Foundations of Machine Learning: Week 1: Random Variables and Discrete distributions

Professor Christopher Yau
christopher.yau@manchester.ac.uk

October 21, 2020

Foundations of Machine Learning:Week 1: Random Variables and Discrete distributions

Foundations of Machine Learning:
Week 1: Random Variables and Discrete distributions

Professor Christopher Yau
christopher.yau@manchester.ac.uk
October 21, 2020

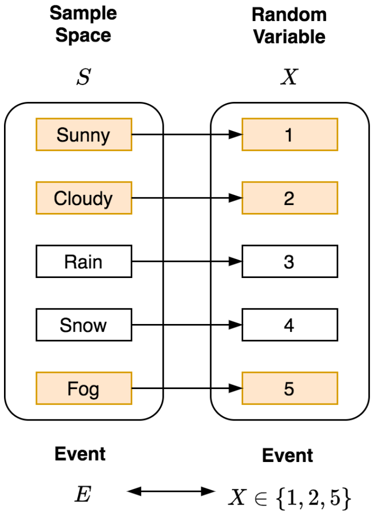
Random Variables

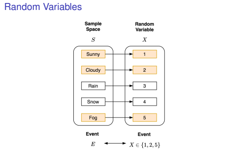
A **random variable** (r.v.) is a numeric quantity whose values map to the possible outcomes of an experiment.

Random Variables
A **random variable** (r.v.) is a numeric quantity whose values map to the possible outcomes of an experiment.

In this lecture we will be talking about random variables and discrete probability distributions.
So far, we have referred to the **sample space** and **events** directly but it would be more convenient to deal with numeric quantities.
A **random variable** (r.v.) is a numeric quantity whose values map to the possible outcomes of an experiment.

Random Variables





Consider the following example, suppose I have an experiment in which I am interested in the weather. The sample space consists of all possible weather types but instead of dealing directly with the textual description of each type, we map each possibility to a number.

The random variable X then takes on 5 possible values.

An event is therefore any subset of values that X can take.

For example, the event E that it is sunny, cloudy or foggy is the same as the random variable X taking the values 1, 2 or 5.

Random Variables

It is notational convention to denote a random variable by an uppercase letter X and the possible values it can take on by a lowercase letter x.

Hence, $P(X = x)$ means *"the probability that the random variable X takes on the value x"*.

Sometimes, we might just say $P(x)$ to mean the same thing.

Random Variables

It is notational convention to denote a random variable by an uppercase letter X and the possible values it can take on by a lowercase letter x.

Hence, $P(X = x)$ means *"the probability that the random variable X takes on the value x"*.

Sometimes, we might just say $P(x)$ to mean the same thing.

In terms of mathematical notation, it is conventional (but not the rule) to use an upper case letter to denote the random variables but the values it can take (if not explicitly given) on by its lowercase letter.
So we can write P of capital X equals little x to mean the probability that the random variable capital X takes on the value little x.
Sometimes we can use a shorthand, P of little x to mean the same thing.

Numerical Example

A family has two children.

What is the probability that both are boys given that at least one is a boy?

Numerical Example
A family has two children.
What is the probability that both are boys given that at least one is a boy?

Let's re-examine a previous problem using this idea of random variables.
A family has two children.
What is the probability that both are boys given that at least one is a boy?

Solution

Let X denote the number of boys:

$GG \Rightarrow X = 0$, $GB, BG \Rightarrow X = 1$, $BB \Rightarrow X = 2$

We want:

$$P(X = 2 | X \geq 1) = \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{1/4}{3/4} = 1/3$$

Using complement law:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 1/4 = 3/4$$

This is the same answer as before but using random variables we have simplified the notation.

Solution
Let X denote the number of boys:
$GG \Rightarrow X = 0$, $GB, BG \Rightarrow X = 1$, $BB \Rightarrow X = 2$
We want:
$$P(X = 2 | X \geq 1) = \frac{P(X = 2 \cap X \geq 1)}{P(X \geq 1)} = \frac{1/4}{3/4} = 1/3$$
Using complement law:
$$P(X \geq 1) = 1 - P(X = 0) = 1 - 1/4 = 3/4$$
This is the same answer as before but using random variables we have simplified the notation.

