Probability Theory

CS771: Introduction to Machine Learning

Purushottam Kar

Announcements

Reminder: Quiz 2 30 August, 6PM (venue L20 as before)

Reminder: Assignment 1 deadline 01 September, 11:59PM IST

Note: correction in hinge loss calculations in eval.py (assignment 1) as well as lec 7-8, lec 9 lecture notebooks

Please be careful with update equations

- *The above mistake does not affect the ML algo itself but it results in us plotting y axis of convergence curves wrongly – may mislead/confuse us*
- *Several queries on Piazza also seem to go away the moment a careful look is given to derivation of equations (e.g. tiny things like factors of 2 etc)*

Recap of Last Lecture

Looked at the k-means clustering algo, k-means++ initializer

Studied how clustering can help accomplish a lot of things

- *Make LwP a much more powerful classifier*
- *Make* any *ML algo more powerful by identifying subpopulations in data*
- *Be used to construct stumps for constructing decision trees*
- *Reduce number of features (dimensionality reduction)*

Started looking at probability theory

- ***Sample space****: set of* all *possible samples or outcomes (even unlikely ones)*
- ***Event****: a description of some facts about outcomes that we find useful*
- ***Random variables (r.v.)****: a way to express facts about outcomes as numbers*
  - May think of random variables as *numerical features* describing a sample/outcome
  - Random variables can take discrete (categorical) values or even continuous values

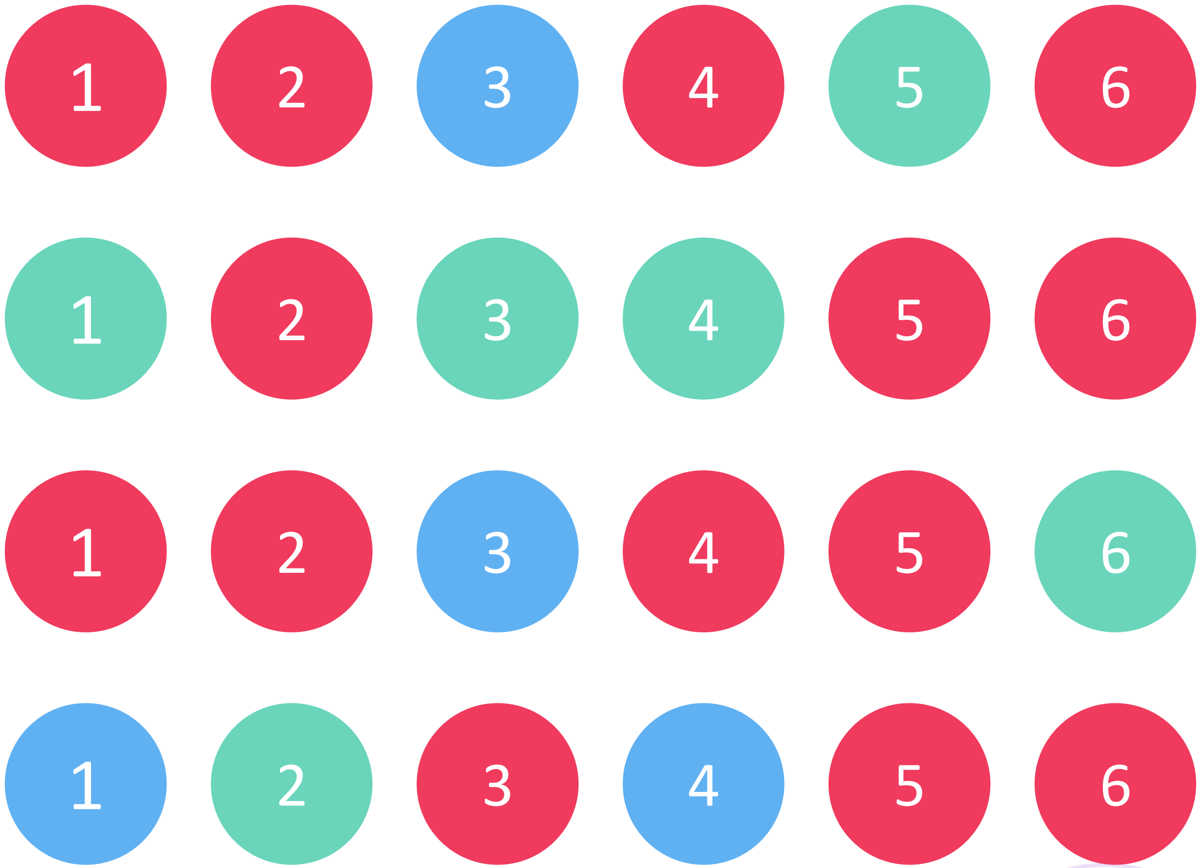
Probability as Proportions

Sample/Outcome: pick one ball

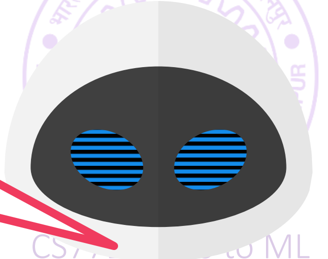
Sample space: $\{\mathrm{R},\mathrm{G},\mathrm{B}\} \times [6]$

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

Toy setting to help us understand 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

Sample/Outcome: pick one ball

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

Define two random variables (r.v.)

