Assignmnet_4_Aparna Bharathi Suresh

1. (15%) Chapter 6, exercise 2

Assignment-4

1. Chapter-6, Ex-2

a) $P(\text{vomiting} = \text{true}) = \frac{6}{10} = 0.6$

b) $P(\text{Headache} = \text{false}) = \frac{3}{10} = 0.3$

c) $P(\text{Headache} = \text{true}, \text{Vomiting} = \text{false}) = \frac{1}{10} = 0.1$

d) $P(\text{vomiting} = \text{false} \mid \text{Headache} = \text{true}) = \frac{1}{7} = 0.1428$

e) $P(\text{Meningitis} \mid \text{Fever} = \text{true}, \text{Vomiting} = \text{false})$

$\Rightarrow P(\text{Meningitis} = \text{true} \mid \text{Fever} = \text{true}, \text{vomiting} = \text{False})$

$$= \frac{1}{4}$$

$$= 0.25$$

$\Rightarrow P(\text{Meningitis} = \text{false} \mid \text{Fever} = \text{true}, \text{vomiting} = \text{false})$

$$= \frac{3}{4}$$

$$= 0.75$$

$P(\text{Meningitis} \mid \text{Fever} = \text{true}, \text{vomiting} = \text{false}) = (0.25, 0.75)$

2. (15%) Chapter 6, exercise 3

2. Chapter- 6, Ex-3

In target feature there are 3 classes.

OK ⇒ 4 records
Settler ⇒ 5 records
Solids ⇒ 4 records

13 records

Prior Probability:

$$P(Status = OK) = \frac{4}{13} = 0.3077$$

$$P(Status = Settler) = \frac{5}{13} = 0.3846$$

$$P(Status = Solids) = \frac{4}{13} = 0.3077$$

Probability Distribution fn for Normal Distribution.

$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

P(SS-IN | Status = OK):

$$\mu = \frac{168 + 156 + 176 + 256}{4}$$

$$= 189$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(168-189)^2+(156-189)^2+(176-189)^2+(256-189)^2}{3}}$$

$$= \sqrt{2062.666}$$

$$= 45.42$$

P(SED-IN|OK):

$$\mu = \frac{3+3+3.5+3}{4}$$

$$= 3.125$$

$$\sigma = \sqrt{\frac{(3-3.125)^2+(3-3.125)^2+(3.5-3.125)^2+(3-3.125)^2}{3}}$$

$$= \sqrt{0.0625}$$

$$= 0.25$$

P(COND-IN|Status=OK):

$$\mu = \frac{1814+1358+2200+2070}{4}$$

$$= 1860.5$$

$$\sigma = \sqrt{\frac{(1814-1860.5)^2 + (1358-1860.5)^2 + (2200-1860.5)^2 + (2070-1860.5)^2}{3}}$$

$$= \sqrt{137939.6667}$$

$$= 371.4$$

P(SS-OUT|Status = OK):

$$M = \frac{15+14+16+27}{4}$$

$$= 18$$

$$\sigma = \sqrt{\frac{(15-18)^2 + (14-18)^2 + (16-18)^2 + (27-18)^2}{3}}$$

$$= \sqrt{36.6666}$$

$$= 6.0553$$

P(SED-OUT|Status = OK):

$$M = \frac{0.001+0.01+0.005+0.2}{4}$$

$$= 0.054$$

$$\sigma = \sqrt{\frac{(0.001-0.054)^2 + (0.01-0.054)^2 + (0.005-0.0054)^2 + (0.2-0.054)^2}{3}}$$

$$= \sqrt{9.4873 \times 10^{-3}}$$

$$= 0.09740$$

P (COND - OUT | Status = OK):

$$\mu = \frac{1879 + 1425 + 2140 + 2700}{4}$$

$$= 2036$$

$$\sigma = \sqrt{\frac{(1879-2036)^2 + (1425-2036)^2 + (2140-2036)^2 + (2700-2036)^2}{3}}$$

$$= \sqrt{283227.333}$$

$$= 532.1910$$

P (SS-IN | Status = Settler) =

$$\mu = \frac{230 + 116 + 242 + 242 + 174}{5}$$

$= 200.8$

$$\sigma = \sqrt{\frac{(230-200.8)^2 + (116-200.8)^2 + (242-200.8)^2 + (242-200.8)^2 + (174-200.8)^2}{4}}$$

