Assignment 1

Aki Vehtari et al.

1 General information

The exercises here refer to the lecture 1/BDA chapter 1 content, not the course infrastructure quiz. This assignment is meant to test whether or not you have sufficient knowledge to participate in the course. The first question checks that you remember basic terms of probability calculus. The second exercise checks you recognise the most important notation used throughout the course and used in BDA3. The third-fifth exercise you will solve some basic Bayes theorem questions to check your understanding on the basics of probability theory. The 6th exercise checks on whether you recall the three steps of Bayesian Data Ananlysis as mentioned in chapter 1 of BDA3. The last exercise walks you through an example of how we can use models to generate distributions for outcomes of interest, applied to a setting of a simplified Roulette table.

This quarto document is not intended to be submitted, but to render the questions as they appear on Mycourses to be available also outside of it.

The exercises constitute 80% of the Quiz 1 grade.

We prepared a quarto template specific to this assignment to help you get started. You still need to fill in your answers on Mycourses! You can inspect this and future templates

- as a qmd file,
- as a rendered html file

⚠ Setup

This is the template for assignment 1. You can download the qmd-file or copy the code from this rendered document after clicking on `</> Code` in the top right corner.
Please replace the instructions in this template by your own text, explaining what you are doing in each exercise.
The following will set-up `markmyassignment` to check your functions at the end of the notebook:

```
library(markmyassignment)
assignment_path = paste("https://github.com/avehtari/BDA_course_Aalto/",
"blob/master/tests/assignment1.yml", sep="")
set_assignment(assignment_path)
```

```
Assignment set:
assignment1: Bayesian Data Analysis: Assignment 1
The assignment contain the following (3) tasks:
- p_red
```

```
- p_box
- p_identical_twin
```

2 1. Basic probability theory notation and terms

This can be trivial or you may need to refresh your memory on these concepts (see, e.g. Aalto course *First Course in Probability and Statistics*). Match each of the following terms to the list of labeled pasted below:

1. probability
2. probability mass (function)
3. probability density (function)
4. probability distribution
5. discrete probability distribution
6. continuous probability distribution
7. cumulative distribution function (cdf)
8. likelihood

Question 1.1

Match the following terms with the correct definition: Note that the answers order and set of possible answers is the same for questions 1.1 - 1.8. Check the BDA chapter 1, the lecture slides, and Wikipedia if you are uncertain about the terms below.

3 2. Notation

This task check whether you recognise the following math symbols.

Question 2.1

Match the following notation with the correct definition:
$\sim$

Question 2.2

Match the following notation with the correct definition:
$\propto$

Question 2.3

Match the following notation with the correct definition:
$E[\cdot]$

Question 2.4

Match the following notation with the correct definition:
$p(y|\theta)$

4 3. Bayes' theorem 1

A group of researchers has designed a new inexpensive and painless test for detecting lung cancer. The test is intended to be an initial screening test for the population in general. A positive result (presence of lung cancer) from the test would be followed up immediately with medication, surgery or more extensive and expensive test. The researchers know from their studies the following facts:

- Test gives a positive result in 98% of the time when the test subject has lung cancer.
- Test gives a negative result in 96% of the time when the test subject does not have lung cancer.
- In general population approximately one person in 1000 has lung cancer.

Here are some probability values that can help you figure out if you copied the right conditional probabilities from the question:

- P(Test gives positive | Subject does not have lung cancer) = 4%
- P(Test gives positive and Subject has lung cancer) = 0.098%
  - this is also referred to as the joint probability of test being positive and the subject having lung cancer

Your end goal for this exercise is to calculate the probability of having cancer given a positive test result: P(cancer|positive).

Question 3.1

Which quantity in Bayes' Theorem does P(cancer|positive) represent?

Question 3.2

What is the probability of the test having a positive result, given that the test subject has cancer (P(B|A))?

Question 3.3

What is the probability of having cancer (P(A))?

Question 3.4

What is the probability of having a positive test (P(B))?

Question 3.5

Using your previous answers, what is the probability of having cancer given a positive test?

If you use pen and paper, it may help to draw pictures as follows (see also assignment_instructions#fig-workflow):

See Figure 1 for illustration of parts of Bayesian workflow.

Here are some probability values that can help you figure out if you copied the right conditional probabilities from the question.

