

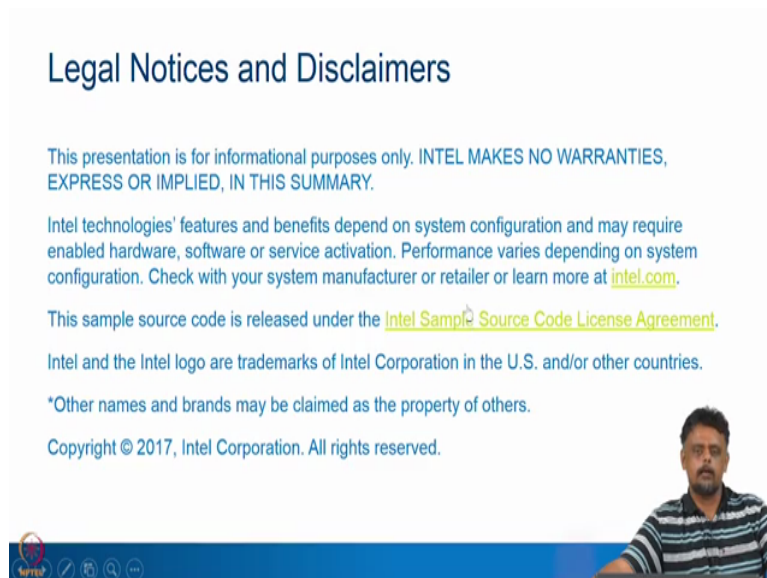
Machine Learning for Engineering and Science Applications
Professor Dr. Ganapathy Krishnamurthi
Department of Engineering Design
Indian Institute of Technology, Madras
Naïve Bayes

(Refer Slide Time: 00:14)



Hello and welcome back in this video we will look at the naive based classifier which is another surprised learning paradigm

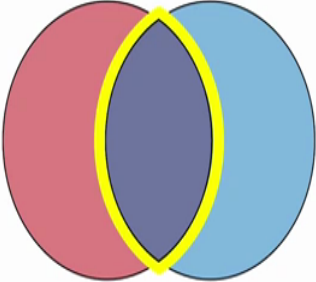
(Refer Slide Time: 00:22)




So all the slides are courtesy of Intel software and we would be using the have such.

(Refer Slide Time: 00:27)

Probability Basics



- Single event probability:
 $P(X), P(Y)$
- Joint event probability:
 $P(X, Y)$
- Conditional probability:
 $P(X|Y), P(Y|X)$
- Joint and conditional relationship:
$$P(X, Y) = P(Y|X) * P(X) = P(X|Y) * P(Y)$$

MPTEL 

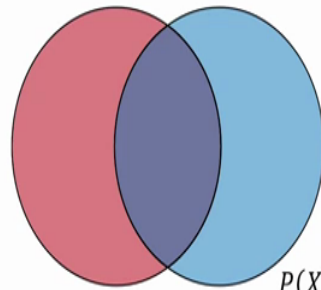
So we will start off with some probability basics some of these would have been covered by in some of previous videos but just for the sake of some continuity, so we use Venn diagrams to denote the space of events X , so the probability of event X is given by P of X which is the circle highlighted in yellow and we have another event Y which is again given by the Venn diagram circle right here, here and again highlighted in yellow.

So we have two events X and Y the joint probability of their occurrence is denoted by P of X comma Y and of course the single even probabilities are P of X and P of Y respectively again here in the Venn diagram it is the region of intersection. The conditional probability again is given P of X given Y is basically the region here if you go back and we see that it is the parts of X that are also in Y .

So that denotes the conditional probability P of X given Y what is the probability that event X occurs given that Y is occurred? Similarly we can define conditional probability P of Y given X ok. So the joint and conditional probabilities are related to each other so the joint probability P of X comma Y is the conditional probability P of Y given X times P of X or equivalently P of X given Y times P of Y , this relationship you should have seen before.

