Tutorial 4 : Probability – Based learning – corrections

1. The table below gives details of symptoms that patients presented and whether they were suffering from meningitis.

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Using this dataset, calculate the following probabilities:

(a) $P(\text{VOMITING} = \textit{true})$

This can be calculated easily by counting:
$P(\text{VOMITING} = \textit{true}) = \frac{6}{10} = 0.6$

(b) $P(\text{HEADACHE} = \textit{false})$

This can be calculated easily by counting:
$P(\text{HEADACHE} = \textit{false}) = \frac{3}{10} = 0.3$

(c) $P(\text{HEADACHE} = \textit{true}, \text{VOMITING} = \textit{false})$

This can be calculated easily by counting:
$P(\text{HEADACHE} = \textit{true}, \text{VOMITING} = \textit{false}) = \frac{1}{10} = 0.1$
Or using the product rule:
$P(\text{HEADACHE} = \textit{true}, \text{VOMITING} = \textit{false}) = P(\text{HEADACHE} = \textit{true} \mid \text{VOMITING} = \textit{false}) \times P(\text{VOMITING} = \textit{false}) = \frac{1}{4} \times \frac{4}{10} = 0.1$

(d) $P(\text{VOMITING} = \textit{false} \mid \text{HEADACHE} = \textit{true})$

This can be calculated easily by counting:
$P(\text{VOMITING} = \textit{false} \mid \text{HEADACHE} = \textit{true}) = \frac{1}{7} = 0.1429$

(e) $P(\text{MENINGITIS} \mid \text{FEVER} = \textit{true}, \text{VOMITING} = \textit{false})$

This can be calculated easily by counting. First,
$P(\text{MENINGITIS} = \textit{true} \mid \text{FEVER} = \textit{true}, \text{VOMITING} = \textit{false}) = \frac{1}{4} = 0.25.$
Then,
$P(\text{MENINGITIS} = \textit{false} \mid \text{FEVER} = \textit{true}, \text{VOMITING} = \textit{false}) = \frac{3}{4} = 0.75$
So,
$\mathbf{P}(\text{MENINGITIS} \mid \text{FEVER} = \textit{true}, \text{VOMITING} = \textit{false}) = \langle 0.25, 0.75 \rangle$

2. Predictive data analytics models are often used as tools for process quality control and fault detection. The task in this question is to create a naive Bayes model to monitor a wastewater treatment plant.[1] The table below lists a dataset containing details of activities at a wastewater treatment plant for 14 days. Each day is described in terms

1. The dataset in this question is inspired by the Waste Water Treatment Dataset that is available from the UCI Machine Learning repository (**?**) at archive.ics.uci.edu/ml/machine-learning-databases/water-treatment. The creators of this dataset reported their work in **?**.

of six descriptive features that are generated from different sensors at the plant. SS-IN measures the solids coming into the plant per day; SED-IN measures the sediment coming into the plant per day; COND-IN measures the electrical conductivity of the water coming into the plant.[2] The features SS-OUT, SED-OUT, and COND-OUT are the corresponding measurements for the water flowing out of the plant. The target feature, STATUS, reports the current situation at the plant: *ok*, everything is working correctly; *settler*, there is a problem with the plant settler equipment; or *solids*, there is a problem with the amount of solids going through the plant.

ID	SS -IN	SED -IN	COND -IN	SS -OUT	SED -OUT	COND -OUT	STATUS
1	168	3	1,814	15	0.001	1,879	ok
2	156	3	1,358	14	0.01	1,425	ok
3	176	3.5	2,200	16	0.005	2,140	ok
4	256	3	2,070	27	0.2	2,700	ok
5	230	5	1,410	131	3.5	1,575	settler
6	116	3	1,238	104	0.06	1,221	settler
7	242	7	1,315	104	0.01	1,434	settler
8	242	4.5	1,183	78	0.02	1,374	settler
9	174	2.5	1,110	73	1.5	1,256	settler
10	1,004	35	1,218	81	1,172	33.3	solids
11	1,228	46	1,889	82.4	1,932	43.1	solids
12	964	17	2,120	20	1,030	1,966	solids
13	2,008	32	1,257	13	1,038	1,289	solids

(a) Create a naive Bayes model that uses probability density functions to model the descriptive features in this dataset (assume that all the descriptive features are normally distributed).

The prior probabilities of each of the target feature levels are

$$P(\text{STATUS} = ok) = \frac{4}{13} = 0.3077$$

$$P(\text{STATUS} = settler) = \frac{5}{13} = 0.3846$$

$$P(\text{STATUS} = ok) = \frac{4}{13} = 0.3077$$

2. The conductivity of water is affected by inorganic dissolved solids and organic compounds, such as oil. Consequently, water conductivity is a useful measure of water purity.

To create the probability density functions required by the model, we simply need to fit a normal distribution to each feature for each level of the target. To do this, we calculate the mean and standard deviation for each feature for the set of instances where the target takes a given value. The table below lists the normal probability distributions fitted to each descriptive feature and target level.

