probability of Density Normal Distribution $= p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$\boxed{\log e^{x} = x}$

Assuming that training examples are mutual independent, give h we can write $p(D|h)$ as the product of various $p(d_i|h)$ that can be shown as

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h)$$

$p(D_i|h)$ can be written as a normal distribution with variance σ^2 and mean $\mu = f(x_i)$ or $h(x_i)$

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-\mu)^2}$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-h(x_i))^2}$$

$\log(ab) = \log a + \log b$

constant we can discard

$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2}\right)$$

independent of h

$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

$$\underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

// maximizing -ve term same minimi +ve term.

$$h_{ML} = \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

$$h_{ML} = \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

ML = least squared error

01/2024

imp. Maximum Likelihood hypothesis for predicting Probabilities.

learner -> Learn to predict probabilities.

Medical data set X -> instance.

(it contains symptoms of Disease

Target func^n $f(x) = 1$ => survives

$f(x) = 0$ => Dies

It is a probabilistic func^n.

set of patients 92% survives 8% dies.

$f' : X \rightarrow [0,1]$ => $P(f(x) = 1) = 0.92$

$f' = P(f(x) = 1) = 0.92$

what is the p that patient survive / $P(f(x) = 1)$

f' => Machine should learn this.

* The objective is to learn to predict probability. suppose we have a medical data set instance space X represents patience ~~having~~ in terms of symptoms of a disease. Target function $f(x) = 1$, if the patient survive the disease and $f(x) = 0$, if he doesn't.
Here f is a probabilistic function. Suppose we have a collection of patients exibhiting the same set of symptoms, we might find that 92% survives & 8% do not. We want to find out What is the probability that $f(x) = 1$ for this, we learn a target function $f' : X \rightarrow [0,1]$ such that $f'(x) = P(f(x) = 1)$

Based on our example $f'(x) = 0.92 = P(f(x) = 1)$
To learn this function f' we desire a maximum likelihood hypothesis for f'.

because it is indepent we can write in product form.

$D = (x_1 d_1 \; x_2 d_2 \ldots\ldots x_i d_i)$

For this, we must obtain an expression for $P(D/h)$

$$P(D|h) = \prod_{i=1}^{m} P(x_i d_i / h) \rightarrow (1)$$

Let us assume that training data D, is of the form $D = \{<x_1 d_1> <x_2 d_2> \ldots\ldots <x_m d_m>$ where $d_i = 0$ or 1 for $f(x_i)$.

Treating both x_i and d_i as random variables and assuming that every training example is independent of other.
we can write $P(D|h)$ as (1)

(only target or d_i depends on hypothesis not the patient details or x_i hypo = {survive, dies}.)

we assume that the probability of any particular instance x_i is independent of hypothesis h, when x is independent of h and survival of the patient alone i.e d_i depends on h we can rewrite (1) as below

$$P(D|h) = \prod_{i=1}^{m} P(d_i / h, x_i) \times P(x_i) \rightarrow (2)$$

Now what is the prob $P(d_i / h, x_i)$ of observing $d_i = 1$ for a single instance x_i, given that hypothesis h holds.

h is our hypothesis regarding ~~this~~ target func^n which computes this probability.
Therefore $P(D_i = 1 / h(x_i)) = h(x_i)$ and in general b:

$$P(d_i|h, x_i) = \begin{cases} h(x_i) & \text{if } d_i = 1 \\ 1-h(x_i) & \text{if } d_i = 0 \end{cases} \rightarrow (3)$$

In order to substitute (3) in to (2) for $P(D|h)$ we reexpress 3 as follows, that is found

$$P(d_i|h, x_i) = h(x_i)^{d_i} \cdot (1-h(x_i))^{1-d_i} \rightarrow (4)$$

substitute eq (4) in eq 2

$$P(D|h) = \prod_{m}^{m} (1-h(x_i))^{1-d_i} \times P(x_i) \rightarrow (5)$$

Now we write an expression for max Likeli hood hypothesis.

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} h(x_i)^{d_i} (1-h(x_i))^{1-d_i} \times P(x_i) \rightarrow (6)$$

$P(x_i)$ is independent of h so we will drop it, for rest of the equation by applying log transformation.

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} d_i \cdot h(x_i) + (1-d_i) \log(1-h(x_i)) - (7)$$

Equation 7 describes the Quantity that must be maximised inorder to obtain the maximum likelihood hypothesis for the given problem setting.

* imp Naive Bayes Classifier.

objective: predict class of the new sample (same as KNN)

It is a highly practical bayesian learning method. The objective of Naive Bayes Classifier is to predict the class of the new sample. It Applies to learning task where each instance x is described by Conjunction of Attribute values and where the target function $f(x)$ can take on any value from some finite set of V where we can represent dataset as a conjunction of attributes.