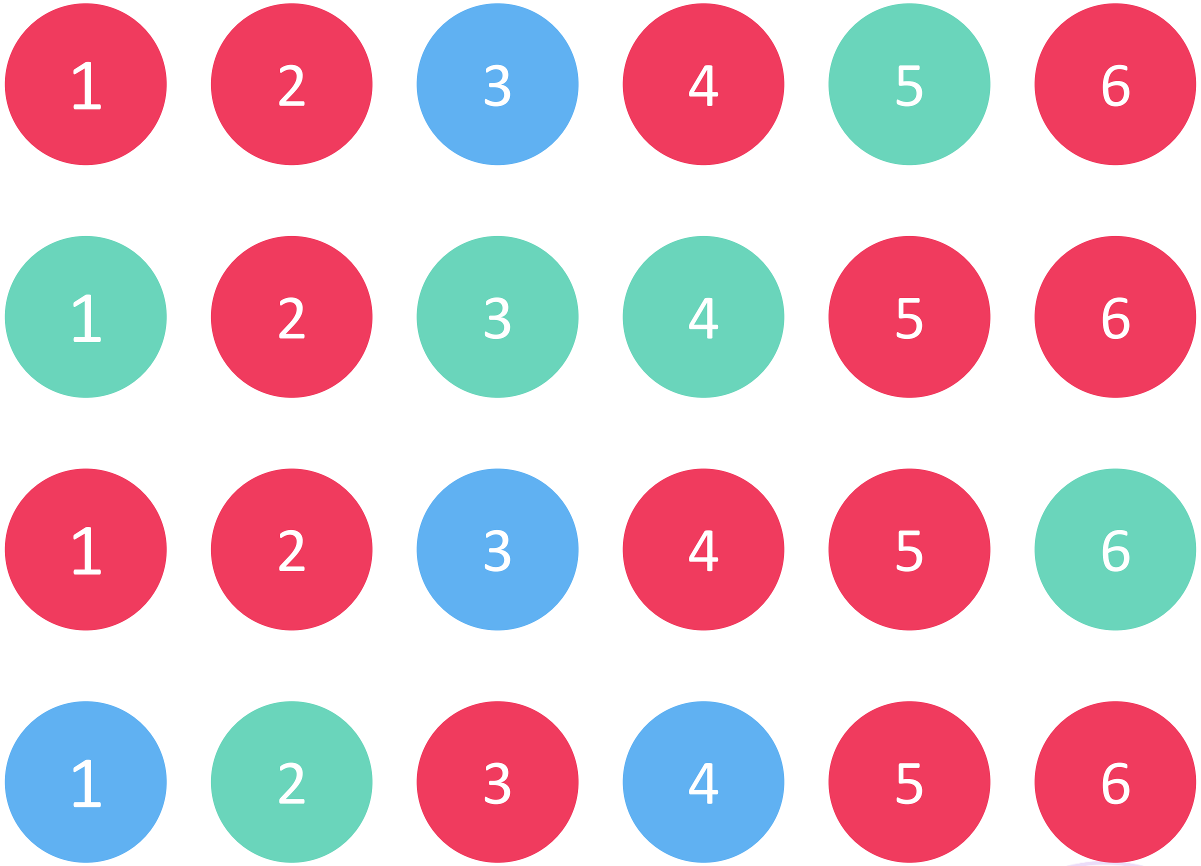
$X \triangleq$ number on the ball $\in [6]$

$Y \triangleq$ colour of the ball

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

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

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



Probability as Proportions

Sample/Outcome: pick one ball

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

Define two random variables (r.v.)

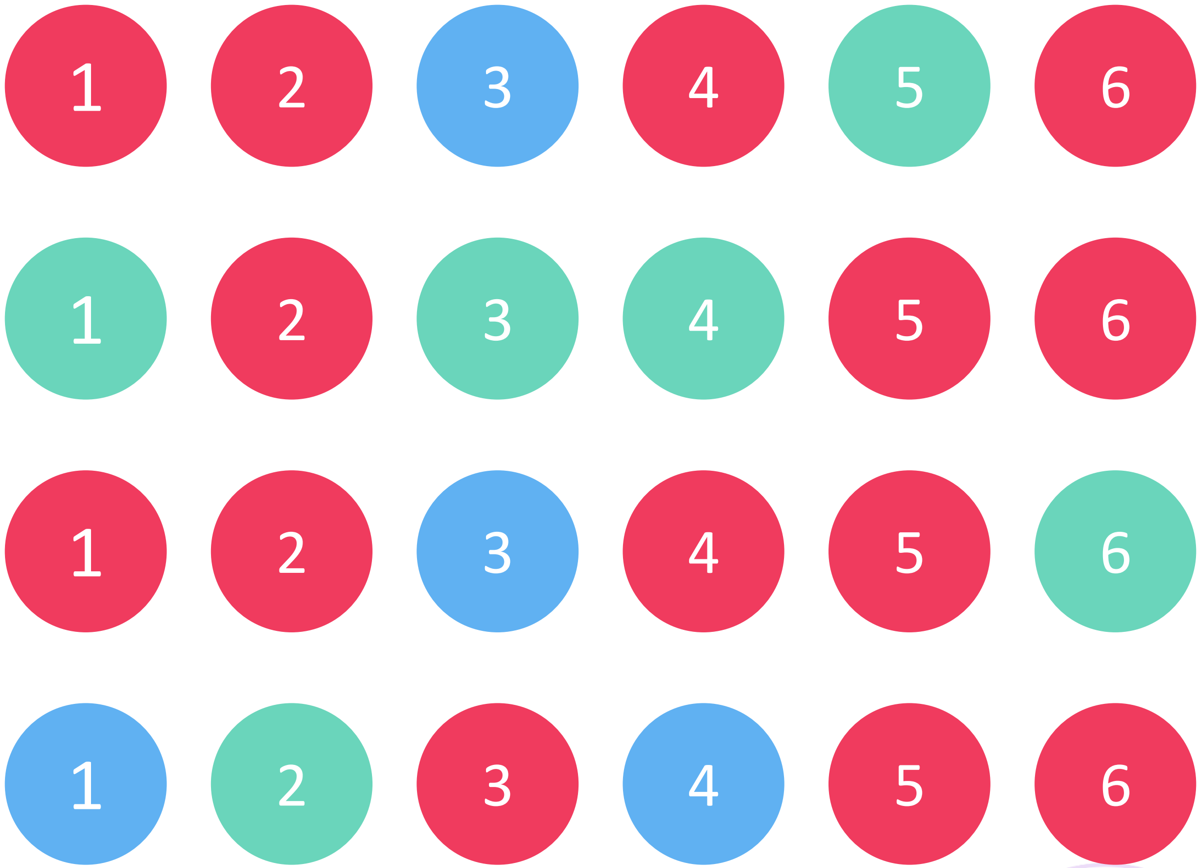
$X \triangleq$ number on the ball $\in [6]$