$P(\text{SS-IN} \mid ok)$	=	$N\left(x, \mu = 189, \sigma = 45.42\right)$
$P(\text{SED-IN} \mid ok)$	=	$N\left(x, \mu = 3.125, \sigma = 0.25\right)$
$P(\text{COND-IN} \mid ok)$	=	$N\left(x, \mu = 1{,}860.5, \sigma = 371.4\right)$
$P(\text{SS-OUT} \mid ok)$	=	$N\left(x, \mu = 18, \sigma = 6.06\right)$
$P(\text{SED-OUT} \mid ok)$	=	$N\left(x, \mu = 0.054, \sigma = 0.10\right)$
$P(\text{COND-OUT} \mid ok)$	=	$N\left(x, \mu = 2{,}036, \sigma = 532.19\right)$
$P(\text{SS-IN} \mid settler)$	=	$N\left(x, \mu = 200.8, \sigma = 55.13\right)$
$P(\text{SED-IN} \mid settler)$	=	$N\left(x, \mu = 4.4, \sigma = 1.78\right)$
$P(\text{COND-IN} \mid settler)$	=	$N\left(x, \mu = 1{,}251.2, \sigma = 116.24\right)$
$P(\text{SS-OUT} \mid settler)$	=	$N\left(x, \mu = 98, \sigma = 23.38\right)$
$P(\text{SED-OUT} \mid settler)$	=	$N\left(x, \mu = 1.018, \sigma = 1.53\right)$
$P(\text{COND-OUT} \mid settler)$	=	$N\left(x, \mu = 1{,}372, \sigma = 142.58\right)$
$P(\text{SS-IN} \mid solids)$	=	$N\left(x, \mu = 1{,}301, \sigma = 485.44\right)$
$P(\text{SED-IN} \mid solids)$	=	$N\left(x, \mu = 32.5, \sigma = 11.96\right)$
$P(\text{COND-IN} \mid solids)$	=	$N\left(x, \mu = 1{,}621, \sigma = 453.04\right)$
$P(\text{SS-OUT} \mid solids)$	=	$N\left(x, \mu = 49.1, \sigma = 37.76\right)$
$P(\text{SED-OUT} \mid solids)$	=	$N\left(x, \mu = 1{,}293, \sigma = 430.95\right)$
$P(\text{COND-OUT} \mid solids)$	=	$N\left(x, \mu = 832.85, \sigma = 958.31\right)$

(b) What prediction will the naive Bayes model return for the following query?

SS-IN = 222, SED-IN = 4.5, COND-IN = 1,518, SS-OUT = 74 SED-OUT = 0.25, COND-OUT = 1,642

The calculation for STATUS = *ok*:

$P(ok)$	=	0.3077		
$P(\text{SS-IN} \mid ok)$	=	$N\left(222, \mu = 189, \sigma = 45.42\right)$	=	0.0068
$P(\text{SED-IN} \mid ok)$	=	$N\left(4.5, \mu = 3.125, \sigma = 0.25\right)$	=	4.3079×10^{-7}
$P(\text{COND-IN} \mid ok)$	=	$N\left(1{,}518, \mu = 1{,}860.5, \sigma = 371.4\right)$	=	0.0007
$P(\text{SS-OUT} \mid ok)$	=	$N\left(74, \mu = 18, \sigma = 6.06\right)$	=	1.7650×10^{-20}
$P(\text{SED-OUT} \mid ok)$	=	$N\left(0.25, \mu = 0.054, \sigma = 0.10\right)$	=	0.5408
$P(\text{COND-OUT} \mid ok)$	=	$N\left(1{,}642, \mu = 2{,}036, \sigma = 532.19\right)$	=	0.0006

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

The calculation for STATUS = *settler*:

$P(settler)$	=	0.3846		
$P(\text{SS-IN} \mid settler)$	=	$N\left(222, \mu = 200.8, \sigma = 55.13\right)$	=	0.0067
$P(\text{SED-IN} \mid settler)$	=	$N\left(4.5, \mu = 4.4, \sigma = 1.78\right)$	=	0.2235
$P(\text{COND-IN} \mid settler)$	=	$N\left(1{,}518, \mu = 1{,}251.2, \sigma = 116.24\right)$	=	0.0002
$P(\text{SS-OUT} \mid settler)$	=	$N\left(74, \mu = 98, \sigma = 23.38\right)$	=	0.0101
$P(\text{SED-OUT} \mid settler)$	=	$N\left(0.25, \mu = 1.018, \sigma = 1.53\right)$	=	0.2303
$P(\text{COND-OUT} \mid settler)$	=	$N\left(1{,}642, \mu = 1{,}372, \sigma = 142.58\right)$	=	0.0005

$$\left(\prod_{k=1}^{m} P(\mathbf{q}\,[k] \mid settler)\right) \times P(settler) = 1.53837 \times 10^{-13}$$

The calculation for STATUS = *solids*:

$P(solids)$	=	0.3077		
$P(\text{SS-IN} \mid solids)$	=	$N\left(x, \mu = 1{,}301, \sigma = 485.44\right)$	=	6.9496×10^{-5}
$P(\text{SED-IN} \mid solids)$	=	$N\left(x, \mu = 32.5, \sigma = 11.96\right)$	=	0.0022
$P(\text{COND-IN} \mid solids)$	=	$N\left(x, \mu = 1{,}621, \sigma = 453.04\right)$	=	0.0009
$P(\text{SS-OUT} \mid solids)$	=	$N\left(x, \mu = 49.1, \sigma = 37.76\right)$	=	0.0085
$P(\text{SED-OUT} \mid solids)$	=	$N\left(x, \mu = 1{,}293, \sigma = 430.95\right)$	=	1.0291×10^{-5}
$P(\text{COND-OUT} \mid solids)$	=	$N\left(x, \mu = 832.85, \sigma = 958.31\right)$	=	0.0003

$$\left(\prod_{k=1}^{m} P(\mathbf{q}\,[k] \mid solids)\right) \times P(solids) = 1.00668 \times 10^{-21}$$

Recall that because we are using the heights of the PDFs rather than calculating the actual probabilities for each feature taking a value, the score of each target level is a relative ranking and should not be interpreted as a probability.

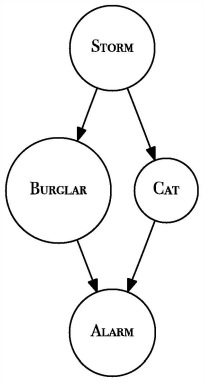
That said, the target level with the highest ranking is STATUS = *settler*. This indicates that there was a problem with the plant's settler equipment on the day of the query.

3. The following is a description of the causal relationship between storms, the behavior of burglars and cats, and house alarms:

Stormy nights are rare. Burglary is also rare, and if it is a stormy night, burglars are likely to stay at home (burglars don't like going out in storms). Cats don't like storms either, and if there is a storm, they like to go inside. The alarm on your house is designed to be triggered if a burglar breaks into your house, but sometimes it can be set off by your cat coming into the house, and sometimes it might not be triggered even if a burglar breaks in (it could be faulty or the burglar might be very good).

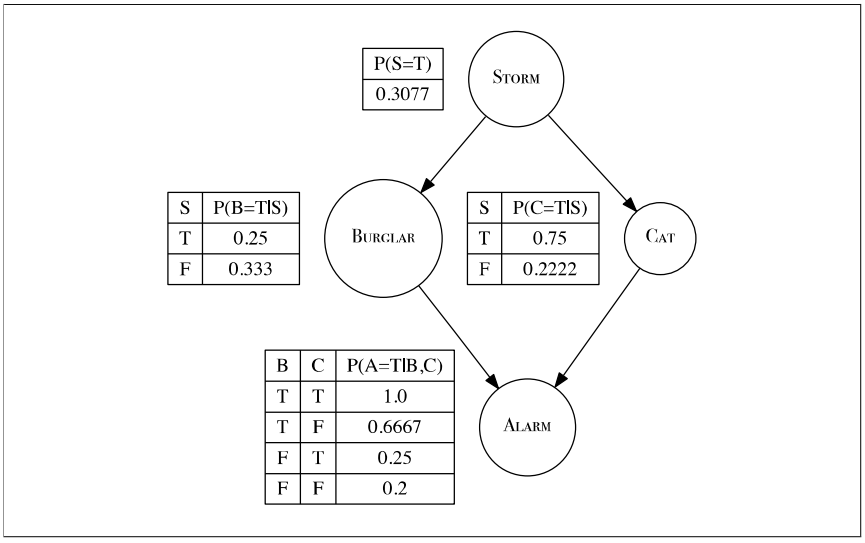
(a) Define the topology of a Bayesian network that encodes these causal relationships.

The figure below illustrates a Bayesian network that encodes the described causal relationships. Storms directly affect the behavior of burglars and cats, and this is reflected by links from the storm node to the burglar and cat nodes. The behavior of burglars and cats both affect whether the alarm goes off, and hence there are links from each of these nodes to the alarm node.



(b) The table below lists a set of instances from the house alarm domain. Using the data in this table, create the conditional probability tables (CPTs) for the network you created in Part (a) of this question.

ID	STORM	BURGLAR	CAT	ALARM
1	false	false	false	false
2	false	false	false	false
3	false	false	false	false
4	false	false	false	false
5	false	false	false	true
6	false	false	true	false
7	false	true	false	false
8	false	true	false	true
9	false	true	true	true
10	true	false	true	true
11	true	false	true	false
12	true	false	true	false
13	true	true	false	true



(c) What value will the Bayesian network predict for ALARM, given that there is both a burglar and a cat in the house but there is no storm?

Because both the parent nodes for ALARM are known, the probability distribution over ALARM is independent of the feature STORM. Consequently, we can read the relevant probability distribution over ALARM directly from the conditional probability table for the ALARM node. Examining the conditional probability table, we can see that when BURGLAR = *true*, and CAT =

true, then ALARM = *true* is the MAP prediction. In other words, the network would predict that the alarm would sound in this situation.