$= \sqrt{3039.2}$

$= 55.1289$

P(SED-IN|Status = Settler):

$$\mu = \frac{5+3+7+4.5+2.5}{5} = 4.4$$

$$\sigma = \sqrt{\frac{(5-4.4)^2 + (3-4.4)^2 + (7-4.4)^2 + (4.5-4.4)^2 + (2.5-4.4)^2}{4}}$$

$= 1.78$

P(COND-IN|Status = Settler):

$$\mu = \frac{1410+1238+1315+1183+1110}{5}$$

$= 1251.2$

$$\sigma = \sqrt{\frac{(1410-1251.2)^2+(1238-1251.2)^2 + (1315-1251.2)^2 + (1183-1251.2)^2 + (1110-1251.2)^2}{4}}$$

$$= 116.2441$$

P(SS-OUT | Status = Settler):

$$\mu = \frac{131+104+104+78+73}{5}$$

$$= 98$$

$$\sigma = \sqrt{\frac{(131-98)^2+(104-98)^2+(104-98)^2+(78-98)^2+(73-98)^2}{4}}$$

$$= \sqrt{546.5}$$

$$= 23.3773$$

P(SED-OUT | Status = Settler):

$$\mu = \frac{3.5+0.06+0.01+0.02+1.5}{5}$$

$$= 1.018$$

$$\sigma = \sqrt{2.3409}$$

$= 1.53$

P(COND-OUT | Status = Settler):

$$\mu = \frac{1575 + 1221 + 1434 + 1374 + 1256}{5}$$

$= 1372$

$\sigma = \sqrt{20329.0564}$

$= 142.58$

P(SS-IN | Status = Solids):

$$\mu = \frac{1804 + 1228 + 964 + 2008}{4}$$

$= 1301$

$\sigma = \sqrt{235651.9936}$

$= 485.44$

P(SED-IN | Status = Solids):

$\mu = 32.5$

$\sigma = 11.96$

P(COND-IN | Status = Solids):

$\mu = 1621$

$\sigma = 453.04$

P(SS-OUT | Status = Solids):

$\mu = 49.1$

$\sigma = 37.76$

P(SED-OUT | Status = Solids):

$\mu = 1,293$

$\sigma = 430.95$

P(COND-OUT | Status = Solids):

$\mu = 832.85$

$\sigma = 958.31$

2.b)

b) Query:

SS-IN => 222, SED-IN = 4.5,
COND-IN = 1518, SS-OUT = 74
SED-OUT = 0.25, COND-OUT = 1642

For Status = OK:

$P(OK) = 0.3077$

$P(SS\text{-}IN \mid OK)$:

$$N(222, \mu = 189, \sigma = 45.42)$$

$$= \frac{1}{45.42\sqrt{2\pi}} e^{-\left(\frac{(222-189)^2}{2\times(45.42)^2}\right)}$$

$$= 8.7834 \times 10^{-3} e^{-0.2639}$$

$$= 0.0067$$

$P(SED\text{-}IN \mid OK)$:

$$N(4.5, \mu = 3.125, \sigma = 0.25)$$

$$= \frac{1}{0.25\sqrt{2\pi}} e^{-\frac{(4.5-3.125)^2}{2\times(0.25)^2}}$$

$$= 1.5957 \times e^{-15.125}$$

$$= 4.3077 \times 10^{-7}$$

$P(COND\text{-}IN \mid OK)$:

