18CSE751 – Introduction to Machine Learning

Department of Computer Science and Engineering

Nitte Meenakshi Institute of Technology, Bangalore

Naïve Bayes Classifier

1. Consider the below dataset on "Transportation mode".

Gender	Cars Owned	Travel Cost	Income Level	Transport Mode
Male	None	Cheap	Low	Bus
Male	One	Cheap	Medium	Bus
Female	None	Cheap	Low	Bus
Male	One	Cheap	Medium	Bus
Female	One	Expensive	High	Car
Male	Two	Expensive	Medium	Car
Female	Two	Expensive	High	Car
Female	One	Cheap	Medium	Train
Male	None	Standard	Medium	Train
Female	One	Standard	Medium	Train

By using the **naïve Bayes classifier**, classify the given test data

Gender	Cars Owned	Travel Cost	Income Level	Transport Mode
Male	None	Cheap	Medium	Bus
Female	Two	Expensive	High	Car

P(Male,TM=Bus)	0.75	P(Male,TM=Car)	0.333333333
P(Female,TM=Bus)	0.25	P(Male,TM=Train)	0.333333333
P(CarsOwned=None,TM=Bus)	0.5		
P(CarsOwned=Two,TM=Bus)	0		
P(TravelCost=Cheap,TM=Bus)	1		
P(TravelCost=Expensive,TM=Bus)	0		
P(IncomeLevel=High,TM=Bus)	0		
P(IncomeLevel=Medium,TM=Bus)	0.25		
P(Female,TM=Car)	0.66666667		
P(Female,TM=Train)	0.66666667		
0	P(CarsOwned=Two,TM=Train)	0	
0.333333333	P(CarsOwned=Two,TM=Car)	0.66666667	
0	P(TravelCost=Cheap,TM=Train)	0.33333333	
1	P(TravelCost=Expensive,TM=Train)	0	
0.666666667	P(IncomeLevel=High,TM=Train)	0	
0.333333333	P(IncomeLevel=Medium,TM=Train)	1	

Course Coordinator: Dr.Vani Vasudevan

P(CarsOwned=None,TM=Car)	0
P(CarsOwned=None,TM=Train)	0.66666667
P(TravelCost=Cheap,TM=Car)	0.33333333
P(TravelCost=Expensive,TM=Car)	0
P(IncomeLevel=High,TM=Car)	0
P(IncomeLevel=Medium,TM=Car)	1

P(Male,TM=Bus)*P(None,TM=Bus)*P(Cheap,TM=Bus)*P(Medium,TM=Bus) *P(Bus)	0.09375*0.4
P(Male,TM=Car)*P(None,TM=Car)*P(Cheap,TM=Car)*P(Medium,TM=Car) *P(Car)	0*0.3
P(Male,TM=Train)*P(None,TM=Train)*P(Cheap,TM=Train)*P(Medium,TM=Train)*P(Train)	0.035937*0.4

P(Female,TM=Bus)*P(Two,TM=Bus)*P(Expensive,TM=Bus)*P(High,TM=Bus)* P(Bus)	0
P(Female,TM=Car)*P(Two,TM=Car)*P(Expensive,TM=Car)*P(High,TM=Car) *P(Car)	0.300763*0.3
P(Female,TM=Train)*P(Two,TM=Train)*P(Expensive,TM=Train)*P(High,TM=Train)*P(Train)	0

2. Consider the below dataset on “car theft”.

ID	Colour	Type	Mileage	Origin	Stolen?
1	Red	Sports	18	Domestic	Yes
2	Red	Sports	27	Domestic	No
3	Red	Sports	45	Domestic	Yes
4	Yellow	Sports	89	Domestic	No
5	Yellow	Sports	25	Imported	Yes
6	Yellow	SUV	32	Imported	Yes
7	Yellow	SUV	74	Domestic	No
8	Red	SUV	24	Imported	No
0	Red	Sports	15	Domestic	Yes
10	Green	Sports	34	Imported	Yes
11	Green	SUV	32	Domestic	No
12	Green	SUV	67	Imported	No