Lets use a random variable X to denote the number of boys.
X can takes values 0, 1 or 2.
Note that in this set up, boy-girl, and girl-boy map to the same number $X = 1$.
We want to compute the probability that X equals 2 given X is greater than 1. So from the chain rule we know we need the probability of X equals 2 and X greater than which is simply the probability that X equals or 1/4 and the probability that X greater than 1 which is 3/4.
This gives us the same answer as before but with much abbreviated notation.

Probability Mass Function

Discrete random variables take on a finite or *countable*[1] number of values.

The set of probabilities associated with each possible value of the random variable is known as a **probability mass function** (pmf).

Definition

If a discrete r.v. X can take on n possible values $x_1, x_2, \ldots, x_n$ then the pmf is given by:

$$f(x_i) = p(X = x_i),\ i = 1, \ldots, n.$$

where pmf must sum to one, i.e. $\sum_{i=1}^{n} f(x_i) = 1$, and all $0 \leq f(x_i) \leq 1$ for all i (note - n can be infinity).

[1] Countable means every possible value of the random variable can be associated with a unique element from the natural numbers.

Probability Mass Function

Discrete random variables take on a finite or *countable*[1] number of values.

The set of probabilities associated with each possible value of the random variable is known as a **probability mass function** (pmf).

Definition

If a discrete r.v. X can take on n possible values $x_1, x_2, \ldots, x_n$ then the pmf is given by:

$f(x_i) = p(X = x_i),\ i = 1, \ldots, n.$

where pmf must sum to one, i.e. $\sum_{i=1}^{n} f(x_i) = 1$, and all $0 \leq f(x_i) \leq 1$ for all i (note - n can be infinity).

[1] Countable means every possible value of the random variable can be associated with a unique element from the natural numbers.

Discrete random variables take on a finite or countable number of values. Each value is associated with a probability of that value occurring and we refer to the collection of those probabilities as the probability mass function. The probability mass function must fulfil our usual axioms. The probabilities must be between 0 and 1 inclusive and the sum of the probabilities must equal one.

Bernoulli distribution

Definition

A binary-valued r.v. X comes from a **Bernoulli distribution** if its pmf is given by:

$$p(X=0) = 1-p,$$
$$p(X=1) = p,$$

where p is a parameter referred to as the success probability.

Bernoulli distribution

Definition

A binary-valued r.v. X comes from a **Bernoulli distribution** if its pmf is given by:

$$p(X=0) = 1-p,$$
$$p(X=1) = p,$$

where p is a parameter referred to as the success probability.

There are specific forms of discrete probability distributions that we will be interested in.
The simplest is the Bernoulli distribution which says that if a random variable X can take on one of two values, i.e. it is binary, then the probability of one outcome is given by a parameter p and the probability of the other outcome is $1-p$.
The parameter p is sometimes called a success probability because the binary outcomes are often used in applications where the outcomes represent some form of success or failure.

Binomial distribution

Definition

A binary-valued r.v. X comes from a **Binomial distribution** if its pmf is given by:

$$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where p is a parameter referred to as the success probability and k is the number of successes in n binary outcome experiments.

Note - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the *binomial coefficient* which determines the number of ways of having k successes in n experiments, e.g. getting two heads in three coin flips can be obtained by (HHT, HTH, THH) or $\binom{3}{2} = 3$.[2]

[2]You can use the function which reads something like ${}_nC_k$ on a handheld scientific calculator to compute this.

Binomial distribution
Definition
A binary-valued r.v. X comes from a **Binomial distribution** if its pmf is given by:
$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
where p is a parameter referred to as the success probability and k is the number of successes in n binary outcome experiments.
Note - $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the *binomial coefficient* which determines the number of ways of having k successes in n experiments, e.g. getting two heads in three coin flips can be obtained by (HHT, HTH, THH) or $\binom{3}{2} = 3$.[2]
[2]You can use the function which reads something like ${}_nC_k$ on a handheld scientific calculator to compute this.

