National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 5

Generative Models for Discrete Data. Classical Probabilistic Models

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

Content

- Discrete distributions;
- Continuous distributions;
- Joint probability distributions;
- Characteristics of distributions.

Common distributions

Discrete distributions

- Binomial and Bernoulli distributions
- Multinomial and multinoulli distributions
- Poisson distribution
- Empirical distribution

Continuous distribution

- Gaussian (normal) distribution
- Degenerate distribution
- Student's t distribution
- Laplace distribution
- Gamma distribution
- Beta distribution
- Pareto distribution

Joint probability distribution

- Multivariate Gaussian
- Multivariate Student t distribution
- Dirichlet distribution

Discrete distributions (1/4). Binomial and Bernoulli distributions

Toss a coin n times. Let $X \in \{0,1,2 \cdots n\}$ be the number of heads. If the probability of head is θ then X has ***binomial distribution***:

$$\boxed{X \sim Bin(k|n,\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k},}$$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!\,k!}$$

is binomial coefficient (number of ways to choose k items from n).

Let $X \in \{0,1\}$ be a binary random variable (success or fail). Then X has ***Bernoulli distribution***:

$$\boxed{X \sim Ber(x|\theta) = \theta^{\mathbb{I}(x=1)} (1-\theta)^{\mathbb{I}(x=0)},}$$

where

$$\mathbb{I}(x) = \begin{cases} 1 & x = true; \\ 0 & x = false; \end{cases}$$

is indicator function.

Discrete distributions (2/4). Multinomial and multinulli distributions

Outcomes of tossing K-sided die. Then random vector $\mathbf{x} = (x_1, x_2 \cdots x_K)$ has ***multinomial distribution***:

$$\boxed{Mu(\mathbf{x}|n,\theta) = \binom{n}{x_1 \cdots x_K} \prod_{j=1}^{K} \theta_j^{x_j}},$$

where

$$\binom{n}{x_1 \cdots x_K} \triangleq \frac{n!}{x_1!\, x_2! \cdots x_K!}$$

is multinomial coefficient.

Let $\mathbf{x} = \big(\mathbb{I}(x=1), \cdots, \mathbb{I}(x=K)\big)$ be a binary random vector. Then $\mathbf{x}$ has ***Multinoulli distribution***:

$$\boxed{Mu(x|1,\theta) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}}$$

Discrete distributions (3/4). Poisson distribution

The $X \in \{0,1,2 \cdots\}$ has ***<u>Poisson distribution</u>*** with parameter $\lambda > 0$ if its pmf is:

$$\boxed{X \sim Pos(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}}$$

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents.

pmf – probability mass function.

Discrete distributions (4/4). Empirical distribution

Given a set of data, $\mathcal{D} = \{x_1, \cdots, x_N\}$, the ***empirical distribution*** is:

$$\boxed{p_{emp}(A) \triangleq \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}(A)\,,}$$

where $\delta_x(A) -$ Dirac measure, defined by

$$\delta_x(A) = \begin{cases} 0, & x \notin A; \\ 1, & x \in A. \end{cases}$$

Continuous distributions (1/8). Gaussian (normal) distribution

Probability distribution function for ***Gaussian distribution*** is given by

$$\boxed{X \sim \mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},}$$

where $\mu = \mathbb{E}[X]$ – the mean; $\sigma^2 = var[X]$ – variance; $\sqrt{2\pi\sigma^2}$ – the normalization constant.

If $\boxed{p(X = x) = \mathcal{N}\left(x|\mu, \sigma^2\right) = \mathcal{N}(0,1)}$ then X has ***standard normal distribution***.

The ***central limit theorem*** tells us that sums of independent random variables have approximately Gaussian distribution.

