EE 126 Probability and Random Processes
January 30, 2018 Continuous Probability

1 Reading

3.1. Continuous Random Variables and PDFs

A random variable is called **continuous** if there is a nonnegative function f_X, called the **probability density function** (PDF) of X, such that

$$P(X \in B) = \int_B f_X(x)dx$$

for every subset B of the real line. The probability that the value of X falls within an interval $[a, b]$ should be

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

and can be interpreted as the area under the PDF's curve. Note that any single value a has $P(X = a) = 0$, so including or excluding the endpoints of an interval has no effect on its probability (e.g. $P(a \leq X \leq b) = P(a < X < b)$).

Also, if it is to be called a PDF, f_X must be nonnegative ($f_X(x) \geq 0$ for every x) and have the normalization property

$$\int_{-\infty}^{\infty} f_X(x)dx = P(-\infty < X < \infty) = 1$$

If δ is very small, then $P([x, x + \delta]) \approx f_X(x) \cdot \delta$.

3.1.1. Expectation

The expectation of a continuous random variable X is just like the discrete case, except the PMF is replaced by the PDF and summation is replaced by integration. $\mathbb{E}[X]$ can still be interpreted as the "center of gravity" of the PDF, or as the anticipated average value of X in a large number of independent repetitions of the experiment.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx$$

The expected value rule for a function $g(X)$ has the form

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx$$

The variance of X is defined by

$$var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x)dx \geq 0$$

3.1.2. Exponential Random Variable

An **exponential random variable** X has a PDF of the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where λ is a positive parameter characterizing the PDF. The probability that X exceeds a certain value decreases exponentially, as for any $a \geq 0$ we have $P(X \geq a) = e^{-\lambda a}$. This can be a good model for the amount of time until an incident of interest takes place, such as a meteorite landing anywhere in the Sahara desert.

The mean and the variance of an exponential random variable can be calculated to be $\mathbb{E}[X] = \frac{1}{\lambda}$ and $var(X) = \frac{1}{\lambda^2}$.

3.2. Cumulative Distribution Functions

While PMFs are only for *discrete* random variables and PDFs are only for *continuous* random variables, CDFs (**cumulative distribution functions**) are for all. The CDF of a random variable X is denoted by F_X and provides the probability $P(X \leq x)$. In particular, for every x we have

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(t)dt, & \text{if } X \text{ is continuous} \end{cases}$$

3.2.1. Properties of a CDF

- F_X is monotonically nondecreasing: if $x \leq y$, then $F_X(x) \leq F_X(y)$.
- $F_X(x)$ tends to 0 as $x \to -\infty$, and to 1 as $x \to \infty$.
- If X is discrete, then $F_X(x)$ is a piecewise constant function of x.
- If X is continuous, then $F_X(x)$ is a continuous function of x.
- If X is discrete and takes integer values, the PMF and the CDF can be obtained from each other by summing or differencing:

$$F_X(k) = \sum_{i=-\infty}^{k} p_x(i)$$
$$p_X(k) = P(X \leq k) - P(X \leq k-1) = F_X(k) - F_X(k-1)$$

- If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$
$$f_X(x) = \frac{dF_X}{dx}(x)$$

Because the CDF is defined for any type of random variable, it provides a convenient means for exploring relationships between continuous and discrete random variables, which might otherwise lie in distinct universes.

3.3. Normal Random Variables

A **normal** (aka **Gaussian**) random variable is a continuous random variable with a PDF of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

where $\mathbb{E}[X] = \mu$ and $var(X) = \sigma^2$. Note that the PDF is symmetric around μ. The **standard normal** is defined as the normal random variable with mean 0 and variance 1. Its CDF is denoted by Φ:

$$\Phi(y) = P(Y \leq y) = P(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2}dt$$

One important property of normal random variables is that they are preserved by linear transformations. If X is a normal random variable with mean μ and variance σ^2, then the random variable $Y = aX + b$ is also normal, with mean $E[Y] = a\mu + b$ and variance $var(Y) = a^2\sigma^2$.

