

Data Mining
Prof. Pabitra Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 16
Bayes Classifier-V

We quickly recapitulate our discussion on the naive Bayes classifier.

(Refer Slide Time: 00:33)

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
tomato	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

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

$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$

$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$

$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$

$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$

$P(A|M)P(M) > P(A|N)P(N)$
 \Rightarrow Mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

What we did in the naive Bayes classifier was the following. We made it made the independence assumption; that means, if we have 2 variables A and B if they are independent, the joint probability is product of that individual probabilities.

So, as you know that if you have 2 both A and B, we call it a joint probability and if the individual A and B we call them marginal probabilities the marginals. So, for example, in this particular case with if this 4 attributes give birth can fly live in water have legs are independent, we have probability this give birth equal to yes.

(Refer Slide Time: 01:24)

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	no	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

$P(B=Y, F=N, W=Y, L=N | M)$
 $= P(B=Y|M) \times P(F=N|M) \times P(W=Y|M) \times P(L=N|M)$

$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$

$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$

$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$

$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$

$P(A|M)P(M) > P(A|N)P(N)$
 \Rightarrow Mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Let me I will write here B equal to y. And can fly equal to no. And so, this comma is an and actually live in water w equal to yes, and have legs L equal to no. This probability given say mammal if these attributes; that means, whether an species gives birth to a young. Does not depend on whether it can fly.

Similarly, it does not depend on whether it lives in water or have legs. So, if the attributes are independent of one another; that means, what value one takes no way depends on what value the other variable other attribute takes. Then we can write this joint probability as probability give birth yes fly no, what are yes legs no, equal to product of their marginal probabilities.

That means, probability birth equal to yes given mammal into probability fly equal to no given mammal probability water equal to yes given mammal into probability legs equal to no given mammal.

Note that we can write this only when this attributes are independent, but if we can write it this way, then definitely there is an advantage. So, what is the advantage. The problem we faced in the in finding out the probability of the joint 4 attributes taking on certain combination of values is that there are only few examples out of all these 7 mammals who have exactly this 4 combination.

So, we have a kind of absence of training set. We have not seen what happens these. We have not encountered before what happens in this case, but each of these individual cases that is it is mammal and give birth equal to yes. It is mammal and fly equal to no, each of these individual cases they are not so rare, they are not so rare.

So, if you find out actually you can see that out of 7 mammals 6 give birth. Similarly, 6 cannot fly 2 lives in water. So, in other words, I get a better estimate of the joint probability by expressing it as the product of the marginal probabilities. Which I can again I remind again that which I can only do if they are independent.

And once you estimate this rest of the thing is same as that of the map classifier. Check compare the 2 classes whichever is higher put it that class. In fact, in practice we to in extreme case what happen can happen is that still, you see the one thing about this it is that I am this probability as a product of 4 probabilities.

So, even if one of these 4 probabilities goes to becomes turns out to be 0, entire thing becomes 0, it is a product. So, it may happen that one of the probability is small, but I want this small thing to be represented not by 0, but by a small epsilon value, small value. So, that is the idea of smoothing (Refer Time: 06:25) idea of smoothing.

(Refer Slide Time: 06:30)

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

Original : $P(A_i | C) = \frac{N_{ic}}{N_c}$

Laplace : $P(A_i | C) = \frac{N_{ic} + 1}{N_c + c}$

m - estimate : $P(A_i | C) = \frac{N_{ic} + mp}{N_c + m}$

Handwritten notes:
 Email Classification Problem
 An email is represented by the words it contains. Attributes: w_1, w_2, \dots, w_m . Classes: Spam/non-spam.
 c: number of classes
 p: prior probability
 m: parameter

So, if my actual estimate is this, and if this is 0, I do not make it 0 I add a small term one. And divide it by C. So, this entire that thing will be small, but not 0. So, it will be small,

but not 0. And the it will not collapse by after product entire thing. So, in practical application you have to often use this.

So, I will all this looks very fine this is a very attractive. In fact, this is one of the commonly used also. So, I will maybe give you an popular example. So, it is like this. I have I by the words it contains. So, these are the so-called attributes.

