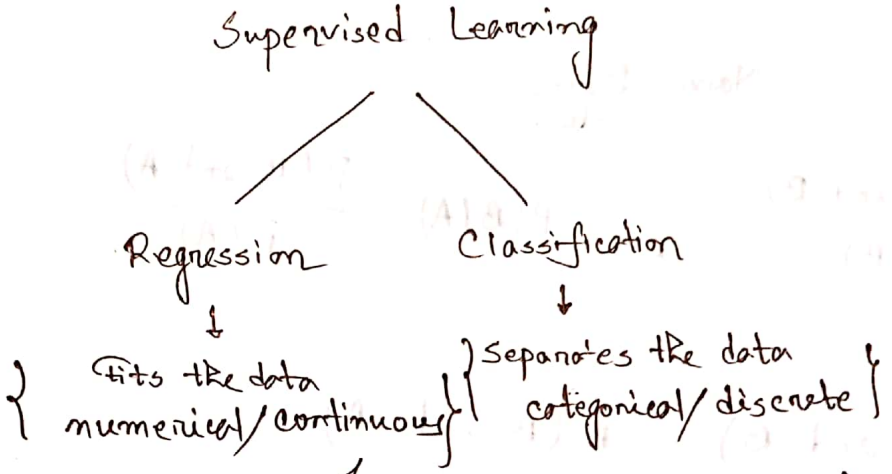
Machine Learning

Supervised Learning: Is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training example.



Classification is a process of categorizing a given set of data into classes. It can be performed both structured and unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

Regression is basically a statistical approach to find the relationship between variables. This is used to predict the outcome of an event based on the relationship between variables obtained from the data-set.

Scatter Plot: A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

Decision Boundary or **Decision Surface** is a hypersurface that partitions the underlying vector space into two sets, one for each class. When a Decision Surface is a straight line we call it linear.

Naive Bayes

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \qquad P(B|A) = \frac{P(B \text{ and } A)}{P(A)}$$

We know,

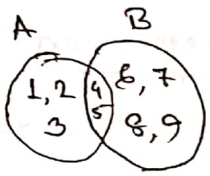
$$P(A \text{ and } B) = P(B \text{ and } A)$$

$$\Rightarrow P(A/B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$\Rightarrow \frac{P(A/B) \cdot \cancel{P(B)}}{\cancel{P(B)}} = \frac{P(B/A) \cdot P(A)}{P(B)} \quad [\text{Divided both side by } P(B)]$$

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

Example: $A = \{1, 2, 3, 4, 5\}$; $B = \{4, 5, 6, 7, 8, 9\}$



Now, $P(A) = \frac{5}{9}$

$P(B) = \frac{6}{9}$

$P(B|A) = \frac{2}{5}$

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

$$= \frac{2/\cancel{5} \cdot \cancel{5}/9 \cdot 9}{\cancel{6}/\cancel{9} \cdot 9}$$

$$= \frac{2}{\cancel{6}} = \frac{1}{3} \quad \text{(Ans)}$$

Example: A particular study showed that 12% of men will likely develop prostate cancer at some point in their lives. A man with prostate cancer has a 95% chance of a positive test result from a medical screening exam. A man without prostate cancer has a 6% chance of getting a false positive test result. What is the probability that a man has cancer given that he has a positive test result.

Answer: $P(\text{cancer}) = 12\% = 0.12$

$P(\text{positive} \mid \text{cancer}) = 95\% = 0.95$

$P(\text{positive} \mid \neg\text{cancer}) = 6\% = 0.06$

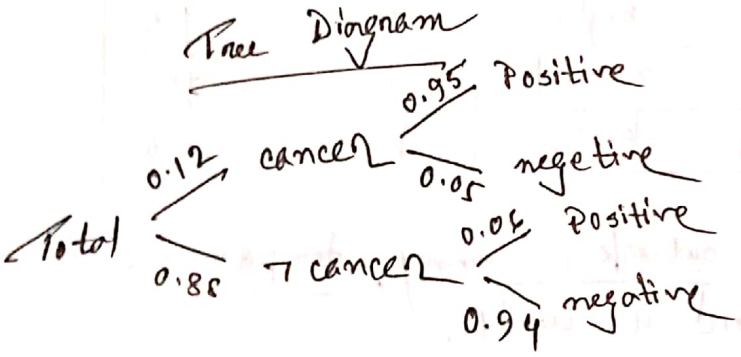
$P(\text{cancer} \mid \text{positive}) = ?$