This means that we can "standardize" any normal random variable X (with mean μ and variance σ^2) by defining a new random variable $Y = (X - \mu)/\sigma$, which will then have mean 0 and variance 1. Then we can calculate X's CDF using the standard normal table:

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Y \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Normal random variables are important because they're good for modeling the additive effect of many independent factors. The sum of a large number of i.i.d. random variables, whether they are individually normal or not, has an approximately normal CDF!

3.4. Joint PDFs of Multiple Random Variables

Two continuous random variables associated with the same experiment are **jointly continuous** and can be described in terms of a **joint PDF** $f_{X,Y}$ if $f_{X,Y}$ is a nonnegative function that satisfies

$$P((X,Y) \in B) = \iint\limits_{(x,y)\in B} f_{X,Y}(x,y)dx\,dy$$

for every subset B of the two-dimensional plane. In the particular case where B is a rectangle of the form $B = \{(x,y)|a \leq x \leq b, c \leq y \leq d\}$, we have

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)dx\,dy$$

We can view $f_{X,Y}(a,c)$ as the "probability per unit area" in the vicinity of (a,c).

The marginal PDF f_X of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

while the marginal PDF f_Y of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

If we fix some subset S of the two-dimensional plane, the corresponding uniform joint PDF on S is defined to be

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\text{area of } S} & \text{if } (x,y) \in S \\ 0 & \text{otherwise} \end{cases}$$

while for any set $A \subset S$, the probability that (X,Y) lies in A is

$$P((X,Y) \in A) = \frac{\text{area of } A}{\text{area of } S}$$

If X and Y are two random variables associated with the same experiment, we define their joint CDF by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dt\,ds$$

The PDF can be recovered from the CDF by differentiating:

$$f_{X,Y}(x,y) = \frac{\partial F_{X,Y}}{\partial x \partial y}(x,y)$$

3.4.1. Expectation

It is still the case that $\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dx\,dy$, and that $\mathbb{E}[aX+bY+c] = a\mathbb{E}[X]+b\mathbb{E}[Y]+c$.

3.5. Conditioning

3.5.1. Conditioning a Random Variable on an Event

The conditional PDF of a continuous random variable X, given an event A with $P(A) > 0$, is defined as a nonnegative function $f_{X|A}$ that satisfies

$$P(X \in B|A) = \int_B f_{X|A}(x)dx$$

for any subset B of the real line. If we condition on an event of the form $\{X \in A\}$, with $P(X \in A) > 0$, we have

$$P(X \in B|X \in A) = \frac{P(X \in B, X \in A)}{P(X \in A)} = \frac{\int_{A\cap B} f_X(x)dx}{P(X \in A)}$$

which yields

$$f_{X|\{X\in A\}}(x) = \begin{cases} f_X(x)/P(X \in A) & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

If $A_1, A_2, ..., A_n$ are disjoint events that form a partition of the sample space, and $P(A_i) > 0$ for all i, then

$$f_X(x) = \sum_{i=1}^{n} P(A_i)f_{X|A_i}(x)$$

3.5.2. Conditioning One Random Variable on Another

Let X and Y be continuous random variables with joint PDF $f_{X,Y}$. For any y with $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Here, it is best to view y as a fixed number and consider $f_{X|Y}(x|y)$ as a function of the single variable x.