So, basically, I will take all the words w_1 that English has w_n . And an email let me call it e one is represented as. So, that this word is present I write one other is 0. So, this binary vector is a representation of an email.

And my talks is from this binary representation, I have to classify emails into 2 classes spare or non-spare. So, how I do it is that I collect lot of spam mail. I apply just like I did in the previous example count how many times w_1 is yes, count how many times w_2 is yes. Multiply them using the Bayes assumption. And get the total probability it is a spam, on a and similarly the total probability. It is a non-spam, and whichever is higher I put it into that class.

The underlying assumption is the probability of what is appearing is independent of each other. So, maybe it will happen after I do this maybe the spam emails, will have a higher probability of having words like lottery and dollar and so on. And maybe non-spam will have a higher probability of other words say meeting and so on. It is pretty successful also.

(Refer Slide Time: 10:51)

Conditional Independence

- Event A and B are **conditionally independent given C** in case
$$\Pr(AB|C) = \Pr(A|C)\Pr(B|C)$$

Handwritten: $P(AB) = P(A)P(B) = 0.5 \times 0.5$
- A set of events $\{A_i\}$ is conditionally independent given C in case
$$\Pr(\bigcup_i A_i | C) = \prod_i \Pr(A_i | C)$$

Handwritten: $P(A|B) = P(A)$

IT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

All right, but there is no guarantee that this independence holds; that means, there is no guarantee that p of A B 2 attributes equal to, but maybe there is a less strict condition; condition known as conditional independent independence. So, independence is; this C is not there.

So, conditional independence says that individually you cannot write this, but I can write this once I have another variable C and I have observe the value of that variable. So, in other words I can write probability of A B given C is probability of A given C into probability of B given C . So, in more general terms, if you have I variables A_1 to A_I probability of that union given C is the product probability of their marginals given C . So, this is less strict.

So, you can think of it is like this. Suppose, I am tossing 2 coins unbiased coins. And just 2 outcomes either one or 0 head or tail. And if the A is probability the first coin is coin is tossed gives head B is the probability that second coin toss gives head and other combinations.

So, you see they are independent because what is the probability together both will be head I can write as the probability of one is head and the other is head. So, it will be if they are unbiased it will be 0.5 into 0.5 there is actually another way of writing. This this actually you can write it as probability if A is independent of B probability of A given B

is just probability of a using the definition of probability you can derive it by the way. So, this cross is not there. So, if they are independent they are equal actually.

One way of looking at this is that B gives no information about A knowing what the value of B is does not help me predict what A is. When I toss 2 coins that is the case what is the outcome of one coin does not help me predict the outcome of the other coin. But if there is a third variable which helps then I call it as conditionally. Independent let me take this example tell you.

(Refer Slide Time: 15:15)

Example

$P(A|B|C) = P(A)P(B|C)$
CI

Let the two events be the probabilities of persons A and B getting home in time for dinner, and the third event is the fact that a snow storm hit the city. While both A and B have a lower probability of getting home in time for dinner, the lower probabilities will still be independent of each other. That is, the knowledge that A is late does not tell you whether B will be late. (They may be living in different neighborhoods, traveling different distances, and using different modes of transportation.) However, if you have information that they live in the same neighborhood, use the same transportation, and work at the same place, then the two events are NOT conditionally independent.

