Lecture 10 - Naive Bayes

We move to Naive Bayes, which is a simple yet quite powerful algorithm.

Probability Recap:

Marginal probability: $P(x)$

Joint probability: $P(x, y)$

Conditional probability: $P(x|y)$

Law of total probability: $P(x) = \sum_y P(x, y)$

Conditional probability: $P(x|y) = \frac{P(x, y)}{P(y)}$

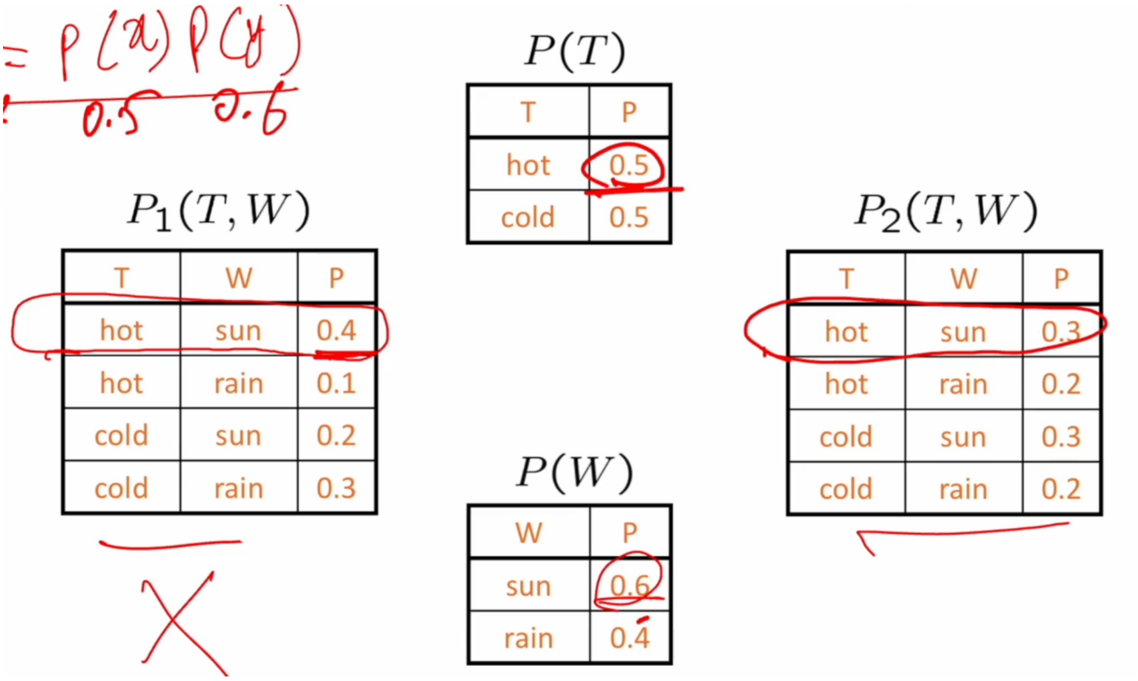
Product rule: $P(x, y) = P(x|y)P(y)$

Bayes' rule: $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$

Chain rule: $P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdots = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$

What is **independence**? Two random variables are independent if $\forall x, y$: $P(x, y) = P(x)P(y)$. In words, whether you observe y or not, it will not impact x. Recall that $P(x, y) = P(x|y)P(y)$, so this is essentially saying that $P(x|y) = P(x)$. The joint probability will decompose into two separate marginal probabilities. Factors into a product of simpler distributions.

$$\forall x, y: P(x|y) = P(x)$$



$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T, W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

Which one is independent? Check to see if $P_1(T, W) = P(T)\,P(W)$ and the same for $P_2(T, W)$. Clearly, we can see that for at least one of them, $P_1(T, W) = 0.4 \neq 0.5 \times 0.6$. Check for P_2: For all of them, we can see that $P_1(T, W) = P(T)P(W)$. Therefore, P_2 is independent.