(d) What value will the Bayesian network predict for ALARM, given that there is a storm but we don't know if a burglar has broken in or where the cat is?

In this case, the values of the parents of the target feature are unknown. Consequently, we need to sum out both the parents for each value of the target. The network would calculate the probability of the event ALARM = *true* as follows:

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

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

$$\begin{aligned}
&= (P(a \mid b, c) \times P(b \mid s) \times P(c \mid s) \times P(s)) \\
&+ (P(a \mid b, \neg c) \times P(b \mid s) \times P(\neg c \mid s) \times P(s)) \\
&+ (P(a \mid \neg b, c) \times P(\neg b \mid s) \times P(c \mid s) \times P(s)) \\
&+ (P(a \mid \neg b, \neg c) \times P(\neg b \mid s) \times P(\neg c \mid s) \times P(s)) \\
&= (1.0 \times 0.25 \times 0.75 \times 0.3077) + (0.6667 \times 0.25 \times 0.25 \times 0.3077) \\
&+ (0.25 \times 0.75 \times 0.75 \times 0.3077) + (0.2 \times 0.75 \times 0.25 \times 0.3077) \\
&= 0.125324
\end{aligned}$$

$$P(a \mid s) = \frac{P(a,s)}{P(s)} = \frac{0.125324}{0.3077} = 0.4073$$

This implies that $P(\text{ALARM} = \textit{false}) = 0.5927$, so in this instance, ALARM = *false* is the MAP level for the target, and this is the prediction the model will return.

* 4. The table below lists a dataset containing details of policyholders at an insurance company. The descriptive features included in the table describe each policy holders' ID, occupation, gender, age, type of insurance policy, and preferred contact channel. The preferred contact channel is the target feature in this domain.

ID	OCCUPATION	GENDER	AGE	POLICY TYPE	PREF CHANNEL
1	lab tech	female	43	planC	email
2	farmhand	female	57	planA	phone
3	biophysicist	male	21	planA	email
4	sheriff	female	47	planB	phone
5	painter	male	55	planC	phone
6	manager	male	19	planA	email
7	geologist	male	49	planC	phone
8	messenger	male	51	planB	email
9	nurse	female	18	planC	phone

(a) Using **equal-frequency binning**, transform the AGE feature into a categorical feature with three levels: *young*, *middle-aged*, *mature*.

There are 3 bins and 9 instances. So using equal-frequency binning we know that there will be 3 instances in each bin. In the table below we have ordered the instances in ascending order by AGE

ID	OCCUPATION	GENDER	AGE	POLICY TYPE	PREF CHANNEL
9	nurse	female	18	planC	phone
6	manager	male	19	planA	email
3	biophysicist	male	21	planA	email
1	lab tech	female	43	planC	email
4	sheriff	female	47	planB	phone
7	geologist	male	49	planC	phone
8	messenger	male	51	planB	email
5	painter	male	55	planC	phone
2	farmhand	female	57	planA	phone

If we put the first 3 instances into the *young* bin, the next 3 instances into the *middle-aged* bin, etc. we end up with the dataset in the table below.

ID	OCCUPATION	GENDER	AGE	POLICY TYPE	PREF CHANNEL
9	nurse	female	young	planC	phone
6	manager	male	young	planA	email
3	biophysicist	male	young	planA	email
1	lab tech	female	middle-aged	planC	email
4	sheriff	female	middle-aged	planB	phone
7	geologist	male	middle-aged	planC	phone
8	messenger	male	mature	planB	email
5	painter	male	mature	planC	phone
2	farmhand	female	mature	planA	phone

The thresholds for the different bins are calculated as the mid-point between the AGE values of the two instances on either side of the boundary. So, the threshold between the *young* and *middle-aged* bins would be the mid point between $\mathbf{d}_3 with \text{AGE} = 21$ and $\mathbf{d}_1 with \text{AGE} = 43$:

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

Likewise, the threshold between the *middleaged* and *mature* bins would be the mid point between $\mathbf{d}_7 with \text{AGE} = 49$ and $\mathbf{d}_8 with \text{AGE} = 51$:

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

(b) Examine the descriptive features in the dataset and list the features that you would exclude before you would use the dataset to build a predictive model. For each feature you decide to exclude, explain why you have made this decision.

As is always the case the ID feature should not be used as a descriptive feature during training. However, in this example there is another feature that should be removed from the dataset prior to training. The OCCUPATION feature has different and unique levels for each instance in the dataset. In other words, the OCCUPTAITON feature is equivalent to an id for each instance. Consequently, it should also be removed from the dataset prior to training a model.

(c) Calculate the probabilities required by a **naive Bayes model** to represent this domain.

To train a naive Bayes model using this data, we need to compute the prior probabilities of the target feature taking each level in its domain, and the conditional probability of each feature taking each level in its domain condi-

tioned for each level that the target feature can take. The table below lists the probabilities required by a naive Bayes model to represent this domain.