(Refer Slide Time: 02:07)

Bayes Theorem Derivation



- Use conditional and joint relationship:

$$P(Y|X) * P(X) = P(X|Y) * P(Y)$$

- To invert conditional probability:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$P(X) = \sum_Z P(X, Z) = \sum_Z P(X|Z) * P(Z)$$



So given this rule we can then equate P of Y given X times P of X is the same as P of X given Y times P of Y , so that is the joint conditional (prob) conditional and how the conditional and joint probabilities are related, ok. So if you invert the probability so if you can make use of this relationship right here and if you can bring the P of X to the numerator on the denominator on the right hand side, so we get this relation is P of Y given X the conditional probability of Y given X in terms of conditional probability X given Y and the individual probabilities X and Y , ok.

Similarly then the denominator P of X can also be derived from by marginalizing the joint probability over some event Z ok, so P of X is submission over Z emission over the joint of the event set P of X joint probability of P of X comma Z or also be written as conditional probability P of X given a Z times P of Z and of course we marginalize over Z which is a submission over set, here we are to given choose Z such that they are (itse) we can choose that in terms of a mutually exclusive of possibilities.

So for instance if you are looking at treatment a diagnostic test we can say Z is the event that test is positive and test is negative ok, so two options in Z test is positive and test is negative and X is the basically the if you can think of X as the event that you have the disease ok. So that way you can make a individual choice of Z and you can marginalize over Z to get the probability P of X in the denominator, ok.

(Refer Slide Time: 03:50)

Bayes Theorem

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$



So this relationship that you have seen so far is the Bayes theorem which are you should have been introduced to where the conditional probability P of Y given X is written in terms of P of X given Y and the individual probabilities P of Y and X, ok. So this is this can be interpreted in the context of if you are you know surprise unsupervised learning or supervised learning this specifically in supervised learning this P of Y given X is referred to as a posterior probability P of x given Y is the likelihood P of Y is the prior and the denominator is known as the evidence, ok.

So typically this is the one that is a very difficult to calculate because we see that it involves submission over marginalizing over another variable that typically turns out to be an intractable calculation in most cases, ok. So there are ways of avoiding this and it is what we will see later on at least for the name Bayes we will see why we can avoid this, ok. So in the context of whatever we have seen so far I think I might have mentioned when we talked about probability distribution the idea of likelihood ok, so where we calculate likelihood of obtaining data that is exactly here right there ok and later on and we looked at when we get to look at map maximum if you are sure your Bayesian regression we will once again revisit these concepts, ok.

(Refer Slide Time: 05:14)

Naïve Bayes Classification

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)}$$

$$posterior = \frac{likelihood * prior}{evidence}$$



So like you mentioned earlier calculating this denominator is the difficult task and most of them they are not we try to not calculate it and in fact we will just ignore it as some constant in most problems and so this whatever you calculate as posterior it will not exactly be a probability you might have to normally it is not normalized, so this is the difficult thing to calculate and we try to get over this calculation in most problems that we encounter, ok.

(Refer Slide Time: 05:44)

Training Naïve Bayes

- For each class (C), calculate probability given features (X)

$$P(C|X) = P(X|C) * P(C)$$

Class Feature



So what we do now is we will in the context of I know learning supervised learning classification we will rewrite the Bayes rule and see how we can be used for classifying, so the naive Bayes classifier that is how it draws it is name from. So we replace Y with the class label C , ok (06:07) the class label C and of course this is again Bayes rule for that, for that

class of course like I mentioned earlier we have not calculating the evidence in this case because it will involve or sum which is difficult to calculate, ok.

So the P of Y is P of X is something we are not taking into account here we will see later because we see that what you are done here is we are trying to calculate P of C given X, so the probability of the class given X so the class is basically you know if you have a classification problem it will be multi class or binary class problems feature is basically a re input data points training data.

So the probability of class given X is basically your classification right that is we are trying to classify the classify based on input feature X, so what we are (cal) so this P of C given X so if you have three classes we can write 3 i, 3 such expressions so C 1, C 2 and C 3 and in each of them the denominator is going to be a same P of X the is going to be the same, so when you are comparing them this code the output P of C given X up to a multi same multiplicative constant we can compare them, so we do not actually have to explicitly calculate P of X right.