$$N(1518, \mu = 1860.5, \sigma = 371.4)$$

$$= \frac{1}{371.4\sqrt{2\pi}} e^{-\frac{(1518-1860.5)^2}{2\times(371.4)^2}}$$

$= 0.0007$

$P(SS\text{-}OUT \mid OK) = N(74, \mu = 18, \sigma = 6.06)$

$$= \frac{1}{6.06\sqrt{2\pi}} e^{-\frac{(74-18)^2}{2\times(6.06)^2}}$$

$= 1.7650 \times 10^{-20}$

$P(SED\text{-}OUT \mid OK) = N(0.25, \mu = 0.054, \sigma = 0.10)$

$$= \frac{1}{0.10\sqrt{2\pi}} e^{-\frac{(0.25-0.054)^2}{2\times(0.10)^2}}$$

$= 0.5408$

$P(COND\text{-}OUT \mid OK) = N(1642, \mu = 2036, \sigma = 532.19)$

$$= \frac{1}{532.19\sqrt{2\pi}} e^{-\frac{(1642-2036)^2}{2\times(532.19)^2}}$$

$= 0.0006$

$$\left(\prod_{k=1}^{m} P(q[k] \mid OK)\right) \times P(OK) = 3.41577 \times 10^{-36}$$

For Status = Settler:

$$P(\text{Settler}) = 0.3846$$

$$P(\text{SS-IN}\mid\text{Settler}) = N(222, \mu = 200.8, \sigma = 55.13)$$

$$\frac{1}{55.13\sqrt{2\pi}} e^{-\frac{(222-200.8)^2}{2\times(55.13)^2}}$$

$$= 0.0067$$

$$P(\text{SED-IN}\mid\text{Settler}) = N(4.5, \mu = 4.4, \sigma = 1.78)$$

$$= 0.2235$$

$$P(\text{COND-IN}\mid\text{Settler}) = N(1518, \mu = 1251.2, \sigma = 116.24)$$

$$= 0.0002$$

$$P(\text{SS-OUT}\mid\text{Settler}) = N(94, \mu = 98, \sigma = 23.38)$$

$$= 0.0101$$

$$P(\text{SED-OUT}\mid\text{Settler}) = N(0.25, \mu = 1.018, \sigma = 1.53)$$

$$= 0.2303$$

$$P(\text{COND-OUT}\mid\text{Settler}) = N(1642, \mu = 1372, \sigma = 142.58)$$

$$= 0.0005$$

$$\left(\prod_{k=1}^{m} P(q[k] \mid \text{Settler})\right) \times P(\text{Settler})$$

$$= 1.53837 \times 10^{-13}$$

For Status = Solids:

$P(\text{Solids}) = 0.3077$

$P(\text{SS-IN} \mid \text{Solids}) = N(x, \mu = 1301, \sigma = 485.44)$

$$\frac{1}{485.44\sqrt{2\pi}} e^{-\frac{(222-1301)^2}{2 \times (485.44)^2}} = 6.9430 \times 10^{-5}$$

$P(\text{SED-IN} \mid \text{Solids}) = N(4.5, \mu = 32.5, \sigma = 11.96)$

$= 0.0022$

$P(\text{COND-IN} \mid \text{Solids}) = N(1518, \mu = 1621, \sigma = 453.04)$

$= 0.0009$

$P(\text{SS-OUT} \mid \text{Solids}) = N(74, \mu = 49.1, \sigma = 37.76)$

$= 0.0085$

$P(\text{SED-OUT} \mid \text{Solids}) = N(0.25, \mu = 1293, \sigma = 430.98)$

$= 1.0291 \times 10^{-5}$

$$P(\text{COND-OUT} \mid \text{Solids}) = N(1.642, \mu = 832.80, \sigma = 958.31)$$

$$= 0.0003$$

$$\left(\prod_{k=1}^{m} P(q[k] \mid \text{Solids})\right) \times P(\text{Solids})$$

$$= 1.00668 \times 10^{-21}$$

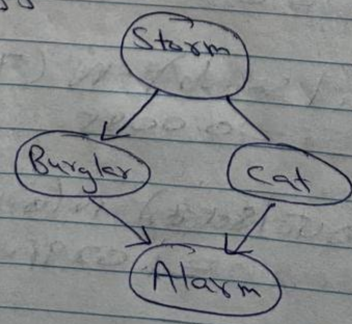
The score of each target class is a relative ranking.

The target with highest rank is Status = Settler. i.e there was a problem with the plant's settler equipment.

