

Probability Refresher

CS771: Introduction to Machine Learning

Purushottam Kar

What is Probability

2

Depends on whom we ask this question

A statistician may claim probability is a way of measuring how frequently does something happen

“If I recommend an iPhone to 1000 female customers aged 25-30 years, roughly 600 of them will make a purchase”

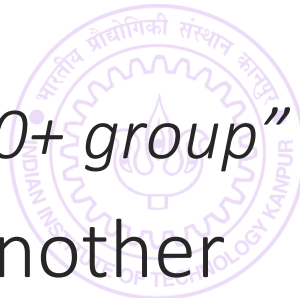
A logician may claim that probability is a way of measuring the amount of uncertainty in a certain statement

“If John makes a credit card transaction worth more than ₹10,000, then there is a 70% chance it is fraudulent since he never spends so much”

A measure theoretician may claim probability is a way of assigning positive scores in a way so that two scores can be easily compared

“This customer is more likely to be in the 20-30 age group than the 50+ group”

Machine Learning subscribes to all these views in one way or another



Sample Space

3

Denotes an exhaustive enumeration of all possible outcomes that either have happened or *could* happen (even if extremely unlikely)

A Toy RecSys Problem: our website has 10 products on sale. Users visit our website, browse and are shown one ad. Depending on their experience, they either purchase one of the 10 products or don't purchase anything. We record gender, age of customer and how many seconds they spend on the website.

Sample Space: $\{M, F, T\} \times \mathbb{N} \times \mathbb{N} \times \{A_0, \dots, A_9\} \times \{P_0, \dots, P_9, \emptyset\}$

Gender

Age

Time Spent

Ad Shown

Purchase

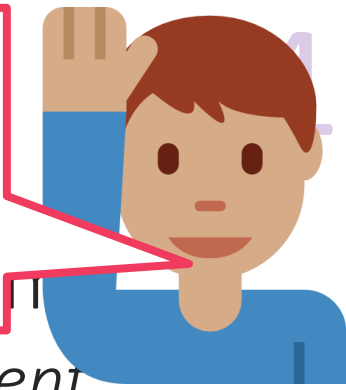
Sample spaces are often infinite in size in real settings since they enumerate all possibilities, even very unlikely ones



Even

An e

Notice that events may choose to precisely describe certain aspects and neglect others e.g. in the events given here, some events are very specific about the gender of the user, others are not. Some are concerned about whether a purchase was made, others are merely tracking time spent etc.



“A male user in age group 20-30 years visiting our website” *is an event*

“A female user being shown an ad for a P2 (a laptop)” *is an event*

“A user buying something that was shown as an ad” *is an event*

“A user buying something that was not shown as an ad” *is an event*

“A user spending more than 200 seconds on the website” *is an event*

ML can be used to do several useful things

Tell us how frequently does an event occur/if one event more likely than other

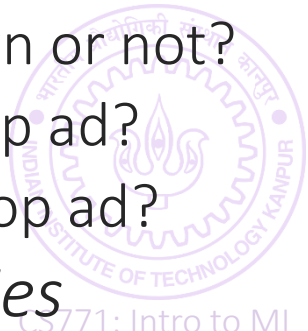
What fraction of male customers aged 35-40 purchase P6 (a phone) if shown an ad?

What fraction of female customers purchase P2 (a laptop) whether ad shown or not?

Is it more likely that a purchase will be made if I show a mobile ad or a laptop ad?

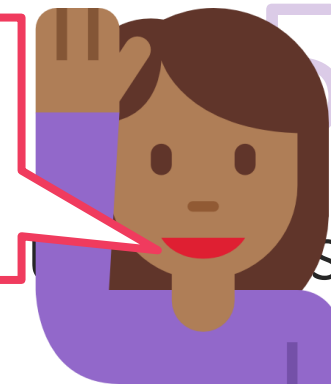
Is it more likely that a 20-25 year old will purchase if I show a mobile vs laptop ad?

Tell us how confident is the ML algorithm while giving the above replies



Ran

I could have also defined a random variable such that $S = 1$ if purchase made (whether or not on ad shown) and $S = 0$ otherwise. What I define as a random variable (or even an event) is totally up to my creativity



Rand

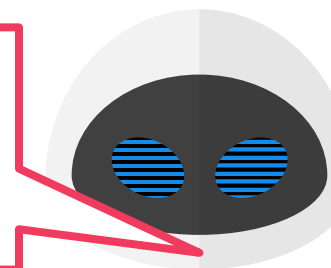
as numbers so that we can do math with them

Rand

Cat

Nu

We had earlier seen that features for data points in ML may be categorical, numerical etc. This is no accident. Random variables can indeed be seen as giving us “features” for the outcome of the experiment



Numerical (Continuous): $Z = \text{number of seconds spent on the website}$

Indicator: $W = 1$ if purchase made on ad shown, $W = 0$ otherwise

Example Outcome: A male customer aged 25 years spent 18 minutes on our website but did not purchase the product whose ad was shown

$X = 2, Y = 25, Z = 1080, W = 0$

Can arrange many random variables as vectors too e.g. $[2, 25, 1080, 0]$



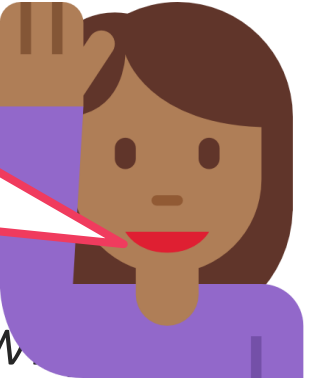
Probability Distribu

For starters, a 120 year old human being is almost certainly a woman not a man



For the purpose of ML, a probability distribution serves two purposes

Give In this case, we are interested in getting samples of female customers who are shown an ad for P6. For example, if such customers are more likely to buy P6 then we would like $W = 1$ more frequently for these samples too!



Note that random variables can be used to define events too e.g. $W = 1$ event (that a purchase was made on the product whose ad was shown)

Generate a sample outcome

We want outcomes that are more likely to be generated more often than those that are rare e.g. “120 year old man who is shown an ad for P8, spent 1000 seconds but did not purchasing anything” is not a very likely outcome

Sometimes, we may be interested in a sample outcome with some restrictions e.g. we can request for a random “female customer who is shown an ad for P6”. In this case, we only want outcomes that satisfy the above but would like those that are more likely, to be generated more often



Getting Started

7

Sample space: $\{R, G, B\} \times [6]$

$$\mathbb{P}[R] = \frac{14}{24} = \frac{7}{12}$$

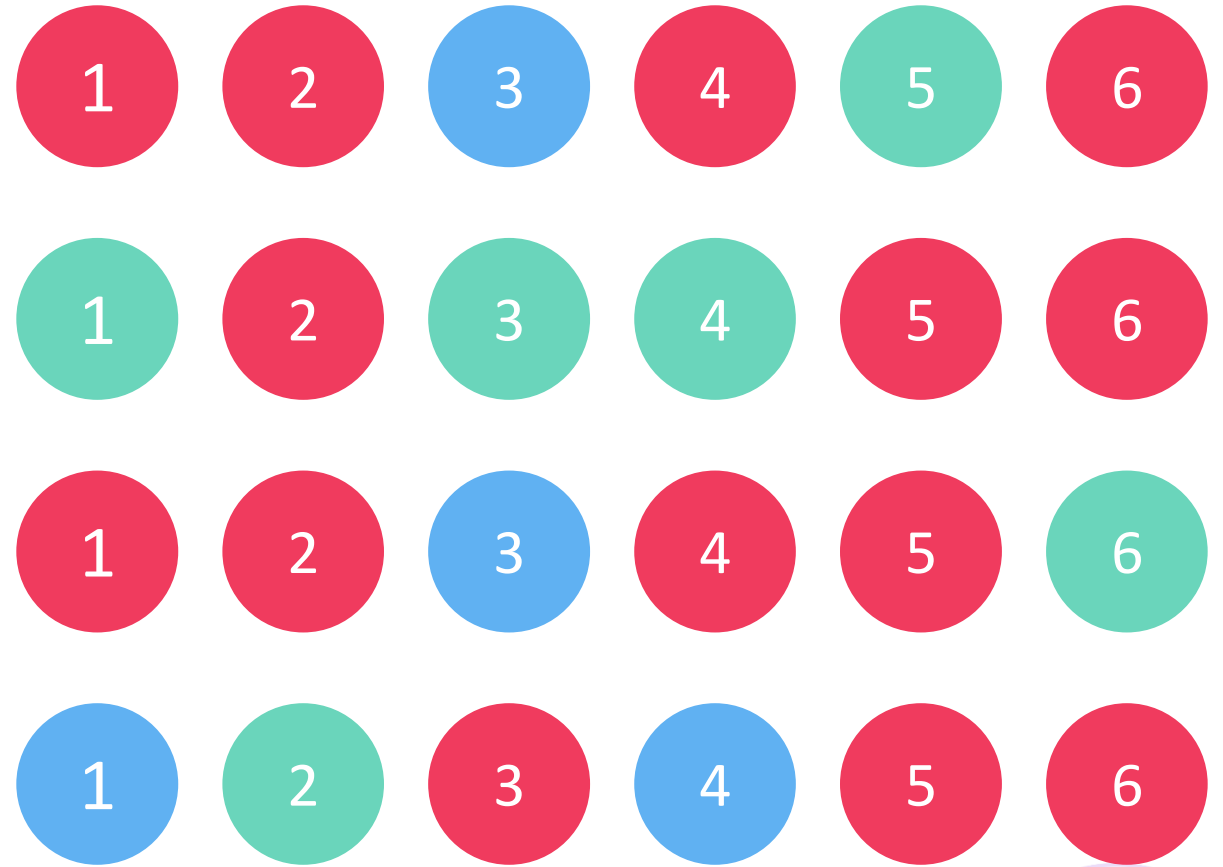
$$\mathbb{P}[B] = \frac{4}{24} = \frac{1}{6}$$

$$\mathbb{P}[G] = \frac{6}{24} = \frac{1}{4}$$

Note: $\mathbb{P} \geq 0$ always

$$\mathbb{P}[1] = \frac{1}{6} = \mathbb{P}[2] = \dots = \mathbb{P}[6]$$

$$\mathbb{P}[R \wedge 5] = \frac{3}{24} = \frac{1}{8}$$



Initially, to get used to things, it is good to think of probability in terms of *proportions* or *frequency*