$P(phone)$	$=$	0.56	$P(email)$	$=$	0.44
$P(\text{GENDER} = female \mid phone)$	$=$	0.6	$P(\text{GENDER} = female \mid email)$	$=$	0.25
$P(\text{GENDER} = male \mid phone)$	$=$	0.4	$P(\text{GENDER} = male \mid email)$	$=$	0.75
$P(\text{AGE} = young \mid phone)$	$=$	0.2	$P(\text{AGE} = young \mid email)$	$=$	0.5
$P(\text{AGE} = middle\text{-}aged \mid phone)$	$=$	0.4	$P(\text{AGE} = middle\text{-}aged \mid email)$	$=$	0.25
$P(\text{AGE} = mature \mid phone)$	$=$	0.4	$P(\text{AGE} = mature \mid email)$	$=$	0.25
$P(\text{POLICY} = planA \mid phone)$	$=$	0.2	$P(\text{POLICY} = planA \mid email)$	$=$	0.5
$P(\text{POLICY} = planB \mid phone)$	$=$	0.2	$P(\text{POLICY} = planB \mid email)$	$=$	0.25
$P(\text{POLICY} = planC \mid phone)$	$=$	0.6	$P(\text{POLICY} = planC \mid email)$	$=$	0.25

(d) What target level will a **naive Bayes model** predict for the following query:

$$\text{GENDER} = female,\ \text{AGE} = 30,\ \text{POLICY} = planA$$

The first step in calculating this answer is to bin the AGE feature. We know from part (a) of this question that the threshold between the *young* and *middle-aged* bins is 32. The value for the AGE feature in the query is less then 32 so it is mapped to the *young* bin. This results in the query being defined as

$$\text{GENDER} = female,\ \text{AGE} = young,\ \text{POLICY} = planA$$

The calculation for $P(\text{CHANNEL} = phone \mid \mathbf{q})$ is

$P(phone)$	$=$	0.56
$P(\text{GENDER} = female \mid phone)$	$=$	0.6
$P(\text{AGE} = young \mid phone)$	$=$	0.2
$P(\text{POLICY} = planA \mid phone)$	$=$	0.2

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid phone)\right) \times P(phone) = 0.01344$$

The calculation for $P(\text{CHANNEL} = email \mid \mathbf{q})$ is

$P(email)$	$=$	0.44
$P(\text{GENDER} = female \mid email)$	$=$	0.25
$P(\text{AGE} = young \mid email)$	$=$	0.5
$P(\text{POLICY} = planA \mid email)$	$=$	0.5

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid email)\right) \times P(email) = 0.0275$$

The target level with the highest ranking is CHANNEL = *email*, and this is the prediction returned by the model.

* 5. Imagine that you have been given a dataset of 1,000 documents that have been classified as being about *entertainment* or *education*. There are 700 *entertainment* documents in the dataset and 300 *education* documents in the dataset. The tables below give the number of documents from each topic that a selection of words occurred in.

Word-document counts for the *entertainment* dataset

fun	is	machine	christmas	family	learning
415	695	35	0	400	70

Word-document counts for the *education* dataset

fun	is	machine	christmas	family	learning
200	295	120	0	10	105

(a) What target level will a **naive Bayes model** predict for the following query document: "*machine learning is fun*"?

A naive Bayes model will label the query with the target level that has the highest probability under the assumption of conditional independence between the evidence features. So to answer this question, we need to calculate the probability of each target level given the evidence and assuming conditional independence across the evidence.

To carry out these calculations, we need to convert the raw document counts into conditional probabilities by dividing each count by the total number of documents occurring in each topic:

w_k	Count	$P(w_k \mid entertainment)$
fun	415	$\frac{415}{700} = .593$
is	695	$\frac{695}{700} = .99$
learning	35	$\frac{35}{700} = .05$
machine	70	$\frac{70}{700} = .10$

w_k	Count	$P(w_k \mid education)$
fun	200	$\frac{200}{300} = .667$
is	295	$\frac{295}{300} = .983$
learning	120	$\frac{120}{300} = .40$
machine	105	$\frac{105}{300} = .35$

We can now compute the probabilities of each target level:

$$\begin{aligned} P(entertainment \mid \mathbf{q}) &= P(entertainment) \times P(machine \mid entertainment) \\ &\quad \times P(learning \mid entertainment) \\ &\quad \times P(is \mid entertainment) \\ &\quad \times p(fun \mid entertainment) \\ &= 0.7 \times 0.593 \times 0.99 \times 0.5 \times 0.1 \\ &= 0.00205 \end{aligned}$$

$$\begin{aligned} P(education \mid \mathbf{q}) &= P(education) \times P(machine \mid education) \\ &\quad \times P(learning \mid education) \\ &\quad \times P(is \mid education) \\ &\quad \times p(fun \mid education) \\ &= 0.3 \times 0.667 \times 0.983 \times 0.4 \times 0.35 \\ &= 0.00275 \end{aligned}$$

As $P(education \mid \mathbf{q}) > P(entertainment \mid \mathbf{q})$, the naive Bayes model will predict the target level of *education*.