If we perform a number of Bernoulli experiments and look at the number of successes and failures across those experiments then the distribution of those successes obeys something known as the Binomial distribution.
Formally, it says that if the success probability of each Bernoulli experiment is p and that experiments is repeated n times, the probability that we will see k successes is given by this probability mass function.
This probability mass function says that the probability of k successes, if the experiments are identical and independent, is p^k whilst the remaining $n - k$ experiments must be failures with probability $(1 - p)^{n-k}$.
However, there are a number of different arrangements of the successes and failures and to compute those we use the binomial coefficient which tells us the number of ways of having k successes in n experiments.
The binomial coefficient can be computed using a special function on calculators or in programming languages or by hand for small numbers.

Binomial distribution

Example

A machine learning prediction algorithm has a probability $p = 0.9$ of predicting the outcome of a sports match correctly.

What is the probability that it will get three predictions correct in 5 independent matches?

Solution

Let X denote the number of correct predictions then:

$$P(X = 3) = \binom{5}{3} 0.9^3 (1 - 0.9)^{5-3} = 0.0729$$

Binomial distribution

Example

A machine learning prediction algorithm has a probability $p = 0.9$ of predicting the outcome of a sports match correctly.

What is the probability that it will get three predictions correct in 5 independent matches?

Solution

Let X denote the number of correct predictions then:

$$P(X = 3) = \binom{5}{3} 0.9^3 (1 - 0.9)^{5-3} = 0.0729$$

Lets look at an example.
A machine learning prediction algorithm has a probability $p = 0.9$ of predicting the outcome of a sports match correctly.
What is the probability that it will get three predictions correct in 5 independent matches?
Let X denote the number of correct predictions, there are $n = 5$ experiments so using the binomial distribution we can compute the probability of three correct predictions in five matches. First, we use the binomial coefficient to compute the number of ways of having 3 successes in 5 predictions then we multiply by the probability of three successes 0.9^3 and then the probability of two failures $(1 - 0.9)^2$.

Geometric distribution

Definition

If a r.v. X comes from a **Geometric distribution**, its pmf is given by:

$$p(X = x) = (1 - p)^{x-1} p,$$

where p is a success probability parameter and X is a positive-valued integer.

It is useful for modelling the first occurrence of an outcome after repeated identical trials.

Geometric distribution
Definition
If a r.v. X comes from a **Geometric distribution**, its pmf is given by:
$p(X = x) = (1 - p)^{x-1} p,$
where p is a success probability parameter and X is a positive-valued integer.
It is useful for modelling the first occurrence of an outcome after repeated identical trials.

A related discrete probability distribution to the binomial distribution is the geometric distribution where now, instead of counting the number of successes in a fixed number of experiments, we count how many experiments we must do until we meet the first success.
So the probability mass function is given by the probability of $x - 1$ failures which is $(1 - p)^{x-1}$ times the probability of the first success p where x is the number of experiments until the first success.

Geometric distribution

Example

A person attempts a driving test repeatedly until they are able to pass the exam.

If their probability of success in each test is constant at $p = 0.8$, what is the probability:

1. that they will pass on the third attempt?

$$p(X = 3) = (1 - 0.8)^2 0.8 = 0.032$$

2. take three or more attempts?

$$\begin{aligned} p(X \geq 3) &= 1 - p(X = 1) - p(X = 2), \\ &= 1 - 0.8 - (1 - 0.8)0.8 = 0.04 \end{aligned}$$

Geometric distribution

Example

A person attempts a driving test repeatedly until they are able to pass the exam.

If their probability of success in each test is constant at $p = 0.8$, what is the probability:

1. that they will pass on the third attempt?

$$p(X = 3) = (1 - 0.8)^2 0.8 = 0.032$$

2. take three or more attempts?

$$\begin{aligned} p(X \geq 3) &= 1 - p(X = 1) - p(X = 2), \\ &= 1 - 0.8 - (1 - 0.8)0.8 = 0.04 \end{aligned}$$

Lets look at an example.
A person attempts a driving test repeatedly until they are able to pass the exam.
If their probability of success in each test is constant at $p = 0.8$, what is the probability that they will pass on the third attempt?
We can model this using the Geometric distribution. For a pass on the third attempt, there must be two failures with probability $(1-0.8)^2$ and then a success with probability 0.8.
Now if want to find the probability that a pass will occur after three or more attempts, we can use the complement law and compute one minus the probability that you pass after one and two attempts.