Probability as Proportions

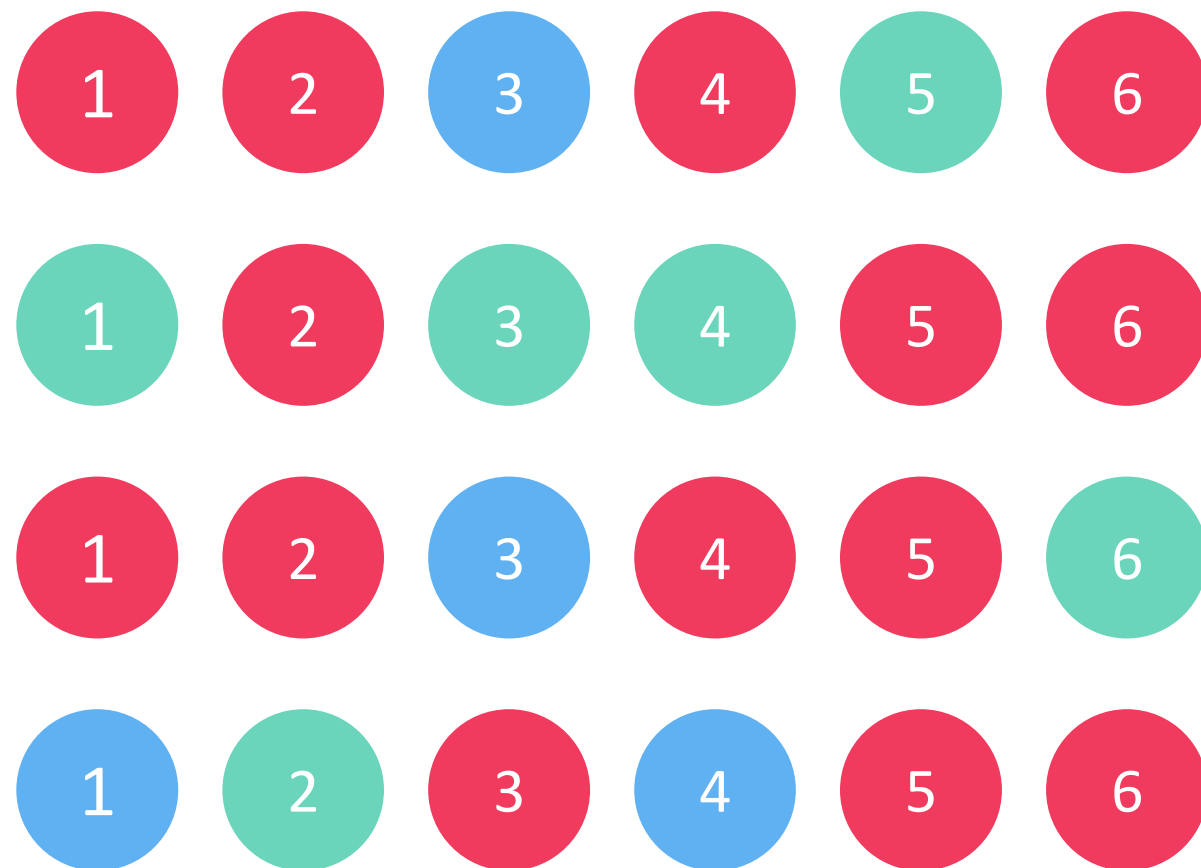
8

Sample/Outcome: pick one ball

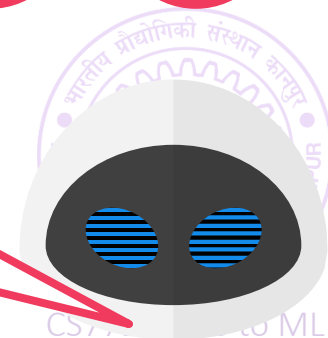
Sample space: $\{R, G, B\} \times [6]$

Assume that picking any ball is equally likely. In other words, the probability of picking any ball is $\frac{1}{24}$ since there are only 24 balls

Use this toy setting to get comfortable with concepts since in this case, probability of some event is simply the proportion of the outcomes when that happens



For now, we will only look at discrete random variables (categorical/numeric)



Probability as Proportions

9

Sample/Outcome: pick one ball

Sample space: $\{R, G, B\} \times [6]$

Define two random variables (r.v.)

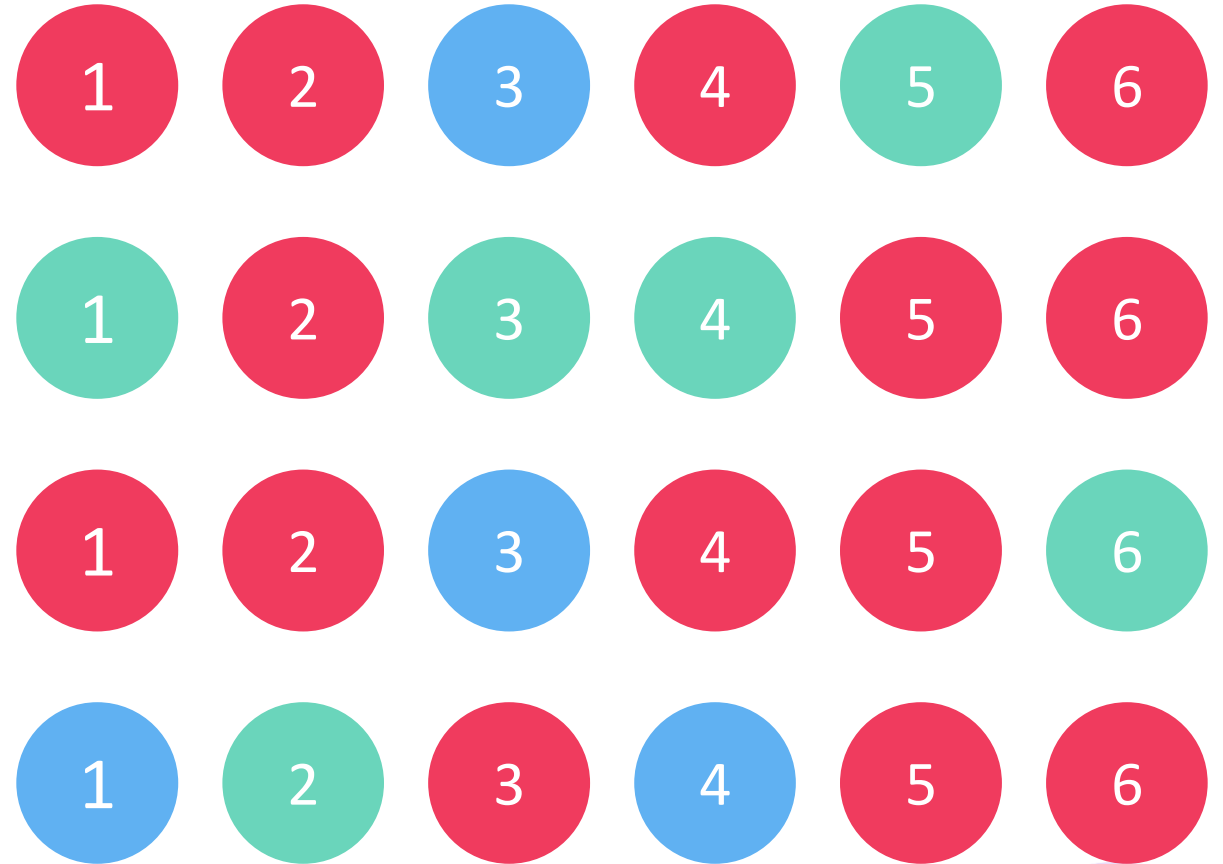
$X \triangleq$ number on the ball $\in [6]$

$Y \triangleq$ colour of the ball

$\{R = 1, G = 2, B = 3\}$

$\mathbb{P}[X = 1] \triangleq$ proportion of samples for which we have $X = 1$

$$\mathbb{P}[X = 1] = \frac{4}{24} = \frac{1}{6}$$



Probability as Proportions

10

Sample/Outcome: pick one ball

Sample space: $\{R, G, B\} \times [6]$

Define two random variables (r.v.)

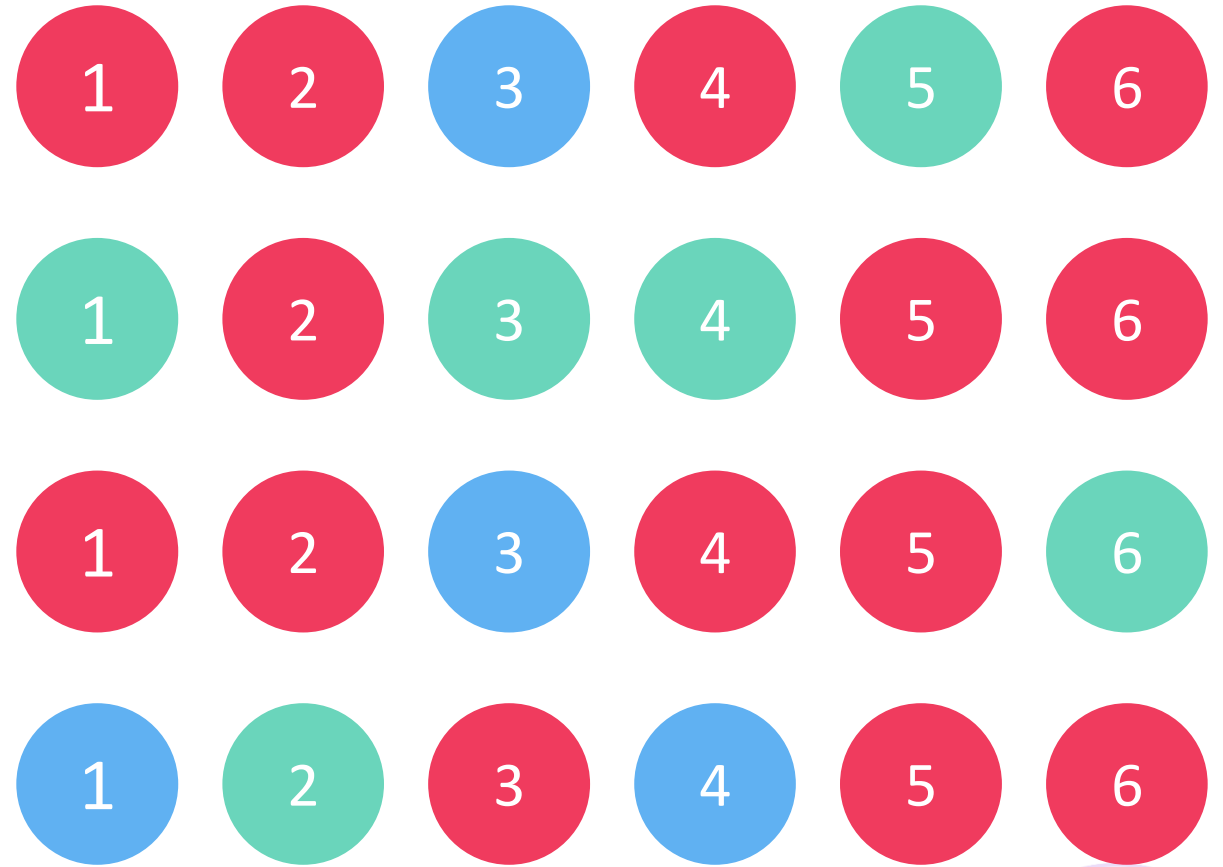
$X \triangleq$ number on the ball $\in [6]$