(b) What target level will a **naive Bayes model** predict for the following query document: "*christmas family fun*"?

Because the word *christmas* does not appear in any document of either topic, both conditional probabilities for this word will be equal to 0:
$P(christmas \mid entertainment) = 0$ and $P(christmas \mid education) = 0$.

Consequently, the probability for both target levels will be 0, and the model will not be able to return a prediction.

(c) What target level will a **naive Bayes model** predict for the query document in Part (b) of this question, if **Laplace smoothing** with $k = 10$ and a vocabulary size of 6 is used?

The table below illustrates the smoothing of the posterior probabilities for $P(word \mid entertainment)$.

Raw Probabilities	$P(christmas \mid entertainment)$	=	0
	$P(family \mid entertainment)$	=	0.5714
	$P(fun \mid entertainment)$	=	0.5929
Smoothing Parameters	k	=	10
	$count(entertainment)$	=	700
	$count(christmas \mid entertainment)$	=	0
	$count(family \mid entertainment)$	=	400
	$count(fun \mid entertainment)$	=	415
	$\lvert Domain(vocabulary)\rvert$	=	6
Smoothed Probabilities	$P(christmas \mid entertainment) = \frac{0+10}{700+(10\times6)}$	=	0.0132
	$P(family \mid entertainment) = \frac{400+10}{700+(10\times6)}$	=	0.5395
	$P(fun \mid entertainment) = \frac{415+10}{700+(10\times6)}$	=	0.5592

The smoothing of the posterior probabilities for $P(word \mid education)$ is carried out in the same way:

Raw Probabilities	$P(christmas \mid education)$	=	0
	$P(family \mid education)$	=	0.5714
	$P(fun \mid education)$	=	0.5929
Smoothing Parameters	k	=	10
	$count(education)$	=	300
	$count(christmas \mid education)$	=	0
	$count(family \mid education)$	=	10
	$count(fun \mid education)$	=	200
	$\lvert Domain(vocabulary)\rvert$	=	6
Smoothed Probabilities	$P(christmas \mid entertainment) = \frac{0+10}{300+(10\times6)}$	=	0.0278
	$P(family \mid entertainment) = \frac{10+10}{300+(10\times6)}$	=	0.0556
	$P(fun \mid entertainment) = \frac{200+10}{300+(10\times6)}$	=	0.5833

We can now compute the probabilities of each target level:

$$\begin{aligned} P(entertainment \mid \mathbf{q}) &= P(entertainment) \\ &\quad \times P(christmas \mid entertainment) \\ &\quad \times P(family \mid entertainment) \\ &\quad \times P(fun \mid entertainment) \\ &= 0.7 \times 0.0132 \times 0.5395 \times 0.5592 \\ &= 0.0028 \end{aligned}$$

$$\begin{aligned} P(education \mid \mathbf{q}) &= P(education) \times P(christmas \mid education) \\ &\quad \times P(family \mid education) \times P(fun \mid education) \\ &= 0.3 \times 0.0278 \times 0.0556 \times 0.5833 \\ &= 0.0003 \end{aligned}$$

As $P(entertainment \mid \mathbf{q}) > P(education \mid \mathbf{q})$, the model will predict a label of *entertainment* for this query.

* 6. A **naive Bayes model** is being used to predict whether patients have a high risk of stroke in the next five years (STROKE=*true*) or a low risk of stroke in the next five years (STROKE=*false*). This model uses two continuous descriptive features AGE and WEIGHT (in kilograms). Both of these descriptive features are represented by

probability density functions, specifically normal distributions. The table below shows the representation of the domain used by this model.

$P(Stroke = true) = 0.25$	$P(Stroke = false) = 0.75$
$P(\text{AGE} = x \mid Stroke = true)$	$P(\text{AGE} = x \mid Stroke = false)$
$\approx N\begin{pmatrix} x, \\ \mu = 65, \\ \sigma = 15 \end{pmatrix}$	$\approx N\begin{pmatrix} x, \\ \mu = 20, \\ \sigma = 15 \end{pmatrix}$
$P(\text{WEIGHT} = x \mid Stroke = true)$	$P(\text{WEIGHT} = x \mid Stroke = false)$
$\approx N\begin{pmatrix} x, \\ \mu = 88, \\ \sigma = 8 \end{pmatrix}$	$\approx N\begin{pmatrix} x, \\ \mu = 76, \\ \sigma = 6 \end{pmatrix}$