3. (15%) Chapter 6, exercise 4

3: Chapter-6, Ex-4

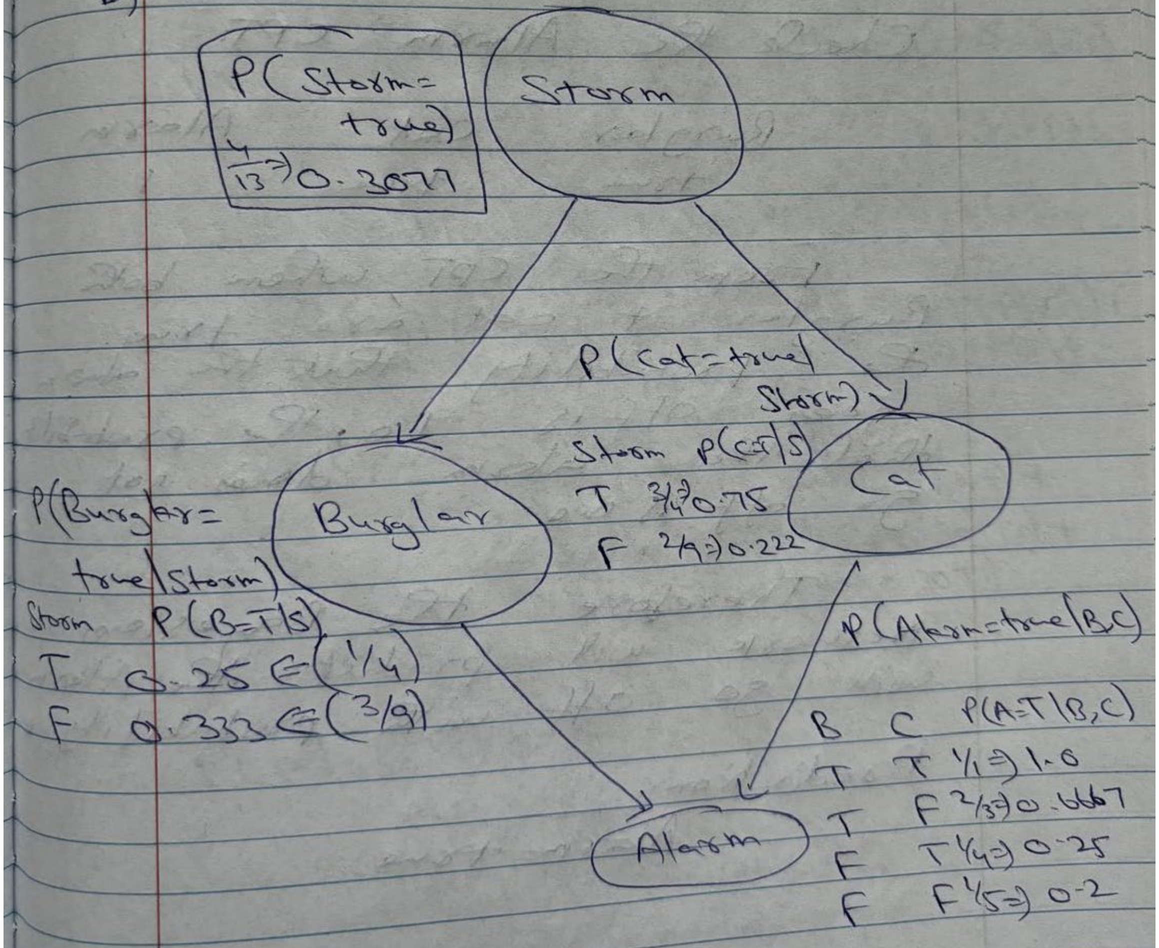
a) topology:



Storm influences both Burglar and Cat.

Burglar + cat influences Alarm.

b)



c) Given:

Burglar = true
Cat = true
Storm = false

Check the Alarm CPT:

Burglar	Cat	Alarm
true	true	1

From the CPT, when both Burglar + cat are true the probability that the alarm goes off is 1.0, the probability that the alarm does not go off is 0.0.