$Y \triangleq$ colour of the ball

$\{R = 1, G = 2, B = 3\}$

$\mathbb{P}[Y = 2] \triangleq$ proportion of
samples for which we have $Y = 2$

$$\mathbb{P}[Y = 2] = \frac{6}{24} = \frac{1}{4}$$



Probability beyond Proportions

11

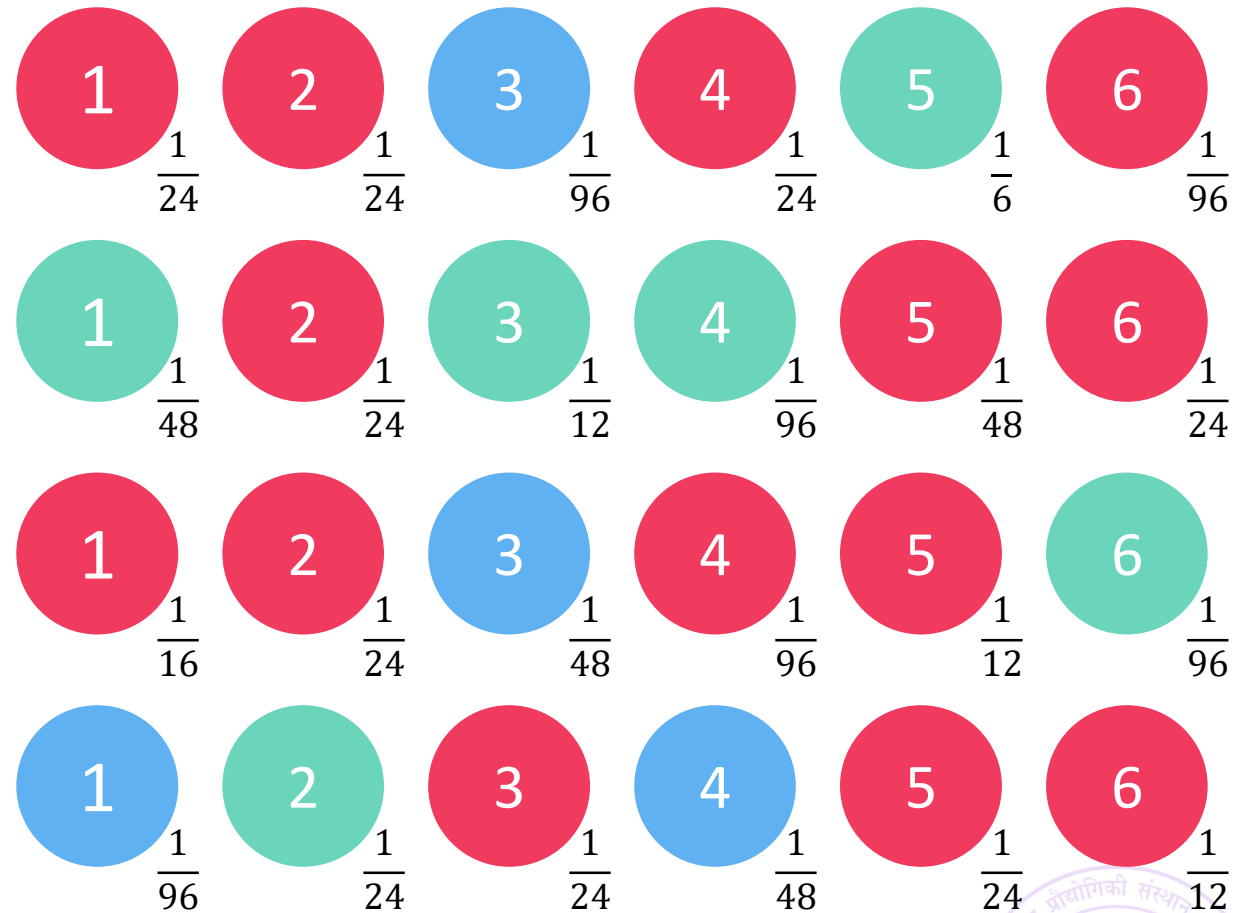
Suppose now that not all samples are equally likely, i.e. not all balls are equally likely to be picked

$\mathbb{P}[Y = 2] \triangleq$ sum of probabilities of samples for which $Y = 2$

$$\mathbb{P}[Y = 2] = \frac{1}{48} + \frac{1}{24} + \frac{1}{12} + \frac{1}{96} + \frac{1}{6} + \frac{1}{96} = \frac{1}{3}$$

$\mathbb{P}[X = 1] \triangleq$ sum of probabilities of samples for which $X = 1$

$$\mathbb{P}[X = 1] = \frac{1}{24} + \frac{1}{48} + \frac{1}{16} + \frac{1}{96} = \frac{13}{96}$$



Rules of Probability

12

Let Ω denote the sample space (set of all possible outcomes)

Let X be any (discrete) random variable and let S_X denote the set of (numerical) values X could possibly take (even unlikely values)

The set S_X is called the support of the random variable X

In previous example, $S_X = [6], S_Y = [3]$

For any outcome $\omega \in \Omega$, let $X(\omega)$ denote

For example, $X(3) = 3$ and $Y(3) = 2$ (Y

p_ω is the probability with which an outcome ω happens. E.g $p_3 = \frac{1}{12}$

For any value $x \in S_X$ (i.e. any valid value), we de

$$\mathbb{P}[X = x] = \sum_{\omega \in \Omega: X(\omega)=x} p_\omega$$

Sometimes we use lazy notation to denote $\mathbb{P}[x] \triangleq \mathbb{P}[X = x]$



Rules of Probability

13

No matter how we define our random variable, if it is discrete valued, then the following must hold

For all $x \in S_X$, we must have $\mathbb{P}[X = x] \geq 0$

If $\mathbb{P}[X = x] = 0$ then we say x is an impossible value for random variable X

If $\mathbb{P}[X = x] = 1$ then we say that X almost surely takes the value x

We must have $\sum_{x \in S_X} \mathbb{P}[X = x] = 1$

Another way of saying that when we get a sample, the random variable X must take some valid value on that sample, it cannot remain undefined!

It is a different thing that we (e.g. the ML algo) may not know what value X has taken on that sample, but there must be some hidden value it did take

An immediate consequence of the above two rules is that we must have, for $x \in S_X$, we must have $\mathbb{P}[X = x] \leq 1$



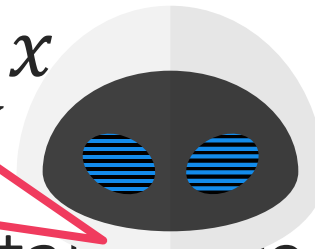
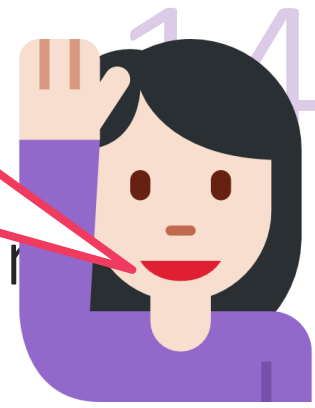
Probability

Note that this will always give us values $x \in S_X$ and that too in a way so that if $\mathbb{P}_X[x]$ is large, we will get that value x more likely than a value \tilde{x} for which $\mathbb{P}[\tilde{x}] \approx 0$

A fancy name for a function that tells us the probability of the

That is correct. E.g. in our toy setting (where not all samples are equally likely),

$\mathbb{P}[R] = \frac{29}{48}, \mathbb{P}[G] = \frac{16}{48}, \mathbb{P}[B] = \frac{3}{48}$ so if we sample $y \sim \mathbb{P}_Y$ (recall that Y encodes color) then we are almost twice likely to get $Y = 1$ than $Y = 2$. There is a comparatively much smaller chance that we would get $Y = 3$



any x
for X
— take care

Often the blackboard letter P i.e. \mathbb{P} used to denote PMF i.e. $\mathbb{P}[x] \triangleq \mathbb{P}[X = x]$

Sometimes may write $\mathbb{P}_X[\cdot]$ to emphasize that this PMF for X and not some Y

Sometimes, $\mathbb{P}[X]$ is also used to refer to the PMF of the random variable X

Sampling from a PMF: $x \sim \mathbb{P}[X]$ or $x \sim \mathbb{P}_X$ or even $X \sim \mathbb{P}[X]$ means that we generated an outcome $\omega \in \Omega$, e.g. 3 according to the probability distribution and are looking at $X(\omega)$



Joint Probability

15

$\mathbb{P}[X = 1 \wedge Y = 1] \triangleq$ proportion of samples for which we have both $X = 1$ and $Y = 2$

