Learning classifiers based on Baye's rule

$X \rightarrow < x_1, x_2, x_3 \dots x_n > \rightarrow$ Assume each of $x_1, x_2 \dots$ are boolean variables

$Y \rightarrow$ Boolean variable

$f: X \rightarrow Y \Rightarrow$ Given X predict $Y \Rightarrow P(Y/X)$

$$P(Y=y_i|X=x_k) = \frac{\overbrace{P(X=x_k|Y=y_i)}^{①}P(Y=y_i)}{\sum_j P(X=x_k|Y=y_j)P(Y=y_j)}$$ → (As per Bayes theorem)

Estimating number of possibilities:-

If $n \rightarrow$ vector $P(X=x_k | Y=y_i) \Rightarrow P(x_1 / x_2, x_3 \dots x_n, y_i) * \underbrace{P(x_2, x_3 \dots x_n * y_i)}$

$\Rightarrow P(x_2/x_3, \dots x_n y_i) * P(x_3 \dots x_n y_i)$

2^n possible vector combinations for X

2 possible values for y_i

Total $\rightarrow 2^{n+1}$ parameters must be estimated

$n=3 \rightarrow P(x = x_k | y = y_i) \quad (x_k = < x_1, x_2, x >$

$P(x_1 / x_2, x_3, y_i) * P(x_2, x_3, y_i)$

$\Rightarrow P(x_1 / x_2, x_3, y_i) * P(x_2 / x_3, y_i)$

$\Rightarrow P(x_1 / x_2, x_3, y_i) * P(x_2 / x_3, y_i) * P(x_3 / y_i) * P(y_i)$

Assume this is the only training data

x_1	x_2	x_3	y_i
0	1	0	0
1	0	1	1

Given $\rightarrow (0,0,0)$

$P(0 / x_2 = 0, x_3 = 0, y_i = 0) \rightarrow = 0$

$P(x_2 = 0 / x_3 = 0, y_i = 0) \rightarrow = 0$

* you need to have atleast 2^n datapoints.
* If $n=30 \rightarrow 2^{30} \rightarrow 500$ billion datapoints

Assuming $X \rightarrow$ Binary

It is really hard to calculate probability from training data as we donot get that many datapoints in the training data. And if one probability return zero entire multiplication will return zero.

Naive Bayes

Naive bayes make a naive assumption that the features are conditionally independent

Definition: Given three sets of random variables X, Y and Z, we say X is **conditionally independent** of Y given Z, if and only if the probability distribution governing X is independent of the value of Y given Z; that is

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Coming to our previous e.g.:-

$$x_K = \langle x_1, x_2, x_3 \rangle$$

$$P(x = x_K | y = y_i)$$

$$\underbrace{P(x_1 | x_2, x_3, y_i)} * P(x_2 | x_3, y_i) * P(x_3 | y_i) * P(y_i)$$

Assuming conditional independent

$$P(x_1 | y_i) * P(x_2 | y_i) * P(x_3 | y_i) * P(y_i) \Rightarrow \boxed{P(y_i) \prod_{j=1}^{n} P(x_j | y_i)} \rightarrow (2)$$

with this assumption we only need to calculate 2n probabilities

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Sub → (2) in the above equation

$$P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

To convert this into a classification problem we need to find the maximum probability.

$$Y \leftarrow \arg\max_{y_k} \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

As the denominator doesnot change based on the value of y_K, it is redundant to calculate everytime.

$$Y \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i | Y = y_k)$$

Understanding Train and test phase in NB

Sample data

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$$Y \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i | Y = y_k)$$

$y_k \rightarrow \{yes, No\}$

$x_i \rightarrow \{outlook, temp, humidity, windy\}$

outlook $\rightarrow$ {Rainy, overcast, sunny}; temp $\rightarrow$ {hot, mild, cool}; humidity $\rightarrow$ {high, normal}; windy $\rightarrow$ {True, False}

$$* \; P(x_i = rainy \,/\, y = yes) = \frac{P(x_i = rainy \text{ and } y = yes)}{P(y = yes)}$$

$$= \frac{2/14}{9/14} = \frac{2}{9}$$

In training, we calculate all the possible combinations of probability

$$P(y = yes) = \frac{9}{14} \qquad P(y = No) = \frac{5}{14}$$

We can repeat the process for all the data.

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

Train time complexity:-

$$O(n * d * c)$$

n → no of data points; d → features; c → classes

why n?

In worst case a feature can have n-different possibilities

Space complexity

$$O(n * d * c)$$

Store $n*d*c$ cells (from lookup tables)

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

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

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Testing phase

Given a point $x_q \to$ <overcast, hot, high, True> $\to$ play Golf yes/NO??

$$p(y=yes/x_q) = p(yes) * p(overcast/yes) * p(hot/yes) * p(high/yes) * p(True/yes)$$

$$= \frac{9}{14} \times \frac{4}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.3456$$

$$p(y=NO/x_q) = p(NO) * p(overcast/NO) * p(hot/NO) * p(high/NO) * p(True/NO)$$

$$= \frac{5}{14} \times 0 \times \ldots = 0$$

$p(y=yes/x_q) > p(y=NO/x_q) \to x_q \to$ Target yes.

Test time complexity

$O(\underset{\text{dimensions}}{d} * \overset{\text{no. of classes}}{c})$

Wait!! $\to$ Just because one feature resulted in zero probability.

$\downarrow$

entire posterior probability came to zero.

$\downarrow$

This is not fair, what if the other features are important?

To solve this problem, we have laplace smoothing.

$$p(overcast/NO) = \frac{\text{no. of points where Outlook is overcast and playgolf = NO}}{\text{no. of points where play golf is NO}} = \frac{0}{5} = 0$$

$\downarrow$ modified

$$\boxed{= \frac{\text{no. of points where Outlook is overcast and playgolf = NO} + \alpha}{\text{no. of points where play golf is NO} + c\alpha}}$$

$\alpha \to$ hyper-param

$c \to$ no. of classes

here ($c=2$)

$$= \frac{0+1}{5+2\times(1)} = \frac{1}{7} = 0.14 \ (\neq 0)$$

Here $\alpha \to$ hyperparam

α is small $\to$ overfitting

α is large $\to$ underfitting

As you can see the example on the right higher values of alpha gives us almost a constant value $\left(\frac{1}{n}\right)$

(without Laplace smoothing)

$p(x/y)$	$\alpha = 1$	$\alpha = 100$
$2/3$	$\frac{2+1}{3+2} = \frac{3}{5}$	$\frac{2+100}{3+200} = \frac{102}{203} = 0.5$
1	$\frac{1+1}{1+2} = \frac{2}{3}$	$\frac{1+100}{1+200} = \frac{101}{201} \simeq 0.5$