CS 5135/6035 Learning Probabilistic Models

Lecture 5: Continuous Probability Distributions[1]

Gowtham Atluri

September 6, 2018

[1] These slides cover material from Chapter 5 of the book, Scheaffer, Richard L., and Linda Young. Introduction to Probability and its Applications, 2009.

Part a. Characterizing Continuous Probability Distributions

- Continuous random variables
- Probability density function
- Cumulative distribution function
- Expectation (Mean) and Variance

Continuous random variables

- Discrete random variables
  - Number of coin flips that turned heads in n trials (finite:$1, 2, \ldots n$)
  - Number of crimes reported in Cincinnati (countably-infinite: $0, 1, 2, \ldots \ldots$)
- Many random variables seen in practice have more than a countable collection of possible values
  - Weight of adult patients visiting a clinic (80 to 300 lbs)
  - Diameters of machine rods in an industrial process (1.2 to 1.5 cm)
  - Proportions of impurities in ore samples (0.10 to 0.80)
- These random variables can take on any value in an interval of real numbers
- No guarantee that all values will we found in a sample if one looks long enough
  - no value can be ruled out as a possible observation
- As these r.v. have a continum of possible values, they are called continuous r.vs.

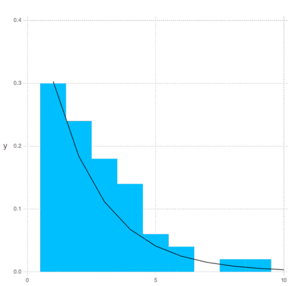
Probability assignment for Continuous random variables

- An experimenter is measuring the life length x of a transistor
  - x has infinite number of possible values
- We cannot assign a positive probability to each possible outcome
  - and ensure that they sum to 1
- We can assign positive probabilities to invtervals of real numbers
  - in a manner consistent with probability axioms

Example

- Let x be a r.v. capturing the life lengths of batteries of a certain type
- Data collected from 50 batteries are as follows:

```
## 10×5 Array{Float64,2}:
##  0.406  0.538  5.587  0.023  3.491
##  0.685  0.234  0.517  0.225  2.921
##  4.778  4.025  3.246  1.514  1.624
##  1.725  3.323  2.33   3.214  0.334
##  8.223  2.92   1.064  3.81   4.49
##  2.343  5.088  2.563  3.334  1.267
##  1.401  1.458  0.511  2.325  1.702
##  1.507  1.064  3.246  0.333  2.634
##  0.294  0.774  2.33   7.514  1.849
##  2.23   0.761  1.064  0.968  0.186
```



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

Example

- The function $f(x) = \frac{1}{2}e^{-x/2}, x > 0$ seems to fit the data well
- This function serves as a mathematical model for relative frequency of the data
- Before purchasing a battery of this type, we want to know
  - the probability that it will last longer than 400 hours

$$\int_4^{\infty} \frac{1}{2}e^{-x/2}dx = -e^{-x/2}|_4^{\infty} = -e^{-\infty/2} + e^{-4/2} = e^{-2} = 0.135$$

- this is similar to the fraction of life times that exceed 4: $8/50 = 0.16$

Example

- The function $f(x) = \frac{1}{2}e^{-x/2}, x > 0$ seems to fit the data well
- This function serves as a mathematical model for relative frequency of the data
- Before purchasing a battery of this type, we want to know
  - the probability that it will last longer than 400 hours

$$\int_4^\infty \frac{1}{2}e^{-x/2}dx = -e^{-x/2}|_4^\infty = -e^{-\infty/2} + e^{-4/2} = e^{-2} = 0.135$$

- this is similar to the fraction of life times that exceed 4: $8/50 = 0.16$
- More importantly, the mathematical model can provide satisfactory answers that cannot be otherwise obtained
  - E.g., probability x is greater than 9

$$\int_9^\infty \frac{1}{2}e^{-x/2}dx = 0.011$$

Probability Density Function

A function $f(x)$ that models the relative frequency behavior of the continuous valued data is called **probability density function** (pdf). For $f(x)$ to be a **pdf** , the following criteria should be met:

- $f(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $p(a \leq x \leq b) = \int_a^b f(x)dx$

Probability Density Function

A function $f(x)$ that models the relative frequency behavior of the continuous valued data is called **probability density function** (pdf). For $f(x)$ to be a **pdf** , the following criteria should be met:

- $f(x) \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $p(a \leq x \leq b) = \int_a^b f(x)dx$

A few things to note:

- $p(x = a) = \int_a^a f(x)dx = 0$
  - The $p(x = 4.97) = 0$ does not rule out 4.97 as a possible length
  - chance of observing this particular length is quite small
- $p(a \leq x \leq b) = p(a < x \leq b) = p(a \leq x < b) = p(a < x < b)$
  - only for continuous random variables

Cumulative distribution function (cdf)

- The distribution function for a random variable x is defined as $cdf(b) = p(x \leq b)$.
- If x is continuous with pdf $f(x)$, then

$$cdf(b) = \int_{-\infty}^{b} f(x)dx$$

Cumulative distribution function (cdf)

- The distribution function for a random variable x is defined as $cdf(b) = p(x \leq b)$.
- If x is continuous with pdf $f(x)$, then

$$cdf(b) = \int_{-\infty}^{b} f(x)dx$$

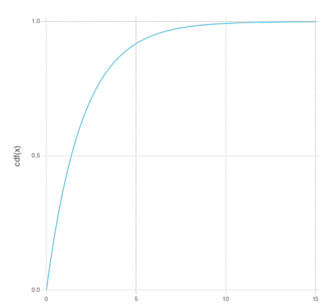
- Example: For a pdf

$$f(x) = \begin{cases} \frac{1}{2}e^{-x/2} & x > 0 \\ 0 & otherwise \end{cases}$$

$$cdf(b) = \int_0^b \frac{1}{2}e^{-x/2}dx = -e^{-x/2}|_a^b = \begin{cases} 1 - e^{-b/2}, & b > 0 \\ 0, & b \leq 0 \end{cases}$$

Cumulative distribution function (cdf) plot

$$cdf(b) = \int_0^b \frac{1}{2}e^{-x/2}dx = -e^{-x/2}|_a^b = \begin{cases} 1 - e^{-b/2}, & b > 0 \\ 0, & b \leq 0 \end{cases}$$



Cumulative distribution function (cdf)

- CDF of a continuous r.v. is continuous over the whole real line
  - in contrast to a discrete r.v., which is a step function
- Four properties of distribution function are:
  - $\lim_{x \to -\infty} cdf(x) = 0$
  - $\lim_{x \to \infty} cdf(x) = 1$
  - cdf(x) is a non-decreasing function
    - if $a < b$, $cdf(a) \leq cdf(b)$
  - cdf(x) is righ-hand continuous $\lim_{h \to 0+} cdf(x+h) = cdf(x)$
- In addition, the cdf of a continuous r.v. is also left-hand continuous
  - $\lim_{h \to 0-} cdf(x+h) = cdf(x)$
- pdf can also be derived from cdf

$$f(x) = \frac{d}{dx} cdf(x), \quad x \in \mathbb{R}$$

Expectation (Mean) and Variance of a continuous r.v.

Expected value of a continuous r.v. x that has a pdf $f(x)$ is given by

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

If x is a continuous r.v. with pdf $f(x)$, and if $g(x)$ is any real-valued function of x, then

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Expectation (Mean) and Variance of a continuous r.v.

Expected value of a continuous r.v. x that has a pdf $f(x)$ is given by

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

If x is a continuous r.v. with pdf $f(x)$, and if $g(x)$ is any real-valued function of x, then

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Variance of a continuous r.v. x that has a pdf $f(x)$ is given by

$$\sigma^2 = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = E(x^2) - \mu^2$$

Example for computing mean (μ) for continuous r.v.

For a given teller in a bank, let x denote the proportion of time, out of a 40-hour work week, that he is directly serving customers. Suppose that x has a pdf

$$f(x) = \begin{cases} 3x^2, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Find the mean propertion of time during a 40-hour workweek the teller directly serves customers.