$Y \triangleq$ colour of the ball

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

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

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



Probability beyond Proportions

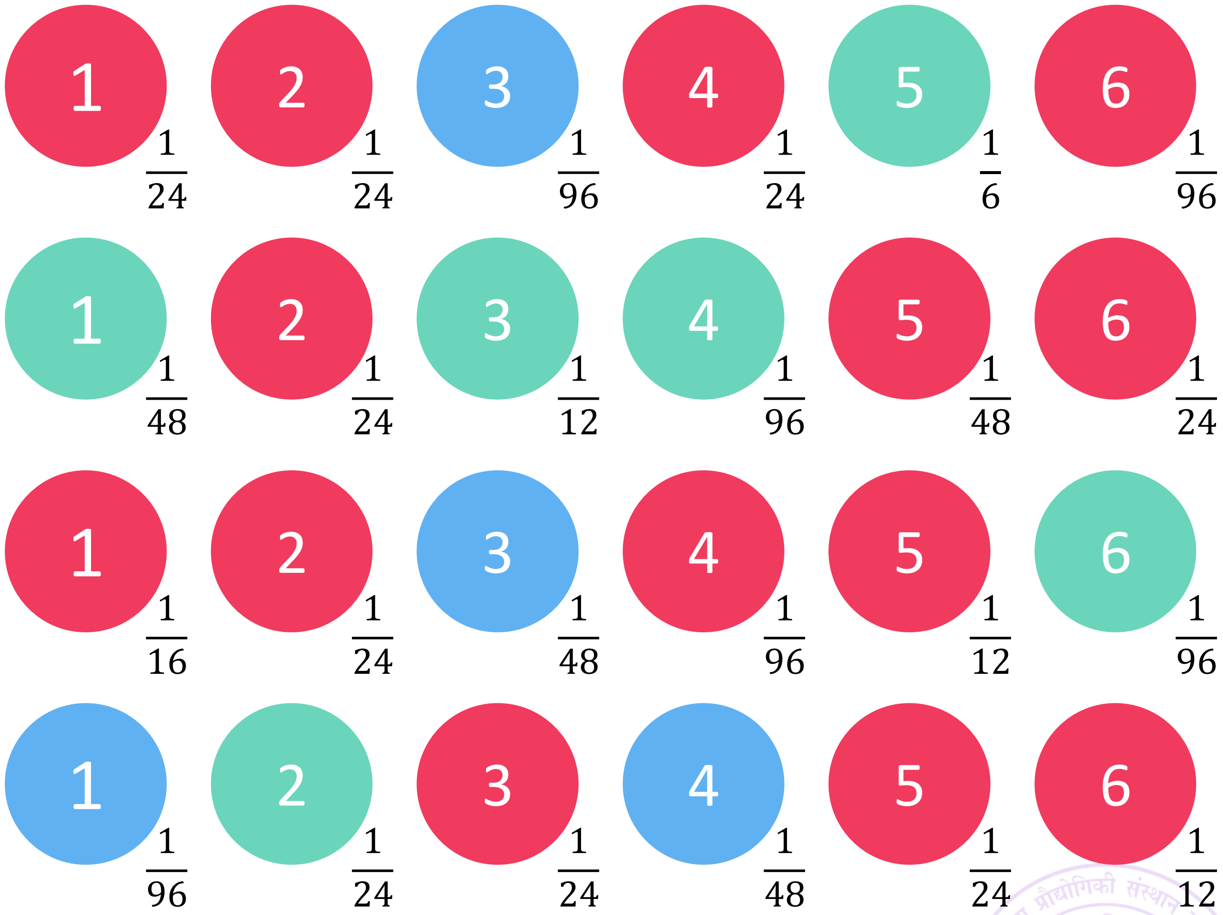
Suppose now that not all samples are equally likely, i.e. not all balls are equally likely to be picked

$\mathbb{P}[Y = 2] \triangleq$ 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

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 often 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 value of X on that outcome

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

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

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

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

Rules of Probability

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 often 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$

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

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 Mass Function (PMF)

A fancy name for a function that tells us the probability of the random variable taking any particular value

For a discrete random variable X, its PMF $f(\cdot)$ tells us, for any $x \in S_X$, what is the probability of X taking the value x i.e. $f(x) = \mathbb{P}[X = x]$

Warning: papers/books often use lazy or confusing notation – 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 $X(\omega)$

Probability M

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 r

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 $Y = 3$

ny x

[X

Warning: papers/books often use lazy or confusing notation – 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