Note that

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)dy$$

and

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x|y)dx$$

3.5.3. Conditional Expectation

Let X and Y be jointly continuous random variables, and let A be an event with $P(A) > 0$. Then the conditional expectation of X given the event A is defined by

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} xf_{X|A}(x)dx$$

and the conditional expectation of X given that $Y = y$ is defined by

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx$$

The expected value rule and the total expectation theorem are both the same as their discrete counterparts, except that sums are replaced by integrals and PMFs are replaced by PDFs. For example,

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i]$$

if $A_1, ..., A_n$ are disjoint events that form a partition of the sample space, and similarly

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y = y] f_Y(y) dy$$

3.5.4. Independence

Two continuous random variables X and Y are independent if their joint PDF is the product of the marginal PDFs, i.e. $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y. This is the same as the condition $f_{X|Y}(x|y) = f_X(x)$ for all y with $f_Y(y) > 0$ and all x.

Independence implies that $F_{X,Y}(x, y) = F_X(x)F_Y(y)$. The converse is also true. Also, if X and Y are independent then we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$, and $var(X + Y) = var(X) + var(Y)$.

3.6. The Continuous Bayes' Rule

The continuous Bayes' rule is defined as

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(t)f_{Y|X}(y|t)dt}$$

If N is a discrete random variable, we have

$$f_Y(y)P(N = n \mid Y = y) = p_N(n)f_{Y|N}(y|n)$$

which leads to the formulas

$$P(N = n \mid Y = y) = \frac{p_N(n)f_{Y|N}(y|n)}{f_Y(y)} = \frac{p_N(n)f_{Y|N}(y|n)}{\sum_i p_N(i)f_{Y|N}(y|i)}$$

and

$$f_{Y|N}(y|n) = \frac{f_Y(y)P(N = n \mid Y = y)}{p_N(n)} = \frac{f_Y(y)P(N = n \mid Y = y)}{\int_{-\infty}^{\infty} f_Y(t)P(N = n \mid Y = t)dt}$$

with similar formulas existing for $P(A \mid Y = y)$ and $f_{Y|A}(y)$.

3.7. Summary and Discussion

A **continuous uniform** random variable over the interval $[a, b]$ has the PDF

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

Its mean is $\mathbb{E}[X] = (a + b)/2$, and its variance is $var(X) = (b - a)^2/12$.

An **exponential** random variable with the parameter λ has the PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and the CDF

$$f_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Its mean is $\mathbb{E}[X] = 1/\lambda$, and its variance is $var(X) = 1/\lambda^2$.

A **normal** random variable with parameters μ and σ^2 has the PDF $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$, the mean $\mathbb{E}[X] = \mu$, and the variance $var(X) = \sigma^2$.

2 Lecture

Entropy

If X is a discrete random variable, then $H(X) = \mathbb{E}[\log_2(\frac{1}{P(X)})]$ (in units of "bits/symbol") is the **Shannon entropy**. If we flip a coin with two heads and we get a heads, do we learn anything? No! We knew it would be heads. So in this case the entropy would be 0.

In general, entropy answers the question "how surprised are we to see what we're seeing?" The *surprise in seeing* $X = k$ is defined to be $\log(\frac{1}{P(X=k)})$. If we have a distribution where A occurs with probability 0.8, B occurs with probability 0.15, and C occurs with probability 0.05, then we will be less surprised if we observe A.

Example. Let X be a random variable which takes on the characters A, B, C, and D with probability 1/2, 1/4, 1/8, and 1/8 respectively.

In this case $H(X) = \sum_k P(X = k) \log_2 \frac{1}{P(X=k)} = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + (\frac{1}{8}\log_2 8) \cdot 2 = 1.75$ bits/symbol. This is the fundamental compression limit for sending information (on average)!

Say we were sending the string $A\ B\ A\ C\ D\ A\ A\ B$. If we naively used two bits to encode each symbol, it would cost us 16 bits in total. However, since we know that A occurs more frequently, we should make it less costly. If we follow the encoding scheme $A = 0$, $B = 10$, $C = 110$, and $D = 111$, then the encoded stream of $A\ B\ A\ C\ D\ A\ A\ B$ becomes 01001101110010, or 14 bits for an average of 1.75 bits/symbol. So there is a bit of improvement! It turns out that this is as good as it gets, on average.