(Refer Slide Time: 07:21)

Training Naïve Bayes: The Naïve Assumption

- For each class (C), calculate probability given features (X)

$$P(C|X) = P(X|C) * P(C) / P(X)$$

↑
P(C, x)
P(C, x)

→ same for all classes

n-features
- Difficult to calculate joint probabilities produced by expanding for all features

$$P(C|X) = P(X_1, X_2, \dots, X_n|C) * P(C)$$



$$= P(X_1|X_2, \dots, X_n, C) * P(X_2, \dots, X_n|C) * P(C)$$

chain rule of prob.

P(C, X₂ | X₃, X_n, C)

P(X₃... X_n | C)

*P(A, B) = P(A|B) * P(B)*

So given this based on Bayes rule we have this expression P of C given X is P of X given C times P R C we are ignoring P of X because it is the same for all classes alright, so what the idea is we have to calculate, evaluate this for P of C 1, C 2 let us say a 2 class problem and we just compare these two numbers so of course since we are not normalized by P of X it will not exactly be a probability value we can call this as score, and we compare this course we get in whichever is the higher score we assign it to that class, the input data point to that class ok.

So how do we actually go about doing that right so let us just look at the calculation so we want to calculate $P(C \text{ given } X)$ ok, so X is not in this case it is not 1 variable one dimensional right, it is n dimensional, ok n dimensional so which means that there are n features ok. So we can explicitly write down this $P(X)$ this is what is written explicitly here $P(X_1 \text{ to } X_n \text{ and given } C)$ times $P(C)$ and then we apply Bayes the chain rule of probability to this particular expression $P(X_1 \text{ to } X_n \text{ given } C)$ which then we can write it in this form.

So $P(C)$ so if you if you look at it, it is $P(X_1 \text{ to } X_n)$ you have taken X_1 as I say it is individual variable and then group these as 1 variable $X_2 \text{ to } X_n$ is 1 variable so then we can again once again use the Bayes rule to rewrite this as $P(X_1 \text{ to } X_n \text{ given } C)$ sorry we can use the joint probability how we write the joint probability in terms of conditional probabilities you can use that rule to write down this $P(X_1 \text{ to } X_n \text{ given } C)$ as $P(X_1 \text{ given } X_2 \text{ to } X_n \text{ and } C)$ times $P(X_2 \text{ to } X_n \text{ given } C)$, ok.

So which is the way we write how we learn to write the joint probability in terms of the conditional probability if you recall I will use the A and B instead of X and Y just to we wrote $P(A \text{ given } B)$ as $P(A \text{ and } B \text{ given } P)$ sorry given P of A given B times $P(B \text{ given } P)$ that is exactly what we did here but then there is a conditional probability here which is conditioned on C the class and of course we have taken that it account when it wrote it up, ok.

So this so if we can do this successfully right so we have done this for X_1 now X_1 is here again we will take this and we can write it as $P(X_2 \text{ given } X_3 \text{ to } X_n \text{ comma } C)$ times $P(X_3 \text{ to } X_n \text{ given } C)$ we can write this expression like that right, so we can keep writing this way and decompose it so but the problem is doing this calculation it is hard because then we have to do so many terms if it 100 features you literally have to expand and then we have to calculate this probability conditioned on you know calculate the probability of one feature conditioned on the others all right.

(Refer Slide Time: 11:10)

Training Naïve Bayes: The Naïve Assumption

- For each class (C), calculate probability given features (X) $P(C|X) = P(X|C) * P(C)$
- Solution:** assume all features independent of each other $P(C|X) = P(X_1|C) * P(X_2|C) * P(X_n|C) * P(C)$
- This is the "naïve" assumption $P(C|X) = P(C) \prod_{i=1}^n P(X_i|C)$