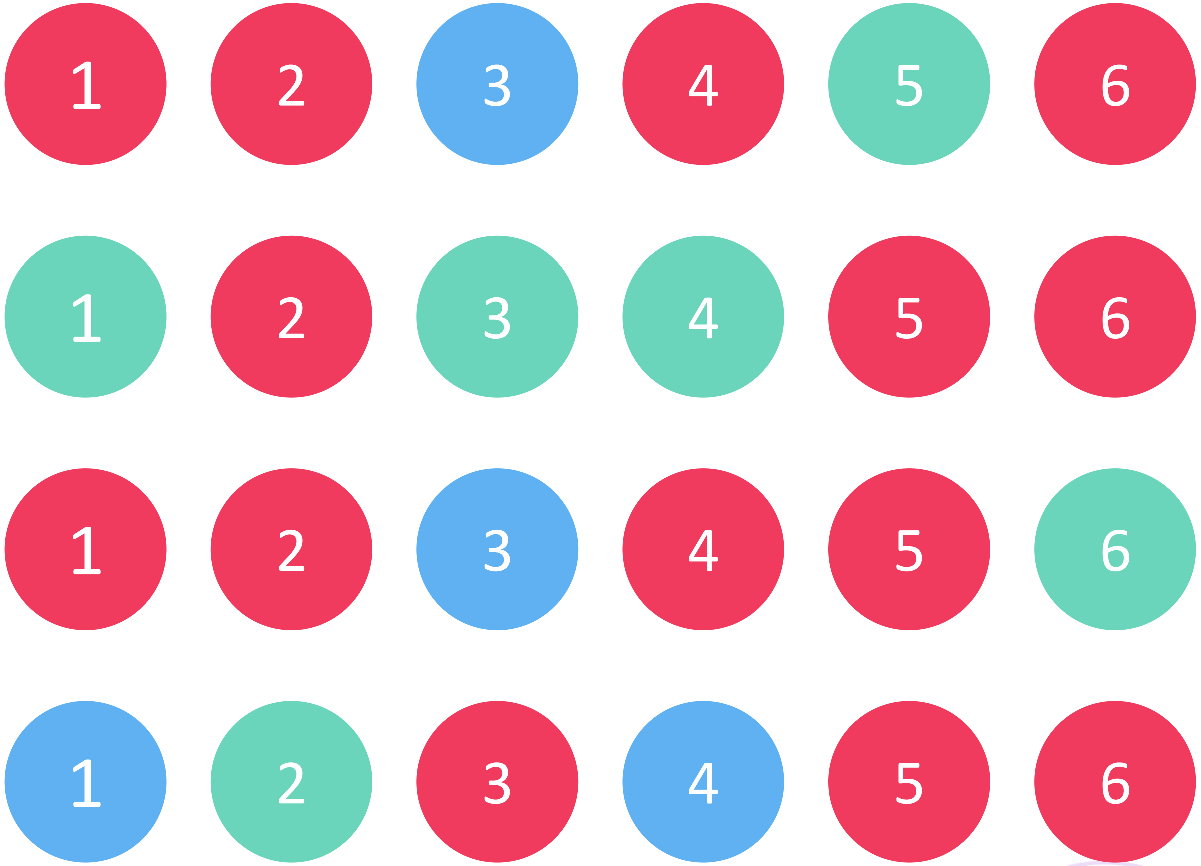
$\mathbb{P}[X = 1 \wedge Y = 1] \triangleq$ proportion of samples for which we have both $X = 1$ and $Y = 2$

Let us look at uniform case first

$$\mathbb{P}[X = 1 \wedge Y = 1] = \frac{2}{24} = \frac{1}{12}$$

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

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



Joint Probability

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

Let us look at uniform case first

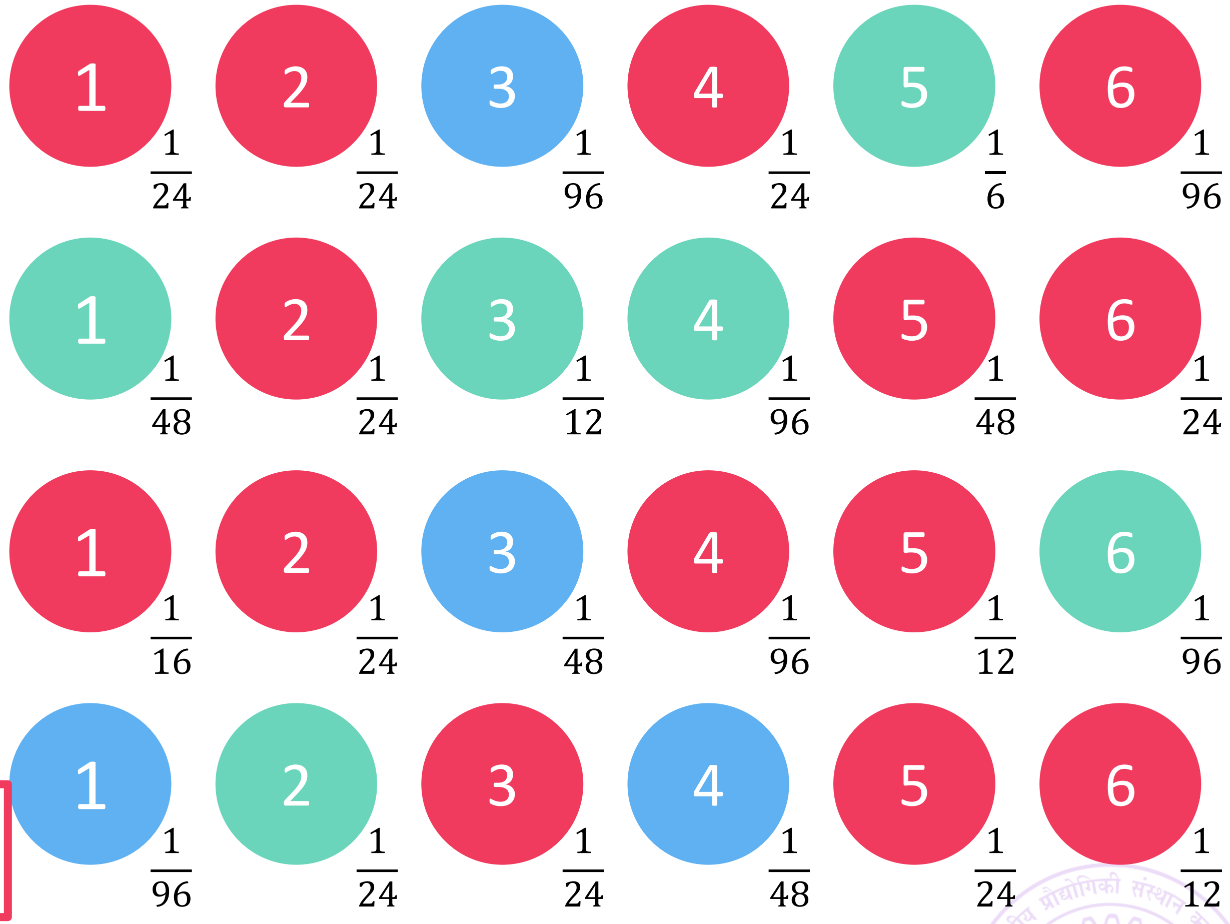
$$\mathbb{P}[X = 1 \wedge Y = 1] = \frac{1}{24} + \frac{1}{16} = \frac{5}{48}$$