$P(A|B|C) \neq P(A)P(B|C)$
Not CI

NPTEL ONLINE CERTIFICATION COURSES
IIT KHARAGPUR

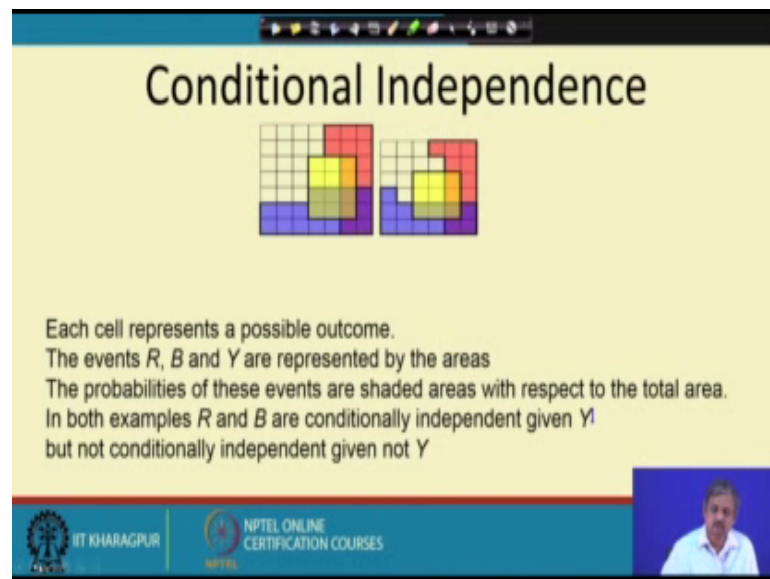
Let there be 2 persons A and B. And the events be they are both of A one of them comes from A comes home for dinner. And becomes for term dinner and the third event is that C is that a snow storm hits the city. So, in this case both have low probabilities of getting home, but they are still independent. So, whether a is late does not tell you whether B is late.

So, if they are say to quite different they are not related in any way. So, they live far away from each other then they are independent even if there is no or no storm, in from where the outcome of C does not affect. So, one does not a outcome of a whether a has come home early does not help me predict whether B will come home.

So, in this case they are independent. So, I write I can write probability. For the first case I can write that. But suppose they we know that they stay in the same neighbourhood and

take the same transport. Then and I can use the information about A to predict B. If I know a is late I know that well the B will also be late because they use the same transport. So, they are no longer independent. So, in the second case, first case is conditionally independent C i the second case is not C i pictorially this will look something like this.

(Refer Slide Time : 18:04)



So, I have 3 events red, blue and yellow. And each cell is a possibility. So, all these whenever the blue event happens all these blue cells get a colour, and for red and for yellow. So, in this case you see if y is given R and B are independent if I know y if I know y has happened I can tell something about relative probabilities of A and B R and B in this case, alright.

So, this is an example so, this conditional independence that this fact that probability let me write as $x \perp y$; conditional independence is written by this notation, that x some perpendicular y given z .

(Refer Slide Time: 19:04)

CI: Conditional Independence

- Variables are rarely independent but we can still leverage local structural properties like CI.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
 - The following should hold for all x, y, z
 - $P(X=x \mid Z=z, Y=y) = P(X=x \mid Z=z)$
 - $P(Y=y \mid Z=z, X=x) = P(Y=y \mid Z=z)$
 - $P(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z) P(Y=y \mid Z=z)$

CI
 $P(X \perp Y \mid Z) = P(X \mid Z) P(Y \mid Z)$
 $X \perp Y \mid Z - CI$
 $X \perp Y$ Independence

We call these factors : very useful concept !!

So, this is the symbol for independence. So, plane independence is x independent of y this is just independence x, y, z are random variables. And this is conditional independence.

So, if the conditional independence holds, then I can write the following 3 thing. I can write this x equal to x means x at taking on certain value. I can symmetrically write this. So, x independent of y means y is also independent of x. And of course, the definition I can write this.

(Refer Slide Time : 20:48)

Exercise: Conditional independence

p(smarter ^ study ^ prep)	smarter		-smarter	
	study	-study	study	-study
prepared	.432	.16	.084	.008
-prepared	.048	.16	.036	.072

$P(P_r, S_m, S_t)$

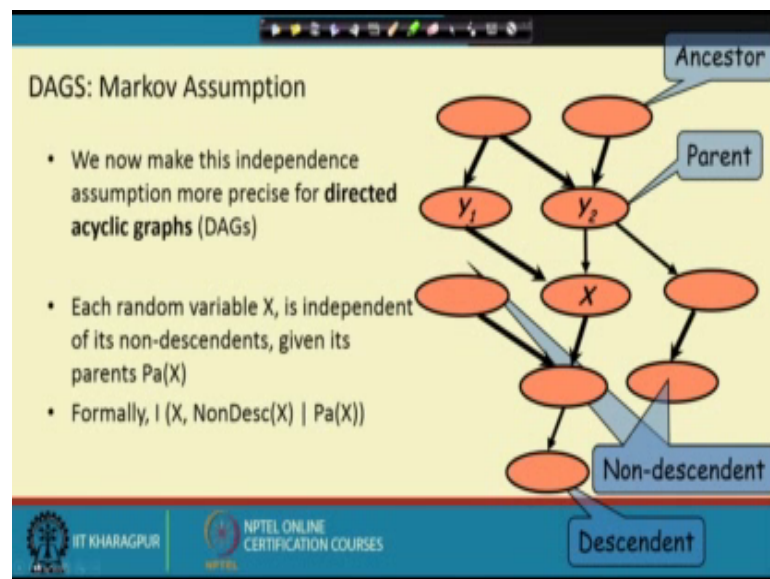
- Queries:
 - Is smarter conditionally independent of prepared, given study?
 - Is study conditionally independent of prepared, given smarter?