What is **conditional independence**? X is conditionally independent of Y given Z iff $\forall x, y, z$: $P(x|z, y) = P(x|z)$. Conditional independence does not imply full independence.

Naive Bayes Model:

- Model: A single unobserved variarble Y directly infleunces a number of observed variables X_i
- Assumptions: All observations are conditionally independent of each other given Y.

  In other words, $P(X_1|Y,X_2) = P(X_1|Y)$, and so on.
- Let's just start with one observed variable. Joint probability: $P(Y,X) = P(X|Y)P(Y)$
- Posterior probability: $P(Y|X) = \frac{P(Y,X)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$

Let's move on to two random variables, X_1 and X_2.

- Joint probability (using the chain rule): $P(Y,X_1,X_2) = P(Y)P(X_1|Y)P(X_2|X_1,Y)$

  Now, if we use the conditional independence assumption, we know that $P(X_2|X_1,Y) = P(X_2, Y)$. Therefore, the joint becomes:

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

- Posterior probability:

$$P(Y|X_1,X_2) = \frac{P(Y,X_1,X_2)}{P(X_1,X_2)} = \frac{P(Y)P(X_1|Y)P(X_2|Y)}{P(X_1,X_2)}$$

What if we have three random variables? We just keep expanding the same thing.

- Joint probability:

$$P(Y,X_1,X_2,X_3) = P(Y)P(X_1|Y)P(X_2|Y)P(X_3|Y)$$

- Posterior probability:

$$P(Y|X_1,X_2,X_3) = \frac{P(Y,X_1,X_2,X_3)}{P(X_1,X_2,X_3)} = \frac{P(Y)P(X_1|Y)P(X_2|Y)P(X_3|Y)}{P(X_1,X_2,X_3)}$$

And so on and so forth...

To generalrize:

- Joint probability:

$$P(Y,X_1,X_2,\ldots,X_n) = P(Y)P(X_1|Y)P(X_2|Y)\cdots P(X_n|Y)$$

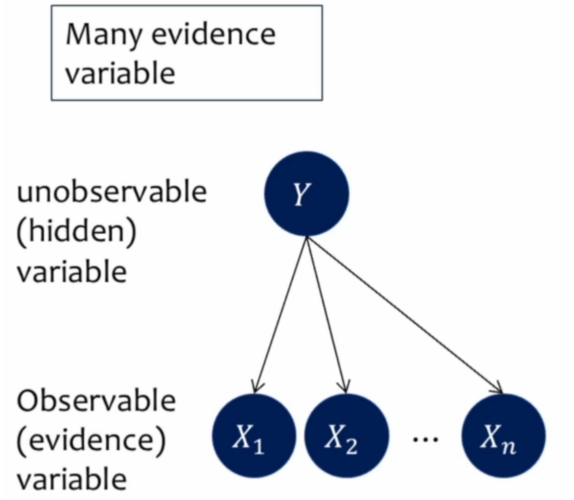
$$= P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

- Posterior probability:

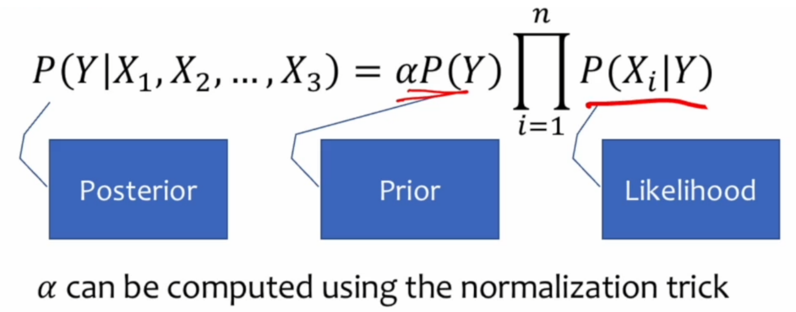
$$P(Y|X_1,X_2,\ldots,X_n) = \frac{P(Y)P(X_1|Y)P(X_2|Y)\cdots P(X_n|Y)}{P(X_1,X_2,\ldots,X_n)}$$

$$= \alpha P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

The α can be calculated using the normalization trick, which I'm still not sure I fully understand, but no worries. It is just the denominator.