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

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

Still zero ☺

If not all samples are equally likely, then we similarly look at the sum of probabilities of all samples where $X = 1 \wedge Y = 1$ etc etc …



A PMF for the Joint Distribution?

The joint probabilities also form a valid distribution

For any $x \in S_X, y \in S_Y$*, we have* $\mathbb{P}[X = x, Y = y] \triangleq \mathbb{P}[X = x \wedge Y = y] \geq 0$

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

Proof*: Recall that we defined* $\mathbb{P}[X = x] = \sum_{\omega \in \Omega: X(\omega) = x} p_\omega$ *where* p_ω *is the probability with which an outcome* ω *happens. Thus, we are interested in all samples where* $X(\omega) = x$*. Another way of saying this is that we are interested in all samples where* $X(\omega) = x$ *but we do not care what value* $Y(\omega)$ *takes i.e.*

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

Thus, we have $\sum_{\omega \in \Omega: X(\omega) = x} p_\omega = \sum_{y \in S_Y} \sum_{\omega \in \Omega: X(\omega) = x \wedge Y(\omega) = y} p_\omega$

Thus, we have $\mathbb{P}[X = x] = \sum_{y \in S_Y} \mathbb{P}[X = x, Y = y]$

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

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

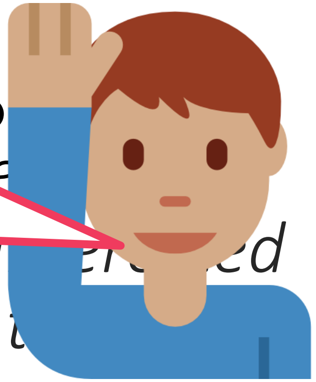
A

The

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

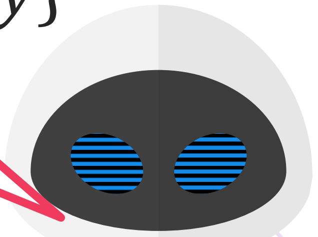
The PMF for this joint distribution would simply be a function that takes two inputs, namely $x \in S_X$ and $y \in S_Y$ and gives us $\mathbb{P}[X = x \wedge Y = y]$. Often the notation $\mathbb{P}[X, Y]$ or $\mathbb{P}_{X,Y}[\cdot]$ is used to refer to this joint distribution



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

$= y\}$

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 $\left(X(\omega), Y(\omega)\right)$