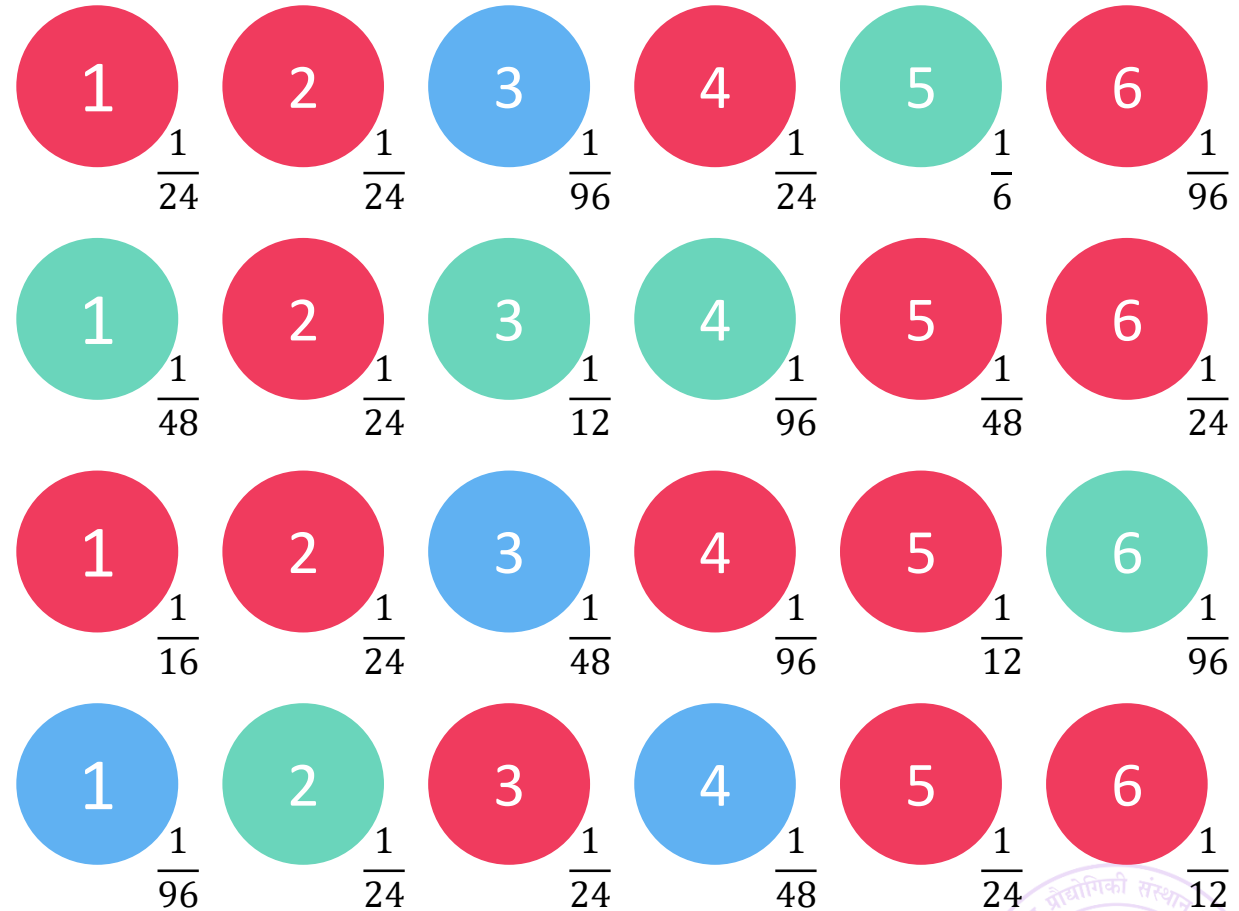
Let us look at uniform case first

$$\mathbb{P}[X = 1 \wedge Y = 1] = \frac{1}{24} + \frac{1}{16} = \frac{5}{48}$$

Notation: $\mathbb{P}[X = 1, Y = 1]$ means the same as $\mathbb{P}[X = 1 \wedge Y = 1]$

$$\mathbb{P}[X = 2 \wedge Y = 3] = 0$$

If not all samples are equal zero ☺, then we similarly look at the sum of probabilities of all samples where $X = 1 \wedge Y = 1$ etc etc ...



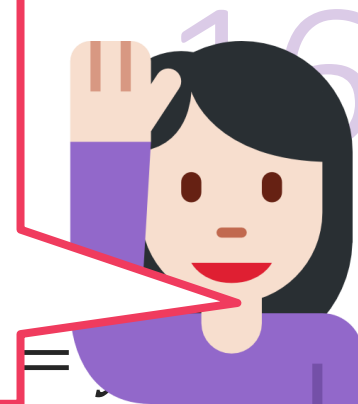
Still

zero ☺



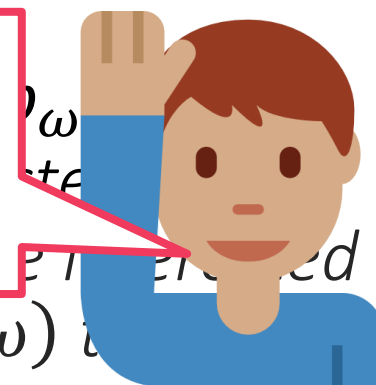
A

Keep in mind that this result holds for *any two* (or even more than two) r.v.s no matter how they are defined. This result holds even if the two r.v.s are clones of each other! This is so because the proof of this result never uses facts such as Y uses color in its definition and X does not etc. Even if both X, Y were defined using color of the ball, this result would still be true



The sum of probabilities over all values of x, y add up to 1 too

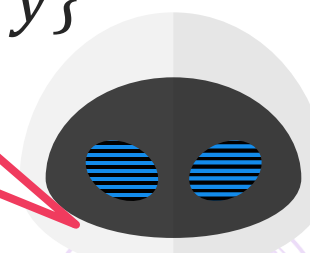
The PMF for this joint distribution is simply a function that takes two inputs, namely $x \in S_X$ and $y \in S_Y$ and gives us $\mathbb{P}[X = x \wedge Y = y]$. Often the notation $\mathbb{P}[X, Y]$ or $\mathbb{P}_{X,Y}[\cdot]$ is used to refer to this joint distribution



in all samples where $X(\omega) = x$ but we do not care what value $Y(\omega)$ takes

Just as before, we can sample from this PMF except in this case, the PMF would return back two numbers instead of one i.e. $(x, y) \sim \mathbb{P}_{X,Y}$ since what the PMF would do is obtain an outcome $\omega \in \Omega$ and simply return $(X(\omega), Y(\omega))$

$= y\}$



However, since $\sum_{x \in S_X} \mathbb{P}[X = x] = 1$, we conclude that we must also have

$$\sum_{x \in S_X} \sum_{y \in S_Y} \mathbb{P}[X = x, Y = y] = 1$$



Joint Distributions on more R.V.s

17

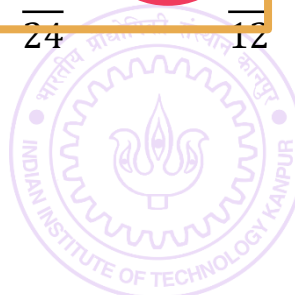
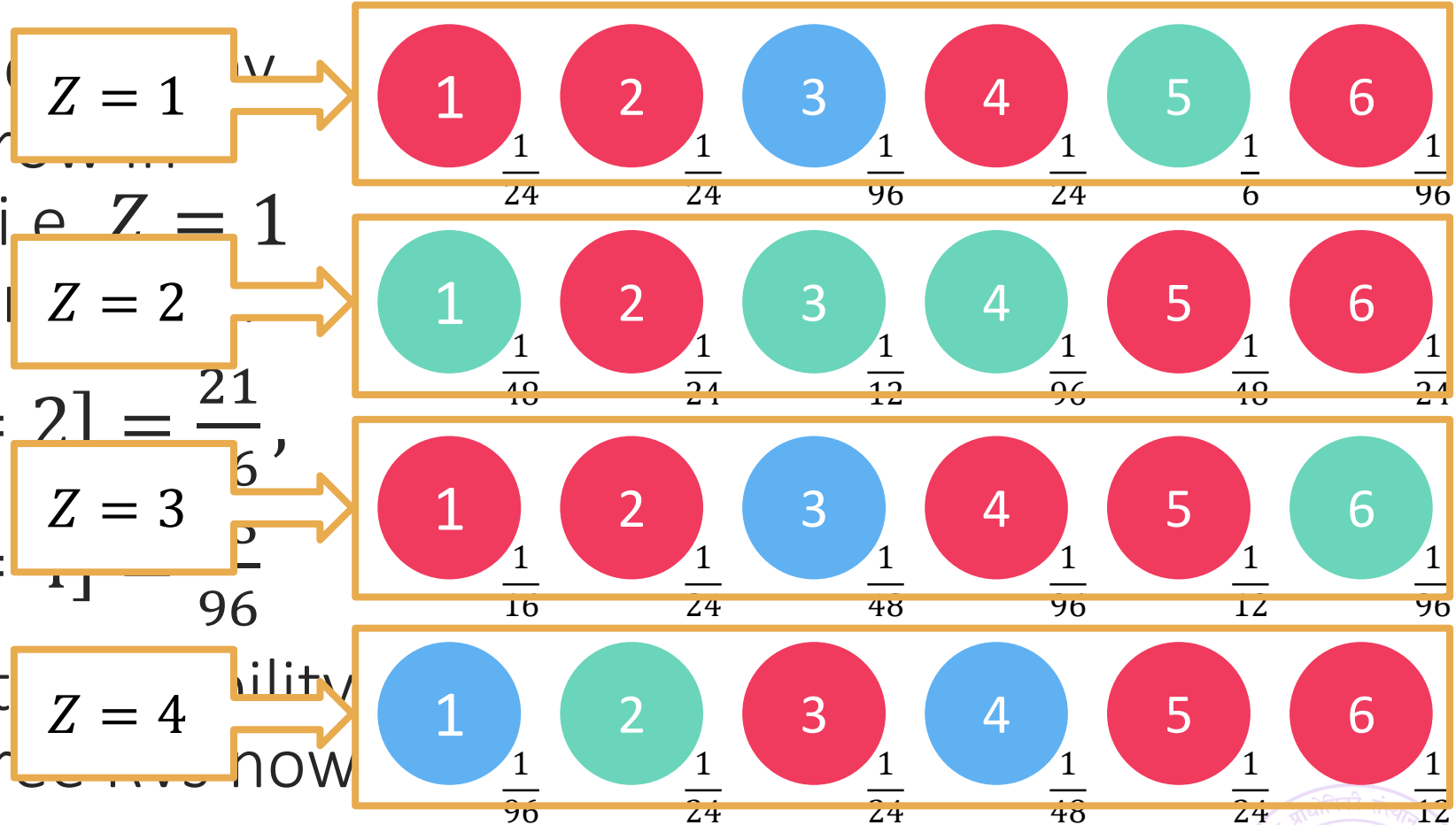
Suppose we had another
 $Z \in [4]$ indicating the row in which the ball is listed i.e. $Z = 1$ if the ball is in the first row

$$\mathbb{P}[Z = 1] = \frac{30}{96}, \mathbb{P}[Z = 2] = \frac{21}{96},$$

$$\mathbb{P}[Z = 3] = \frac{22}{96}, \mathbb{P}[Z = 4] = \frac{23}{96}$$

We could define a joint distribution on the three r.v.s now

$$\mathbb{P}[x, y, z] = \mathbb{P}[X = x, Y = y, Z = z]$$



Marginal Probability

18

When we had only two RVs (namely X, Y) we looked at how they behave at the same time (by looking at $\mathbb{P}_{X,Y}$) or how they behaved on their own (by looking at $\mathbb{P}_X, \mathbb{P}_Y$)