Many evidence variable
unobservable (hidden) variable
Y
Observable (evidence) variable
X_1
X_2
…
X_n

These broken down probabilities are generally easier to obtain. This is the whole point of the Naive Bayes model. For inference:



$$P(Y|X_1, X_2, \ldots, X_3) = \alpha P(Y) \prod_{i=1}^{n} P(X_i|Y)$$

α can be computed using the normalization trick

It is important to note that we are making an assumption that all the observations are conditionally independent of each other given Y. Let's look at an example: 2-test cancer.

Variables:

- C: cancer
- t_1 and t_2: the two tests that we have.

To find the posterior probability:

$$P(c|t_1,t_2) = \alpha P(c)P(t_1|c)P(t_2|c)$$

Two likelihoods, one for each of the two tests. The prior and likelihood values are:

- $P(c) = 0.01$
- $P(t_1|c) = 0.9, P(t_1|\sim c) = 0.2$
- $P(t_2|c) = 0.9, P(t_2|\sim c) = 0.2$

Based on this, we want to find the probability that the patient has cancer given that both the tests came out positive.

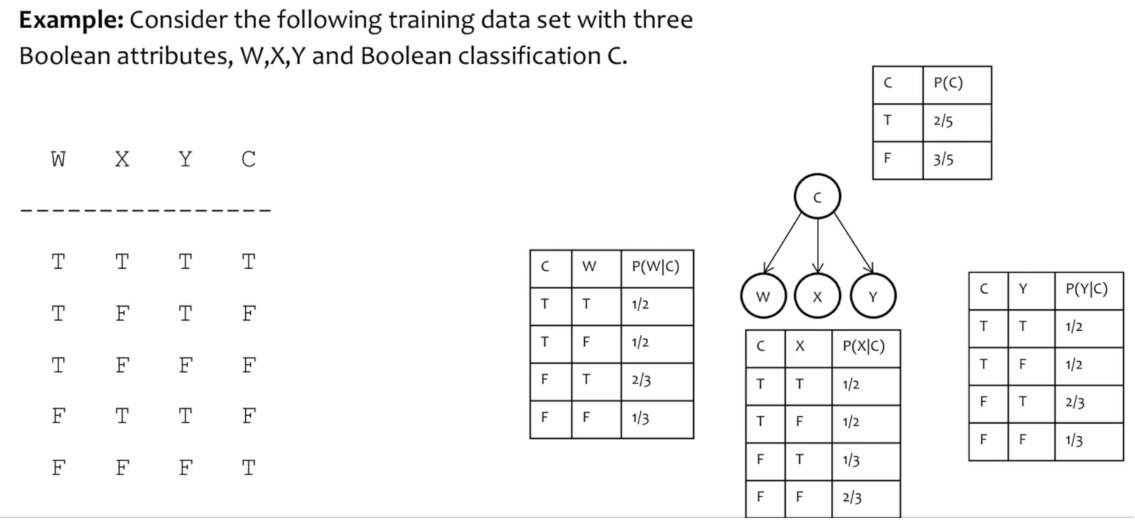
$$\begin{aligned} P(c|t_1,t_2) &= \alpha P(c)P(t_1|c)P(t_2|c) \\ &= \alpha(0.9)(0.9)(0.01) \\ &= \alpha(0.0081) \\ P(\sim c|t_1,t_2) &= \alpha P(\sim c)P(t_1|\sim c)P(t_2|\sim c) \\ &= \alpha(0.2)(0.2)(0.99) \\ &= \alpha(0.0396) \end{aligned}$$

Now you have to do the normalization, which is simply:

$$\frac{0.0081}{0.0081+0.0396}=0.1698\approx 17\%$$

$$\frac{0.0396}{0.0081+0.0396}=0.830\approx 83\%$$

Therefore, the probability of someone having cancer given that both tests came out positive is 17%. We can assume that the two tests are independent of each other. We only need to know 5 values here. Let's say we did not use a naive Bayes approach and did a full joint instead. In that case, we have 3 random variables, so we need to know $2^3=8$ probability values. Now, how do we learn the models' priors and likelihoods? Based on training data. Take a look at this simplified example:



Example: Consider the following training data set with three Boolean attributes, W,X,Y and Boolean classification C.