Thus, we have $\mathbb{P}[X = x] = \sum_{y \in S_Y} \mathbb{P}[X = x, Y = y]$

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

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

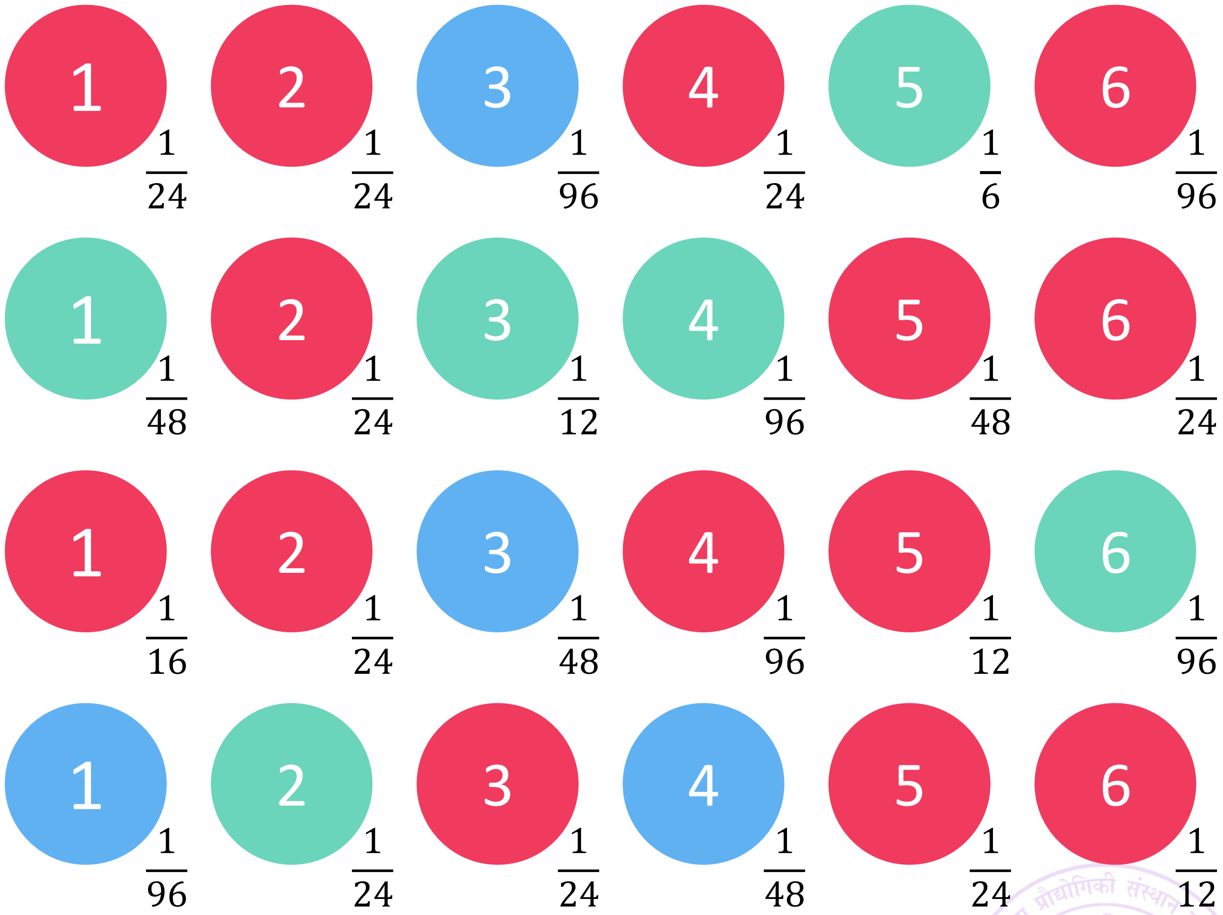
Joint Distributions on more R.V.s

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

$$\mathbb{P}[Z = 1] = \frac{30}{96}, \mathbb{P}[Z = 2] = \frac{21}{96},$$
$$\mathbb{P}[Z = 3] = \frac{22}{96}, \mathbb{P}[Z = 4] = \frac{23}{96}$$

We could define a joint probability distributions on the three RVs now

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



🔴 1 $\frac{1}{24}$	🔴 2 $\frac{1}{24}$	🔵 3 $\frac{1}{96}$	🔴 4 $\frac{1}{24}$	🟢 5 $\frac{1}{6}$	🔴 6 $\frac{1}{96}$
🟢 1 $\frac{1}{48}$	🔴 2 $\frac{1}{24}$	🟢 3 $\frac{1}{12}$	🟢 4 $\frac{1}{96}$	🔴 5 $\frac{1}{48}$	🔴 6 $\frac{1}{24}$
🔴 1 $\frac{1}{16}$	🔴 2 $\frac{1}{24}$	🔵 3 $\frac{1}{48}$	🔴 4 $\frac{1}{96}$	🔴 5 $\frac{1}{12}$	🟢 6 $\frac{1}{96}$
🔵 1 $\frac{1}{96}$	🟢 2 $\frac{1}{24}$	🔴 3 $\frac{1}{24}$	🔵 4 $\frac{1}{48}$	🔴 5 $\frac{1}{24}$	🔴 6 $\frac{1}{12}$

Joint Distributions on more R.V.s

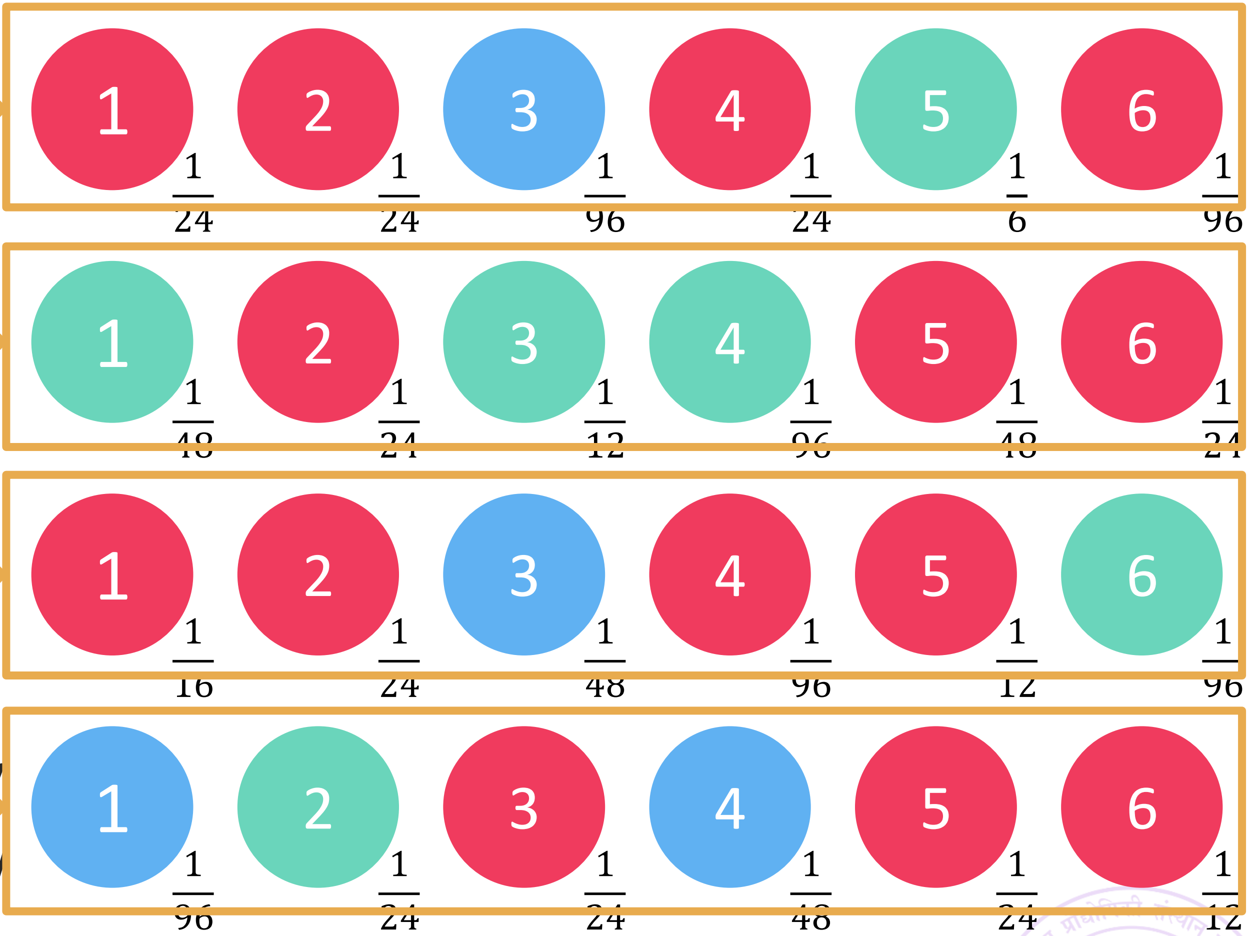
Suppose we had anoth v $Z \in [4]$ indicating the r which the ball is listed i.e. $Z = 1$ if the ball is in the first r