Continuous Probability

In many (actually most) settings, a continuous sample space is more natural than a discrete one. Think of velocity, temperature, time, distance, intensity... most things in the real world are actually continuous values!

Continuous Random Variables (CRVs)

Continuous random variables are defined on an interval instead of at a certain point. $f_X(x)$ is no longer the probability, but the probability *density.* X is continuous if there exists a nonnegative function f_X such that $P(X \in B) = \int_B f_X(x)dx$ is well-defined for every interval B of $\mathbb{R}$. Also, $\int_{-\infty}^{\infty} f_X(x)dx$ must be 1.

We can be sloppy about inequality versus equality! $P(X < a)$ and $P(X \leq a)$ are the same thing.

Let's take the probability

$$P(X \in [x, x + \epsilon]) = \int_x^{x+\epsilon} f_X(t)dt \approx f_X(x) \cdot \epsilon$$

The **probability density function** (PDF) $f_X(x)$ is the ratio

$$\frac{P(X \in [x, x+\epsilon])}{\epsilon}$$

i.e. it is probability per unit length (hence the name density). Note that $f_X(x)$ is not a probability!

The **cumulative distribution function** (CDF) gives us $P(X \leq x)$. Suddenly we can have both discrete and continuous random variables living together. Formally, $F_X(x) = P(X \leq x)$. We have $F_X(\infty) = 1$, and $F_X(-\infty) = 0$. Also, if X is continuous we have $f_X(s) = \frac{d}{dx} F_X(x)$. If X is discrete, $P(X = k) = F_X(k) - F_X(k-1)$.

X and Y are called **independent** if $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events for all $x, y \in \mathbb{R}$. In particular, $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \implies F_{XY}(x, y) = F_X(x)F_Y(y)$. If X and Y are independent, so are deterministic functions of X and Y, i.e. $f(X)$ and $g(Y)$.

The **mean** of a random variable is $\mathbb{E}[X] = \int_{-infty}^{\infty} x f_X(x) dx$, as long as the integral converges absolutely. Note that $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$.

Tail-Sum Formula
If $X \geq 0$, $\mathbb{E}[X] = \sum_0^\infty P(X \geq x)dx = \sum_0^\infty P(X > x)dx = \sum_0^\infty [1 - F_X(x)]dx$.

Popular CRVs

- The uniform random variable $X \sim U[a, b)$. Its PDF is $f_X(x) = 1/(b-a)$. Its mean is $\mathbb{E}[X] = \int_a^b \frac{x}{b-a} dx = (a+b)/2$. Its variance is $var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (b-a)^2/12$.
- The exponential random variable is defined by its PDF of $f_X(x) = \lambda e^{-\lambda} x$ for $x > 0$. It has an exponential dropoff, i.e. it goes exponentially to 0 at the rate of λ. Its CDF is $F_X(x) = \int_{-\infty}^{\infty} f_X(x)dx = 1 - e^{-\lambda x}$ for $x \geq 0$. The most pleasing representation of the exponential is $P(X > x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$ for $x \geq 0$.

  The exponential's mean is $\mathbb{E}[X] = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$ (using the tail-sum-formula). The exponential's variance is $var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

  The exponential is a continuous version of the geometric. They both share amnesia; they don't remember (i.e. they both have the memoryless property). If $X \sim \text{Exp}(\lambda)$, then $P(X > t+s | X > t) = P(X > s)$ (where $t, s > 0$). X can be thought of as an arrival time, notably measured in continuous time instead of a discrete number of coin flips. The exponential can answer the question "what is the probability that [something of interest] will arrive in the next X seconds?"

  Proof of the memoryless property. $P(X > t+s | X > t) = P(X > t+s, X > t)/P(X > t) = P(X > t+s)/P(X > t) = e^{-\lambda(t+s)}/e^{-\lambda t} = e^{-\lambda s} = P(X > s)$.

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific books. Athena Scientific, 2002.