Therefore the Bayesian Network will predict the Alarm will go off with probability 1.0.

Prediction:

Alarm = true

d) Probability of Alarm = True:

$$P(a|s) = \frac{P(a,s)}{P(s)} = \sum_{i,j} \frac{P(a,B_i,C_j,S)}{P(S)}$$

$$\sum_{i,j} P(a,B_i,C_j,S) = \sum_{i,j} P(a|B_i,C_j) \times P(B_i|S) \times P(C_j|S) \times P(S)$$

$$= (P(a|b,c) \times P(b|s) \times P(c|s) \times P(s)) + P(a|b,\neg c) \times P(b|s) \times P(\neg c|s) \times P(s) + P(a|\neg b,c) \times P(\neg b|s) \times P(c|s) \times P(s)) + P(a|\neg b,\neg c) \times P(\neg b|s) \times P(\neg c|s) \times P(s))$$

$$= (1.0 \times 0.25 \times 0.75 \times 0.3077) + (0.6667 \times 0.25 \times 0.25 \times 0.3077) + (0.25 \times 0.75 \times 0.75 \times 0.3077) + (0.2 \times 0.75 \times 0.25 \times 0.3077)$$

$$= 0.125324$$

$$P(a|s) = \frac{P(a,s)}{P(s)}$$

$$= \frac{0.12534}{0.3077} = 0.4073$$

$\therefore P(\text{Alarm} = \text{false}) = 1 - 0.4073$
$= 0.5927$

Model will predict Alarm = false.

4. (15%) Chapter 6, exercise 5

4. Chapter-6, Ex-5

a) Sort the age:

18
19
21 } Young

43
47
49 } Middle-aged

51
55
57 } Mature

Threshold between young and middle aged:

$$\frac{21+43}{2} = 32$$

Threshold between middle-aged & mature:

$$\frac{49+51}{2} = 50$$

Transforming the age into categorical

ID	Occupation	Gender	Age	Policy	Pref
9	Nurse	Female	young	C	Phone
6	Manager	Male	young	A	email
3	Biophysicist	Male	young	A	email
1	labtech	Female	Middle	C	email
4	Sheriff	Female	Middle	B	Phone
7	geologist	Male	Middle	C	Phone
8	Messenger	Male	Mature	B	email
5	Painter	Male	Mature	C	Phone
2	farmhand	Female	Mature	A	Phone

b) Discriptive features to exclude:

ID → has unique value for each row
Occupation → has unique Value for each row

c) To train Naive Bayes Model we need:

Prior Probabilities of target
Conditional Probabilities of each feature

$$P(\text{phone}) = \frac{5}{9} = 0.5556$$

$$P(\text{email}) = \frac{4}{9} = 0.4444$$

Conditional Probabilities:

P(Gender=female|phone) = 0.6
P(Gender=male|phone) = 0.4
P(Age=young|phone) = 0.2
P(Age=MiddleAge|phone) = 0.4
P(Age=Mature|phone) = 0.4
P(Policy=PlanA|Phone) = 0.2
P(Policy=planB|phone) = 0.2
P(policy=PlanC|phone) = 0.6

P(Gender=female|Email) = 0.~~2~~5
P(Gender=male|email) = 0.75
P(Age=young|email) = 0.5
P(Age=Middle-Age|email) = 0.25
P(Age=Mature|email) = 0.25
P(Policy=PlanA|~~phone~~ email) = 0.5
P(policy=planB|email) = 0.25
P(Policy=PlanC|email) = 0.25

d) Query:

Gender=female, Age=30, Policy=PlanA

From Part A Age 30 is under young.

P(Channel = phone | q):

P(phone) = 0.56
P(Gender = female | phone) = 0.6
P(Age = young | phone) = 0.2
P(Policy = PlanA | phone) = 0.2

$$\left(\prod_{k=1}^{m} P(q[k] \mid phone)\right) \times P(phone)$$

$= 0.6 \times 0.2 \times 0.2 \times 0.56$

$= 0.01344$

P(channel = email | q):

$$\left(\prod_{k=1}^{m} P(q[k] \mid email)\right) \times P(email)$$

$= 0.25 \times 0.5 \times 0.5 \times 0.4444$

$= 0.027775$

Highest rank is Channel = email

This is the prediction

5. (15%) Chapter 6, exercise 6

5. Chapter-6, Ex-6

a) Machine Learning is fun.