Example for computing mean (μ) for continuous r.v.

For a given teller in a bank, let x denote the proportion of time, out of a 40-hour work week, that he is directly serving customers. Suppose that x has a pdf

$$f(x) = \begin{cases} 3x^2, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Find the mean propertion of time during a 40-hour workweek the teller directly serves customers.

$$\begin{aligned} E(x) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_0^1 x(3x^2) dx \quad = \quad \int_0^1 3x^3 dx \\ &= 3[\frac{x^4}{4}]_0^1 \quad = \quad \frac{3}{4} \quad = \quad 0.75 \end{aligned}$$

On average, teller spends 75% of his time each week directly serving customers.

Example for computing variance (σ^2) for continuous r.v.

A continuous r.v. x has a pdf

$$f(x) = \begin{cases} 3x^2, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Find the variance of the propertion of during a 40-hour workweek the teller directly serves customers.

$\sigma^2 = E(x^2) - [E(x)]^2$

Example for computing variance (σ^2) for continuous r.v.

A continuous r.v. x has a pdf

$$f(x) = \begin{cases} 3x^2, & 0 \le x \le 1 \\ 0, & \textit{otherwise} \end{cases}$$

Find the variance of the proportion of during a 40-hour workweek the teller directly serves customers.

$\sigma^2 = E(x^2) - [E(x)]^2$

$$\begin{aligned} E(x^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx \\ &= \int_0^1 x^2(3x^2)dx \quad = \quad \int_0^1 3x^4 dx \\ &= 3[\frac{x^4}{4}]_0^1 \quad = \quad \frac{3}{5} \quad = \quad 0.6 \end{aligned}$$

$$\sigma^2 = E(x^2) - [E(x)]^2 = 0.6 - (0.75)^2 = 0.0375$$

Part b. Standard Distributions

- Uniform distribution
- Exponential distribution
- Gamma distribution
- Normal/Gaussian distribution
- Beta distribution
- Weibull distribution

Uniform distribution

- Consider an experiment that consists of observing events in a certain time frame
  - arrival of buses at a bus stop
  - telephone calls coming into a switchboard during a specified period
- We are interested in the probability distribution of the actual time of occurrence (x) of the event
  - A bus arrived between 8:00 and 8:10
- A simple model assumes that x is equally likely to lie in any subinterval (of length d)
  - irrespective of where it lies with (a, b)
- This assumption leads to the **uniform probability distribution**

Uniform distribution

Probability density function

Uniform distribution has a pdf

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \textit{elsewhere} \end{cases}$$

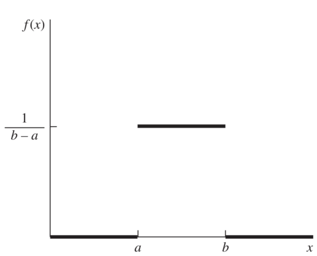
Cumulative distribution function

Uniform distribution has a cdf

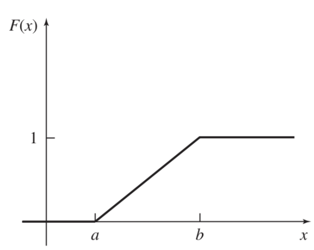
$$cdf(x) = \begin{cases} 0, & x < a \\ \int_a^x \frac{1}{b-a}dt = \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Uniform distribution (plots)

Probability density function $f(x)$



Cumulative distribution function $cdf(x)$



Relationship between Uniform and Poisson

- Suppose that the # events that occur in an interval – say, $(0, t)$ – has a Poisson distribution
- If exactly one of these events is known to have occurred in the interval (a, b)
  - then the conditional probability distribution of the actual time of occurrence for this event is uniform over (a, b)

Mean of Uniform distribution

$$\begin{aligned}
E(x) &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_a^b x(\frac{1}{b-a})dx \\
&= (\frac{1}{b-a}) \int_a^b x dx \\
&= (\frac{1}{b-a}) \frac{x^2}{2}|_a^b \\
&= (\frac{1}{b-a})(\frac{b^2-a^2}{2}) \\
&= \frac{a+b}{2}
\end{aligned}$$