Continuous distributions (2/8). Degenerate distribution

In the limiting case where $\sigma^2 \to 0$, the Gaussian becomes an infinitely tall and infinitely thin "spike" centered at μ:

$$\boxed{\lim_{\sigma^2 \to 0} \mathcal{N}(x|\mu, \sigma^2) = \delta(x - \mu),}$$

where $\delta(\cdot)$ – Dirac delta function, defined as

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

such that

$$\int_{-\infty}^{+\infty} \delta(x) dx = 1.$$

Continuous distributions (3/8). Student's *t* distribution

Distribution robust to outliers is the ***Student distribution*** with probability distribution function :

$$\boxed{\mathcal{T}(x|\mu,\sigma^2,v) \propto \left[1+\frac{1}{v}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{v+1}{2}}},$$

where $\mu = \mathbb{E}[X]$ – the mean; $\sigma^2 > 0$ – scale parameter; $v > 0$ – degrees of freedom.

To ensure finite variance $v > 2$; for $v \gg 5$ Student distribution rapidly approaches Gaussian distribution and loses its robustness properties.

If $v = 1$ distribution is known as ***Cauchy*** or ***Lorentz distribution***.

Continuous distributions (4/8). Laplace distribution

<u>Laplace distribution</u> or ***<u>double sided exponential distribution</u>*** has probability distribution function:

$$\boxed{Lap(x|\mu,b) \triangleq \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}},$$

where $\mu -$ location parameter; $b\ (b>0) -$ scale parameter.

Laplace distribution is more robust to outlier than Gaussian distribution and puts more probability density at 0, which is useful way to encourage sparsity in a model.

Continuous distributions (5/8). Gamma distribution

Gamma distribution is defined in terms of the shape $a > 0$ and the rate $b > 0$:

$$\boxed{Ga(x|a,b) \triangleq \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}.}$$

where $\Gamma(a)$ – gamma function:

$$\Gamma(a) \triangleq \int_0^{+\infty} u^{x-1} e^{-u} du\,.$$

Inverse gamma distribution defined by

$$\boxed{IG(x|a,b) \triangleq \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-b/x}.}$$

Continuous distributions (6/8). Special cases of gamma distribution

Exponential distribution is defined by $Expon(x|\lambda) \triangleq Ga(x|1, \lambda)$. Distribution describes the times between Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate λ.

Erlang distribution is the same as the Gamma distribution where a is integer – $Erlang(x|\lambda) = Ga(x|a, \lambda), a \in \mathbb{Z}$.

Chi-squared distribution is the sum of squared Gaussian random variables and is defined as $\chi^2(x|v) = Ga\left(x|\frac{v}{2}, \frac{1}{2}\right)$.

Continuous distributions (7/8).
Beta distribution

<u>Beta distribution</u> has support over the interval $[0; 1]$ and is defined as follow:

$$\boxed{Beta(x|a,b) \triangleq \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}}$$

where $B(p,q)$ – beta function:

$$B(p,q) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Continuous distributions (8/8). Pareto distribution

The ***Pareto distribution*** is defined as follow:

$$\boxed{Pareto(x|k,m) \triangleq km^k x^{-(k+1)} \mathbb{I}(x \geq m)}$$

As $k \to +\infty$, the distribution approaches $\delta(x-m)$. Distribution has the long (heavy) tails and it is widely used for modeling the power-low dependencies.

Joint probability distributions (1/3). Multivariate Gaussian

The ***multivariate Gaussian (normal) distribution*** in D dimensions defined as:

$$\boxed{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}.$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ – the mean vector; $\boldsymbol{\Sigma} = cov[\mathbf{x}]$ – the $D \times D$ covariance matrix.

Precision (concentration) matrix is just the inverse covariance matrix:

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}.$$

Joint probability distributions (2/3). Multivariate Student distribution

The ***multivariate Student's t distribution*** in $\underline{D}$ dimensions defined as:

$$\boxed{\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, v) \triangleq \frac{\Gamma(v/2 + D/2)}{\Gamma(v/2)} \cdot \frac{|\boldsymbol{\Sigma}|^{-1/2}}{v^{D/2}\pi^{D/2}} \cdot \left[1 + \frac{1}{v}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]^{-\frac{v+D}{2}}}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ – the mean vector; $\boldsymbol{\Sigma}$ – the scale matrix.