We know,

$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer}) \, P(\text{cancer})}{P(\text{positive})}$$

we don't have this

Tree Diagram



In this case,

$P(\text{positive}) = P(\text{cancer and positive}) + P(\neg\text{cancer and positive})$

$= (0.12)(0.95) + (0.88)(0.06)$

$= 0.1668$

$\therefore P(\text{cancer} \mid \text{positive}) = \frac{(0.95)(0.12)}{0.1668}$

$= 0.683$

Example:

DATASET

	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	cold cool	normal	false	yes
6	rainy	cold cool	normal	true	no
7	overcast	cold 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

Question: Can you play when the outlook is sunny, temperature is cool, humidity is high and it's windy.

Solution: $P(Yes) = \frac{9}{14}$; $P(No) = \frac{5}{14}$

Outlook				
	Yes	No	P(Yes)	P(No)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature				
	Yes	No	P(Yes)	P(No)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity				
	Yes	No	P(Yes)	P(No)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind				
	Yes	No	P(Yes)	P(No)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Probability that we can play the game :

⇒ P (Outlook = Sunny | Play = Yes) = 2/9

⇒ P (Temperature = Cool | Play = Yes) = 3/9

⇒ P (Humidity = High | Play = Yes) = 3/9

⇒ P (Wind = Strong | Play = Yes) = 3/9

⇒ P (Play = Yes) = 9/14

Probability we can't play a game :

⇒ P (Outlook = Sunny | Play = No) = 3/5

⇒ P (Temperature = Cool | Play = No) = 1/5

⇒ P (Humidity = High | Play = No) = 4/5

⇒ P (Wind = Strong | Play = No) = 3/5

⇒ P (Play = No) = 5/14

So

$P(X \mid Play = Yes)\, P(Play = Yes)$

$= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053$

Similarly,

$P(X \mid Play = No)\, P(Play = No)$

$= (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206$

Here,

$P(X) = P(\text{Outlook} = \text{Sunny}) * P(\text{Temperature} = \text{Cool}) * P(\text{Humidity} = \text{High}) * P(\text{Wind} = \text{Strong})$

$= (5/14) * (4/14) * (7/14) * (6/14)$

$= 0.02186$

Dividing the results by this value!

$P(\text{Play} = \text{Yes} | X) = 0.0053 / 0.02186 = 0.2424$

$P(\text{Play} = \text{No} | X) = 0.0206 / 0.02186 = 0.9421$

Since, $0.9421 > 0.2424$, the answer is "No".

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

Likelihood - which describes how well the model predicts the data

Prior Probability - which describes the degree to which we believe the model accurately describes reality based on all of our prior information

Posterior Probability - which represents the degree to which we believe the given model accurately describes the situation given the available data and all of our prior information

Normalizing constant - the constant that makes the posterior density integrate to one

Mathematical Solution:

We know,

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

With feature vector x,

$$X = (x_1, x_2, x_3, \ldots, x_n)$$

Assume that all features are mutually independent,

$$P(y|x) = \frac{P(x_1|y) \cdot P(x_2|y) \cdots P(y)}{P(x)}$$

Select class with highest probability

$$y = \text{argmax}_y\, P(y|x) = \text{argmax}_y \frac{P(x_1|y)\,P(x_2|y) \cdots P(x_n|y) \cdot P(y)}{P(x)}$$

$$y = \text{argmax}_y\, P(x_1|y)\,P(x_2|y) \ldots P(x_n|y) \cdot P(y)$$

$$= \text{argmax}_y \log(P(x_1|y)) + \log(P(x_2|y)) + \ldots \log(P(x_n|y)) + \log(P(y))$$

$$= \text{argmax}_y \log(P(y)) \prod_{i=1}^{n} \log(P(x_i|y))$$

Gaussian Naive Bayes

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \cdot \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The parameter σ_y and μ_y are estimated using maximum likelihood.