CSC6780 - Data Science; Assignemt 7

Jesse Annan || ID: 002708111

March 29, 2023

Question 1

$$\mathbf{P}(event) = \frac{\#coinflips \;\; \mathbf{C}ombination \;\; \#exactheads}{2^{\#coinflips}}$$

(a) Let A:= the probability that exactly two of them will get heads.

$$\mathbf{P}(A) = \frac{3\mathbf{C}2}{2^3} = \frac{3}{8}$$

(b) Let B:= the probability that exactly eight eight of them will get heads.

$$\mathbf{P}(B) = \frac{20\mathbf{C}8}{2^{20}} \approx 0.12013435$$

(c) Let C:= the probability that at least four of them will get heads.

$$\begin{aligned}\mathbf{P}(C) &= 1 - \mathbf{P}(\leq 3 \text{ people getting heads}) \\ &= 1 - \left[\frac{20\mathbf{C}0}{2^{20}} + \frac{20\mathbf{C}1}{2^{20}} + \frac{20\mathbf{C}2}{2^{20}} + \frac{20\mathbf{C}3}{2^{20}}\right] \\ \mathbf{P}(C) &\approx 0.99871159\end{aligned}$$

Question 2

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

Let **H**:= HEADACHE, **F**:= FEVER, **V**:= VOMITING,
$\mathbb{M}$:= MENINGITIS, $\mathbf{M}_T$:= MENINGITIS = True, $\mathbf{M}_F$:= MENINGITIS = False

(a) $\mathbf{P}(\mathbf{V}= \text{True}) = \frac{6}{10}$

(b) $\mathbf{P}(\mathbf{H}= \text{True}) = \frac{3}{10}$

(c)

$$\begin{aligned}\mathbf{P}(\mathbf{H}= \text{True},\ \mathbf{V}= \text{False}) &= \mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{H}= \text{True}) \\ &\quad \times \mathbf{P}(\mathbf{H}= \text{True}) \\ &= \frac{1}{7} \times \frac{7}{10} \\ \mathbf{P}(\mathbf{H}= \text{True},\ \mathbf{V}= \text{False}) &= \frac{1}{10}\end{aligned}$$

(d) $\mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{H}= \text{True}) = \frac{1}{7}$

(e) $\mathbf{P}(\mathbb{M} \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) = (\mathbf{P}(\mathbf{M}_T \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) , \mathbf{P}(\mathbf{M}_F \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}))$

$$\begin{aligned}\mathbf{P}(\mathbf{M}_T \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) &= \frac{\mathbf{P}(\mathbf{F}= \text{True}, \mathbf{V}= \text{False} \mid \mathbf{M}_T) \times \mathbf{P}(\mathbf{M}_T)}{\mathbf{P}(\mathbf{F}= \text{True}, \mathbf{V}= \text{False})} \\ &= \frac{\mathbf{P}(\mathbf{F}= \text{True} \mid \mathbf{M}_T) \times \mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{F}= \text{True}, \mathbf{M}_T) \times \mathbf{P}(\mathbf{M}_T)}{\mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{F}= \text{True}) \times \mathbf{P}(\mathbf{F}= \text{True})} \\ &= \frac{1/3 \times 1 \times 3/10}{1 \times 4/10}\end{aligned}$$

$\mathbf{P}(\mathbf{M}_T \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) = 0.25$

$$\begin{aligned}\mathbf{P}(\mathbf{M}_F \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) &= \frac{\mathbf{P}(\mathbf{F}= \text{True}, \mathbf{V}= \text{False} \mid \mathbf{M}_F) \times \mathbf{P}(\mathbf{M}_F)}{\mathbf{P}(\mathbf{F}= \text{True}, \mathbf{V}= \text{False})} \\ &= \frac{\mathbf{P}(\mathbf{F}= \text{True} \mid \mathbf{M}_F) \times \mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{F}= \text{True}, \mathbf{M}_F) \times \mathbf{P}(\mathbf{M}_F)}{\mathbf{P}(\mathbf{V}= \text{False} \mid \mathbf{F}= \text{True}) \times \mathbf{P}(\mathbf{F}= \text{True})} \\ &= \frac{3/7 \times 1 \times 7/10}{1 \times 4/10}\end{aligned}$$