Training Naïve Bayes: The Naïve Assumption

- For each class (C), calculate probability given features (X) $P(C|X) = P(X|C) * P(C) / p(x)$ *Same for all classes*
- Difficult to calculate joint probabilities produced by expanding for all features $P(C|X) = P(X_1, X_2, \dots, X_n|C) * P(C)$
 $= P(X_1|X_2, \dots, X_n, C) * P(X_2, \dots, X_n|C) * P(C)$
chain rule of prob. p(x₂ | x₁, x_n, c)
 $p(A, B) = p(A|B) * p(B)$ *n-features*



So this is hard so then what we do is to make this so called Naive Bayes approximation which says that all the features are independent of each other so if you go back so if you look at this let us just look at write down this term here $P(X_1|X_2 \dots X_n \text{ and } C)$, so then what happens is $P(X_1|X_2 \dots X_n \text{ and } C)$ ok since we assume that all features are independent of each other then this is the same as $P(X_1|C)$ ok because it does not depend on X_2 takes.

So this is the Naive Bayes approximation or the assumption if you can call it this is an assumption because you will have seen in real data features are typically not independent there is always some correlation or some kind of relationship between the features but then

we just ignore those and then we independently calculate these quantities ok to get to the probability of a class given a feature X ok.

So like so if you have a multi class problem we calculate this probability for each one of those classes and assign in to the largest class of the largest probability for a given new data point X, ok.



So this can be written in product form and like this it is probability of C of the class prior progress is called the prior probability of the class multiplied by i product of i equal to 1 to N P of X i given C ok. So this comes about by making the assumption that all the features are independent of each other so that way the dependencies of a one feature on the others goes away and that is why we are able to write it in this simplified form, ok.

(Refer Slide Time: 13:03)

Training Naïve Bayes

- For each class (C), calculate probability given features (X) $P(C|X) = P(X|C) * P(C)$
- Class assignment is selected based on *maximum a posteriori* (MAP) rule $\underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(X_i|C_k)$

Means select potential class with largest value



So as I mentioned earlier the way to do class assignment is what we call the map rule or the maximum a posterior rule since we calculate for every class k if there are capital K classes we calculate this probability of C 1 given X 2 probability of C k given X calculates ok numbers will be calculated and the data point will be assigned to the class with the highest probability, so that is the argmax there ok, this is (13:39) this is the (13:40) of the Naive Bayes classifier, ok so as my name select potential class with largest value, ok.

(Refer Slide Time: 13:47)

The Log Trick

- Multiplying many values together causes computational instability (underflows)

$$\underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(X_i | C_k)$$

- Work with log values and sum the results

$$\log(P(C_k)) + \sum_{i=1}^n \log(P(X_i | C_k))$$



It is no longer a probability because we have not normalized by P of X, so it is just some number it is called a score you have added the score and it is assigned to the class with the largest score. So multiplying so many numbers because in this case let us say we have a thousand features or hundred thousand features this can cause overflow problems under flow problems etcetera especially since you are multiplying with probabilities and cause underflow problems.

So then you just calculate the log of that so deciding in another score that you can calculate so the product gets if you take the log of this whole expression then the product gets converted to a to a sum here ok, the we can do that.

(Refer Slide Time: 14:28)

Example: Predicting Tennis With Naïve Bayes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



So let us see how we can look at we have looked at this in the I think we did the binary decision tree algorithm, so we look at how do predicting tennis with Naïve Bayes right remember we have from a bunch of days we have the following features Outlook, Temperature, Humidity and Wind and based on though outlook temperature humidity and wind we decide whether to play tennis or not, so it is a 0, 1 problem, ok.

So then let us see how we can use naive Bayes to deal with this, ok.

(Refer Slide Time: 15:01)

Example: Training Naïve Bayes Tennis Model

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	2/5
Weak	6/9	3/5

Create probability lookup tables based on training data

The Log Trick

- Multiplying many values together causes computational instability (underflows)

$$\underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(X_i | C_k)$$

- Work with log values and sum the results

$$\log(P(C_k)) + \sum_{i=1}^n \log(P(X_i | C_k))$$

So example how do you are trained the Naive Bayes classifier ok, so the probability of play equal to Yes is 9 over 14 we have 14 data points probability of play equal to No is 5 divided by 14 ok, so this is your P of C remember from previous slide this how we calculate the

probability of the class right and we calculate the probability of that particular feature given the class right.