Mean lies at the midpoint of the interval

Variance of Uniform distribution

$\sigma^2(x) = E[(x-\mu)^2] = E(x^2) - \mu^2$

we have $\mu = \frac{a+b}{2}$, we need to compute $E(x^2)$

$$\begin{aligned}
E(x^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx \\
&= \int_a^b x^2(\frac{1}{b-a})dx \\
&= (\frac{1}{b-a}) \int_a^b x^2 dx \\
&= (\frac{1}{b-a}) \frac{x^3}{3}|_a^b \\
&= (\frac{1}{b-a})(\frac{b^3-a^3}{3}) \\
&= \frac{a^2+ab+b^2}{3}
\end{aligned}$$

$$\begin{aligned}
\sigma^2(x) &= E(x^2) - \mu^2 \\
&= \frac{a^2+ab+b^2}{3} - (\frac{a+b}{2})^2 \\
&= \frac{1}{12}[4(a^2+ab+b^2) - 3(a+b)^2] \\
&= \frac{1}{12}(b-a)^2
\end{aligned}$$

Variance depends only on the length of the interval

Computing probability based on Uniform distribution: example

Delivery time (t) for a water pump at a farmer's location is uniformly distributed over the interval from 1 to 4 days. What is the probability that the pump does not arrive within the first three days.

$$f(x) = \begin{cases} \frac{1}{4-1} = \frac{1}{3} & 1 \leq x \leq 4 \\ 0 & elsewhere \end{cases}$$

$$p(t) = \int_{-\infty}^{\infty} \frac{1}{3} dt$$

$$\begin{aligned}
p(t > 3) &= \int_3^{\infty} \frac{1}{3} dt \\
&= \int_3^4 \frac{1}{3} dt \\
&= \frac{1}{3} t|_3^4 \\
&= \frac{1}{3}(4-3) \\
&= \frac{1}{3}
\end{aligned}$$

Computing probability based on Uniform distribution: Julia example

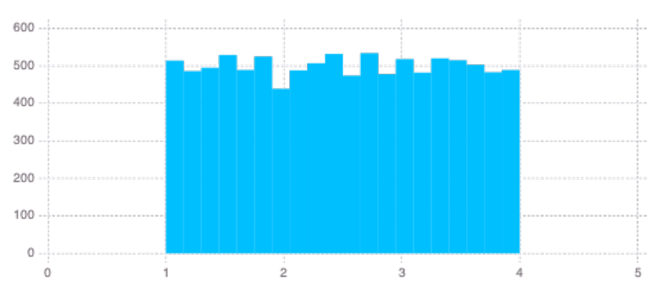
Delivery time (t) for a water pump at a farmer's location is uniformly distributed over the interval from 1 to 4 days. What is the probability that the pump does not arrive within the first three days.

```
using Distributions;
d = Uniform(1, 4);
#computing p(t>3) = 1-cdf(3)
1-cdf(d,3)
```

```
## 0.33333333333333337
```

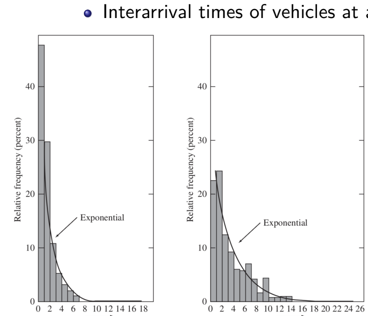

Julia example for sampling from a uniform distribution

```
d = Uniform(1, 4);
sample = rand(d,10000);
myplot = plot(x=sample,Geom.histogram(bincount=20),
              Coord.Cartesian(xmin=0, xmax=5));
draw(PNG("./figs/uniform_sample.png", 6inch, 3inch), myplot);
```