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

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

We could define a joint ility distributions on the thr how

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



Marginal Probability

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

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

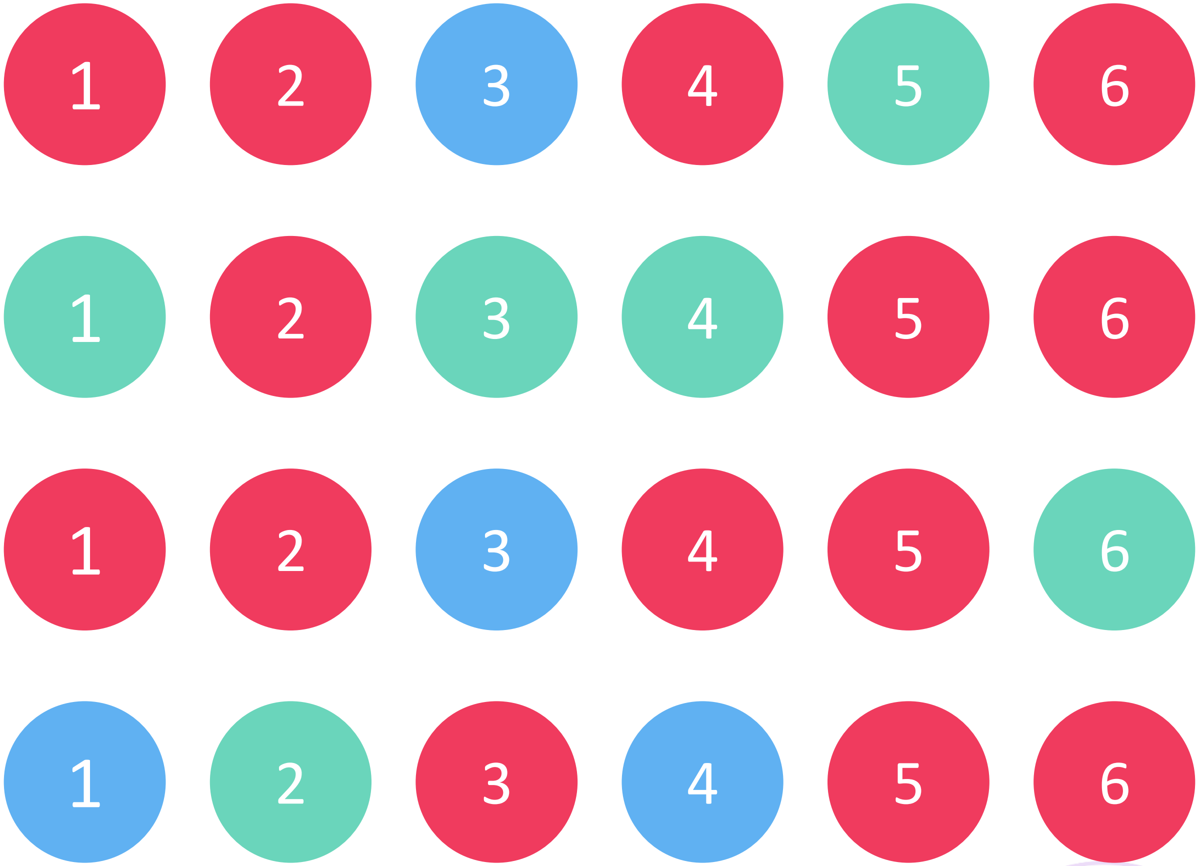
Perhaps one of the most important concepts w.r.t ML applications

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 Probability

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

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

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

Dividing numerator and denominator by possible number of samples i.e. 24

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

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 Probabil

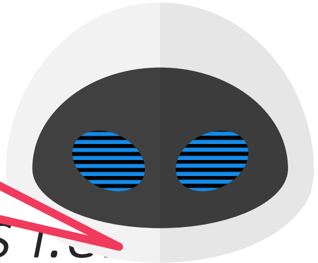
What if $\mathbb{P}[X = 2]$ happens to be 0? Won't we get a divide-by-zero error?