So how do we do that so the outlook takes value sunny, overcast and rainy ok, so if you go back and look at the counts so out of the 9 times so when we say play is yes so given that the class is that we play we decide to play twice it is sunny, 4 times it is overcast and 3 times it is rainy, so that probability we can just calculate as a frequency right. Similarly out of the 5 times when we decide not to play when the class is 0, 3 times it is sunny, 0 times is overcast, 2 times it is rainy we will see how to deal with this 0 later but that is right for now, ok.

Similarly we can do that for temperature none for the 9 times we say we decide to play twice it is hot, 4 times it is mild and 3 times it is cool, so the probability is that, ok. So this is basically if you go back and show you the this, this we are calculating these numbers probability of X_i given C_k so what does it mean probability this we are trying to calculate probability of temperature equal to hot when play equal to yes that is this column.

Similarly this column is probability of temperature equal to hot when play equal to no, it is not exactly the whole column this is basically this particular row that is what we are calculating similarly the probability of temperature equal to mild when play equal to yes is what we calculate here and with probability of temperature equal to cool and play equal to yes is what we calculate here.

Similarly so that is the so here the class is play equal to (yes or no) yes or no corresponds to C equal to 1 play equal to no corresponds to C equal to 0, ok and these are the exercise that we are talking about, ok. So we can calculate these tables for the (train) based on the training data right, so we can talk about these conditional probability tables based on a training data. Similarly for each class classification C equal to 1, C equal to 0 we can look at humidity and say probability humidity equal to i given class C equal to 1 here so on and so forth, ok.

So one good exercise is to go back and see if we can do this calculation yourself ok, so that is the way to I do this calculation right. So similarly for wind, wind is strong or weak so probability of wind equal to strong given play equal to yes which means here again corresponds is equal to 1 and this corresponds to C equal to 0 ok these are the calculations that you have to do.

So what we do in Naïve Bayes model is we work with the given training data and estimate these conditional probabilities just by doing the frequency of occurrence ok, so that is the only calculation that we do for the naive based model.

(Refer Slide Time: 18:55)

Example: Predicting Tennis With Naïve Bayes

Predict outcome for the following:

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

$$P(\text{yes}|\text{sunny, cool, high, strong}) = P(\text{sunny}|\text{yes}) * P(\text{cool}|\text{yes}) * P(\text{high}|\text{yes}) * P(\text{strong}|\text{yes}) * P(\text{yes})$$

$$P(\text{no}|\text{sunny, cool, high, strong}) = P(\text{sunny}|\text{no}) * P(\text{cool}|\text{no}) * P(\text{high}|\text{no}) * P(\text{strong}|\text{no}) * P(\text{no})$$

$$P(C=x) = P(C) \prod_{i=1}^M P(C|X_i) \Rightarrow P(\text{yes})$$



Example: Training Naïve Bayes Tennis Model

$P(\text{Play}=\text{Yes}) = 9/14$

$P(\text{Play}=\text{No}) = 5/14$

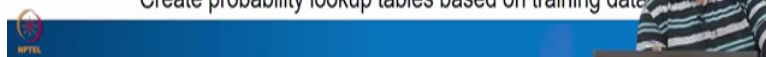
Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	2/5
Weak	6/9	3/5

Create probability lookup tables based on training data



So then what happens when new data point comes in, so new data point is a new set of features so we have probability or we have X equal to outlook we have to give outlook, temperature, humidity and wind and we have to for this X outlook happens to be sunny, temperature happens to be cool, humidity happens to be high and windy is equal to strong ok. So then the way to calculate this probability we want to calculate probability of Class C equal to 1, C equal to 1 given X sorry this is basically probability of C equal to 1 given X this is probability of C equal to 0 given X, ok.

So how do we do this we go back to the Naïve Bayes formulation, so if you remember correctly it is probability of C class of the class times the product of i equal to 1 to N probability of class given feature that is our formula this is the probability of C given X this is what we did. So the probability of C in this case probability of C equal to 1 for the first row, probability of yes which is we calculated that as 9 over 14 if you remember there are 9 data points which are.