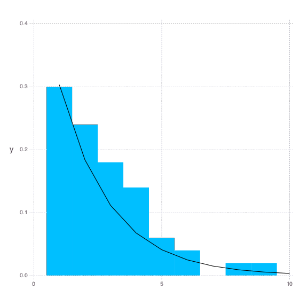
Exponential distribution

- Exponential cruve seems to fit many r.vs in engineering and the sciences
    - Interarrival times of vehicles at a fixed point on the freeway



(a) Sample no. 1 (VOL. 2400 veh/h)

(b) Sample no. 1 (VOL. 1000 veh/h)



Exponential distribution

Probability density function

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x \geq 0 \\ 0, & elsewhere \end{cases}$$

where the parameter θ is a constant (θ>0)
θ is the rate at which the curve decreases.

Cumulative distribution function

$$cdf(x) = \begin{cases} 0, & x < 0 \\ \int_0^x \frac{1}{\theta} e^{-x/\theta} dx = -e^{-x/\theta}|_0^x = 1 - e^{-x/\theta}, & x \geq 0 \end{cases}$$

Gamma function $\Gamma(\alpha)$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

- This gamma function has some interesting properties:
  - $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
  - $\Gamma(n) = (n-1)!$

$$\Gamma(\alpha + 1) = \int_0^\infty x^\alpha e^{-x} dx$$

Using integration by parts, let

$$u = x^\alpha \text{and} \qquad v = -e^{-x}$$
$$du = \alpha x^{\alpha-1} \qquad dv = e^{-x} dx$$

Gamma function $\Gamma(\alpha)$

$$\Gamma(\alpha + 1) = \int_0^\infty x^\alpha e^{-x} dx$$

Using integration by parts, let

$$u = x^\alpha \qquad \text{and} \qquad v = -e^{-x}$$
$$du = \alpha x^{\alpha-1} \qquad dv = e^{-x} dx$$

Then

$$\begin{aligned} \Gamma(\alpha + 1) &= -x^\alpha e^{-x}|_0^\infty + \int_0^\infty \alpha x^{\alpha-1} e^{-x} dx \\ &= -(0-0) + \alpha \int_0^\infty x^{\alpha-1} e^{-x} dx \\ &= \alpha\Gamma(\alpha) \end{aligned}$$

Gamma function $\Gamma(\alpha)$

Introducing β

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx$$

for positive constants α and β can be evaluated by introducing $y = x/\beta$ or $x = \beta y$, where $dx = \beta dy$.

$$\int_0^\infty (\beta y)^{\alpha-1} e^{-y} (\beta dy) = \beta^\alpha \int_0^\infty y^{\alpha-1} e^{-y} dy = \beta^\alpha \Gamma(\alpha)$$

Mean of Exponential Distribution

$$\begin{aligned} E(x) &= \int_{-\infty}^\infty x f(x) dx \\ &= \int_0^\infty x (\frac{1}{\theta}) e^{-x/\theta} dx \\ &= \frac{1}{\theta} \int_0^\infty x e^{-x/\theta} dx \\ &= \frac{1}{\theta} \Gamma(2) \theta^2 \\ &= \theta \end{aligned}$$

Parameter θ is actually the mean of the distribution.

Variance of Exponential Distribution

$\sigma^2 = E(x^2) - \mu^2$

$$\begin{aligned} E(x) &= \int_{-\infty}^\infty x^2 f(x) dx \\ &= \int_0^\infty x^2 (\frac{1}{\theta}) e^{-x/\theta} dx \\ &= \frac{1}{\theta} \int_0^\infty x^2 e^{-x/\theta} dx \\ &= \frac{1}{\theta} \Gamma(3) \theta^3 \\ &= 2\theta^2 \end{aligned}$$

It follows that

$$\begin{aligned} \sigma^2 &= E(x^2) - \mu^2 \\ &= 2\theta * 2 = \theta^2 \\ &= \theta^2 \end{aligned}$$

θ is the std as well as the mean.

Probabilistic reasoning usnig Exp. Dist.: Example

Problem: A sugar refinery has three processing plants that processes raw sugar in bulk. The amount of sugar that one plant can process in one day can be modeled as having an exponential distribution with a mean of 4 tons for each of the three plants. If the plants operate independently, find the probability that exactly two of the three plants will process more than 4 tons on a given day.

