(x) \, \mathrm{d}x = 1$$

2.3.4 Exponential Power Distribution

A random variable X is said to have an exponential power distribution with parameter μ, σ, q if its probability density function is :

$$f(x) = \frac{1}{2^{\frac{q+1}{q}}\Gamma(\frac{q+1}{q})\sigma}\, e^{\left(-\frac{1}{2}|\frac{x-\mu}{\sigma}|^q\right)}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, $q > 0$.

This family includes the normal distribution when $q = 2$ and it includes the Laplace distribution when $q = 1$: $f(x) = \frac{1}{4\sigma}e^{-\frac{|x-\mu|}{2\sigma}}$

To validate $\int f(x) = 1$, the following formulas may help. For $a > 0, p > 0$,

$$\begin{aligned}
\int_0^\infty x^{p-1}e^{-ax}\mathrm{d}x &= a^{-p}\Gamma(p) \\
\int_0^\infty x^{-(p+1)}e^{-ax^{-1}}\mathrm{d}x &= a^{-p}\Gamma(p) \\
\int_0^\infty x^{p-1}e^{-ax^2}\mathrm{d}x &= \frac{1}{2}a^{-\frac{p}{2}}\Gamma(\frac{p}{2}) \\
\int_0^\infty x^{-(p+1)}e^{-ax^{-2}}\mathrm{d}x &= \frac{1}{2}a^{-\frac{p}{2}}\Gamma(\frac{p}{2})
\end{aligned}$$

More generally, for $a > 0, p > 0$,

$$\begin{aligned}
\int_0^\infty x^{p-1}e^{-ax^q}\mathrm{d}x &= \frac{1}{q}a^{-\frac{p}{q}}\Gamma(\frac{p}{q}) \\
\int_0^\infty x^{-(p+1)}e^{-ax^{-q}}\mathrm{d}x &= \frac{1}{q}a^{-\frac{p}{q}}\Gamma(\frac{p}{q})
\end{aligned}$$

2.3.5 Gamma Distribution

A random variable X that is gamma-distributed is denoted by $X \sim \text{Gamma}(r, \frac{\alpha}{2})$,

$$f(x) = \frac{\alpha^r}{2^r\Gamma(r)}x^{r-1}e^{-\frac{\alpha x}{2}}$$

when $r = 1$, it's exponential distribution. If $X_i \sim \text{Gamma}(r_i, \alpha)$, then $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n r_i, \alpha)$.

Inverse Gamma Distribution

A random variable X that is inverse gamma-distributed is denoted by $X \sim \text{Inv-Gamma}(r, \frac{\beta}{2})$,

$$f(x) = \frac{\beta^\tau}{2^\tau\Gamma(\tau)}x^{-(\tau+1)}e^{-\frac{\beta}{2x}}, \tau = -r$$

2.3.6 Generalized Inverse Gaussian Distribution

A continuous random variable X is said to have generalized inverse Gaussian distribution(GIG) with parameters α, β, r, if the probability density function of X is given by:

$$f(x) = \frac{(\alpha/\beta)^{r/2}}{2K_r(\sqrt{\alpha\beta})} x^{r-1} e^{-(\alpha x+\beta/x)/2}, x > 0$$

where K_r is a modified Bessel function of second kind with index r, $\alpha > 0$, $\beta > 0$.

Properties of Bessel Function

(1) $K_r(u) = K_{-r}(u)$

(2) $K_{r+1}(u) = 2\frac{r}{u}K_r(u) + K_{r-1}(u)$

(3) $K_{1/2}(u) = K_{-1/2}(u) = \sqrt{\frac{\pi}{2u}}e^{-u}$

(4) $u \to 0, K_r(u) \sim \begin{cases} \frac{1}{2}\Gamma(r)\left(\frac{u}{2}\right)^{-r} & r > 0 \\ \ln u & r = 0 \end{cases}$

(5) $u \to \infty, K_r(u) \sim \sqrt{\frac{\pi}{2u}}e^{-u}$

Inverse Gaussian Distribution

Specially, when $r = -\frac{1}{2}$,

$$\begin{aligned} f(x) &= \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp(\sqrt{\alpha\beta}) x^{-\frac{3}{2}} \exp(-\frac{\alpha x + \beta x^{-1}}{2}) \\ &= \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp\frac{-\lambda(x-\mu)^2}{2\mu^2 x} \end{aligned}$$

where $\alpha = \lambda/\mu^2, \beta = \lambda$.
For the case $r = \frac{1}{2}$,

$$f(x) = \left(\frac{\alpha}{2\pi}\right)^{\frac{1}{2}} \exp(\sqrt{\alpha\beta}) x^{-\frac{1}{2}} \exp(-\frac{\alpha x + \beta x^{-1}}{2})$$

Note that $\int_0^\infty x^{-\frac{1}{2}} \exp(-\frac{\alpha x+\beta x^{-1}}{2})dx = (\frac{2\pi}{\alpha})^{\frac{1}{2}} \exp(-\sqrt{\alpha\beta})$, because $\int_0^\infty x^{-\frac{1}{2}} \exp(-\frac{a^2x+b^2x^{-1}}{2})dx = \frac{\sqrt{2\pi}}{a}\exp(-|ab|)$.

