Random Variables & Discrete Probability Distributions

CS 3630

Random Variables

A ***random variable*** is a mapping from outcomes to real numbers, $X: \Omega \to \mathbb{R}$.

For example, we can map our categories to integers:

- Cardboard $\longrightarrow$ 0
- Paper $\longrightarrow$ 1
- Can $\longrightarrow$ 2
- Scrap Metal $\longrightarrow$ 3
- Bottle $\longrightarrow$ 4

- We typically use upper case letters, e.g., X, to denote a random variable, and lower-case letters, e.g., x_i, to denote the values taken by X.
- In our example, $X \in \{0,1,2,3,4\}$ indicates that X is a random variable that can take values from the set $\{0,1,2,3,4\}$.

Probability Mass Functions (pmf's)

- When a random variable takes its values from a finite (or possibly countably infinite) set, it is called a ***discrete random variable***.
- The probability distribution for a discrete random variable is typically defined as a ***probability mass function (pmf)***.
- For random variable X, the pmf is defined as

$$p_X(x) \triangleq P(X = x)$$

For our example,

- $p_X(0) = 0.20 \quad \mathit{cardboard}$
- $p_X(1) = 0.30 \quad \mathit{paper}$
- $p_X(2) = 0.25 \quad \mathit{can}$
- $p_X(3) = 0.20 \quad \mathit{scrap\ metal}$
- $p_X(4) = 0.05 \quad \mathit{bottle}$

As we will soon see, random variables can be very useful for outcomes that are naturally associated to real numbers, e.g., roll of a die, weight of a person, or, in our case, cost of applying an action.

Using pmf's

Even for this example, where categories don't naturally have numerical semantics, we can use the pmf to answer interesting questions.

For example, what is the probability that an object is a paper product?

- Paper products correspond to paper and cardboard, $X \in \{0,1\}$:

$$P(X \in \{0,1\}) = p_X(0) + p_X(1) = 0.5$$

Alternatively, we could write:

$$P(X \in \{0,1\}) = P(X \leq 1)$$

This form, $P(X \leq \alpha)$ turns out to be very useful.

Cumulative Distribution Function

The **Cumulative Distribution Function (CDF)** is defined as

$$F_X(\alpha) = P(X \leq \alpha) = \sum_{x_i \leq \alpha} p_X(x_i)$$

If we order the x_i's, such that $x_0 < x_2 \ldots < x_n$ we can write this as:

$$F_X(\alpha) = P(X \leq \alpha) = \sum_{i=0}^{k-1} p_X(x_i)$$

when we choose k such that $x_{k-1} \leq \alpha < x_k$.

CDF for our trash categories

It is straightforward to compute the CDF for the random variable (r.v.) associated to various trash categories:

$$F_X(\alpha) = P(X \leq \alpha) = \sum_{i=0}^{k-1} p_X(x_i)$$

r.v. $\boldsymbol{x}$	$\boldsymbol{p_X(x)}$
0	0.20
1	0.30
2	0.25
3	0.20
4	0.05

Category ($\boldsymbol{\omega}$)	**r.v. $\boldsymbol{x}$**	$\boldsymbol{F_X(\alpha)}$
Cardboard	0	
Paper	1	
Cans	2	
Scrap Metal	3	
Bottle	4	

Work out these five values

CDF for our trash categories

It is straightforward to compute the CDF for the random variable (r.v.) associated to various trash categories:

$$F_X(\alpha) = P(X \leq \alpha) = \sum_{i=0}^{k-1} p_X(x_i)$$

r.v. x	$\boldsymbol{p_X(x)}$
0	0.20
1	0.30
2	0.25
3	0.20
4	0.05

Category (ω)	**r.v. x**	$\boldsymbol{F_X(\alpha)}$
Cardboard	0	$P(X \leq 0) = 0.20$, $\boldsymbol{\alpha} = 0$
Paper	1	$P(X \leq 1) = 0.50$, $\boldsymbol{\alpha} = 1$
Cans	2	$P(X \leq 2) = 0.75$, $\boldsymbol{\alpha} = 2$
Scrap Metal	3	$P(X \leq 3) = 0.90$, $\boldsymbol{\alpha} = 3$
Bottle	4	$P(X \leq 4) = 1.00$, $\boldsymbol{\alpha} = 4$

Simulation by sampling

- It is often useful to simulate robotic systems. In our case, we might like to simulate the arrival of trash to our sorting system, such that it accurately reflects the prior distribution?
- How can we generate a sequence of samples, say $\omega_1, \omega_2, \ldots, \omega_n$, such that $\omega_i = cardboard$ for approximately 20% of the samples, $\omega_i = paper$ for approximately 30% of the samples, etc.?
- Sadly, most programming languages do not include library functions to sample from arbitrary probability distributions.
- Happily, the is almost always a random number generator that generates a random sample from the unit interval, $x \sim U(0,1)$.
- The notation $x \sim U(0,1)$ indicates that x is a number chosen at random from the interval $[0,1]$, and that all possible outcomes are equally likely.

➢Let's see how to use this…

Simulation by sampling

Suppose we generate the samples $s_1 = 0.97$ and $s_2 = 0.29$

Category (ω)	r.v. x	$p_X(x)$	$F_X(\alpha), \alpha = 0, 1, 2, 3, 4$
Cardboard	0	0.20	0.20
Paper	1	0.30	0.50
Cans	2	0.25	0.75
Scrap Metal	3	0.20	0.95
Bottle	4	0.05	1.00

- Note that $\boldsymbol{F_X(x_3)} = 0.95 < (s_1 = 0.97) \leq 1 = \boldsymbol{F_X(x_4)}$.
- The probability that this occurs is exactly 0.05, since the probability of $x \in [a, b] = (b - a)$ for the uniform distribution on $[0,1]$.
- $P(bottle) = 0.05$ …. Return category bottle.

- Similarly, $\boldsymbol{F_X(x_0)} = 0.20 < (s_2 = 0.29) \leq 0.50 = \boldsymbol{F_X(x_1)}$
- The probability that this occurs is exactly 0.30.
- $P(paper) = 0.30$ … Return category paper.

We can generalize this to develop an algorithm that draws a sample from an arbitrary distribution.

1. Generate a sample $x \sim U(0,1)$.
2. Determine k such that $F_X(x_{k-1}) < x \leq F_X(x_k)$.
3. Select category ω_k

References

Definitions, descriptions and sample code available in the

Introduction to Robotics and Perception book at http://roboticsbook.org

- Chapter 1: Introduction

- Chapter 2: Trash Sorting Robot