Poisson distribution

Definition

If a r.v. X comes from a **Poisson distribution** then its pmf is given by:

$$p(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!},$$

where λ is a rate parameter and X is a positive-valued integer.

It expresses the probability of a given number of events x occurring in a fixed interval (of, say, time or space) if these events occur with a known constant mean rate λ and independently of the time since the last event.

Poisson distribution

Definition
If a r.v. X comes from a **Poisson distribution** then its pmf is given by:
$$p(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!},$$
where λ is a rate parameter and X is a positive-valued integer.

It expresses the probability of a given number of events x occurring in a fixed interval (of, say, time or space) if these events occur with a known constant mean rate λ and independently of the time since the last event.

The final discrete probability distribution I will introduce is the Poisson distribution.
This is used when we want to model the number of events X occurring in a fixed interval often of time or space.
The rate of these events is assumed to be constant and occur with a rate which here we denote by λ.

Poisson distribution

Example

If patients arrive in a hospital emergency department at an average rate of 0.5 per hour.

What is the probability that there will be 3 patients in a 6 hour shift?

Let X denote the number of patients arriving during a 6 hour shift and $\lambda = 6 \times 0.5 = 3$ patients/shift be the rate of patients arriving per shift.

$$p(X = 3) = \frac{\lambda^3 \exp(-\lambda)}{3!} = \frac{3^3 \exp(-3)}{3 \times 2 \times 1} = 0.224,$$

Poisson distribution

Example

If patients arrive in a hospital emergency department at an average rate of 0.5 per hour.

What is the probability that there will be 3 patients in a 6 hour shift?

Let X denote the number of patients arriving during a 6 hour shift and $\lambda = 6 \times 0.5 = 3$ patients/shift be the rate of patients arriving per shift.

$$p(X = 3) = \frac{\lambda^3 \exp(-\lambda)}{3!} = \frac{3^3 \exp(-3)}{3 \times 2 \times 1} = 0.224,$$

Lets illustrate by examining an example.
If patients arrive in a hospital emergency department at an average rate of 0.5 per hour.
What is the probability that there will be 3 patients in a 6 hour shift?
If we let X denote the number of patients arriving during a 6 hour shift then the rate at which patients arrive per shift is $6 \times 0.5 = 3$.
Thus, the probability that three patients arrive in a six hour shift is given by the Poisson probability mass function with rate 3 and $X = 3$.

Expectations

Expectations allow us to calculate summaries of a random variable.

Definition

If a discrete r.v. X has a pmf $f(x)$ then the expected value $\mathbb{E}[g(X)]$ of any function g of X is given by:

$$\mathbb{E}[g(X)] = \sum_{i=1}^{\infty} g(x_i) f(x_i)$$

Example:

x	1	2	3
$g(x) = 2.5X$	2.5	5	7.5
$f(x)$	0.2	0.3	0.5
$f(x)g(x)$	0.5	1.5	3.75

So $\mathbb{E}[g(x)] = 0.5 + 1.5 + 3.75 = 5.75$

Expectations
Expectations allow us to calculate summaries of a random variable.
Definition
If a discrete r.v. X has a pmf $f(x)$ then the expected value $\mathbb{E}[g(X)]$ of any function g of X is given by:
$$\mathbb{E}[g(X)] = \sum_{i=1}^{\infty} g(x_i) f(x_i)$$
Example:
So $\mathbb{E}[g(x)] = 0.5 + 1.5 + 3.75 = 5.75$

Given a probability model, we often want to compute summaries of our random quantities or expected values.

The expected value or expectation of a function g of a random variable X is given by the weighted sum over all possible values of the function of the random variable where the weight is given by the probability of that value occurring.

Lets consider a quick example where a random variable X has a probability mass function where it has 0.2, 0.3 and 0.5 probability of taking the values 1, 2 and 3.

