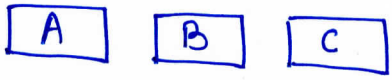
BAYES CLASSIFIER

Prior Probability:- It is the probability of event before occurrence of new data.

Posterior Probability:- It is the updated (revised) probability that considers additional new information.

e.g:- let we have 3 pieces of land.



Out of these three pieces of land, one has oil inside the surface.

∴ Probability of getting oil from land C before digging starts is $\boxed{P(C) = 1/3}$. This is called as Prior Probability.

Now, let land A is digged & we didn't get oil. Hence now the probability of getting oil from land C becomes $\boxed{P(C) = 1/2}$ as now only B & C are remaining & we got information that A doesnot have oil. This is called as Posterior Probability.

BAYES RULE:

It gives posterior probability if Prior probabilities are know & class conditional probability is known.

It is given by

$$P(A|B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

e.g:- In a class of students, 60% are boys & 40% are girls. All boys carry black color mobiles whereas girls carry black color & white color mobiles in equal numbers. Let a student is selected randomly & found that the mobile carried was black. what is the probability that the student is girl ?

→ $P(\text{Boys}) = 0.6 \quad P(\text{Girls}) = 0.4$

$P(\text{Black} | \text{Girl}) = 0.5 \quad P(\text{Black}) = 0.8$

$\therefore P(G | \text{Black}) = ?$

$\therefore P(G | \text{Black}) = \dfrac{P(\text{Black} | \text{Girl}) \cdot P(\text{Girl})}{P(\text{Black})}$

$= \dfrac{0.5 \times 0.4}{0.8}$

$$\boxed{P(G | \text{Black}) = 0.25}$$

On similar Basis, let us say we have two classes C_1 & C_2.

From Bayes Rule.

$$P(w_i | x) = \frac{P(x | w_i) \cdot P(w_i)}{P(x)}$$

where

$P(x) \xrightarrow{\text{prob.}}$ event has occured.

let us say we want to check, the given vector x (feature vector) belongs to which class. For this we check what is the probability of given vector x (vector term is used because it may have more than one variable) lying in either class C_1 & in class C_2

i.e $P(C_1|x)$ → probability of given x lying in class C_1,

$P(C_2|x)$ → -11- class C_2.

whichever probability is more, that x vector belongs to that class.

i.e if $P(C_1|x) > P(C_2|x)$

x belongs to C_1

or if $P(C_1|x) < P(C_2|x)$

x belongs to C_2.

Now by Bayes theorem. $P(C_1|x)$ can be determined by

$$P(C_1|x) = \frac{P(x|C_1) \cdot P(C_1)}{P(x)}.$$

&

$$P(C_2|x) = \frac{P(x|C_2) \cdot P(C_2)}{P(x)}$$

∴ comparing

$$P(C_1|X) > P(C_2|X)$$

i.e

$$\frac{P(X/C_1)\cdot P(C_1)}{P(X)} > \frac{P(X/C_2)\cdot P(C_2)}{P(X)}$$

The denominator $P(X)$ can be cancelled out as the result of comparison doesnot change without it. It is called as normalizing Constant.

Thus for comparison we can use only

$$P(X|C_1)\cdot P(C_1) > P(X|C_2)\cdot P(C_2)$$

X belongs to C_1

&

$$P(X|C_1)\cdot P(C_1) < P(X|C_2)\cdot P(C_2)$$

X belongs to C_2.

Ex:- consider the following data set. There are total 4 variables (outlook, temperature, Humidity & windy). There are two classes play - Yes & play - No Each data instance will have 4 attributes.

$$X = \{x_1, x_2, x_3, x_4\}$$

↓ outlook ↓ Temp ↓ Hum ↓ windy

Notes:- Hence X is called as feature vector.

Now e.g consider 1st instance

$X = \{$ sunny, HOT, high, False $\}$. This belongs to class - play - NO.

Similarly for X = {Rainy, Cool, Normal, False}, the class is Yes. {see data serial no. (5)}.

But we want to find the class of

X = {Rainy, Cool, High, True} i.e this if these are the weather conditions whether to play or not to play. From the data set we cannot determine the class. Hence we have to find its class i.e under such conditions whether to play or not to play. **Data set - DS1**

S.NO	outlook	Temperature	Humidity	windy	Play
1	sunny	Hot	high	False	NO
2	sunny	Hot	high	True	NO
3	overcast	Hot	high	False	Yes
4	Rainy	mild	high	False	yes
5	Rainy	cool	normal	False	yes
6	Rainy	cool	normal	True	NO
7	overcast	cool	normal	True	yes
8	sunny	mild	high	False	NO
9	sunny	cool	normal	False	Yes
10	Rainy	mild	normal	False	Yes
11	sunny	mild	normal	True	yes
12	overcast	mild	high	True	yes
13	overcast	Hot	normal	False	Yes
14	Rainy	mild	high	True	NO