$P(\text{entertainment}) = \frac{700}{1000} = 0.7$

w_k	Count	$P(w_k \mid \text{entertainment})$
Machine	70	$\frac{70}{700} = 0.10$
Learning	35	$\frac{35}{700} = 0.05$
is	695	$\frac{695}{700} = 0.99$
fun	415	$\frac{415}{700} = 0.593$

$P(\text{education}) = \frac{300}{1000} = 0.3$

w_k	Count	$P(w_k \mid \text{education})$
Machine	105	$\frac{105}{300} = 0.35$
Learning	120	$\frac{120}{300} = 0.40$
is	295	$\frac{295}{300} = 0.983$
fun	200	$\frac{200}{300} = 0.667$

P(entertainment|q) = P(entertainment) ×
P(machine|entertainment)
× P(learning|entertainment)
× P(is|entertainment)
× P(fun|entertainment)

$= 0.7 \times 0.583 \times 0.99 \times 0.5 \times 0.1$

$= 0.00205$

P(education|q) = P(education) ×
P(machine|education) ×
P(learning|education) ×
P(is|education) ×
P(fun|education)

$= 0.3 \times 0.667 \times 0.983 \times 0.4 \times 0.35$

$= 0.00275$

P(education|q) is > P(entertainment|q)

So the prediction is education

b) query:

Christmas family fun.
christmas does not appear in both doc.

P(christmas | entertainment) = 0

P(christmas | education) = 0

∴ Probability for both target levels will be 0.

Model will not return a prediction.

c) Smoothing: P(word | entertainment)

Raw Probabilities:

P(christmas | entertainment) = 0

P(family | entertainment) = $\frac{400}{700}$

= 0.5714

$P(\text{fun}|\text{entertainment}) = \frac{415}{700}$

$= 0.59285$

Smoothing Parameters:

$K = 10$

Count(entertainment) = 700

Count(christmas|entertainment) = 0

Count(family|entertainment) = 400

Count(fun|entertainment) = 415

Vocabulary = 6

Smoothed Probabilities:

P(christmas|entertainment):

$$\frac{0+10}{700+(10\times6)} = 0.0132$$

P(family|entertainment):

$$\frac{400+10}{700+(10\times6)} = 0.5395$$

P(fun|entertainment):

$$\frac{415+10}{700+(10\times6)} = 0.5592$$

Smoothing: P(word|education)

Raw
Probabilities:

P(christmas|education) = 0
P(family|education) = $\frac{10}{300}$

= 0.0333

P(fun|education) = $\frac{200}{300}$

= 0.6667

Smoothing Parameters:

k = 10
Count(education) = 300
Count(christmas|education) = 0
Count(family|education) = 10
Count(fun|education) = 200
Vocabulary = 6

Smoothed Probabilities:

P(christmas|education):

$$= \frac{0+10}{300+(10\times6)} = 0.02778$$

P(family|education).

$$= \frac{10+10}{300+(10\times6)} = 0.0556$$

$$P(\text{fun}|\text{education}) = \frac{200+10}{300+(10\times6)}$$

$$= 0.5833$$

Probability.

P(entertainment|q) = P(entertainment) x
P(christmas|entertainment) x
P(family|entertainment) x
P(fun|entertainment)

$$= 0.7\times0.0132\times0.5395\times0.5592$$

$$= 0.0028$$

$P(\text{education} \mid q) = P(\text{education}) \times$
$P(\text{christmas} \mid \text{education}) \times$
$P(\text{family} \mid \text{education}) \times$
$P(\text{fun} \mid \text{education})$

$= 0.3 \times 0.0278 \times 0.0556 \times 0.5833$
$= 0.0003$

$P(\text{entertainment} \mid q) > P(\text{education} \mid q)$

So the model will predict entertainment.