

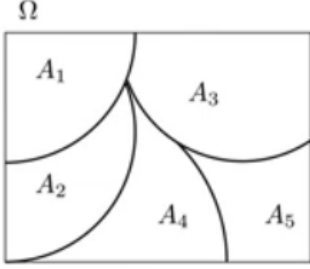
Lecture 16.0

Recap of Probability

Theory

So, we'll start with a quick recap of probability theory. The assignment is also designed to kind of make you just go back and read about these things which you would have done at some point in your life and I'll just quickly go over the first module or the zeroth module, which is a recap of probability theory.

Refer Slide Time: (0: 35)



Axioms of Probability

- For any event A ,

$$P(A) \geq 0$$
- If $A_1, A_2, A_3, \dots, A_n$ are disjoint events (i.e., $A_i \cap A_j = \phi \quad \forall i \neq j$) then

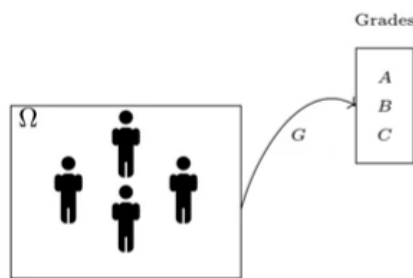
$$P(\cup A_i) = \sum_i P(A_i)$$
- If Ω is the universal set containing all events then

$$P(\Omega) = 1$$

That's, why I said embarrassingly back rate, so axioms of probability, so for any event we know that probably of the event should be greater than equal to zero and, if you have the Universal set which contains all the events in your, all the events, then the probability of the universal set is going to be one. These are the basic axioms of probability.

Refer Slide Time: (0: 52)

Random Variable (intuition)



- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a *random variable* G which each student to one of the 3 possible values
- And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
- Of course, both interpretations are conceptually equivalent



Now, random variables so, so here's the intuition behind random variables, I suppose a student can get one of three possible grades which is A,B,C one way of looking at it is that of all the possible events there are these three events, that the student gets a grade A, all the student gets a grade B, and the student gets a grade C and, there would be students in each of these events and you're trying to find the probability of this even, the other way of looking at it is that you have, this set of students and you have a random variable, which unfortunately is not a variable it's a function actually, which maps each of these students from your set to a particular value weight so, that's what a random variable is? The random variable is actually a function, which Maps your outcomes to your values .Right? So, from for each of these students we have a function which connects them to one of these three possible grades so, that's another way of looking at it so, one way was to think of these grades themselves as even, the other way is to think that you have a set which has a lot of outcomes and for each of these elements of the set you can map them to some value which is a green .Okay? So, we will see why this is the more better way of doing it so, irrespective of the first fee or the second view, everything remains the same the answers that you are going to get if I ask you what is the probability of the grade being a certain value at grade being A or, B or, C whether you take the first week or the second view, the answer is going to remain the same that doesn't matter but, why do we focus on random variables other than the first view is that

Refer Slide Time: (2: 26)

Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - *grade, height, age, etc.*
- We could have one random variable corresponding to each attribute
- And then ask for outcomes (of students) where $Grade = g$, $Height = h$, $Age = a$ and so on

You might be interested in several things about a student. Right? You might be interested in what are the heights, of different students how many of them are short, how many of them are tall and, so, on how many are adults young and, so on it have various things about a student that you could add each of these random variables actually operates on the same set and maps them to different values so, this view is more modular or, more reusable in that sense .Right? You have this set of possible outcomes and for each of them you are trying to map them to certain values and these values could be different it could be grades, height, age, whatnot. Right? Everything could be possible. Right? So, you could have a random variable for each of these quantities that you are interested and then you could ask questions .Right? Give me all the outcomes for which the grade is a certain value, the height is a certain value and the age is a certain value .Right?