(a) What target level will the **naive Bayes model** predict for the following query:

$$\text{AGE} = 45, \text{WEIGHT} = 80$$

$P(Stroke = true) = 0.25$	$P(Stroke = false) = 0.75$
$P(\text{AGE} = 45 \mid Stroke = true)$	$P(\text{AGE} = 45 \mid Stroke = false)$
$\approx N\begin{pmatrix} x = 45, \\ \mu = 65, \\ \sigma = 15 \end{pmatrix}$	$\approx N\begin{pmatrix} x = 45, \\ \mu = 20, \\ \sigma = 15 \end{pmatrix}$
$= 0.0109$	$= 0.0066$
$P(\text{WEIGHT} = 80 \mid Stroke = true)$	$P(\text{WEIGHT} = 80 \mid Stroke = false)$
$\approx N\begin{pmatrix} x = 80, \\ \mu = 88, \\ \sigma = 8 \end{pmatrix}$	$\approx N\begin{pmatrix} x = 80, \\ \mu = 76, \\ \sigma = 6 \end{pmatrix}$
$= 0.0302$	$= 0.0532$

$$P(Stroke = true \mid \text{AGE} = 45, \text{WEIGHT} = 80) = 0.25 \times 0.0109 \times 0.0302 = 0.000082295$$

$$P(Stroke = false \mid \text{AGE} = 45, \text{WEIDGHT} = 80) = 0.75 \times 0.0066 \times 0.0532 = 0.000263340$$

Based on these calculations the model will predict $Stroke = false$ for this patient.

* 7. The table below lists a dataset of books and whether or not they were purchased by an individual (i.e., the feature PURCHASED is the target feature in this domain).

ID	SECONDHAND	GENRE	COST	PURCHASED
1	false	romance	expensive	true
2	false	science	cheap	false
3	true	romance	cheap	true
4	false	science	cheap	true
5	false	science	expensive	false
6	true	romance	reasonable	false
7	true	literature	cheap	false
8	false	romance	reasonable	false
9	frue	science	cheap	false
10	true	literature	reasonable	true

(a) Calculate the probabilities (to four places of decimal) that a **naive Bayes** classifier would use to represent this domain.

A naive Bayes classifier would require the prior probability for each level of the target feature and the conditional probability for each level of each descriptive feature given each level of the target feature:

$P(Purchased = true) = 0.4$
$P(2ndHand = true|Purchased = true) = 0.5$
$P(2ndHand = false|Purchased = true) = 0.5$
$P(Genre = literature \mid Purchased = true) = 0.25$
$P(Genre = romance \mid Purchased = true) = 0.5$
$P(Genre = science \mid Purchased = true) = 0.25$
$P(Price = cheap \mid Purchased = true) = 0.5$
$P(Price = reasonable \mid Purchased = true) = 0.25$
$P(Price = expensive \mid Purchased = true) = 0.25$
$P(Purchased = false) = 0.6$
$P(2ndHand = true|Purchased = false) = 0.5$
$P(2ndHand = false|Purchased = false) = 0.5$
$P(Genre = literature \mid Purchased = false) = 0.1667$
$P(Genre = romance \mid Purchased = false) = 0.3333$
$P(Genre = science \mid Purchased = false) = 0.5$
$P(Price = cheap \mid Purchased = false) = 0.5$
$P(Price = reasonable \mid Purchased = false) = 0.3333$
$P(Price = expensive \mid Purchased = false) = 0.1667$

(b) Assuming conditional independence between features given the target feature value, calculate the **probability** (rounded to four places of decimal) of each outcome (PURCHASED=true, and PURCHASED=false) for the following book:

SECONDHAND=false, GENRE=literature, COST=expensive

The initial score for each outcome is calculated as follows:
$(Purchased = true) = 0.5 \times 0.25 \times 0.25 \times 0.4 = 0.0125$
$(Purchased = false) = 0.5 \times 0.1667 \times 0.1667 \times 0.6 = 0.0083$
However, these scores are not probabilities. To get real probabilities we must normalise these scores. The normalisation constant is calculated as follows:
$\alpha = 0.0125 + 0.0083 = 0.0208$
The actual probabilities of each outcome is then calculated as:

$$P(Purchased = true) = \frac{0.0125}{0.0208} = (0.600961...) = 0.6010$$
$$P(Purchased = false) = \frac{0.0083}{0.0208} = (0.399038...) = 0.3990$$

(c) What prediction would a **naive Bayes** classifier return for the above book?