By Bayes theorem

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

we can ignore $P(X)$

i.e

$$P(C_1|X) = P(X|C_1) P(C_1)$$

& $$P(C_2|X) = P(X|C_2) \cdot P(C_2)$$

we have to compare whether

$$P(X|C_1) \cdot P(C_1) > P(X|C_2) \cdot P(C_2)$$

where

$P(C_1) \rightarrow P(yes)$ let $C_1 \rightarrow$ class yes

$P(C_2) \rightarrow P(NO)$. $C_2 \rightarrow$ class NO

From the data there are total 9 yes & 5 NO out of 14 values

$\therefore$ $P(C_1) = P(yes) = 9/14$

$P(C_2) = P(NO) = 5/14$ } - equn (1).

Now when we have to find $P(X/C_i)$, it is very expensive activity. Because we have to consider all the combinations.

i.e if we consider outlook as sunny then temperature can be hot or mild or cool also for these comb^n again humidity can be high or normal & again for all these comb^n windy can be false or true. Similarly for outlook = overcast again so many combinations. Thus the

complexity again increases if n (variables) increases. In order to simplify the calculations, we assume for each class the attributes are independent. The classifier resulting from this assumption is called as Naive Bayes classifier.

with Naive Bayes classifier we can write

$$P(X|C_i) = \prod_{K=1}^{n} P(x_K|C_i).$$

i.e The product of probabilities of each of values of attributes of X for given class C.

i.e The probability of outlook = sunny for Play = No will be

$$P(x_1 = \text{sunny}|C_2) = 3/5$$

i.e out of 5 NO's there are 3 sunny outlook.

let us verify with Naive classifier the class ~~out~~ for vector $X = \{\text{sunny, hot, high, false}\}$.

→ For this we have to compare

$$P(X|C_1) \cdot P(C_1) \quad \& \quad P(X|C_2) \cdot P(C_2)$$

we know $P(C_1) = 9/14 \quad P(C_2) = 5/14$

Now for class C_1 i.e play = yes

$$P(X|C_1) = ?$$

$$P(X = \text{sunny, hot, high, false}\}|C_1) = ?$$

∴ finding seperate probabilities & then taking their product. finding for class C_1 i.e Yes

i.e $P(x = \text{sunny} | C_1) = 2/9$

∴ there are 9 yes & out of 9 there are 2 sunny.

Similarly

$$P(x = \text{hot} | C_1) = 2/9$$

$$P(x = \text{high} | C_1) = 3/9$$

$$P(x = \text{False} | C_1) = 6/9$$

$$\therefore P(x | C_1) = P(x = \text{sunny} | C_1) * P(x = \text{hot} | C_1) * P(x = \text{high} | C_1) * P(\text{False} | C_1)$$

$$= \frac{2}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{6}{9}$$

$$\Rightarrow P(X | C_1) = 0.0109$$

$$\therefore P(C_1 | X) = P(X | C_1) \cdot P(C_1)$$

$$= 0.0109 \times \frac{9}{14}$$

$$\therefore \boxed{P(C_1 | X) = 0.0070} \quad - (A)$$

ie Probability of X belonging to class C_1 ie Play = Yes is 0.0070.

similarly Now finding for class C_2 ie NO

$$P(x = \text{sunny} | C_2) = 3/5$$

$$P(x = \text{hot} | C_2) = 2/5$$

$$P(x = \text{high} | C_2) = 4/5$$

$$P(x = \text{False} | C_2) = 2/5$$

$$\therefore P(X|C_2) = P(x = \text{sunny}|C_2) * P(x = \text{hot}|C_2) * P(x = \text{high}|C_2) * P(x = \text{False}|C_2)$$

$$= \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{2}{5}$$

$$P(X|C_2) = 0.0768$$

$$\therefore P(C_2|X) = P(X|C_2) \cdot P(C_2)$$

$$= 0.0768 \times \frac{5}{14}$$

$$\boxed{P(C_2|X) = 0.02742} \quad - (B)$$

i.e Probability of X belonging to class C_2 i.e Play = NO is 0.02742.

$\therefore$ From (A) & (B)

$$P(C_2|X) > P(C_1|X)$$

$\therefore$ The result is class Play = NO. i.e for given X vector there should not be any play.

We can see that our calculation is correct as we got the same result what is given in the data set.

But you can find the class for vect

$X = \{$ Rainy, cool, high, True $\}$ by same method. Note the class is not given in the dataset. You have to find it.