W	X	Y	C
T	T	T	T
T	F	T	F
T	F	F	F
F	T	T	F
F	F	F	T

C	P(C)
T	2/5
F	3/5

C	W	P(W|C)
T	T	1/2
T	F	1/2
F	T	2/3
F	F	1/3

C	X	P(X|C)
T	T	1/2
T	F	1/2
F	T	1/3
F	F	2/3

C	Y	P(Y|C)
T	T	1/2
T	F	1/2
F	T	2/3
F	F	1/3

$P(c)=\frac{2}{5}$,$P(\sim c)=\frac{3}{5}$. Simple stuff. For example, $P(W|C)$ is just checking where both of them are true. Note that in joint probabilities, the sum of all should be one. But for conditional, it does not need to sum to one.

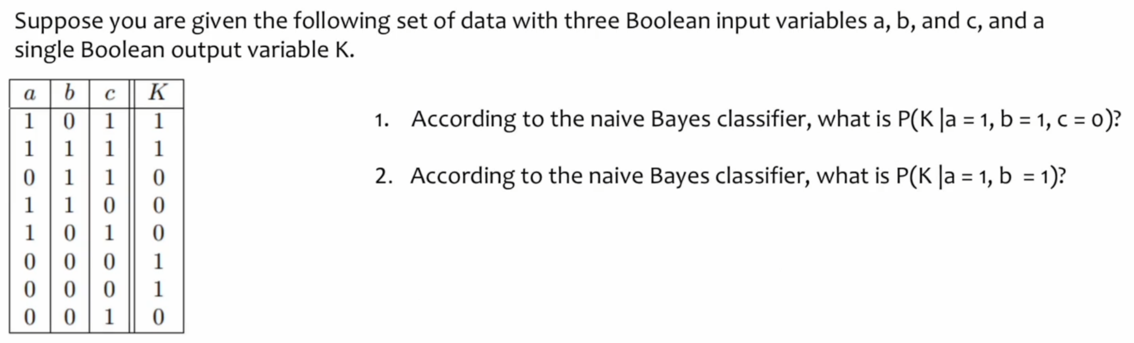
What is the total number of values that we need to know to be able to do Naive Bayes? Total 7. One for $P(C)$, and 2 for each of the other ones. If we were to use the full joint distribution, we would need $2^4=16$.

Maximum Likelihood Estimate: $P(x)=\frac{\text{count}(x)}{N}$. Used to find the probability values that make the training data most probable. This is how we used the truth table on the left.

Now, let's try to find $P(C|W=F,X=T,Y=F)$. This would be equal to:

$$\begin{aligned}
&=\alpha P(C)P(W=F|C)P(X=T|\,C)P(Y=F|\,C)\\
&=\alpha\left(\frac{2}{5}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) && \text{for } C \text{ true}\\
&=\alpha\left(\frac{3}{5}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{1}{3}\right) && \text{for } C \text{ false}\\
&\text{for } C \text{ true: } \alpha 0.05\\
&\text{for } C \text{ false: } \alpha 0.022222\\
&\text{Normalize:}\\
&C \text{ true: } 0.692\\
&C \text{ false: } 0.308
\end{aligned}$$

Let's look at another problem:



Suppose you are given the following set of data with three Boolean input variables a, b, and c, and a single Boolean output variable K.