So, I give you a small exercise. So, I have 3 variables. A student is smart. A student studied, and a student is prepared for an exam. And these are the probability values that means, if a student is prepared and is smart and studied. So, this is actually prepared. And studied and smart that probability is 0.432 and this is prepared and not studied, but smart is 0.16.

So, what we are supposed to do is to figure out 2 things. First thing find out from this margin. Also, I am actually looking at (Refer Time: 22:05) prepared smart study. What are the marginals, which things had up to 1. What are the conditionals that can be there? And what are the conditional independencies? That hold, please do this exercise you will get a better idea. I will put up the solution. All of you please note it down. For the moment right, it done I will put it up in the exercise set also note that it is 3 variables.

So, all combinations 8 combinations it turns out that this kind of conditional. So, this is one example in general there will be many.

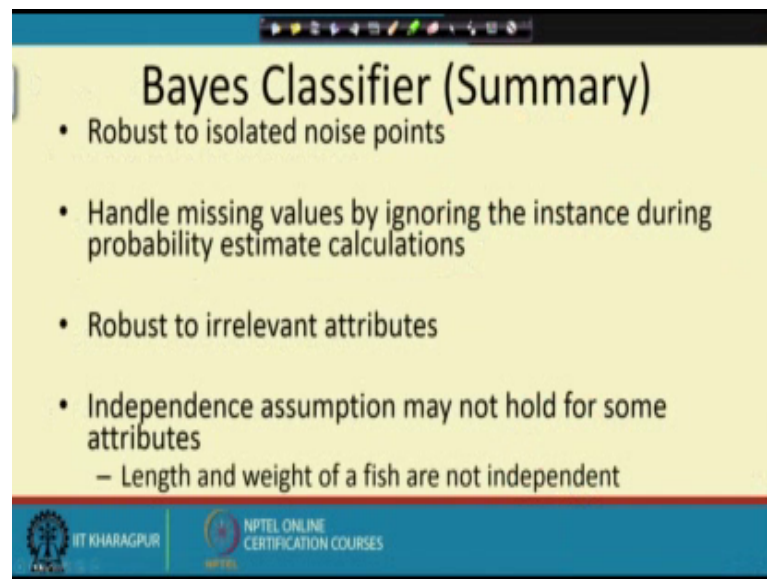
(Refer Slide Time: 23:27)



You can capture using something called a directed acyclic graph, alright. So, every every node is a variable, and the following holds and the following holds. So, I hope you understand what is a de parent ancestor, and non-descendant in a graph. So, this is a directed graph there is a directional arrow edge. And every attribute or variable is a node of the graph.

So, I can represent will see in our next lecture is that I can represent conditional independencies, whatever are present in among the attributes by a directed acyclic graph like this. There is something called a markov assumption which tells that it is it is memory less it forgets other variables. So, I can capture using this and have a more complex type of Bayes classifier.

(Refer Slide Time: 25:17)



Bayes Classifier (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
 - Length and weight of a fish are not independent

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in summary, this is what I have. I have some robust probabilistic classifier. It can handle irrelevant attributes as well as missing values. You can have the independence assumption makes life easy naive Bayes, or you can have a directed acyclic graph capturing conditional independencies. So, this is the summary. In the this is the basic introduction to Bayes classifier, in our next class we will move on to the conditional independencies and Bayesian networks. Not just classifiers.

Thank you for today.