29/01/24

Tasks → <a_1 a_2 a_3 a_4>

if $a_1 \wedge a_2 \wedge a_3 \wedge a_4$ then ES = Yes/No

(finite set collect^n of target values) V = {contains all target values yes no}

"Naive Bayes Classifier always predicts the most probable target value. (V_{MAP})

$$v_{map} = \underset{h \in H}{\text{argmax}}\ P\left(\frac{v_j}{a_1, a_2, a_3, \ldots a_n}\right)$$

A set of training examples of and the new instances describe th

A set of training examples of target function is provided and a new instance is described by the tuple of attribute values <a_1 a_2 a_3 ... a_n> the learner is asked to predict the value of the target class for the new instance. The Bayesian approach for classifying the new approach is to assign the most probable target value denoted as V_{MAP}, given the attribute values <a_1 a_2 ... a_n> that describe the new instance. We can write V_{MAP} as:

$$V_{MAP} = \underset{v \in V}{\text{argmax}}\ P(v_j / a_1\ a_2 \ldots a_n)$$

We can use Baye's theorem to rewrite this expression as follows:

$$V_{MAP} = \underset{v_j \in V}{\text{argmax}}\ \frac{P(a_1\ a_2 \ldots a_n / v_j) \times P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \; P(a_1, a_2 \ldots a_n / v_j) \times P(v_j) \rightarrow (1)$$

we can attempt to calculate the second term in equation (1) i.e it is easy to estimate each of ~~Vmap = arg~~ the $P(v_j)$ by counting the frequency with which each target value ~~is~~ v_j ~~how to~~ occurs in the training data, and we need to know how is to estimate the first term.

The Naive Baye's classifier is based on the ~~we can attempt to calculate~~ assumption that the attribute values are conditionally independent given the target value in other words the assumption is that given the target value of the instance the probability of the conjunction $\langle a_1, a_2 \ldots a_n \rangle$ is just the product of the probabilities of the individual attributes i.e $P(a_1, a_2 \ldots a_n / v_j)$ can be written as $P(a_1, a_2 \ldots a_n / v_j) = \prod_i (a_i / v_j) \rightarrow (2)$

substitute (2) in (1), we have the approach used by Naive bayes Classifier as follows.

$$V_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \; P(v_j) \times \prod_i (a_i / v_j) \rightarrow (3)$$

where V_{NB} denotes the target value o/p by the Naive Bayes classifier.

$$V_{NB} = \underset{v_j \in V}{\text{argmax}}\ P(v_j) \times \prod_i (a_i / v_j)$$

Example-1

* Playtennis dataset
new instance = (outlook = sunny, temp = cool, humidity = high, wind = strong)

$$v_{NB} = \underset{v_j \in \{Yes, No\}}{\text{argmax}}\ P(v_j) \times P(outlook = sunny / v_j) * \cancel{P(temp = cool/v_j) * P(humidity = high/v_j)} * P(wind = strong / v_j)$$

$$v_{NB}(yes) = \underset{v_j \in yes}{\text{argmax}}\ P(yes) * P(sunny / yes) * P(cool/yes) * P(high/yes) \times P(strong / yes)$$

$$= \frac{9}{14} * \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9}$$

$\cancel{(0.642)(0.142)(0.214)(0.214)(0.214)}$

$(0.642)(0.23)(0.34)(0.34)(0.34)$

$= 0.005$

$$V_{NB}(No) = \underset{v_j \in No}{\text{argmax}}\ P(No) * P(Sunny / No) * P(Cool / No) * P(high / No) * P(strong / No)$$

$$= \frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5}$$

$(0.357)(0.6)(0.2)(0.8)(0.6)$

$= 0.02$

CALEES

$V_{NB}(No) > V_{NB}(Yes)$

hence new instance is classified as No.

Example-2

#	Color	legs	height	Smelly	Species
1	white	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	white	2	Short	Yes	H
8	white	2	Tall	No	H

$P(Spe = M) = 4/8$

$P(Spe = H) = \frac{4}{8}$

New instance = Color = Green, legs = 2, Height = Tall, Smelly = No)

$$V_{NB} = \underset{v_j \in (M,H)}{\text{argmax}} P(v_j) \times P(Color = Green / v_j) \times P(legs = 2 / v_j) \times P(Height = Tall / v_j) \times P(Smelly = No / v_j)$$

$$V_{NB} = \underset{v_j \in M}{\text{argmax}} P(M) \times P(Green / M) \times P(2 / M) \times P(Tall / M) \times P(No / M)$$

$$\frac{4}{8} \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$$

$$(0.5)(0.5)(0.25)(0.25)(0.25)$$

$$= 0.0039$$

$$V_{NB} = \underset{v_j \in H}{\text{argmax}} P(H) \times P(Gr / H) P(2 / H) P(Tall / H) P(No / H)$$

$$= \frac{4}{8} \times \frac{1}{4} \times \frac{4}{4} \times \frac{2}{4} \times \frac{3}{4}$$

$$(0.5)(0.25)(1)(0.5)(0.75) = 0.0468$$

$V_{NB}(H) > V_{NB}(M)$

new instance is classified as H.

Example-3

#	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4.	Yellow	Sports	Domestic	No
5.	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes
.	New	Red,	SUV,	Domestic.