$\mathbf{P}(\mathbf{M}_F \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) = 0.75$

$\mathbf{P}(\mathbb{M} \mid \mathbf{F}= \text{True}, \mathbf{V}= \text{False}) = (0.25\ ,\ 0.75)$

Question 3

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) $Bins = 3$
young : $\{18,\ 19,\ 21\}$ *middle-aged* : $\{43,\ 47,\ 49\}$ *mature* : $\{51,\ 55,\ 57\}$
Defining the range of the **AGE** column by averaging the ends from the above three levels:

$$\begin{array}{lcl} & \text{young} & \leq\ 32 \\ 32 < & \text{middle-age} & \leq\ 50 \\ 50 < & \text{mature} & \end{array}$$

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

(b) Take out **"ID" and "OCCUPATION"**
This is because both ID and OCCUPATION have too many unique levels which will create a high dimensionality problem for any model.

(c) Let **G**:= GENDER, **A**:= AGE, **PT**:= POLICY TYPE, **PC**:= PREF CHANNEL

$\mathbf{P}(\mathbf{PC}=\text{email}) = \frac{4}{9}$	$\mathbf{P}(\mathbf{PC}=\text{phone}) = \frac{5}{9}$
$\mathbf{P}(\mathbf{PT}=\text{planA} \mid \mathbf{PC}=\text{email}) = \frac{2}{4}$	$\mathbf{P}(\mathbf{PT}=\text{planA} \mid \mathbf{PC}=\text{phone}) = \frac{1}{5}$
$\mathbf{P}(\mathbf{PT}=\text{planB} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{PT}=\text{planB} \mid \mathbf{PC}=\text{phone}) = \frac{1}{5}$
$\mathbf{P}(\mathbf{PT}=\text{planC} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{PT}=\text{planC} \mid \mathbf{PC}=\text{phone}) = \frac{3}{5}$
$\mathbf{P}(\mathbf{G}=\text{male} \mid \mathbf{PC}=\text{email}) = \frac{3}{4}$	$\mathbf{P}(\mathbf{G}=\text{male} \mid \mathbf{PC}=\text{phone}) = \frac{2}{5}$
$\mathbf{P}(\mathbf{G}=\text{female} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{G}=\text{female} \mid \mathbf{PC}=\text{phone}) = \frac{3}{5}$
$\mathbf{P}(\mathbf{A}=\text{young} \mid \mathbf{PC}=\text{email}) = \frac{2}{4}$	$\mathbf{P}(\mathbf{A}=\text{young} \mid \mathbf{PC}=\text{phone}) = \frac{1}{5}$
$\mathbf{P}(\mathbf{A}=\text{middle-age} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{A}=\text{middle-age} \mid \mathbf{PC}=\text{phone}) = \frac{2}{5}$
$\mathbf{P}(\mathbf{A}=\text{mature} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{A}=\text{mature} \mid \mathbf{PC}=\text{phone}) = \frac{2}{5}$

(d) `Query: GENDER = female, AGE = 30, POLICY = planA`

$\mathbf{P}(\mathbf{PC}=\text{email}) = \frac{4}{9}$	$\mathbf{P}(\mathbf{PC}=\text{phone}) = \frac{5}{9}$
$\mathbf{P}(\mathbf{PT}=\text{planA} \mid \mathbf{PC}=\text{email}) = \frac{2}{4}$	$\mathbf{P}(\mathbf{PT}=\text{planA} \mid \mathbf{PC}=\text{phone}) = \frac{1}{5}$
$\mathbf{P}(\mathbf{G}=\text{female} \mid \mathbf{PC}=\text{email}) = \frac{1}{4}$	$\mathbf{P}(\mathbf{G}=\text{female} \mid \mathbf{PC}=\text{phone}) = \frac{3}{5}$
$\mathbf{P}(\mathbf{A}=\text{young} \mid \mathbf{PC}=\text{email}) = \frac{2}{4}$	$\mathbf{P}(\mathbf{A}=\text{young} \mid \mathbf{PC}=\text{phone}) = \frac{1}{5}$