So we here in this case when you do probabilities you just do the frequency, right relative frequencies so out of 14 data points 9 times you decide to play yes so probability of yes is 9 over 14 ok and for each one of them probability of sunny given that you play tennis probability of cool when you given that it play tennis, probability of high winds when you play tennis proper your strong wind and temperature is higher I think this is a temperature.

So we can look at all these numbers up from the tables that you calculated here ok and we just plug them back in there similarly we can do the same thing for probability of not playing tennis.

(Refer Slide Time: 20:29)

Example: Predicting Tennis With Naïve Bayes

Predict outcome for the following:

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

Feature	Play=Yes	Play=No
Outlook=Sunny	2/9	3/5
Temperature=Cool	3/9	1/5
Humidity=High	3/9	4/5
Wind=Strong	3/9	3/5
Overall Label	9/14	5/14
Probability	0.0053	0.0206



So for instance this is how the table outlines how we calculate it, so for feature outlook equal to sunny these are the conditional probabilities that you evaluate and if you calculate it ok this is the overall label this is these are the P of X i given C and this is P of C you take the product of all of them whichever comes up with the higher score here in this case is point 026 is higher than point 0053, so you just decide that you will play you will not play tennis , ok.

(Refer Slide Time: 21:32)

Example: Predicting Tennis With Naïve Bayes

Predict outcome for the following:

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

Feature	Play=Yes	Play=No
Outlook=Sunny	2/9	3/5
Temperature=Cool	3/9	1/5
Humidity=High	3/9	4/5
Wind=Strong	3/9	3/5
Overall Label	9/14	5/14
Probability	0.0053	0.0206



Example: Training Naïve Bayes Tennis Model

$P(\text{Play}=\text{Yes}) = 9/14$

$P(\text{Play}=\text{No}) = 5/14$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

Create probability lookup tables based on training data

Example: Predicting Tennis With Naïve Bayes

Predict outcome for the following:

$x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

$$P(\text{yes} | \text{sunny}, \text{cool}, \text{high}, \text{strong}) = P(\text{sunny} | \text{yes}) * P(\text{cool} | \text{yes}) * P(\text{high} | \text{yes}) * P(\text{strong} | \text{yes}) * P(\text{yes})$$

$$P(\text{no} | \text{sunny}, \text{cool}, \text{high}, \text{strong}) = P(\text{sunny} | \text{no}) * P(\text{cool} | \text{no}) * P(\text{high} | \text{no}) * P(\text{strong} | \text{no}) * P(\text{no})$$

$$P(x) = P(y) \prod_{i=1}^n P(x_i | y) \Rightarrow P(\text{yes}) = \frac{9}{14}$$

So we can we will address the problem of what happens when you have a category with zero in it right, so what does that mean? So for instance if we go back I think there was fun remember this, so what happens here so here probability of overcast given that play equal to no is 0 there are no incident there are no events there in your data point there are no you do not have any data point with outlook is overcast and when you decide not to play ok, you do not have that data point so then when you construct this empirical probability you get 0.

So what is the problem with that let us say instead of Outlook being sunny I put Outlook is overcast especially here I put overcast then I will be multiplying with the 0 one of these numbers is a 0 right instead of 1 this one will become a 0, so then we get up 0 probability so that does notices make sense, ok. So one way so you will go back here.

(Refer Slide Time: 22:48)

Laplace Smoothing

- Problem:** categories with no entries result in a value of "0" for conditional probability

$$P(C|X) = P(X_1|C) * P(X_2|C) * P(C)$$
- Solution:** add "1" to numerator and denominator of empty categories

$$P(X_1|C) = \frac{1}{\text{Count}(C) + n}$$

$$P(X_2|C) = \frac{\text{Count}(X_2 \& C) + 1}{\text{Count}(C) + m}$$

Example: Predicting Tennis With Naïve Bayes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$P(x|c)$
 $P(\text{outlook} = \text{Sunny} | \text{play} = \text{Yes})$
 $= \frac{2}{9}$