$P(Yes) = \frac{5}{10} = \frac{1}{2}$

$P(No) = \frac{5}{10} = \frac{1}{2}$

$$V_{NB} = \underset{v_j \in Yes}{\text{argmax}}\ P(Yes) \cdot P(Red/Yes) \cdot P(SUV/Yes) \cdot P(Domestic/Yes)$$

$$= \frac{5}{10} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{2}{5}$$

$$(0.5)(0.6)(0.2)(0.4) = 0.024$$

$$V_{NB} = \underset{v_j \in No}{\text{argmax}}\ P(No) \cdot P(Red/No)\, P(SUV/No) \cdot P(Domestic/No)$$

$$= \frac{5}{10} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5}$$

$$= (0.5)(0.4)(0.6)(0.6) = 0.072$$

$V_{NB}(No) > V_{NB}(Yes)$

$\therefore$ Hence new instance is classified as No.

Example - 4

#	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

$P(+) \frac{5}{10}$ $P(-) = \frac{5}{10}$

new inst (0 1 0)

$P(+) \; (0/+) \cdot P(1/+) \quad P(0/+)$

$V_{NB} = \frac{5}{10} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{1}{5}$

$= (0.5)(0.4)(0.2)(0.2) = 0.008$

$V_{NB} = P(-) P(0/-) \; P(1/-) \; P(0/-)$

$(0.5) \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot 0 = 0$

$V_{NB}(+) > V_{NB}(-)$

Hence new instance can be classified as '+'

Example - 5 $\qquad P(+) = \frac{4}{7} \cdot \; P(-) = \frac{3}{7}$

new instance (Japan, Audi, Red, 1990, Sports)

$V_{NB}(+) = \frac{4}{7}\left(\frac{4}{4}\right)(0)(0)\left(\frac{1}{4}\right)(0) = 0$

$V_{NB}(-) = \frac{3}{7}\left(\frac{1}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)(0)\left(\frac{1}{3}\right) = 0$

$V_{NB}(+) = V_{NB}(-) = 0$

So we can choose arbitrarily either + or −

01/02/2024

most imp ★ Baycsion Belief Network.

Consider a set of Random variables $Y_1\ Y_2 \ldots Y_n$ where each variable Y_i can taken on a set of possible values $V(Y_i)$ we define the joint space of the set of variables Y, ~~as the~~ to be the cross product of $V(Y_1) \times V(Y_2) \ldots \times V(Y_n)$
The probability distribution over this joint space is joint " — "

A Bayesian Belief network describes the joint probability distribution for the set of variables.

Conditional Independence.

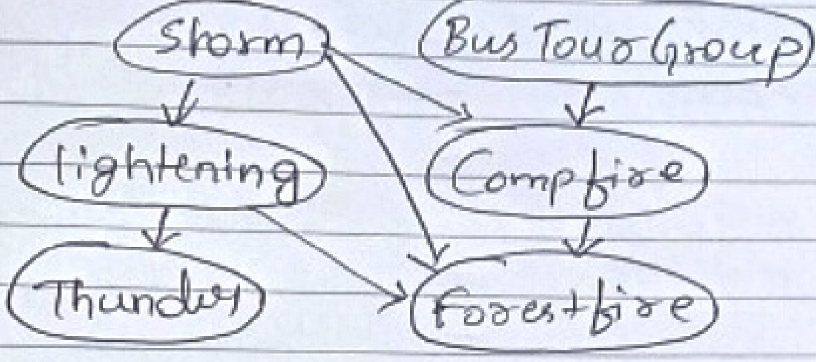
Let x y z be three discrete valued random variables. we say that x is conditionally independent y given z, if the probability distribution learning X is independent of the value of Y given a value for z, i.e if (for all) $\forall$ ~~$x_i\, y_j\, k$~~ x_i, y_j, z_k. if $P(X = x_i / Y = y_i \mid Z = Z_k)$ then $P(X = x_i / Z = Z_k)$
where $x_i \in V(x)$, $y_j \in V(Y)$, $Z_k \in V(Z)$.

We can rewrite this expression as
$$P(X/Y, Z) = P(X/Z)$$

This can also be extended to a set of variables:

$$P(X_1\ X_2 \ldots X_m / Y_1, Y_2 \ldots Y_n, Z_1, Z_2 \ldots Z_l) = P(X_1 \ldots X_m / Z_1 \ldots Z_l)$$

Representation of Bayesian Belief Network.
BBN represents the joint PD for a set of variables



The Bayesian n/w in the figure represents the joint probability Dist over the Boolean Variables, storm, lightining thunder, Bustour Group Campfire and forestfire.

Each variable in the joint space ~~bu~~ is represented by a node in the network. for each variable 2 types of information are specified.

① -> The network arcs represent that the variable is conditionally independent of its non-descendent in the n/w, Given its immediate predecessor in the n/w. we say that x is a descendent of y, if there is a directed path or arc from y to x.

② -> A Condiⁿ Prob table is given for each variable describing the PD for that variable Given the values of its immediate predecessor

2/2/24 In this fig consider the node campfire the n/w nodes and arcs represent. the campfire is conditionally independent of its non-descendents lighting and thunder, given its immedi