a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

1. According to the naive Bayes classifier, what is P(K |a = 1, b = 1, c = 0)?

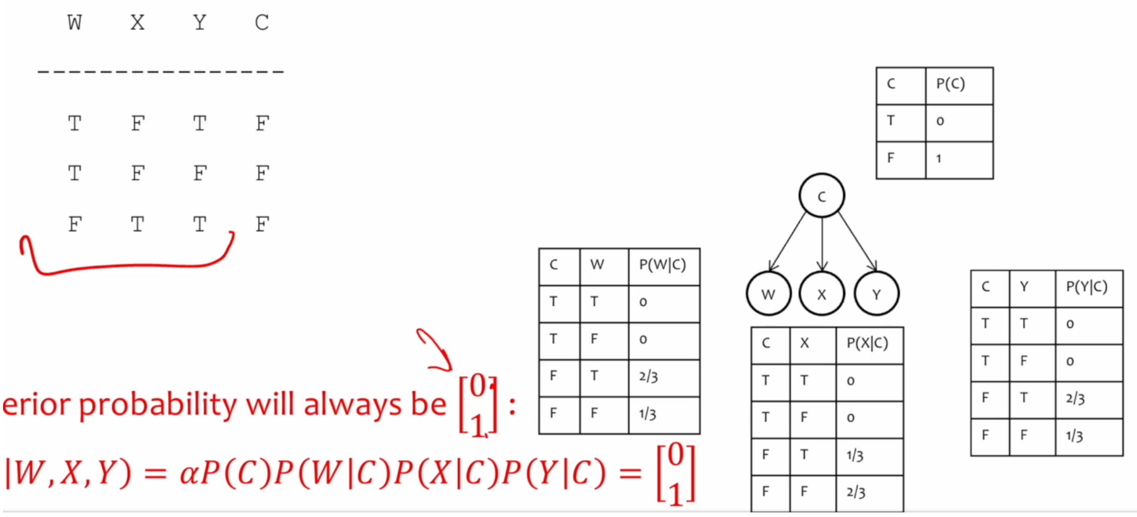
2. According to the naive Bayes classifier, what is P(K |a = 1, b = 1)?

$$\begin{aligned} P(K|a=1,b=1,c=0) \\ =P(K)P(a|K)P(b|K)P(c|K) \\ P(K)=\frac{1}{2} \\ \dots \end{aligned}$$

For the second question, the variable C is missing in the expression. We can ignore the missing variable as it has no effect in the inference. So it would be:

$$P(K|a=1b=1)=P(K)P(a|K)P(b|K)$$

What if we have limited training data or biased training data? We use some sort of regularization, in the case of Naive Bayes models we use Laplace Smoothing.



W	X	Y	C
T	F	T	F
T	F	F	F
F	T	T	F

C	P(C)
T	0
F	1

C	W	P(W|C)
T	T	0
T	F	0
F	T	2/3
F	F	1/3

C	X	P(X|C)
T	T	0
T	F	0
F	T	1/3
F	F	2/3

C	Y	P(Y|C)
T	T	0
T	F	0
F	T	2/3
F	F	1/3

erior probability will always be $\begin{bmatrix}0\\1\end{bmatrix}$:

$|W,X,Y) = \alpha P(C)P(W|C)P(X|C)P(Y|C) = \begin{bmatrix}0\\1\end{bmatrix}$

In this case, we would always get $\begin{bmatrix}0\\1\end{bmatrix}$ because $P(c=T)=0$. Therefore, this is not a very good prior to have. We have to use Laplace smoothing then.

$$P(x)=\frac{\text{count}(x)+\lambda}{N+\lambda|x|}$$

$|x|$ is the number of possible discrete values x can have. If we have a larger λ, then we would have more smoothing, and the probabilities would be closer to uniform distribution. Let $\lambda=1$. Then:

$$P(C)=\frac{\text{count}(c)+\lambda}{N+\lambda|x|}=\frac{3+1}{3+1|2|}=\frac{4}{5}$$

Here, this would be the probability of $C = \text{false}$ because that's all we have for the count. $1 - \frac{4}{5}$ would give us $P(C = \text{true})$. His microphone goes mute here. I don't really understand how this effects the rest of the posterior probability.

We are essentially adding more examples? Look into this further.

The rest of the lecture is him going through a problem set. We'll return to this later as well.