2.3.7 Chi-Squared Distribution

A continuous random variable X is said to have chi-squared distribution, if the probability density function of X is given by:

$$f(x) = \frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}, x > 0$$

Note that $\|\boldsymbol{N}_{i=1,\ldots,k}(0,1)\|^2 \sim \chi_k^2$ (The squared norm of k standard normally distributed variables is a chi-squared distribution with k degrees of freedom)

2.3.8 Beta Distribution

A continuous random variable X is said to have beta distribution, if the probability density function of X is given by:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $0 < x < 1$. The beta function is defined as:

$$\text{Beta}(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

When $\alpha = 1, \beta = 1$, it is uniform distribution on $[0,1]$.

2.3.9 Student's t-distribution

A continuous random variable X is said to have Student's t-distribution($X \sim t_\nu$), if the probability density function of X is given by:

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{\frac{\nu+1}{2}}} \frac{1}{\sqrt{\nu\pi}/\sigma}$$

where ν denotes the degree of freedom. When $\nu = 1$, it is Cauchy distribution:

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

and when $\nu \to \infty$, it is Gaussian distribution:

$$\lim_{\nu\to\infty}(1+\frac{\nu+1}{2\nu}\frac{2}{\nu+1}(\frac{x-\mu}{\sigma})^2)^{-\frac{\nu+1}{2}} = \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

We can prove the Stirling Formula $\lim_{p\to\infty} \frac{\Gamma(p)}{(2\pi)^{\frac{1}{2}} p^{p-\frac{1}{2}} e^{-p}} = 1$.

$$\lim_{\nu\to\infty} \nu^{-\frac{1}{2zx}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}$$

It can be shown that the t-distribution is like an infinite sum of Gaussians, where each Gaussian has a different variance:

$$\int_0^\infty \mathcal{N}(x \mid \mu, (\lambda\tau)^{-1}) \text{Gamma}(\tau \mid \frac{\nu}{2}, \frac{\nu}{2}) = t_\nu(x \mid \mu, \lambda^{-1})$$

This means t-distribution is a scale mixture of normal distribution.It results from compounding a Gaussian distribution with mean μ and unknown precision (the reciprocal of the variance), with a gamma distribution placed over the precision with parameters $r = \nu/2$ and $\alpha/2 = \nu/2$. In other words, the random variable X is assumed to have a normal distribution with an unknown precision distributed as gamma, and then this is marginalized over the gamma distribution.

Example 2.6 *Suppose $X \sim Bernoulli(\theta)$, $\theta \sim Beta(\alpha, \beta)$.*

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta)p(\theta \mid \alpha, \beta) \\ &\propto \theta^x (1-\theta)^{1-x} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto \theta^{x+\alpha-1} (1-\theta)^{\beta-x} \\ &\sim Beta(x+\alpha, \beta-x+1) \end{aligned}$$

We say that beta distribution is the conjugate prior for the Bernoulli distribution. Generally, if the posterior distributions $p(\theta \mid x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Example 2.7 *Suppose $X \sim \mathcal{N}(0, \lambda)$.*

$$f(x) = \frac{1}{\sqrt{2\pi}} \lambda^{-\frac{1}{2}} \exp(-\frac{x^2}{2\lambda})$$

If $\lambda \sim Gamma(r, \alpha/2)$,

$$\begin{aligned} p(\lambda \mid x) &\propto p(x \mid \lambda)p(\lambda \mid r, \alpha/2) \\ &\propto \lambda^{-\frac{1}{2}} \exp(-\frac{x^2}{2\lambda}) \lambda^{r-1} \exp(-\frac{\alpha\lambda}{2}) \\ &\propto \lambda^{r-3/2} \exp(-\frac{1}{2}(\frac{x^2}{\lambda} + \alpha\lambda)) \end{aligned}$$

It is generalized inverse Gaussian distribution, but the prior and posterior are not conjugate distributions.

If $\lambda \sim Inv\text{-}Gamma(\tau, \beta/2)$,

$$\begin{aligned} p(\lambda \mid x) &\propto p(x \mid \lambda)p(\lambda \mid \tau, \beta/2) \\ &\propto \lambda^{-\frac{1}{2}} \exp(-\frac{x^2}{2\lambda}) \lambda^{-(\tau+1)} \exp(-\frac{\beta}{2\lambda}) \\ &\propto \lambda^{-(\tau+1)-1/2} \exp(-\frac{1}{2}(\frac{x^2}{\lambda} + \frac{\beta}{\lambda})) \\ &\sim Inv\text{-}Gamma(\tau + 1/2, \beta + x^2) \end{aligned}$$

Hence, Inv-Gamma is a conjugate prior for the Gaussian distribution with known mean.