Now that we have three RVs (X, Y, Z) we can look at how they behave

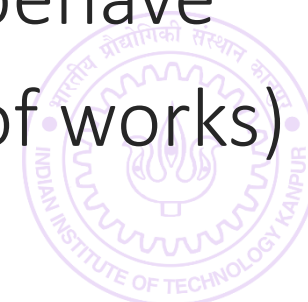
At the same time, by looking at $\mathbb{P}_{X,Y,Z}$

On their own, by looking at $\mathbb{P}_X, \mathbb{P}_Y, \mathbb{P}_Z$

Two at a time, by looking at $\mathbb{P}_{X,Y}, \mathbb{P}_{Y,Z}, \mathbb{P}_{X,Z}$

The distributions $\mathbb{P}_X, \mathbb{P}_Y, \mathbb{P}_Z, \mathbb{P}_{X,Y}, \mathbb{P}_{Y,Z}, \mathbb{P}_{X,Z}$ are called *marginal probability distributions* and they look at how a subset of RVs behave

Marginal distributions are also proper distributions (same proof works) and hence they also have PMFs associated with them



Obtaining Marginal PMF from Joint PMF 19

If we have the joint PMF for a set of RVs, say X, Y, Z , then obtaining the marginal PMF for any subset of these RVs is very simple

Involves a process called *marginalization*: uses the proof we saw earlier

Suppose we are interested in $\mathbb{P}_{X,Z}$ i.e. we don't care about Y

In this case we say that Y has been marginalized out

Earlier argument can be reused to show that for any $x \in S_X, z \in S_Z$

$$\{\omega \in \Omega: X(\omega) = x, Z(\omega) = z\} = \bigcup_{y \in S_Y} \{\omega \in \Omega: X(\omega) = x \wedge Y(\omega) = y \wedge Z(\omega) = z\}$$

This shows that $\mathbb{P}[X = x, Z = z] = \sum_{y \in S_Y} \mathbb{P}[X = x, Y = y, Z = z]$

Similarly, $\mathbb{P}[Z = z] = \sum_{x \in S_X} \sum_{y \in S_Y} \mathbb{P}[X = x, Y = y, Z = z]$

In this case we say that both X and Y have been marginalized out



Conditional Probability

20

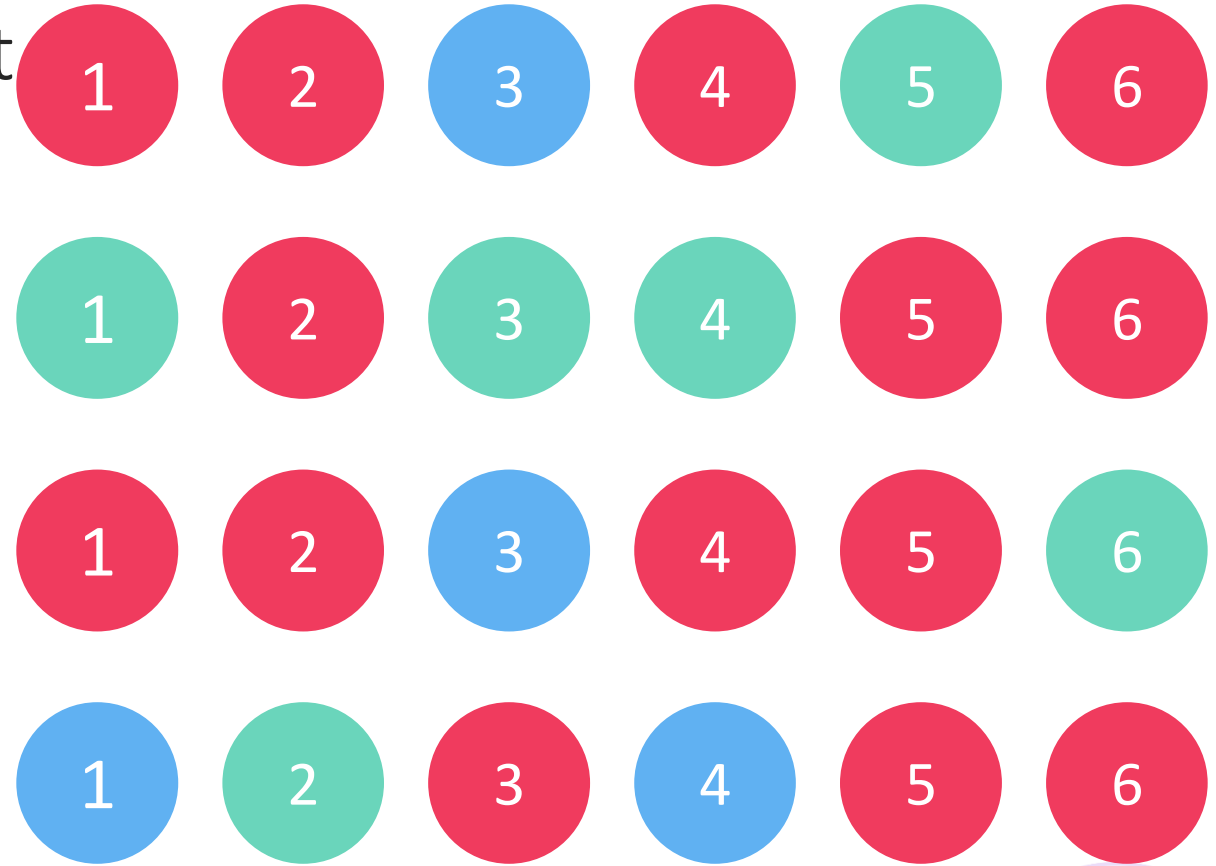
Perhaps one of the most important concepts w.r.t ML applications

Let us look at uniform case first

Notice: if we focus only on balls with number 2 written on them, most (3/4) of those balls are red

Contrast: if the number on the ball is 3, nothing as strong can be said

$\mathbb{P}[Y = 1|X = 2] \triangleq$ proportion of samples with $Y = 1$ among those samples where $X = 2$



In this case $\mathbb{P}[Y = 1|X = 2] = \frac{3}{4}$
and $\mathbb{P}[Y = 1|X = 3] = \frac{1}{4}$



Conditional Probabil

What if $\mathbb{P}[X = 2]$ happens to be 0?
Won't we get a divide-by-zero error?

The previous way of defining the conditional probability is make cumbersome to extend to more general settings

Let us use a different (but equivalent) way of defining conditional probability

$\mathbb{P}[Y = 1$

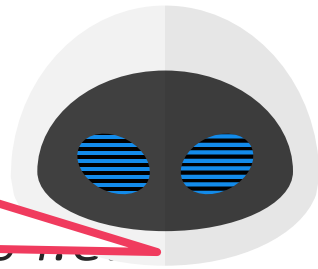
Dividin

Yes, although there are ways to get around this, for this course, we will avoid such cases or else, if convenient, define $\frac{0}{0} = 0$

$$\mathbb{P}[Y = 1|X = 2] \triangleq \frac{\text{proportion of samples with } y=1 \text{ and } x=2}{\text{proportion of samples with } x=2}$$

The above is just another way of saying

$$\mathbb{P}[Y = 1|X = 2] \triangleq \frac{\mathbb{P}[Y=1 \wedge X=2]}{\mathbb{P}[X=2]}$$



Conditional Probability

22

$$\mathbb{P}[Y = 1|X = 2] = \frac{\mathbb{P}[Y=1 \wedge X=2]}{\mathbb{P}[X=2]}$$

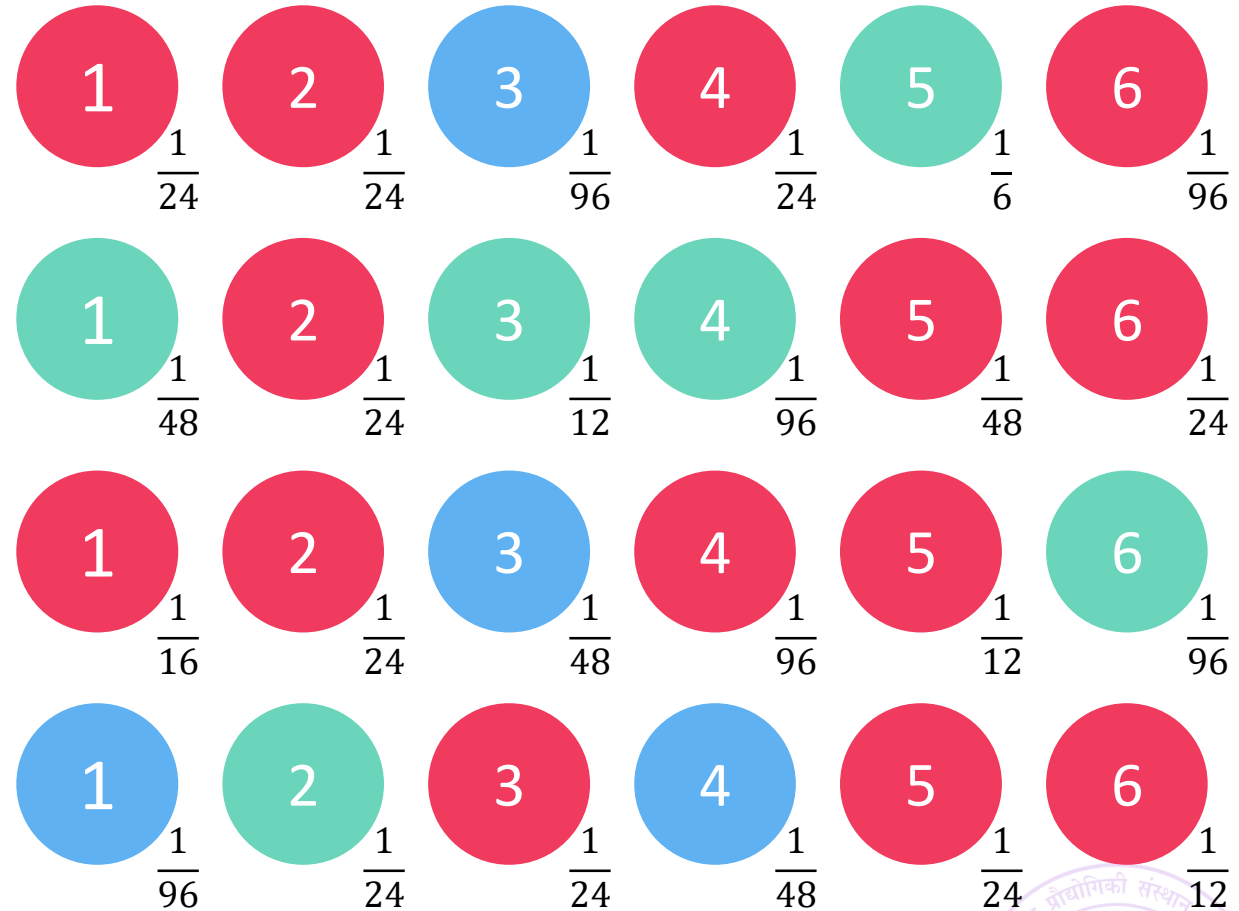
$$= \frac{\frac{1}{24} + \frac{1}{24} + \frac{1}{24}}{\frac{1}{24} + \frac{1}{24} + \frac{1}{24} + \frac{1}{24}} = \frac{3}{4}$$