Refer Slide Time: (3:15)

Random Variable (formal)

- A random variable is a **function** which maps each outcome in Ω to a value
- In the previous example, G (or f_{grade}) maps each student in Ω to a value: A , B or C
- The event $Grade = A$ is a shorthand for the event $\{\omega \in \Omega : f_{Grade} = A\}$

so, the more formal definition is a random variable is a function which Maps each outcome in your Universal set, to a value and the previous example the F grade, which is in shorthand represented as the random variable capital G, is the random variable or the function, which Maps each student to one of these three possible grades a B and C .Right? So, remember random variable is a function it's not a variable I don't know, why it is called a variable but it is called a believer. Okay? And then you could have a random variable, which maps at two ages and, a random variable which maps it to Heights and so on .Right? And the event grade is equal to e is actually a shorthand for the following even, give me all those outcomes from my Universal set for which when I apply the function to this outcomes the answer should be grade A .Right? So, when I say I want the probability of grade equal to A, this is what I actually mean, or if I ask for the set grade equal to A, this is the set that I am looking at maybe one is fine with this. Okay? So, all of you should be comfortable with this definition of random variable this is not my definition just the generic definition address there. Okay?

Refer Slide Time: (4:24)

Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, *weight, height*)
- Or discrete values (for example, *grade, nationality*)
- For this discussion we will mainly focus on discrete random variables

Now, random variable can either be continuous or discrete .Right? So, discrete is the example of grades, where you have grades A, B, C, D and so on, while it's a continuous random variable, height, weight and so on it which can take on any real value it's not discrete. Okay? For this discussion and for the rest of the discussion on this remaining 30%, of the course we'll be focusing only on discrete random variables unless, otherwise mentioned I don't think I'll ever look at continuous random variables you'll only focus on discrete random variables .Right? Okay? So, now that's what a random variable is now that we understand random variables, we can talk about different things related to random variables.

Refer Slide Time: (5:07)

Marginal Distribution

G	$P(G = g)$
A	0.1
B	0.2
C	0.7

- What do we mean by *marginal distribution* over a random variable ?
- Consider our random variable G for grades
- Specifying the marginal distribution over G means specifying

$$P(G = g) \quad \forall g \in A, B, C$$

- We denote this marginal distribution compactly by $P(G)$

The first thing that we can talk about is, marginal distribution so what do we mean by a marginal distribution of a random variable? So, if I ask you, give me a distribution for the grade, the random variable grade, what will you actually give me? What are the marginal distributions in the discrete case actually mean? if I ask you the marginal distribution of a random variable what do you need to actually give me probability of each setting of the random variable .Right? So, for if the random variable can take values A, B, C suppose, the grid can take values A,B,C then you need to give me the table that you see on the whatever side it is the table .Right? The only table which is there. Okay? And we denote this marginal distribution compactly as P of G, so when I say P of G, I actually mean this entire vector or this entire table which is P of G, is equal to A, P of G, is equal to B and ,P of G is equal to C and, so on that's what a marginal distribution, is specifying all the values that the random variable can take probability for all the random values that a random variable I know, this is very elementary but, it's very important for understanding how, many number of parameters do you need to learn in the particular joint distribution or, modular distribution and so on .Right?

Refer Slide Time: (6:26)

Joint Distribution

G	I	$P(G = g, I = i)$
A	High	0.3
A	Low	0.1
B	High	0.15
B	Low	0.15
C	High	0.1
C	Low	0.2

- Consider two random variable G (grade) and I (intelligence $\in \{\mathbf{H}igh, \mathbf{L}ow\}$)
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

$$P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}$$

- We denote this joint distribution compactly by $P(G, I)$

$$P(G | I)$$

3 2



Now what's a Joint Distribution suppose in addition to grade which can take on values A, B, C you also have this random variable intelligence which unfortunately can take only two values in our world which is high or low? Okay? what is a Joint Distribution of our grade and intelligence it's specifying every, is specifying a probability for, every combination of the grade and, so you have this cross product there are three possible values for grades and two possible values for intelligence, for each of these six values, you are going to specify of probably write it so, this table that you see is the Joint Distribution .Right? So, remember that we're always used to saying that Joint Distribution is P of G comma I .Right? But, that means that you have P of G comma I, for every value of G and every value of I that's what you need to specify now again I am repeating this because when I asked you to give me a joint distribution or, learn joint distribution from a data, from a given set of training data, this table is what I expect, I expect you to give me values for all possible combinations of the input variables or the input random variable that's why this is important. Okay? Now what's a conditional distribution so, if I ask you this is what we typically write it, I want P of G given I, what does that mean? How, many values do I need to give you? And again assume that G can take three values and, I can take two values .Right? So, if I ask you that give me this conditional distribution how many values do I need to give you? Six values, it's the same as the Joint Distribution what will I have to give you?

