Machine Learning

Tutorial 4 (Probability-Based Learning)

Exercise1:

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

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	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* (VOMITING = *true*)
(b) *P* (HEADACHE = *false*)
(c) *P* (HEADACHE = *true*, VOMITING = *false*)
(d) *P* (VOMITING = *false* | HEADACHE = *true*)
(e) *P* (MENINGITIS | FEVER = *true*, VOMITING = *false*)

Exercise2:

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. The table below lists a dataset containing details of activities at a wastewater treatment plant for 14 days. Each day is described in terms 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. 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).
(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

Exercise 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.
(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?
(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?

Exercise 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*.
(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.
(c) Calculate the probabilities required by a **naive Bayes model** to represent this domain.
(d) What target level will a **naive Bayes model** predict for the following query:

GENDER = *female*, AGE = 30, POLICY = *planA*

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

Exercise 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.

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

- What target level will the **naive Bayes model** predict for the following query: AGE = 45, WEIGHT = 80

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

Exercise 8:

The following is a description of the causal relationship between storms, the behavior of Jim, Martha, and Wine:

"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*).

(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

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