$$\mathbb{P}[Y = 1|X = 3] = \frac{\mathbb{P}[Y=1 \wedge X=3]}{\mathbb{P}[X=3]}$$

$$= \frac{\frac{1}{24}}{\frac{1}{96} + \frac{1}{12} + \frac{1}{48} + \frac{1}{24}} = \frac{4}{15}$$

$$\mathbb{P}[Y = 3|X = 6] = \frac{\mathbb{P}[Y=3 \wedge X=6]}{\mathbb{P}[X=6]}$$

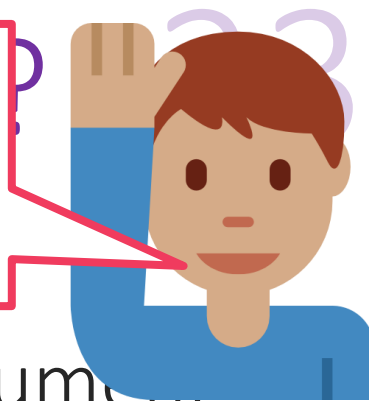
$$= 0$$



A

To sample from $\mathbb{P}[Y|X = x_0]$, we consider the set of only those outcomes $\omega \in \Omega$ in the sample space where $X(\omega) = x_0$, then sample an outcome ω_0 from this set with probability $\frac{p_{\omega_0}}{\mathbb{P}[X=x_0]}$ and then return $Y(\omega_0)$

Con



For any value of $x_0 \in S_X$ we have, by our marginalization argument

$$\sum_{y \in S_Y} \mathbb{P}[Y = y, X = x_0] = \mathbb{P}[X = x_0]$$

$$\text{This implies } \sum_{y \in S_Y} \mathbb{P}[Y = y|X = x_0] = \sum_{y \in S_Y} \frac{\mathbb{P}[Y=y, X=x_0]}{\mathbb{P}[X=x_0]} = 1$$

Thus, we can readily define a PMF for conditional distributions as well that takes in two values $x \in S_X, y \in S_Y$ and gives $\mathbb{P}[Y = y|X = x]$

We can similarly define $\mathbb{P}[X = x|Y = y_0]$ as well

Notation used $\mathbb{P}[Y|X], \mathbb{P}_{Y|X}[\cdot | \cdot]$

May ask for a sample $y \sim \mathbb{P}[Y|X = x_0]$ or $x \sim \mathbb{P}[X|Y = y_0]$ too!



Marginal Conditional Probability????

24

The operations of marginalization and conditioning can be used to define lots of different kinds of PMFs

For example consider $\mathbb{P}[X = x, Y = y | Z = z_0]$

If we marginalize Y out of the PMF for $\mathbb{P}[X = x, Y = y | Z = z_0]$, we will get the PMF for $\mathbb{P}[X = x | Z = z_0]$

$$\mathbb{P}[X = x | Z = z_0] = \sum_{y \in \mathcal{S}_Y} \mathbb{P}[X = x, Y = y | Z = z_0]$$

Can prove the above result using the same marginalization argument

Note that this means that this PMF can be derived from the PMF for the joint distribution i.e. $\mathbb{P}[X = x, Y = y, Z = z]$



Marginal Conditional Probability????

25

The operations of marginalization and conditioning can be used to define lots of different kinds of PMFs

For example consider $\mathbb{P}[Y = y|X = x_0, Z = z_0]$

We can show that this is nothing but

$$\mathbb{P}[Y = y|X = x_0, Z = z_0] = \frac{\mathbb{P}[Y=y \wedge X=x_0|Z=z_0]}{\mathbb{P}[X=x_0|Z=z_0]} = \frac{\mathbb{P}[Y=y \wedge X=x_0 \wedge Z=z_0]}{\mathbb{P}[X=x_0 \wedge Z=z_0]}$$

Try proving this result using the marginalization and conditioning rules

Yet again, this means that this PMF can be derived from the PMF for the joint distribution i.e. $\mathbb{P}[X = x, Y = y, Z = z]$



Rules of Probability

26

Sum Rule (Marginalization Rule) – aka Law of Total Probability

$$\mathbb{P}[x] = \sum_{y \in S_Y} \mathbb{P}[x, y]$$

or more explicitly

Product Rule (Conditionalization Rule)

$$\mathbb{P}[x, y] = \mathbb{P}[x|y] \cdot \mathbb{P}[y]$$

Combine to get $\mathbb{P}[x] = \sum_{y \in S_Y} \mathbb{P}[x|y] \cdot \mathbb{P}[y]$

or more explicitly, $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x|Y = y] \cdot \mathbb{P}[Y = y]$

Chain rule (Iterated Conditioning Rule)

$$\mathbb{P}[x, y, z] = \mathbb{P}[x|y, z] \cdot \mathbb{P}[y|z] \cdot \mathbb{P}[z]$$

We may use the fact that $\mathbb{P}[x, y, z] = \mathbb{P}[y, x, z] = \mathbb{P}[z, x, y] = \dots$ and also that $\mathbb{P}[x|y, z] = \mathbb{P}[x|z, y]$ etc to show that we also have

$$\begin{aligned}\mathbb{P}[x, y, z] &= \mathbb{P}[x|z, y] \cdot \mathbb{P}[z|y] \cdot \mathbb{P}[y] \\ &= \mathbb{P}[y|x, z] \cdot \mathbb{P}[x|z] \cdot \mathbb{P}[z] \\ &= \mathbb{P}[y|z, x] \cdot \mathbb{P}[z|x] \cdot \mathbb{P}[x] \\ &= \mathbb{P}[z|x, y] \cdot \mathbb{P}[x|y] \cdot \mathbb{P}[y] \\ &= \mathbb{P}[z|y, x] \cdot \mathbb{P}[y|x] \cdot \mathbb{P}[x]\end{aligned}$$



Bayes Theorem

27

The foundation of Bayesian Machine Learning

$$\mathbb{P}[Y = y|X = x] = \frac{\mathbb{P}[Y=y, X=x]}{\mathbb{P}[X=x]} \text{ and } \mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X=x, Y=y]}{\mathbb{P}[Y=y]}$$

However $\mathbb{P}[Y = y, X = x]$ and $\mathbb{P}[X = x, Y = y]$ are the same thing

$$\text{Thus, } \mathbb{P}[Y = y|X = x] \cdot \mathbb{P}[X = x] = \mathbb{P}[X = x|Y = y] \cdot \mathbb{P}[Y = y]$$

This gives us

$$\mathbb{P}[Y = y|X = x] = \frac{\mathbb{P}[X=x|Y=y] \cdot \mathbb{P}[Y=y]}{\mathbb{P}[X=x]}$$

Similarly

$$\mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[Y=y|X=x] \cdot \mathbb{P}[X=x]}{\mathbb{P}[Y=y]}$$



Marginal, Joint and Conditional Probability 28

In most settings, we would have defined tons of random variables on our outcomes to capture interesting things about the outcomes

X : what is the gender of the person visiting our website $S_X = \{1,2,3\} = [3]$

Y : what is the age of the person $S_Y = \mathbb{N}$

Z : how many seconds did they spend on our website $S_Z = \mathbb{N}$

A : what ad were they shown $S_A = [10]$

P : what purchase did they make $S_P = [10] \cup \{-1\} = \{-1,1,2, \dots, 10\}$

ML algos like to ask and answer interesting questions about these random variables

Marginal, Joint and Conditional Probability give us the language to speak when asking and answering these questions



Using Probability to do ML

29

Arguably the most interesting random variable of X, Y, Z, A, P is P

Recommendation Systems (RecSys) would like to know what value would P take if we know the values of X, Y, Z, A

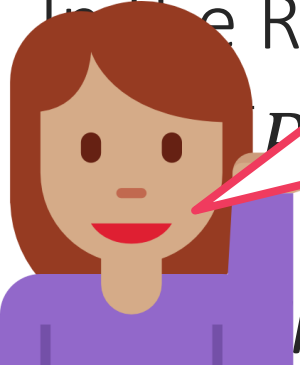
Of these, the website cannot control X, Y, Z but it does control A

The whole enterprise of recommendation and ad placement can be summarized in the following statement

“Given values $x \in S_X, y \in S_Y, z \in S_Z$, find a value of $a \in S_A$ such that $\mathbb{P}[P \neq -1 \mid x, y, z, a]$ is as close to 1 as possible”

ML algos for RecSys can learn distributions (the models for these ML algos are distributions) such that they mimic reality i.e. if the model says $\mathbb{P}[P \neq -1 \mid x, y, z, a] \approx 1$, the user really does buy something

Creating Events from Random Variables?? 30



We can also have events like $\{X < x_0\}$ (that the random variable takes a value less than x_0). We define $\mathbb{P}[X < x_0] \triangleq \sum_{\omega: X(\omega) < x_0} p_{\omega}$.