Example: Training Naïve Bayes Tennis Model

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Play}=\text{No}) = 5/14$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

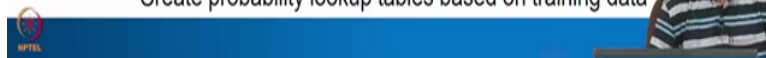
$$p(x|c) \begin{matrix} 1 \rightarrow \text{Yes} \\ 0 \rightarrow \text{No} \end{matrix}$$

$$p(\text{Outlook}=\text{Sunny} | c=\text{Yes}) = \left[\frac{2}{9} \right]$$

$$p(\text{Temp}=\text{Hot} | \text{Play}=\text{Yes}) = \frac{2}{9}$$

$$p(\text{Temp}=\text{Hot} | \text{Play}=\text{No}) = \frac{2}{5}$$

Create probability lookup tables based on training data



So the one way people addressed this one way to addresses is if you at if you get any feature which has like a 0 probability or frequency you can you tend to you can just ignore that feature ok but that is like a very strong decision to make because it might be the more important feature except that maybe you do not just have data points right. So there is something called Laplace Smoothing which is basically you add one to the numerator so you can of increase the probability by assuming some uniform distribution that is what you are doing.

So for instance if this in this case let us say there are two features and that is what is here and one of them feature as 0 frequency of occurrence you have a 0 there ok and you have in this case P of X 1 or C instead of putting a 0 you just 1 over the count that you typically have the denominator of for that particular class plus n ok, so for each one of them you can do a similar thing.

So for instance if so since you have increased this artificially ok you can also do the same thing by for the other class for the other features, so far even for feature to where you actually have data where it is not 0 you have the count for X 2 and C basically this is a number of times that X 2 happens when class is C and then you increase that also by 1, so this is a technique it is called Laplace Smoothing wherein you artificially add to the numerator of that class and that way you make sure that there are no zero multiplications ok, go back and show this probability calculations for this particular data set right.

So how do you calculate these probabilities right, so we saw that just to we will just walk through this so we are trying to calculate if you look if you go back and look at the Naïve

Bayes formula we are trying to calculate these numbers probability of X given category in this case the example we looked at the category is 1 or 0, 1 corresponds to yes 0 corresponds to no, ok.

So in this case we are trying to calculate probability of x given a feature X in this case feature is sunny actually or let me rephrase that so you would be are really clear about how this works. So we want to calculate the probability of the feature sunny sorry we have to I am making the same mistake again so what it is ok in this calculate the probability of feature outlook this is the feature outlook is the feature right, it is the name of a feature and we wanted to take a particular value given that we decide to play ok given that class is 1 that is a probability that we want to calculate.

So the way to calculate that is we compute the number of times, so before it we do the new I have let us do the denominator first so we are looking at the category one class equal to 1, so the number of times in the data set that class equal to 1 occurs is 9 right and among all the data points per class equal to 1 how many times is outlook equal to sunny right that is 2 that is the (prob) how you calculate the probability here based on the training data.

So based on the training data whether it is in this case this is categorical right this particular feature is categorical because it takes on three values and sunny, overcast and rainy ok similarly in fact this in this entire data set all of them are categorical there is no actual continuous variable or anything but that is how you would calculate it, so far it says let us go back so if you look how many times do we play we play we decide to play we say yes category this is equal to 1, 2, 3, 4, 5, 6, 7, 8, and 9 times ok and out of these 2 out of this 9 we have outlook sunny twice if you look at it ok, so both times we decide to play both 2 times when outlook is sunny we decide to play.

So that though that is why the probability of X given C in this case where outlook is sunny given play equal to yes that is what this corresponds to, so this is 2 over 9 ok so this is what we have to calculate for every one of those variables for every class right. So other way to again look at it is if we have a let us say 3 or 4 classes then there will be 3, 4 columns here and for each one of those columns you have to calculate the probability of that particular feature occurring with that particular value ok and make that table.

So Naïve Bayes training basically involves making this probabilistic table so depending upon the variable and your training data you will have to make the figure out a way to calculate the probability.