We want to compute the expected values of a function of X which in this case is simply multiplying X by 2.5. We compute the function values multiply by the respective probabilities and sum to give the expected value of the function $g(x)$.

Expectations

Example:

A route finding algorithm randomly chooses one of three routes A, B, C to send a delivery driver to reach their destination.

The probabilities of selecting each route are given by $0.3, 0.5$ and 0.2 but each route takes $20, 30$ and 15 minutes.

What is the expected journey time for the driver?

Solution:

Let X denote the journey time then the pmf is given by:

$$f(\underbrace{x=20}_{A}) = 0.3, f(\underbrace{x=30}_{B}) = 0.5, f(\underbrace{x=15}_{C}) = 0.2$$

then the expected journey time is given by:

$$\mathbb{E}[X] = 20(0.3) + 30(0.5) + 15(0.2) = 24$$

Expectations
Example:
A route finding algorithm randomly chooses one of three routes A, B, C to send a delivery driver to reach their destination.
The probabilities of selecting each route are given by $0.3, 0.5$ and 0.2 but each route takes $20, 30$ and 15 minutes.
What is the expected journey time for the driver?
Solution:
Let X denote the journey time then the pmf is given by:
$f(\underbrace{x=20}_{A}) = 0.3, f(\underbrace{x=30}_{B}) = 0.5, f(\underbrace{x=15}_{C}) = 0.2$
then the expected journey time is given by:
$\mathbb{E}[X] = 20(0.3) + 30(0.5) + 15(0.2) = 24$

Lets consider an example.
A route finding algorithm randomly chooses one of three routes A, B, C to send a delivery driver to reach their destination.
The probabilities of selecting each route are given by $0.3, 0.5$ and 0.2 but each route takes $20, 30$ and 15 minutes.
What is the expected journey time for the driver?
If we let X denote the journey time then the probability mass function is given by 0.3, 0.5 and 0.2 of the journey taking 20, 30 and 15 minutes respectively.
We want the expected journey time so in this case there is no function of X so we weight each possibility by its probability and sum to get 24 minutes.

Expectations

The variance of a function of a discrete r.v. tells us about the *spread* of the possible values around its expected value.

Definition

If a discrete r.v. X has a pmf $f(x)$ then the variance $\mathbb{V}[g(X)]$ of any function g of X is given by:

$$\begin{aligned}\mathbb{V}[g(X)] &= \mathbb{E}[(g(X) - \mathbb{E}(g(X)))^2], \\ &= \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2\end{aligned}$$

Expectations

The variance of a function of a discrete r.v. tells us about the *spread* of the possible values around its expected value.

Definition

If a discrete r.v. X has a pmf $f(x)$ then the variance $\mathbb{V}[g(X)]$ of any function g of X is given by:

$\mathbb{V}[g(X)] = \mathbb{E}[(g(X) - \mathbb{E}(g(X)))^2], = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2$

One particularly useful expected value is the expectation of the squared distance of the function values from its expected or average value.
This is known as the variance and tells us about the spread of the .
It can be computed in two ways, either by directly calculating the squared differences from the expected value or by computing the expected value of the square of the function and then subtracting the square of the expected value.
Both are mathematically equivalent and can be derived from one another.

Expectations

Example:

Continuing our previous example, what is the variance of the journey time for the driver?

Solution:

$$\mathbb{E}[X^2] = 20^2(0.3) + 30^2(0.5) + 15^2(0.2) = 615$$

so

$$\mathbb{V}[X] = 615 - 24^2 = 39$$

Expectations
Example:
Continuing our previous example, what is the variance of the journey time for the driver?
Solution:
$$\mathbb{E}[X^2] = 20^2(0.3) + 30^2(0.5) + 15^2(0.2) = 615$$
so
$$\mathbb{V}[X] = 615 - 24^2 = 39$$

Lets continue our previous example.
What is the variance of the journey time for the driver?
To compute this, we need to compute the expected value of the square of X which is simply the possible journey times squared and weighted by their respective probabilities and summed.
Since we have already computed the expected value previously, we can then use the expected value of the X-squared and subtract the square of the expected value of X to get 39 for the variance.