This needs to be answered in two steps:

1. Probability that one plant will process more than 4 tons per day
2. Probability that two of the three plants will do this

The first step can be handled using exponential distribution, while the second step can be handled using binomial distribution.

$$p(x > 4) = \int_4^\infty f(x)d(x) = \int_4^\infty \frac{1}{4}e^{-x/4}d(x) = e^{-x/4}|_4^\infty = e^{-1} = 0.37$$

p(two plants use more than 4 tons) $= \binom{3}{2}(0.37)^2(0.63) = 0.26$.

Probabilistic reasoning usnig Exp. Dist.: Julia example

The first step can be handled using exponential distribution, while the second step can be handled using binomial distribution.

$$p(x > 4) = \int_4^\infty f(x)d(x) = \int_4^\infty \frac{1}{4}e^{-x/4}d(x) = e^{-x/4}|_4^\infty = e^{-1} = 0.37$$

```
d1 = Exponential(4);
p_x_gt_4 = 1-cdf(d1,4)
```

```
## 0.36787944117144233
```

p(two plants use more than 4 tons) $= \binom{3}{2}(0.37)^2(0.63) = 0.26$.

```
d2 = Binomial(3,p_x_gt_4);
pdf(d2,2)
```

```
## 0.2566446446062462
```


Probabilistic reasoning using Exp. Dist.: Example

The amount of sugar that a plant can process in one day can be modeled as having an exponential distribution with a mean of 4 tons. How much raw sugar should be stocked for that plant each day so that the chance of running out of product is only 0.05?

Let a denote the amount to be stocked.

$$p(x > a) = \int_a^\infty \frac{1}{4}e^{-x/4}d(x) = e^{-a/4}$$

We want to choose a such that $p(x > a) = e^{-a/4} = 0.05$.

Solving this, $a = 11.98$.

Memoryless property

- Geometric distribution is memoryless among the discrete distributions
- Exponential distribution is the continuous distribution that is memoryless
$$p(x > a + b | x > a) = p(x > b)$$
- This memoryless property sometimes causes concerns about the exponential distribution's usefulness as a model.
    - E.g., the length of time that a light bulb burns may be modeled with an exponential distribution
    - Memoryless property implies that, if a bulb has burned for 10,000 hours, the probability it will burn at least 10 more hours is the same as the probability that the bulb would burn more than 10 hours when new
    - This failure to account for the deterioration of the bulb over time is the property raises question the appropriateness of the exponential model for life-time data

Connection with Poisson distribution

- If the number of events x in a specified area has Poisson dist., the distance between any event and the next event has exp. dist.
- For events occuring according to Poisson dist., the waiting time from occurrence of any event until the next has an exp. distribution.
    - Suppose events are occuring in time according to Poisson dist. with rate of λ events/hour
    - y be the $\#$ events in t hours, with mean λt
    - How long do I have to wait to see the first event occur?
    - x be the length of time until first event.

$$p(x > t) = p[y = 0 \text{ on the interval } (0, t)] = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

$$p(x >\leq t) = 1 - p(x > t) = 1 - e^{-\lambda t}$$

Connection with Poisson distribution

- y be the $\#$ events in t hours, with mean λt
- How long do I have to wait to see the first event occur?
- x be the length of time until first event.

$$p(x > t) = p[y = 0 \text{ on the interval } (0, t)] = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

$$p(x \leq t) = 1 - p(x > t) = 1 - e^{-\lambda t}$$

Notice that $p(x \leq t) = cdf(t)$, has the form of cdf of exponential dist. $\lambda = 1/\theta$.

Differentiating this, we see that pdf of x is

$$f(t) = \frac{d(cdf(t))}{dt} = \frac{d(1 - e^{-\lambda t})}{dt} = \lambda e^{-\lambda t} = \frac{1}{\theta}e^{-t/\theta}, \quad t < 0$$

This is the pdf for exponential distribution.