Refer Slide Time: (8:09)

G	$P(G I = H)$
A	0.6
B	0.3
C	0.1

Conditional Distribution

- Consider two random variable G (grade) and I (intelligence)
- Suppose we are given the value of I (say, $I = H$) then the conditional distribution $P(G|I)$ is defined as

$$P(G = g|I = H) = \frac{P(G = g, I = H)}{P(I = H)} \forall g \in \{A, B, C\}$$

- More compactly defined as

$$P(G|I) = \frac{P(G, I)}{P(I)}$$

or $\underbrace{P(G, I)}_{\text{joint}} = \underbrace{P(G|I)}_{\text{conditional}} * \underbrace{P(I)}_{\text{marginal}}$

So, I'll have to give you these tables, I'll assume that I is equal to I, given that I is equal to I, what are the different properties for P of G, equal to A, B and C and, the other table is given I equal to low, what are the priorities for A G equal to, A, B and C .Right? Okay? And there's some other simple step that this is how, you write the conditional distribution, is the joint distribution, over the marginal distribution .Right? So, this equation actually connects all the things that we have seen so, far the joint distribution, is the conditional distribution, into the marginal distribution, is that fine. Okay? .Right? So, you should be comfortable with if I ask you give me a joint distribution, if I tell you how many values my random variables, can tell me you, can take you should be able to tell me how many parameters I need to specify that distribution that's what this, a basic material is meant to stimulate you to do. Okay?

Refer Slide Time: (9: 04)

Joint Distribution (n random variables)

X_1	...	X_n	$P(X_1, X_2, \dots, X_n)$
...
...
...

$$\sum = 1$$

- The joint distribution of n random variables assigns probabilities to all events involving the n random variables,
- In other words it assigns

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

for all possible values that variable X_i can take

- If each random variable X_i can take two values then the joint distribution will assign probabilities to the 2^n possible events

Fine and what's, the joint distribution of n random variables the table on the next extra table never on the first in all cases then tables should never be on the first what's the joint distribution for n random variables, how many values do I need to give you? If each of these random variables can take K values, how many values will join distribution of K power n . Right? So, far and that's you're used to this because, you have done a lot of logic. Right? where you assume Boolean, variables and for all combinations you try to, write down some truth table and solve it so, it's very similar to that so in other words that assigns P of X_1 , equal to, X_1 , X_2 equal to X_2 , for all possible values that the variable X_i can take. Okay? And if each random variable can take two values you'll have to raise to n by entries in the joint distribution. Okay?

Refer Slide Time: (9: 57)

Joint Distribution (n random variables)

X_1	...	X_n	$P(X_1, X_2, \dots, X_n)$
...
...
...

- The joint distribution over two random variables X_1 and X_2 can be written as,

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

$$\begin{aligned} P(X_1, X_2, \dots, X_n) &= P(X_2, \dots, X_n|X_1)P(X_1) \\ &= P(X_3, \dots, X_n|X_1, X_2)P(X_2|X_1)P(X_1) \\ &= P(X_4, \dots, X_n|X_1, X_2, X_3)P(X_3|X_2, X_1) \\ &\quad P(X_2|X_1)P(X_1) \end{aligned}$$

$$= P(X_1) \prod_{i=2}^n P(X_i|X_1^{i-1}) \quad (\text{chain rule})$$

And the other thing is, just as for two random variables, you could write the joint distribution as a product of a conditional and a marginal, how do you write the joint distribution of n random variables so, I am going to start using some terminology the Joint Distribution of two random variables factorizes as a conditional distribution and, a marginal distribution, what about the Joint Distribution of n random variables? What's the one rule which has stayed with us so, far and once continue to go

into chain rule .Right? So, again we'll have the chain rule here so we have you can assume, that all of these variables are clubbed together so given X_1 and, then probability of X_1 , that's the same as this form .Right? And then just keep doing this recursively, till you get the following the Hyatt variable, depends on all the $I - 1$ variables, before that and you'd have a product of these all .Right? Fine this is known as the chain rule and, you can clearly see that this is just a special case of this form it so, just be very comfortable with the chain rule, this is going to be very important, when you are talking about various things it died directed graphical models, or undirected graphical models, or whatnot .Right? So, it's very essential that you completely understand the chain rule and maybe I'll, get back to later .Okay?