Properties of Expectations

If $g(X)$ is a linear transformation of $aX + b$ then:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

and

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$$

where a and b are some constants.

Properties of Expectations
If $g(X)$ is a linear transformation of $aX + b$ then:
$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$
and
$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$$
where a and b are some constants.

Expectations have certain properties which can be helpful when manipulating these.
If the function is a linear transformation involving some constant a times X plus another constant b then the expected value of the function is simply a times the expected value of X plus b.
With the variance, the variance of the function is given by a^2 times the variance of X.
This is because the variance is computed from the squared distance from the expected value which also results in b not entering into the expression.

Expectations

Example: A discrete r.v. has the following pmf:

x	1	2	3	4
f(x)	0.2	0.1	0.3	0.4

What is the expectation and variance of $2X - 1$?

Solution:

$$\mathbb{E}[X] = 1(0.2) + 2(0.1) + 3(0.3) + 4(0.4) = 2.9$$

$$\mathbb{E}[2X - 1] = 2\mathbb{E}[X] - 1 = 4.8$$

$$\mathbb{E}[X^2] = 1^2(0.2) + 2^2(0.1) + 3^2(0.3) + 4^2(0.4) = 9.7$$

$$\mathbb{V}[X] = 9.7 - 2.9^2 = 1.29$$

$$\mathbb{V}[2X - 1] = 2^2\mathbb{V}[X] = 2^2(1.29) = 5.16$$

Expectations

Example: A discrete r.v. has the following pmf:

x	1	2	3	4
f(x)	0.2	0.1	0.3	0.4

What is the expectation and variance of $2X - 1$?

Solution:

$$\mathbb{E}[X] = 1(0.2) + 2(0.1) + 3(0.3) + 4(0.4) = 2.9$$

$$\mathbb{E}[2X - 1] = 2\mathbb{E}[X] - 1 = 4.8$$

$$\mathbb{E}[X^2] = 1^2(0.2) + 2^2(0.1) + 3^2(0.3) + 4^2(0.4) = 9.7$$

$$\mathbb{V}[X] = 9.7 - 2.9^2 = 1.29$$

$$\mathbb{V}[2X - 1] = 2^2\mathbb{V}[X] = 2^2(1.29) = 5.16$$

Lets look at some examples of how these properties can be used to simplify our computations.
A discrete random variable can take the values 1 to 4 with the following probabilities.
We are asked what is the expectation and variance of $2X - 1$.
Now, we could compute $2X - 1$ are all the values and proceed normally but now we know that for a linear transformation we can do the following instead.
Compute the expectation of X untransformed then multiply by 2 and takeaway 1 to get the expectation of $2X - 1$.
Similarly, for the variance, we compute the variance of X first, then multiply by 2-squared to get the variance of $2X - 1$.
These properties can be useful when one is potentially computing expectations for multiple transformations of the random variables.

Recap

A **random variable** is a numeric quantity used to denote the possible outcomes of a random experiment.

A **discrete** random variables takes on countable number of values only and has an associated **probability mass function** which gives the probability distribution of these values.

There are a number of common discrete probability distributions that are applicable in different situations.

You can compute **expectations** to compute summary statistics for random variables.

Recap

A **random variable** is a numeric quantity used to denote the possible outcomes of a random experiment.

A **discrete** random variables takes on countable number of values only and has an associated **probability mass function** which gives the probability distribution of these values.

There are a number of common discrete probability distributions that are applicable in different situations.

You can compute **expectations** to compute summary statistics for random variables.

A **random variable** is a numeric quantity used to denote the possible outcomes of a random experiment. They allow us to enumerate the sample space.
A **discrete** random variables takes on countable number of values only and has an associated **probability mass function** which gives the probability distribution of these values. Once the probability mass function is given we essentially know "everything" about that random quantity.
There are a number of common discrete probability distributions that are applicable in different situations and you should learn these for assessment purposes.
You can compute **expectations** to compute summary statistics for random variables and functions of random variables.

END LECTURE

END LECTURE