

02

2018

January

Tuesday

December 2017

wk	M	T	W	T	F	S	S
49				1	2	3	
50	4	5	6	7	8	9	10
51	11	12	13	14	15	16	17
52	18	19	20	21	22	23	24
53	25	26	27	28	29	30	31

January 2018

wk	M	T	W	T	F	S	S
01	1	2	3	4	5	6	7
02	8	9	10	11	12	13	14
03	15	16	17	18	19	20	21
04	22	23	24	25	26	27	28
05	29	30	31				

Q.1) We can apply Naive Bayesian classifier to map input tuples into accurate classes

Hence, with the help of the dataset we need to find all the prior & posterior probabilities.

For prior probabilities →

[From dataset]

$$P(\text{on time}) = \frac{14}{20}$$

$$P(\text{late}) = \frac{2}{20}$$

$$P(\text{very late}) = \frac{3}{20}$$

$$P(\text{cancelled}) = \frac{1}{20}$$

Calculating posterior probabilities →
For attribute 'Day' -

$$P(\text{weekday} | \text{on time}) = \frac{9}{14}$$

$$P(\text{weekday} | \text{very late}) = \frac{3}{3}$$

$$P(\text{weekday} | \text{late}) = \frac{1}{2}$$

$$P(\text{weekday} | \text{cancelled}) = \frac{0}{1}$$

February 2018							March 2018								
02	M	T	W	T	F	S	03	Wk	M	T	W	T	F	S	S
04				1	2	3	4	09				1	2	3	4
05	5	6	7	8	9	10	11	10	5	6	7	8	9	10	11
06	12	13	14	15	16	17	18	11	12	13	14	15	16	17	18
07	19	20	21	22	23	24	25	12	19	20	21	22	23	24	25
08	26	27	28					13	26	27	28	29	30	31	
09															

01
2018
January
Wednesday

003-362 | 03-01-2018

03

Similarly calculating and tabulating posterior probabilities for all attributes class

Day	on Time	Late	Very Late	Cancelled
Weekday	9/14	1/2	3/3	0/1
Saturday	2/14	0/2	0/3	1/1
Sunday	1/14	0/2	0/3	0/1
Holiday	2/14	1/2	0/3	0/1

for attribute 'Season' -
class

season	on Time	Late	Very Late	Cancelled
Spring	4/14	0/2	0/3	1/1
Summer	6/14	0/2	0/3	0/1
Autumn	2/14	0/2	1/3	0/1
Winter	2/14	2/2	2/3	0/1

for attribute 'Fog' -
class

Fog	on Time	Late	Very Late	Cancelled
None	5/14	0/2	0/3	0/1
High	4/14	1/2	1/3	1/1
Normal	5/14	1/2	2/3	0/1

for attribute 'Rain' -
class

Rain	on Time	Late	Very Late	Cancelled
None	6/14	1/2	1/3	0/1
Slight	6/14	1/2	0/3	0/1
Heavy	2/14	0/2	2/3	1/1

04

2018

January

Thursday

December 2017						
12	M	T	W	T	F	S
49				1	2	3
50	4	5	6	7	8	9
51	11	12	13	14	15	16
52	18	19	20	21	22	23
53	25	26	27	28	29	30

January 2018						
01	M	T	W	T	F	S
01	1	2	3	4	5	6
02	8	9	10	11	12	13
03	15	16	17	18	19	20
04	22	23	24	25	26	27
05	29	30	31			

For the instance < weekday, winter, high, None >

$$P_{NB}(\text{ontime}) = P(\text{ontime}) \times P(\text{weekday} | \text{ontime}) \\ \times P(\text{winter} | \text{ontime}) \times P(\text{high} | \text{ontime}) \\ \times P(\text{None} | \text{ontime})$$

$$= \frac{14}{20} \times \frac{9}{14} \times \frac{2}{14} \times \frac{4}{14} \times \frac{6}{14} = 0.0079$$

Similarly,

$$P_{NB}(\text{Late}) = \frac{2}{20} \times \frac{1}{2} \times \frac{2}{2} \times \frac{1}{2} \times \frac{1}{2} = 0.0125$$

$$P_{NB}(\text{Very Late}) = \frac{3}{20} \times \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = 0.0111$$

$$P_{NB}(\text{Cancelled}) = \frac{1}{20} \times \frac{0}{1} \times \frac{0}{1} \times \frac{4}{1} \times \frac{0}{1} = 0$$

$P_{NB}(\text{Late})$ is highest, hence correct classification is 'Late'.

Similarly, we can classify any input tuple into accurate class

February 2018							March 2018						
Wk	M	T	W	T	F	S	Wk	M	T	W	T	F	S
01				1	2	3	08				1	2	3
04	5	6	7	8	9	10	10	5	6	7	8	9	10
07	12	13	14	15	16	17	11	12	13	14	15	16	17
10	19	20	21	22	23	24	12	19	20	21	22	23	24
13	26	27	28	29	30	31	13	26	27	28	29	30	31

Q.2) Here, we have to test hypothesis that gender and preferred reading are independent, that is there is no correlation between them.

We can use, χ^2 correlation test with contingency table of size 2×2 (given) and $(2-1) \times (2-1)$ degrees of freedom.

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$o_{ij} \rightarrow$ observed frequency
 $e_{ij} \rightarrow$ expected frequency.

$$\therefore \chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = \underline{\underline{507.9365}}$$

For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at 0.01 significance level is ~~10.828~~ 6.635.

Since our computed value is above this, we can reject that gender and preferred reading are independent and conclude that the two attributes are correlated for the given group of people.