Refer Slide Time: (11: 14)

From Joint Distributions to Marginal Distributions

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

- Suppose we are given a joint distribution over two random variables A, B
- The marginal distributions of A and B can be computed as

$$P(A = a) = \sum_{\forall b} P(A = a, B = b)$$

$$P(B = b) = \sum_{\forall a} P(A = a, B = b)$$

- More compactly written as

$$P(A) = \sum_B P(A, B)$$

$$P(B) = \sum_A P(A, B)$$

So, now from joint distributions to, marginal distributions, suppose I'm given the joint distribution, over two random variables A and, B .Okay? So, the first table that you see here, what kind of a distribution is it? Joint conditional marginal, joint distribution, now from here, I want to find the conditional distribution for A and, B what does that actually mean what am I given? And what am I asking for P of A , P of B , so how do I get the marginal distribution, from the joint distribution sum over what .Okay? Fine so, now first of all if I have to give you the marginal distribution of A , how many values do I need to give you two values that I'm assuming that all my random variables are binary so two values so, from the joint distribution how will I get these two values I'll sum up with two rows, I'll keep the value of a same and sum over the B values and, same for the other great this is again straightforward all of you know? That but just be comfortable with this that you can obtain the marginal distribution, from the joint distribution by, summing over the variables which are not of interest .Right? So, when you want P of A , you will sum over the B 's when you want P of B will sum over these .Okay? So, this is and in general now if I give you a joint distribution of .Okay? This is more compactly .Right? So, this is like for all possible values that B , can take you were going to sum this but compactly this is how we write .Right? We always ignore the value assignment and we just talk about P of a comma B .Okay?

Refer Slide Time: (12: 47)

What if there are n random variables ?

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

- Suppose we are given a joint distribution over n random variables X_1, X_2, \dots, X_n
- The marginal distributions over X_1 can be computed as

$$P(X_1 = x_1) = \sum_{\forall x_2, x_3, \dots, x_n} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

- More compactly written as

$$P(X_1) = \sum_{X_2, X_3, \dots, X_n} P(X_1, X_2, \dots, X_n)$$

Now from here, if you are given n random variables how, are you going to find the marginal distribution from this joint distribution sum over all other variables .Right? So, do you see a problem with the summation you do see a problem with this summation .Right? There's a problem with the basic joint distribution itself, we'll come back to it but we'll focus on these things but if you just kind of vaguely appreciate at this point it's, fine we'll come back to it in a few more slides .Okay? So, even if you are given n random variables and a joint distribution, you can get the marginal distribution, for each of these n random variables by, summing over all those other variables that you don't care about .Okay? Fine and again this is more compactly written as this

Refer Slide Time: (13: 31)

Conditional Independence

- Recall that by Chain Rule of Probability

$$P(X, Y) = P(X)P(Y|X) \quad \leftarrow \quad \rightarrow \quad P(X|Y) = P(X)$$

- However, if X and Y are independent, then

$$P(X, Y) = P(X)P(Y)$$

- Two random variables X and Y are said to be independent if

- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X
- We would expect **Grade** to be dependent on **Intelligence** but independent of **Weight**

what is conditional independence when do I say, that a variable X is independent, of the variable Y in terms of probability what's the equation that you write P of x given by, is equal to P of X , knowing the value of Y does not change your belief about X , that's the English way of saying it .Right? and we denote this as X independent of Y so, just this is a standard notation again and we would expect the grade, to be dependent on intelligence but perhaps not dependent on weight or height or something

this is probably not any connection between them .Okay? And recall that by the chain rule for two variables, we have $P(X, Y)$ is equal to $P(X)$ into $P(Y \text{ given } X)$ so, what will this simplify to so, combination of the chain rule and the independence definition gives you this form for the Joint Distribution of two variables if the variables are independent .Okay? Fine so, that's all the basic stuff from probability that we need, I would encourage you to go back and just be comfortable with all of this and with this. And with this we can now start discussing about Directed Graphical Models.