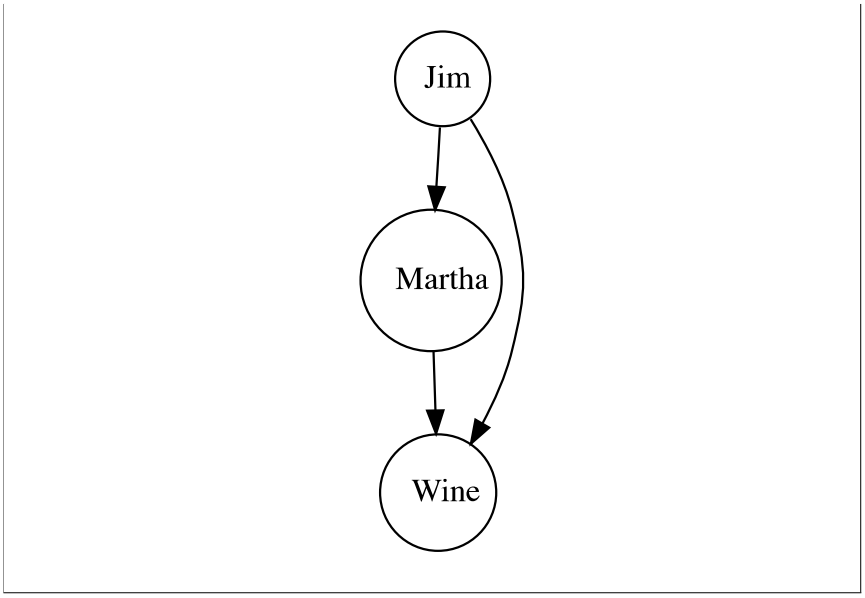
A naive Bayes classifier returns outcome with the maximum a posteriori probability as its prediction. In this instance the outcome PURCHASED=true is the MAP prediction and will be the outcome returned by a naive Bayes model.

* 8. The following is a description of the causal relationship between storms, the behavior of burglars and cats, and house alarms:

Jim and Martha always go shopping separately. If Jim does the shopping he buys wine, but not always. If Martha does the shopping, she buys wine, but not always. If Jim tells Martha that he has done the shopping, then Martha doesn't go shopping, but sometimes Jim forgets to tell Martha, and so sometimes both Jim and Martha go shopping.

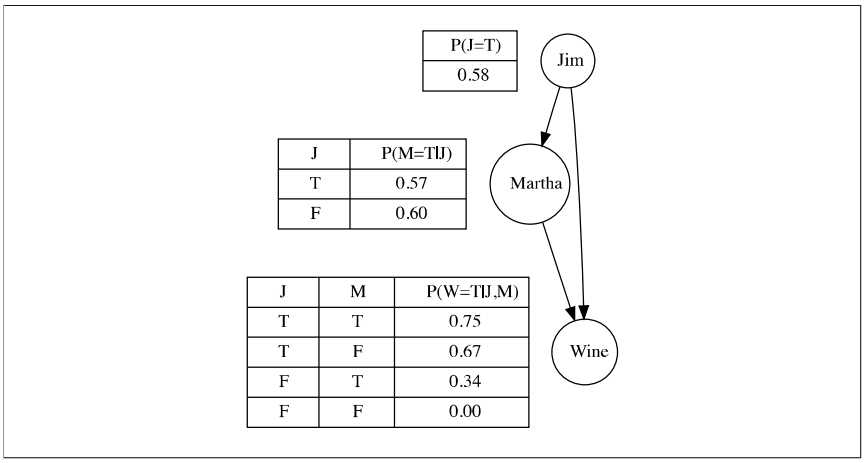
(a) Define the topology of a Bayesian network that encodes these causal relationships between the following Boolean variables: JIM (Jim has done the shopping, *true* or *false*), MARTHA (Martha has done the shopping, *true* or *false*), WINE (wine has been purchased, *true* or *false*).

The figure below illustrates a Bayesian network that encodes the described causal relationships. JIM directly affects the behavior of MARTHA and WINE, and this is reflected by links from the JIM node to the MARTHA and WINE nodes. The MARTHA also affects whether WINE was purchased, and hence there is a link from MARTHA to WINE.



(b) The table below lists a set of instances from the house alarm domain. Using the data in this table, create the conditional probability tables (CPTs) for the network you created in the first part of this question, and round the probabilities to two places of decimal.

ID	JIM	MARTHA	WINE
1	false	false	false
2	false	false	false
3	true	false	true
4	true	false	true
5	true	false	false
6	false	true	true
7	false	true	false
8	false	true	false
9	true	true	true
10	true	true	true
11	true	true	true
12	true	true	false



(c) What value will the Bayesian network predict for WINE if:

JIM=*true* and MARTHA=*false*

Because both the parent nodes for WINE are known, we can read the relevant probability distribution over WINE directly from the conditional probability table for the WINE node. Examining the conditional probability table, we can see that when JIM = *true*, and MARTHA = *false*, then WINE = *true* is the MAP prediction (0.75 versus 0.25). In other words, the network would predict that wine would be purchased in this scenario.

(d) What is the probability that JIM went shopping given that WINE=*true*?

$$
\begin{aligned}
P(\text{JIM} = true \mid \text{WINE} = true) &= \frac{P(\text{JIM} = true, \text{WINE} = true)}{P(\text{WINE} = true)} \\
&= \frac{\sum_{M \in \{T,F\}} P(\text{MARTHA}, \text{JIM} = true, \text{WINE} = true)}{\sum_{M,J \in \{T,F\}} P(\text{MARTHA}, \text{JIM}, \text{WINE} = true)} \\
&= \frac{0.75 + 0.67}{0.75 + 0.67 + 0.34 + 0.00} \\
&= \frac{1.42}{1.76} \\
&= 0.81
\end{aligned}
$$