$\mathbf{P}(\mathbf{PC}=\text{email} \mid \mathbf{PT}=\text{planA}, \mathbf{G}=\text{female}, \mathbf{A}=\text{young}) = \frac{1}{4} \times \frac{2}{4} \times \frac{2}{4} \times \frac{4}{9} = \frac{1}{36} \approx 0.02777\dot{}$
$\mathbf{P}(\mathbf{PC}=\text{phone} \mid \mathbf{PT}=\text{planA}, \mathbf{G}=\text{female}, \mathbf{A}=\text{young}) = \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{5}{9} = \frac{1}{75} \approx 0.01333\dot{}$

`Query Prediction: "email"`

Question 4 size of entertainment = 700 size of education = 300

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) `Query: "machine learning is fun"`

$$
\begin{aligned}
\mathbf{P}(\text{entertainment}) &= \tfrac{700}{1000} \\
\mathbf{P}(\text{machine} \mid \text{entertainment}) &= \tfrac{35}{700} \\
\mathbf{P}(\text{learning} \mid \text{entertainment}) &= \tfrac{70}{700} \\
\mathbf{P}(\text{is} \mid \text{entertainment}) &= \tfrac{695}{700} \\
\mathbf{P}(\text{fun} \mid \text{entertainment}) &= \tfrac{415}{700} \\
\mathbf{P}(\text{education}) &= \tfrac{300}{1000} \\
\mathbf{P}(\text{machine} \mid \text{education}) &= \tfrac{120}{300} \\
\mathbf{P}(\text{learning} \mid \text{education}) &= \tfrac{105}{300} \\
\mathbf{P}(\text{is} \mid \text{education}) &= \tfrac{295}{300} \\
\mathbf{P}(\text{fun} \mid \text{education}) &= \tfrac{200}{300}
\end{aligned}
$$

case 1: $\frac{700}{1000} \times \frac{35}{700} \times \frac{70}{700} \times \frac{695}{700} \times \frac{415}{700} \approx 0.00206179$

case 2: $\frac{300}{1000} \times \frac{120}{300} \times \frac{105}{300} \times \frac{295}{300} \times \frac{200}{300} \approx 0.02753333$

`Query Prediction: "education"`

(b) `Query: "christmas family fun"`

$$
\begin{aligned}
\mathbf{P}(\text{entertainment}) &= \tfrac{700}{1000} \\
\mathbf{P}(\text{christmas} \mid \text{entertainment}) &= 0 \\
\mathbf{P}(\text{family} \mid \text{entertainment}) &= \tfrac{400}{700} \\
\mathbf{P}(\text{fun} \mid \text{entertainment}) &= \tfrac{415}{700} \\
\mathbf{P}(\text{education}) &= \tfrac{300}{1000} \\
\mathbf{P}(\text{christmas} \mid \text{education}) &= 0 \\
\mathbf{P}(\text{family} \mid \text{education}) &= \tfrac{10}{300} \\
\mathbf{P}(\text{fun} \mid \text{education}) &= \tfrac{200}{300}
\end{aligned}
$$

case 1: $\frac{700}{1000} \times 0 \times \frac{400}{700} \times \frac{415}{700} = 0$

case 2: $\frac{300}{1000} \times 0 \times \frac{10}{300} \times \frac{200}{300} = 0$

`Query Prediction: N/A`

(c)