LECTURE NOTES

Campus: PCE Course: BTECH Class/Section: IIIrd year C Date: 23-01-21

Name of Faculty: Praveen Kr. Yadav. Name of Subject: Machine Learning Code: 6CS-02

Date (Prep.): 23-01-21 Date (Del.): 4-2-21 Unit No./Topic: 1 Lect. No: L6-L7

OBJECTIVE: (To be written before taking the lecture (Pl. write in bullet point, the main topics/concepts etc. which will be taught in the lecture)

Naive classifier

KNN Algorithm

IMPORTANT & RELEVANT QUESTIONS:

1. what is Naive classifier? How it will used to classify the dataset.

FEED BACK QUESTIONS (AFTER 20 MINUTES):

1. How to choose the optimal value of K in KNN Algorithm.

OUTCOME OF THE DELIVERED LECTURE: To be written after taking the lecture (Pl. write in bullet points about students' feedback on this lecture, level of understanding of this lecture by students etc.)

REFERENCES: Text/Ref. Book with Page No. and relevant Internet Websites:

Introduction to Machine Learning. Navathe

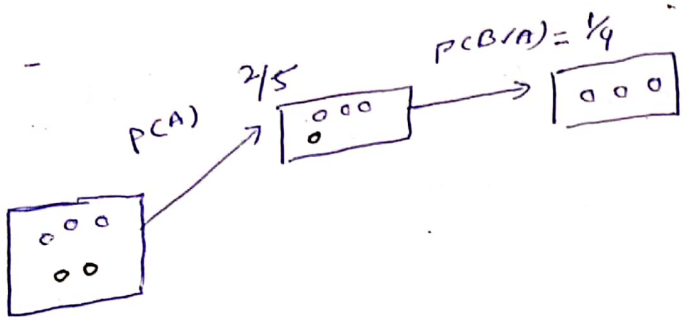
Naive Bayes classifier :-

Bayes Theorem :-

Conditional Probability - $P(A/B) = \frac{P(A \cap B)}{P(B)}$

Indepdent Event

dependent " -



A - event 1

B - event 2

$P(B) = 2/5$

$P(B/A) = 1/4$

$P(A \cap B) = 2/5 * 1/4 = 1/10$

$P(B/A) = 1/10 / 2/5$

$P(B/A) = 1/4$

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

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

$P(A \cap B) = P(B \cap A)$

$P(A \cap B) = P(A/B) * P(B)$

$P(B \cap A) = P(B/A) * P(A)$

$P(A/B) * P(B) = P(B/A) * P(A)$

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

$P(B/A)$ → Likelyhood

$P(A)$ → prior prob.

$P(A/B)$ → posterior prob

$P(B)$ → Marginal prob.

Naive Bayes classifier:-

dataset

$x = \{x_1, x_2, x_3 \ldots x_n\} \{y\}$

features — output labels

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

(A)

$P(y/x_1, x_2 \ldots x_n) = \frac{P(x_1/y) P(x_2/y) P(x_3/y) \ldots P(x_n/y) * P(y)}{P(x_1) P(x_2) P(x_3) \ldots P(x_n)}$

(B)

$= \frac{P(y) \prod_{i=1}^{n} P(x_i/y)}{P(x_1) P(x_2) \ldots P(x_4)}$ (constant)

$P(y/x_1, x_2 \ldots x_n) \propto P(y) \prod_{i=1}^{n} P(x_i/y)$

$y = \arg\max P(y) \prod_{i=1}^{n} (P(x_i/y)$

Outlook	Yes	No	P(Y)	P(N)
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(Y)	P(N)
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%

play		P(Y)	P(N)
yes	9	9/14	5/14
No	5	100%	
Total	14		

Today (Sunny, Hot)
x_1 x_2

$$P(yes/Today) = \frac{P(sunny/yes) * P(Hot/yes) * P(yes)}{P(today)}$$

$$= 2/9 * 2/9 * 9/14 = 0.031$$

$$P(No/Today) = 3/5 * 2/5 * 5/14 = 0.0857$$

$$P(yes) = 0.031 / 0.031 + 0.0857 \approx 0.27$$

$$P(N) = 1 - P(yes) = 0.73$$

O/P = No

K Nearest Neighbours Algorithm:-

KNN is supervised machine learning algorithm that can be used to solve both classification and regression problem.

↳ KNN algorithm assumes that similar things exist in close proximity. (means near to each other). So KNN captures the idea of similarity (distance, proximity or closeness) with some mathematics. (calculating the distance of a new data point with nearest points). ($c^2 = a^2 + b^2$)

↳ KNN classified a data point based on how its neighbours are classified.

Algorithm:-

1. Load the data (store All data)

Prediction Algo^m:-

1. calculate the distance from x to all points in your data.
2. sort the points in your data by increasing distance from x.
3. predict the majority label of the k closest points.

Note- choosing a k will effect what a new point is assigned to.

L→ KNN stores all available cases and classified new cases based on a similarity measure

How do we choose the factor K :-

KNN Algorithm is based on feature similarity. choosing the right value of K is a process called as parameter tuning and it is important for better accuracy.

- To choose a value of K -
  - sqrt(n), where n is the total number of data points.
  - odd value of K is selected to avoid confusion between two classes of data.

eg-

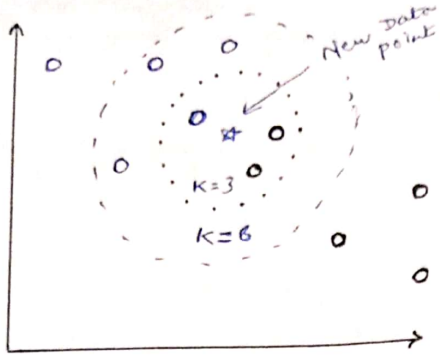
weight	Height	class	ED
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
	170	Normal	2

K=3

x_1	y_1	?
57	170	?

- we have to classify the below set as Normal or underweight.

entropy is zero.



o - class A

o - class B

PAGE NO.

To Find the nearest neighbour, we will calculate Euclidean distance.

$$dis(d) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

So majority neighbours are pointing towards 'normal'. Hence as per KNN Algo'n the class of (57, 170) should be normal.

Summary -

- a positive integer k is specified along with a new sample.
- we select the k-entries in our database which are closest to new sample.
- we find the most common classification of these entries
- This is the classification we give to new sample.