By using the naïve Bayes classifier, classify the given test data

ID	Colour	Type	Mileage	Origin Stolen	
13	Red	SUV	54	Domestic	No
14	Green	Sports	45	Imported	No

Naïve Bayes :

Course Coordinator: Dr.Vani Vasudevan

P(Colour=Red|Yes) = 3/6 = 0.5
P(Colour = Red|No) = 2/6 = 0.33
P(Colour = Yellow|Yes)= 2/6 = 0.33
P(Colour=Yellow|No) =2/6 =.33
P(Colour= Green|No) = 2/6 =.33
P(Colour=Green|Yes) = 1/6 = 0.167
P(Type = Sports|Yes)=5/6 = .83
P(Type = Sports|No) =2/6 =.33
P(Type= SUV|Yes) =1/6 = .167
P(Type=SUV|No) = 4/6 = 0.67
P(Origin=Imported|No) = 2/6 = .33
P(Origin= Imported|Yes) = 3/6 = .5
P(Origin = Domestic|Yes) = 3/6 =.5
P(Origin = Domestic|No)= 4/6 =0.67

For Mileage, Use normal Distribution
If stolen= yes, Sample Mean = 28.16 Sample variance = 66.22
If stolen = no, Sample Mean = 75.6 Sample variance =652.87

P(Mileage = 54 | Yes = $\frac{1}{\sqrt{2*\pi*66.22}} e^{-(\frac{(54-28.16)^2}{2*66.22}}$
= 0.0490*6.4636x10^-3 = 3.1672x10^-4

P(Mileage = 54 | No) = $\frac{1}{\sqrt{2*\pi*652.87}} e^{-(\frac{(54-75.6)^2}{2*652.87}}$
= 0.0156*0.6996 = 0.0109

P(Mileage =45|Yes = $\frac{1}{\sqrt{2*\pi*66.22}} e^{-(\frac{(45-28.16)^2}{2*66.22}}$
= 0.0490*0.1175 = 5.7580x10^-3

P(Mileage=45|No = $\frac{1}{\sqrt{2*\pi*652.87}} e^{-(\frac{(45-75.6)^2}{2*652.87}}$
=0.0156*0.4882 = 7.6153x10^-3

1) Test instance

P(X|stolen=Yes) = P(Colour=Red|stolen=Yes) * P(Type=SUV|stolen=Yes) * P(Mileage =54|Yes) * P(Origin=Domestic|Yes)
= **0.5 * 0.167 * 3.1672x10^-4 *0.5 = 1.322306x10^-5 = 0.00001322**
P(X|stolen=No) = P(Colour=Red|stolen= No) * P(Type=SUV|stolen= No) * P(Mileage =54| No) * P(Origin=Domestic| No)
=0.33*0.67*0.67*0.0109 = 0.0016

Conclusion:
P(X|stolen=No) > P(X|stolen=Yes) ; Hence, Stolen = No

ID	Colour	Type	Mileage	Origin	Stolen
1	Red	SUV	54	Domestic	No

2) Test instance

P(X|Yes) = P(Colour =Green |Yes)*P(Type= Sports |Yes)*P(Mileage=45|Yes) * P(Origin=Imported|Yes)

=0.167*0.83*5.7580x10^-3 *0.5 = 3.991*10^-4 = 0.000399
P(X|No) = P(Colour =Green | No)*P(Type= Sports | No)*P(Mileage=45| No) * P(Origin=Imported| No)
=0.33*0.33*0.33*7.6153x10^-3 = 7.539*10^-3 =0.007539

Conclusion

0.0075v>0.000399; Hence, Stolen = No

ID	Colour	Type	Mileage	Origin	Stolen
2	Green	Sports	45	Imported	No