DATA1001 – Introduction to Data Science and Decisions

Assignment 3

Student details

Surname:

Given names:

Student number:

Tutorial day/time:

Date due: 4pm on 26 October 2018

Date submitted:

Declaration

I declare that this assessment item is my own work, except where acknowledged, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

1. Reproduce this assessment item and provide a copy to another member of the University; and/or
2. Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking)

I certify that I have read and understood the University Rules in respect of Student Conduct.

Signed[1]:

Date:

Mark:

Comments:

[1] Sign in ink if handing in a hard copy; or type your name if handing in an electronic copy.

A bit of Differential Privacy

The Australian Taxation Office would like to find out the proportion of Australians who cheat on their tax returns. They select a random sample (e.g. by tax file numbers) of Australians and ask them if they have ever cheated on their tax return. Of course, they promise that there will be no repercussion if the answer is 'yes', and that they will not record this information against your identity.

If you had cheated on a tax return previously, would you answer truthfully?

Your answer is probably 'No', because you wouldn't trust the ATO to securely manage this information on you. In this assignment, you'll explore a way in which an individual can safely disclose this information.

Procedure: The respondent secretly flips a coin twice. If the first flip shows 'heads', they answer truthfully. Otherwise, they answer 'yes' or 'no' according to the second flip being 'heads' or 'tails'.

The idea is that this way, enough 'randomness' is added to the respondents answer so that they cannot be identified.

Below,

- the (unknown) true proportion of tax return frauds is $\theta \in [0, 1]$
- the variable `fraud` with values 0, 1 denotes whether a respondent is a fraud
- the variable `truth` with values 0, 1 denotes whether a respondent answers truthfully
- the variable `yes` with values 0, 1 denotes whether a respondent answers 'yes' according to the above procedure.

Question 1) The use for the ATO

a) [3 marks]

Draw a tree diagram for the three variables `fraud`, `truth` and `yes`.

b) [3 marks]

Fill in the remaining values in this probability table:

`fraud`	`truth`	`yes`	**P**
0	0	0	?
0	0	1	$\frac{1-\theta}{4}$
0	1	0	?
0	1	1	0
1	0	0	?

fraud	truth	yes	$\mathbf{P}$
1	0	1	?
1	1	0	?
1	1	1	?

Hint: according to your tree diagram,

$$
\begin{aligned}
&\mathbf{P}((\mathtt{fraud} = 0) \cap (\mathtt{truth} = 0) \cap (\mathtt{yes} = 1)) \\
&\quad = \mathbf{P}((\mathtt{fraud} = 0))\mathbf{P}(\mathtt{truth} = 0|\mathtt{fraud} = 0)\mathbf{P}(\mathtt{yes} = 1|(\mathtt{truth} = 0) \cap (\mathtt{fraud} = 0)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad = (1-\theta)(1/2)(1/2) = (1-\theta)/4
\end{aligned}
$$

c) [2 mark]

Show that $\mathbf{P}(\mathtt{yes} = 1) = 1/4 + \theta/2$.

Hint: Follow all paths in your tree diagram that lead to $\mathtt{yes}$=1 and add their probabilities.

d) [1 mark]

The ATO has received 10384 responses, of which 3448 were $\mathtt{yes}$=1. From these numbers, derive an estimate of θ.

Question 2) Is it really safe to participate?

In this question, we'll calculate the probability $\mathbf{P}(\mathtt{fraud} = 1|\mathtt{yes} = 1)$. If this probability is close to 1, then by playing the game and answering yes you will reveal yourself as a tax fraud!

a) [2 mark]

Compute all probabilities $\mathbf{P}((\mathtt{fraud} = i) \cap (\mathtt{yes} = j))$ for all possible values of i and j, and collect them in a table. (This procedure is called computing the "marginal distribution" of fraud and yes.)

Hint: the event $A = (\mathtt{fraud} = 1) \cap (\mathtt{yes} = 0)$ is the disjoint union of $A \cap (\mathtt{truth} = 0)$ and $A \cap (\mathtt{truth} = 1)$. Find the probabilities for these events in your table.

b) [2 mark]

Calculate $\mathbf{P}(\mathtt{fraud} = 1|\mathtt{yes} = 1)$ and $\mathbf{P}(\mathtt{fraud} = 1|\mathtt{yes} = 0)$

c) [2 mark]

What would happen if a biased coin was used in the first toss? (Limit your answer to 3 normal-length sentences.)

Assignment Submission

Due Date: 4pm on 26 October 2018 (That's Friday in Week 13.)

Hand in your assignment to the School Office of the School of Mathematics & Statistics, Level 3, Red Centre Building (Centre Wing).

Late submissions

20% (3 marks) will be deducted at 0, 24, 48, 72, 96 hours after the deadline. Work submitted more than five days late will not be marked.

On Plagiarism

The University regards plagiarism as a form of academic misconduct, and has very strict rules regarding plagiarism. For UNSW policies, penalties, and information to help you avoid plagiarism see: https://student.unsw.edu.au/plagiarism as well as the guidelines in the online ELISE tutorials for all new UNSW students: http://subjectguides.library.unsw.edu.au/elise