Similarly, $\{X \leq x_0\}$, $\{X \geq x_0\}$ are also valid events

are interested in the probability of this event given x, y, z, a
I: events are merely a description of interesting facts about an outcome

Similarly $\{X = 1 \wedge Y = 2\}$ is also an event (that the ball we picked is green and has the number 1 stamped on it)

$\{X = 1\}$ is also an event (that the ball has the number 1 stamped on it)

$\{Y = 2\}$ is also an event (that the ball is green colored)

Given an event, it may happen on certain outcomes, not happen on others
e.g. if $\{Y = 2\}$ then this event will not have taken place if we pick a blue ball

Thus, given any collection of random variables, we can create events out of them and ask interesting questions about the r.v.s



Event Calculus

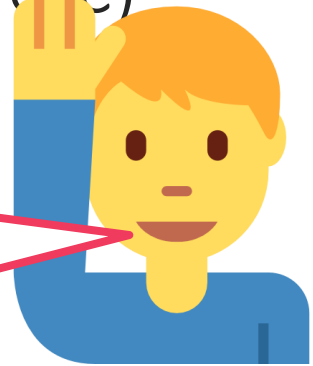
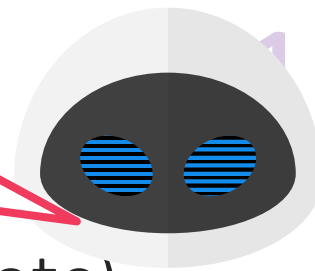
The term *calculus* in general, means a system of rules – it does not necessarily mean differentiation or integration 😊

Let A, B be events (possibly defined using same/different r.v.s etc)

$\neg(A \cup B) = \neg A \cap \neg B$ tell us that saying

“It is not the case that either A happened or B happened or both happened”
is just a funny way of saying

“A did not happen and B did not happen i.e. neither happened”

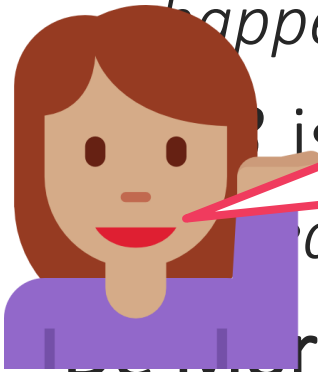


Called the
happened

$\neg(A \cap B) = \neg A \cup \neg B$ tell us that saying

“It is not the case that both A and B happened”
is just another way of saying

“Either A did not happen or B did not happen or both did not happen”



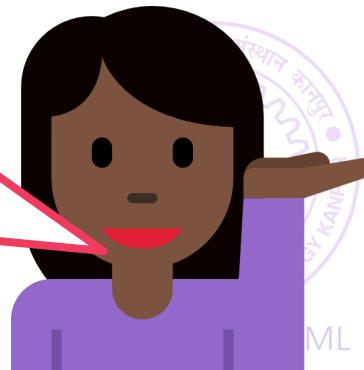
and the

intersection of the two events (both A and B happened)

De Morgan's Law

$$\neg(A \cup B) = \neg A \cap \neg B$$

This means we can be more creative in defining events,
e.g. $\{X \leq x_0 \wedge X \geq x_1\}$ (also written as $\{x_1 \leq X \leq x_0\}$)
or else $\{x_1 \leq X \leq x_0\} \vee \{x_3 \leq X \leq x_2\}$



Example: let $A \equiv \{X = 2\}$, $B \equiv \{Y = 1\}$ be events

$\neg A$ is also an event (number on ball is something other than 2)

$\neg B$ is also an event (the colour of the ball is something other than red)

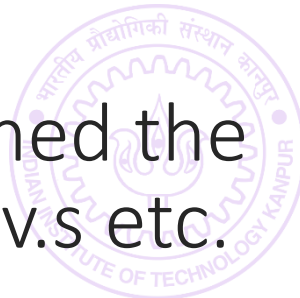
$A \cup B$ is also an event (either I have a red ball or a ball with number 2 written on it or else a red ball with number 2 written on it)

$A \cap B$ is also an event (I have a red ball and the number on it is 2)

De Morgan's Laws: for any two events A, B , we must have

$$\neg(A \cup B) = \neg A \cap \neg B \text{ as well as } \neg(A \cap B) = \neg A \cup \neg B$$

Caution: De Morgan's Laws **always** hold no matter how we defined the events. They do not require events to be defined on separate r.v.s etc.



We can derive an “**intersection rule**” using de-Morgan’s laws and these rules

$$\begin{aligned}\mathbb{P}[A \cap B] &= \mathbb{P}[\neg(\neg(A \cap B))] = 1 - \mathbb{P}[\neg(A \cap B)] = 1 - \mathbb{P}[\neg A \cup \neg B] \\ &= \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[\neg A \cap \neg B] - 1 \text{ (apply union and complement rules)}\end{aligned}$$

Suppose we know probability of two events X and Y

Complement Rule: $\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$

Union Rule: $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

These rules can be proved using a similar proof technique that we used for joint/marginal probability derivations in the last lecture

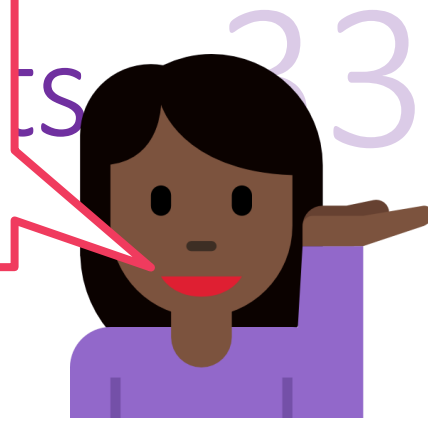
The above allow us to calculate more interesting probabilities

$$\mathbb{P}[X = 1 \vee Y = 2] = \mathbb{P}[X = 1] + \mathbb{P}[Y = 2] - \mathbb{P}[X = 1 \wedge Y = 2]$$

We used only the marginal and the joint probability distributions

$$\mathbb{P}[X \neq 1] = 1 - \mathbb{P}[X = 1]$$

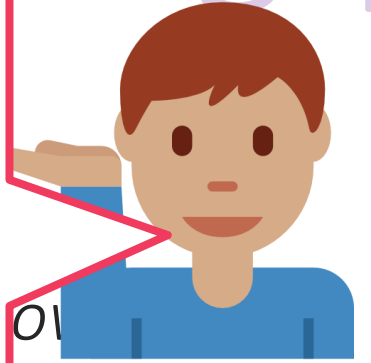
The above rules hold even for conditional probability



In fact, Bayes Theorem applies to events just as well i.e. if A, B are events, then

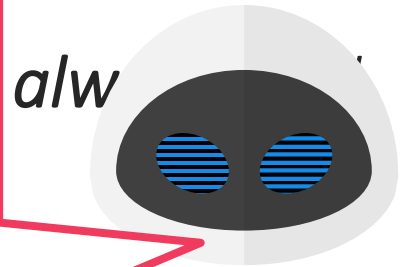
$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$$

Proof: create a random variable for each event e.g. $M = 1$ if event A happens and $M = 0$ if A does not happen. Similarly define a random variable N to tell us whether B happened or not. Now use the Bayes theorem on M, N – done!



The random variables M, N we defined above to tell us whether some event happened or not are called *indicator random variables* since they *indicate* whether an event took place (in which case the r.v. takes value 1) or not (in which case the r.v. takes value 0). **Notation:** $M = \mathbb{I}\{A\}, N = \mathbb{I}\{B\}$.

In general, $\mathbb{I}\{\text{blah}\} = 1$ if blah is true else 0



Conditional Implication Rule: If $C \Rightarrow A$, then $\mathbb{P}[A | C] = 1$. On the other hand, if $C \Rightarrow \neg B$, then $\mathbb{P}[B | C] = 0$.

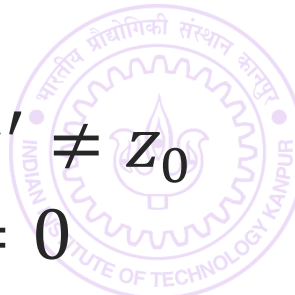
Example: C defined as above and

In the above cases, we will indeed have

Caution: not a standard rule you would find in textbooks

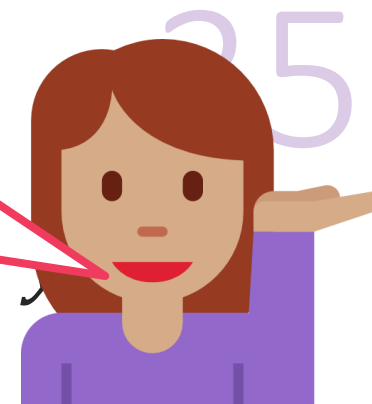
$\{Z = z'\}, z' \neq z_0$

$\mathbb{P}[B | C] = 0$



Independence

$X \perp\!\!\!\perp Y$ means that X and Y will both take values according to their own (marginal) PMFs, happily unmindful of the values the other r.v. is taking!



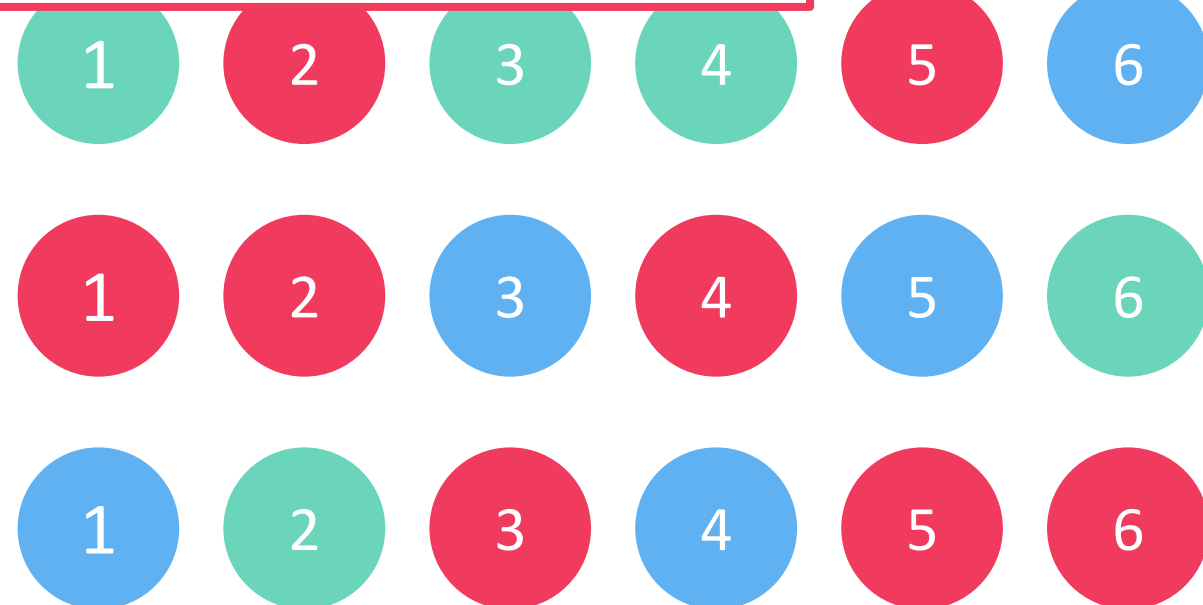
Two r.v.s X, Y are said to be independent if we have $\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$

Can you show that if $X \perp\!\!\!\perp Y$ (i.e. $\mathbb{P}_{Y|X=x_0} = \mathbb{P}_Y$ for all $x_0 \in S_X$) then we must always have $Y \perp\!\!\!\perp X$ (i.e. $\mathbb{P}_{X|Y=y_0} = \mathbb{P}_X$ for all $y_0 \in S_Y$) as well?