Joint probability distributions (3/3). Dirichlet distribution

The ***Dirichlet distribution*** is generalization of beta distribution, defined by:

$$\boxed{Dir(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{d=1}^{D} {x_d}^{\alpha_d - 1} \mathbb{I}(\mathbf{x} \in S_K)}$$

where

$$S_D = \left\{ x: 0 \leq x_d \leq 1, \sum_{d=1}^{D} x_d = 1 \right\}$$

is probability simplex;

$$B(\boldsymbol{\alpha}) \triangleq \frac{\prod_{d=1}^{D} \Gamma(\alpha_d)}{\Gamma(\alpha_0)}, \alpha_0 \triangleq \sum\nolimits_{d=1}^{D} \alpha_d \,,$$

is natural generalization of the beta function to D variable.

Characteristics of discrete and continuous distributions

Distribution	Mean	Variance	Skewness	Kurtosis
Discrete distribution				
Binomial $Bin(k\|n,\theta)$	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$
Bernoulli $Ber(x\|\theta)$	p	$p(1-p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1-6p(1-p)}{p(1-p)}$
Multinomial $Mu(\mathbf{x}\|n,\theta)$	np_i	$np_i(1-p_i)$	$-$	$-$
Multinoulli $Mu(\mathbf{x}\|1,\theta)$	$\mathbf{p}$	$\mathbf{\Sigma}_{ij} = \begin{cases} p_i(1-p_i), i=j \\ -p_i p_j, i \neq j \end{cases}$	$-$	$-$
Poisson $Pos(x\|\lambda)$	λ	λ	$\lambda^{-1/2}$	λ^{-1}

Characteristics of discrete and continuous distributions

Distribution	Mean	Variance	Skewness	Kurtosis
Continuous distribution				
Gaussian $\mathcal{N}(x\|\mu,\sigma^2)$	μ	σ^2	0	0
Student's $\mathcal{T}(x\|\mu,\sigma^2,v)$	$0\ (v>1)$	$\begin{cases}\frac{v}{v-2}, v>2\\ +\infty, v\in(1;2]\end{cases}$	$0\ (v>3)$	$\begin{cases}\frac{6}{v-4}, v>4\\ +\infty, v\in(2;4]\end{cases}$
Laplace $Lap(x\|\mu,b)$	μ	$2b^2$	0	3
Gamma $Ga(x\|a,b)$	a/b	a/b^2	$2/\sqrt{a}$	$6/a$
Beta $Beta(x\|a,b)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$	$\frac{6[(a-b)^2(a+b+1)-ab(a+b+2)]}{ab(a+b+2)(a+b+3)}$
Pareto $Pareto(x\|k,m)$	$\begin{cases}+\infty, k\le 1\\ \frac{km}{k-1}, k>1\end{cases}$	$\begin{cases}+\infty, k\in(0;2]\\ \frac{km^2}{(k-1)^2(k-2)}, k>2\end{cases}$	$\frac{2(k+1)}{k-3}\sqrt{\frac{k-2}{k}}, k>3$	$\frac{6(k^3+k^2-6k-2)}{k(k-3)(k-4)}, k>4$

Characteristics of discrete and continuous distributions

Distribution	Mean	Variance	Skewness	Kurtosis
Joint distribution				
Multivariate Gaussian $\mathcal{N}(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	–	–
Multivariate Student's $\mathcal{T}(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},v)$	$\boldsymbol{\mu}, v > 1$	$\frac{v}{v-2}\boldsymbol{\Sigma}, v > 2$	–	–
Dirichlet $Dir(\mathbf{x}\mid\boldsymbol{\alpha})$	$\frac{\alpha_i}{\sum_k \alpha_k}$	$\frac{\alpha_i(\alpha_0-\alpha_i)}{{\alpha_0}^2(\alpha_0+1)}, \alpha_0 = \sum_k \alpha_k$	–	–

Conclusion

- Common discrete, continuous and joint probability distributions were considered;
- Key parameters of these distributions were presented.