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

$\mathbb{P}[Y = 1$ number of samples with $Y=1$ and $X=2$

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



Dividing numerator and denominator by possible number of samples i.e.

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

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

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

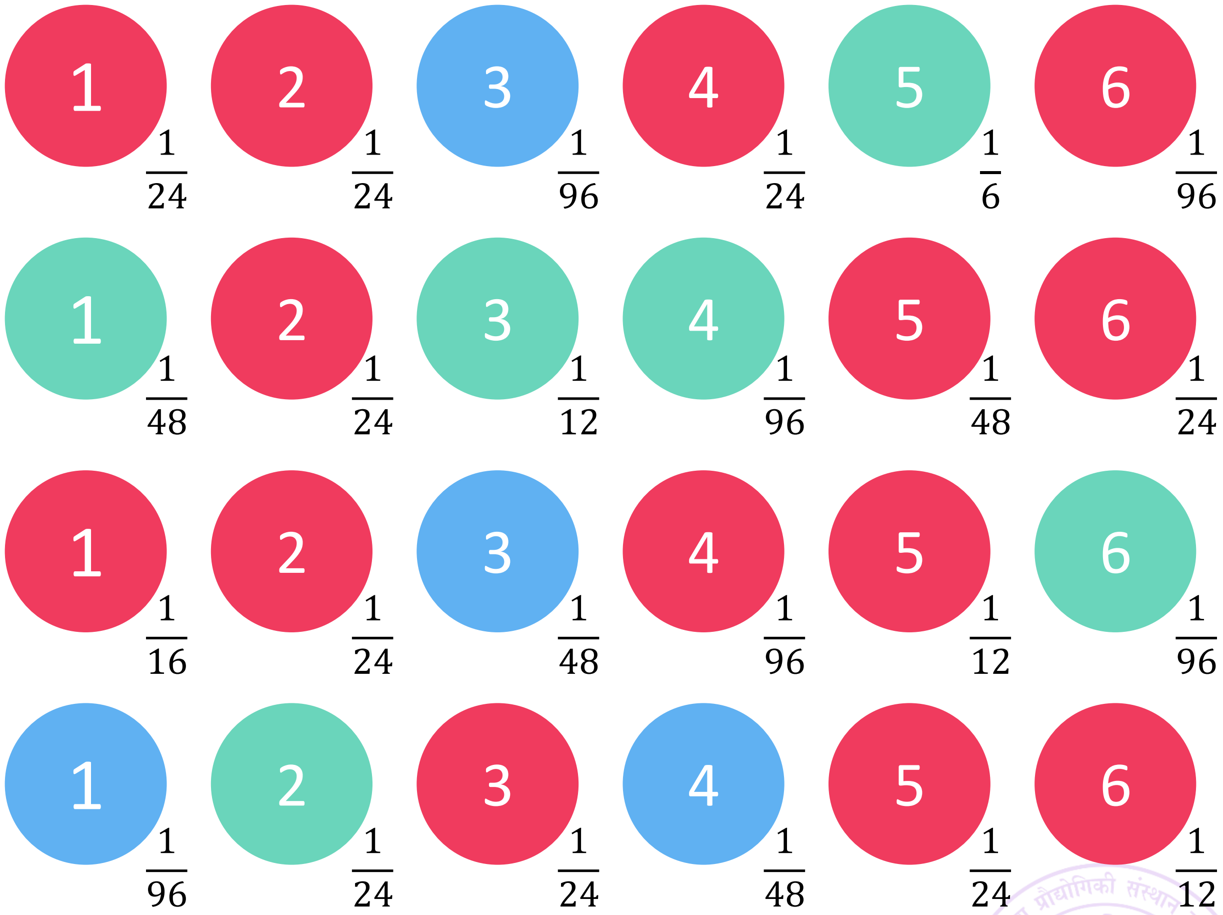
$$= \frac{\frac{1}{24}+\frac{1}{24}+\frac{1}{24}}{\frac{1}{24}+\frac{1}{24}+\frac{1}{24}+\frac{1}{24}} = \frac{3}{4}$$

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

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

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

$$= 0$$



A PMF for the Conditional Distribution?

Conditional probability values also form a distribution

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

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

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

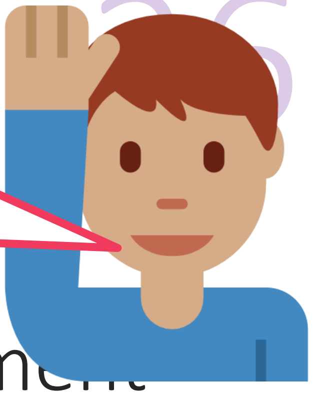
We can similarly define $\mathbb{P}[X = x | Y = y_0]$ as well

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

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

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 $Y(\omega_0)$



Cor

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

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

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

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????

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 $\mathbb{P}[X = x | Z = z_0]$

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

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????

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

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

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

or more explicitly, $\mathbb{P}[X = x] = \sum_{y \in S_Y} \mathbb{P}[X = x \wedge Y = y]$

Product Rule (Conditioning Rule)

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

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

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)

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

Rules of Probability

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

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

or more explicit

Product Rule (C

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

Combine to get

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

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

The foundation of Bayesian Machine Learning

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

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

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* $S_P = [10] \cup \{-1\} = \{-1,1,2,\ldots,10\}$

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

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