Hint: use the definition of independence



$\mathbb{P}_{Y|X=x_0} = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right] = \mathbb{P}_Y$ for
all $x_0 \in [6] = S_X$ i.e. $X \perp\!\!\!\perp Y$
Similarly, we can verify that
 $\mathbb{P}_{X|Y=y_0} = \left[\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right] = \mathbb{P}_X$ for
all $y_0 \in [3] = S_Y$ i.e. $Y \perp\!\!\!\perp X$



Conditional Independence

36

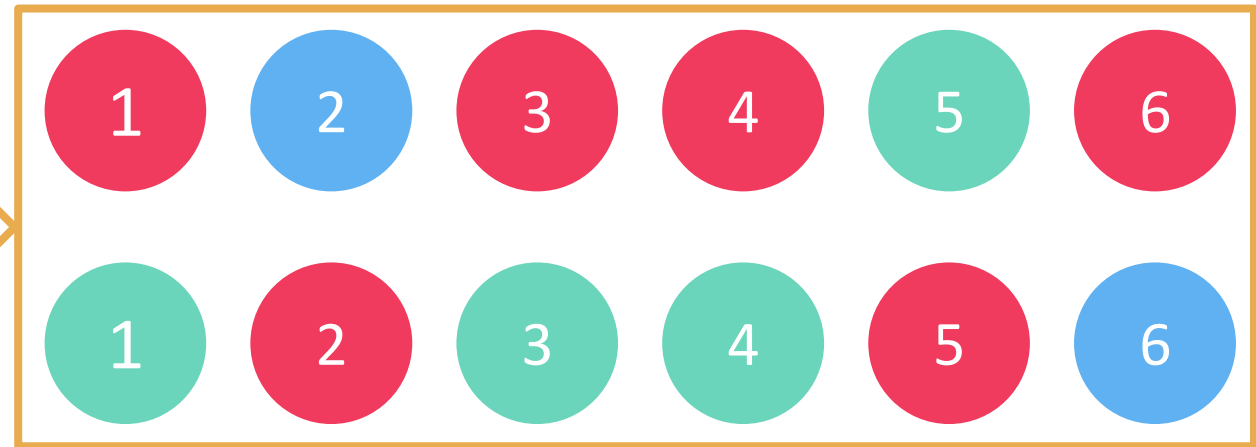
If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y | Z = z] = \mathbb{P}[X = x | Z = z] \cdot \mathbb{P}[Y = y | Z = z]$$

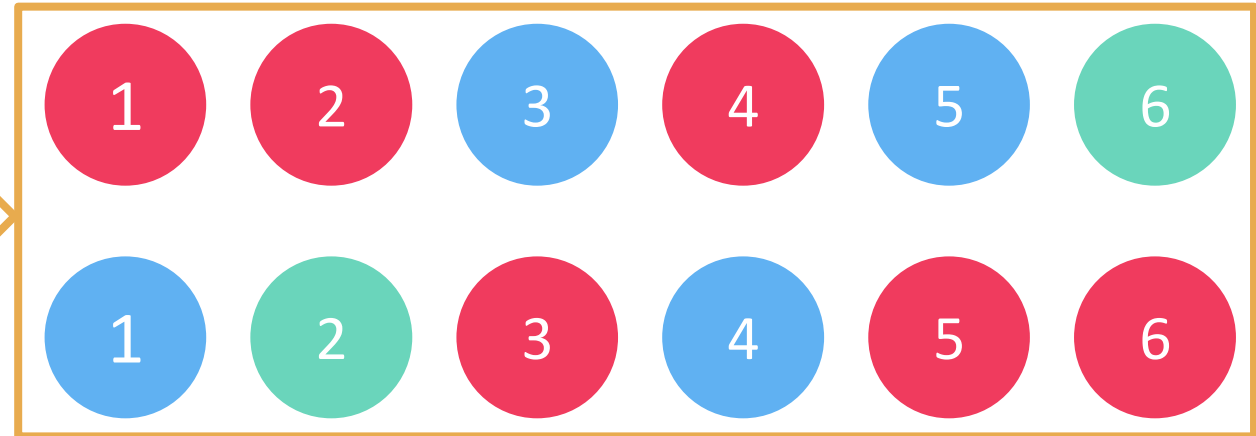
This is the earlier example where we verified that $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp X$. However it is easy to see that we do not have $X \perp\!\!\!\perp Y | Z$ since

$$\mathbb{P}[X = 2, Y = 2 | Z = 1] = 0 \neq \mathbb{P}[X = 2 | Z = 1] \cdot \mathbb{P}[Y = 2 | Z = 1]$$

$Z = 1$



$Z = 2$



Conditional Independence

37

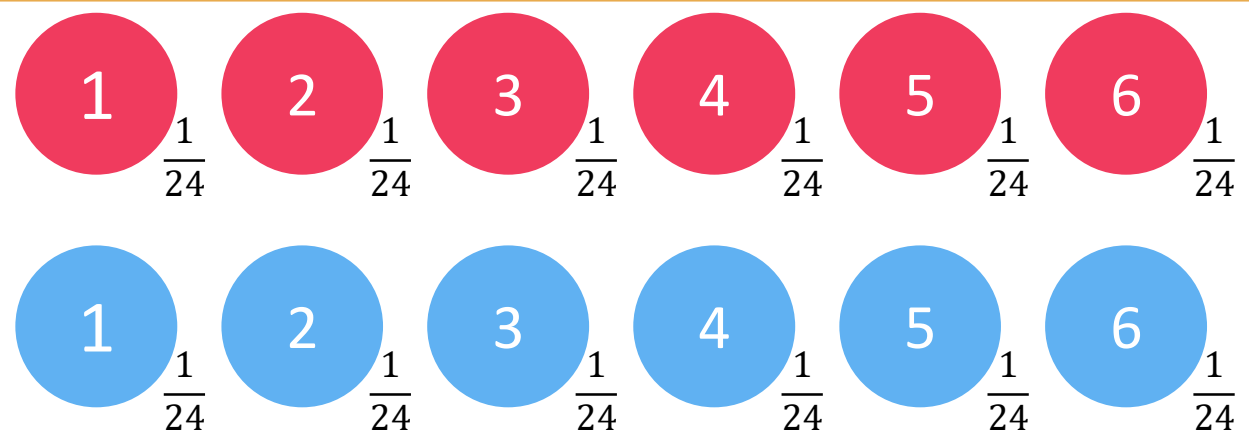
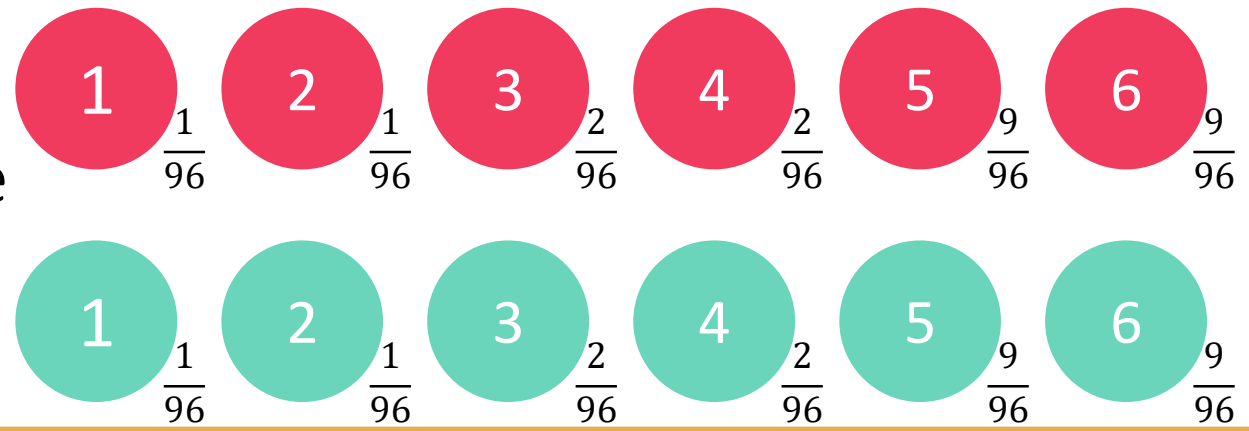
If X, Y, Z are three r.v.s such that for all $x \in S_X, y \in S_Y, z \in S_Z$ we have

$$\mathbb{P}[X = x, Y = y | Z = z] = \mathbb{P}[X = x | Z = z] \cdot \mathbb{P}[Y = y | Z = z]$$

The r.v.s number X and colour Y are not independent here. To note that $\mathbb{P}[Y = 3 | X = 1] = \frac{1}{6} \neq 0.25 = \mathbb{P}[Y = 3]$. However, if we condition on a new random variable Z that distinguishes the first two rows from the last two rows, then X and Y become independent r.v.s but we do have $X \perp\!\!\!\perp Y | Z$

$Z = 1$

$Z = 2$



Conditional Independence

38

Slightly more “practical” examples

If I throw a fair dice twice, the outcome on the first throw in no way influences the second outcome i.e. two outcomes are independent of each other and each takes values in $[6]$. However, if I additionally am told that the sum of the two numbers is 8, then the throws are no longer conditionally independent since for example we have

$$\mathbb{P}[X = 1, Y = 4 \mid Z = 8] = 0 \neq \mathbb{P}[X = 1 \mid Z = 8] \cdot \mathbb{P}[Y = 4 \mid Z = 8]$$

In the above, X is number on first throw, Y denotes second, Z is sum

Sample space is $[6] \times [6]$ i.e. all possible outcomes of two throws

Examples where non-independent r.v.s become conditionally independent are most commonly found in a branch of ML called graphical models.