Now if the variables are with numeric values then the normal distribution PDF should be evaluated

e.g: consider the same previous data set with variables outlook, Temperature, Humidity & windy, only difference is now consider the variable outlook & windy same but temperature are humidity are with numeric values as follows:

Data set - DS2

S.N	Outlook	temperature	Humidity	windy	play
1	sunny	85	85	False	NO
2	sunny	80	90	True	NO
3	overcast	83	86	False	Yes ←
4	Rainy	70	96	False	Yes ←
5	Rainy	68	80	False	Yes ←
6	Rainy	65	70	True	NO
7	overcast	64	65	True	Yes ←
8	sunny	72	95	False	NO
9	sunny	69	70	False	Yes ←
10	Rainy	75	80	False	Yes ←
11	sunny	75	70	True	Yes ←
12	overcast	72	90	True	Yes ←
13	overcast	81	75	False	Yes ←
14	Rainy	71	91	False	NO

Find the class for vector $X = \{$ Rainy, temp = 60, Humidity = 62, windy = False $\}$

we have to compare

$$P(C_1/X) \;\&\; P(C_2/X)$$

i.e

$$P(X/C_1)\cdot P(C_2) \;\&\; P(X/C_2)\cdot P(C_2)$$

we know:

$$P(C_1) = 9/14 \qquad P(C_2) = 5/14.$$

For

$P(X/C_1)$ i.e for class C_1 i.e Play = Yes.

$X = \{$ Rainy, temp = 60, Humidity = 62, windy = false$\}$

$$P(X/C_1) = P(x = \text{Rainy}/C_1) * P(x = \text{temp} = 60/C_1) * P(x = \text{humd} = 62/C_1) * P(x = \text{windy} = \text{false}/C_1)$$

$P(x = \text{Rainy}/C_1) = 3/9$ -①

$P(x = \text{temp} = 60/C_1)$ · Now here it is numeric value hence probability cannot be taken directly. Therefore finding PdF. For different classes.

For C_1 ∵ PdF is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

∴ For variable temp

$$f(x = \text{temp}) = \frac{1}{\sigma(\text{temp})\sqrt{2\pi}}\, e^{-\frac{(x-\mu_{\text{temp}})^2}{2\sigma^2_{\text{temp}}}}$$

$$\mu_{\text{temp}} = \frac{83+70+68+64+69+75+75+72+81}{9} \rightarrow \text{for Yes only.}$$

$$\boxed{\mu_{\text{temp}/C_1} = 73}$$

similarly $\sigma^2_{\text{temp}/C_1} = \frac{(83-73)^2+(70-73)^2+\cdots+(81-73)^2}{8}$

$\therefore \boxed{\sigma_{\text{temp}/C_1} = 6.16}$ -②

similarly

$$\therefore f(x = temp = 60) = \frac{1}{6.16\sqrt{2\pi}} e^{-\frac{(60-73)^2}{2(6.16)^2}}$$

i.e $x = 60$

$$f(x = temp = 60 / c_1) = 0.00698. \quad - (2)$$

similarly finding humidity = 62 for c_1.

we get.

$$f(x = humidity = 62 | c_1) = 0.0096. \quad - (3)$$

$\therefore$

$$P(x|c_1) =$$

$$P(x = windy = false | c_1) = 6/9 \quad - (4)$$

$\therefore$

$$P(x|c_1) = (1) \times (2) \times (3) \times (4)$$

$$= \frac{3}{9} \times 0.00698 \times 0.0096 \times \frac{6}{9}$$

$\therefore$

$$P(X|c_1) = 0.01489$$

$\therefore$

$$P(c_1/x) = P(x|c_1) \times P(c_1)$$

$$= 0.01489 \times \frac{9}{14}$$

$$\boxed{P(c_1/x) = 9.572 \times 10^{-6}} \quad - (A)$$

Now Proceeding similarly way for class $c_2 = NO$ we get

$$P(x = rainy | c_2) = 2/5$$

$$P(x = temp = 60 | c_2) = 0.0094$$

$$P(x = hum = 62 | c_2) = 0.0018$$

$$P(x = windy = false | c_2) = 2/5 = 0.4$$

$$\therefore P(X|C_2) = \frac{2}{5} \times 0.0094 \times 0.0018 \times \frac{2}{5}$$

$$= 2.7072 \times 10^{-6}$$

$$\therefore P(C_2|X) = P(X|C_2) \cdot P(C_2)$$

$$= 2.7072 \times 10^{-6} \times \frac{5}{14}$$

$$\boxed{P(C_2|X) = 0.9668 \times 10^{-6}} \quad - (B)$$

comparing (A) & (B) we get

$P(C_1|X) > P(C_2|X)$

$\therefore$ Given X belongs to class C_1 i.e play = Yes.