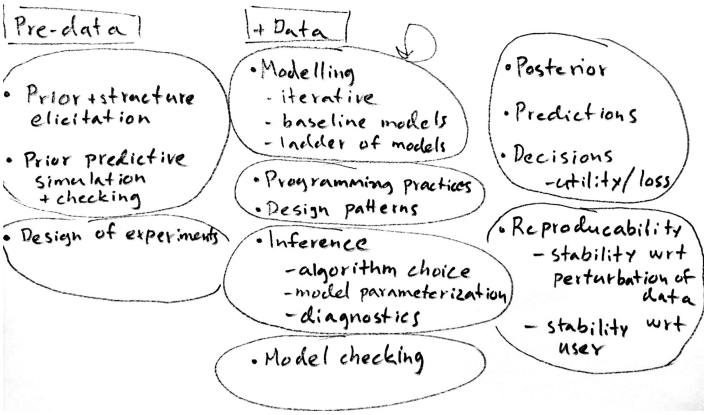

Figure 1: Parts of Bayesian workflow

- P(Test gives positive | Subject does not have lung cancer) = 4%
- P(Test gives positive **and** Subject has lung cancer) = 0.098% this is also referred to as the **joint probability** of *test being positive* and the *subject having lung cancer*.

5 4. Bayes' theorem 2

We have three boxes, A, B, and C. There are

- 2 red balls and 5 white balls in the box A,
- 4 red balls and 1 white ball in the box B, and
- 1 red ball and 3 white balls in the box C.

Consider a random experiment in which one of the boxes is randomly selected and from that box, one ball is randomly picked up. After observing the color of the ball it is replaced in the box it came from. Suppose also that on average box A is selected 40% of the time and box B 10% of the time (i.e. $P(A) = 0.4$).

The following will help you implementing a function to calculate the required probabilities for this exercise. Keep the below name and format for the function to work with `markmyassignment`:

```
boxes_test <- matrix(c(2,2,1,5,5,1), ncol = 2,
    dimnames = list(c("A", "B", "C"), c("red", "white")))

p_red <- function(boxes) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    0.3928571
}

p_box <- function(boxes) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    c(0.29090909,0.07272727,0.63636364)
}
```

Question 4.1

What is the probability of picking a red ball from box A?

Question 4.2

What is the probability of picking a red ball from box B?

Question 4.3

What is the probability of picking a red ball from box C?

Question 4.4

Considering the probabilities of selecting each box, what is the probability of picking a red ball (tolerance of 0.01)?

Question 4.4

If a red ball was picked, calculate the probability that it was picked from (tolerance of 0.04):

6 5. Bayes' theorem 3

Assume that on average fraternal twins (two fertilized eggs and then could be of different sex) occur once in 150 births and identical twins (single egg divides into two separate embryos, so both have the same sex) once in 400 births (**Note!** This is not the true value, see Exercise 1.6, page 28, in BDA3). Assume that an equal number of boys and girls are born on average.

American male singer-actor Elvis Presley (1935 – 1977) had a twin brother who died in birth, your goal is to compute the probability that Elvis was an identical twin.

The R functions below might help you calculating the requited probabilities.

```
fraternal_prob = 1/125
identical_prob = 1/300
```

Keep the below name and format for the function to work with `markmyassignment`:

```
p_identical_twin <- function(fraternal_prob, identical_prob) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    0.4545455
}
```

Question 5.1

What is the probability of having a twin brother, given identical twins (no tolerance)?

Question 5.2

What is the total probability of having a twin brother (either fraternal or identical)? (tolerance of 0.001)

Question 5.3

5.3 What is the probability that Elvis was an identical twin, given that he had a twin brother? (tolerance of 0.001)

7 6. The three steps of Bayesian data analysis

Question 6.1

6.1 Select the three steps of Bayesian data analysis (see BDA3 p. 3):

8 7. A Binomial Model for the Roulette Table

In this course, models are used to explain social and physical data, and we will be able to generate data from our models which we can use for checking how well our model does. In this example, we show how to generate outcomes from a binomial model to explain outcomes of a roulette game (there is a connection to the history of statistics). Suppose a roulette table with only red and black colours. Roulette tables won't be perfect and it's likely that the probability of red vs black is not exactly 0.5 (the tables can have adjustments that are randomized each day to avoid long term bias).

Suppose your model for the tables' ratio of red/black is a Binomial which takes as inputs the number of trials and a probability parameter, theta. Set theta to 0.6 (this is much bigger than what we would expect in real roulette, but makes it easier as a teaching example) and generate a series (for a sequence of 100 equally spaced trial values between 10 and 1000) of red/black ratios. Generate 1000 random draws from your model for each trial value and save the data in a Data frame with columns `Ratios`, `Nsims` and `Trials`. Incomplete code can be found below.

```
# Ratio of red/black
theta <- # declare probability parameter for the binomial model

# Sequence of trials

trials <- seq(#start value of sequence,#end value of sequence,#value for spacing)

# Number of simulation draws from the model
nsims <- # number of of simulations from the binomial model

# Helper function for getting the ratios
binom_gen <- function(trials,theta,nsims){
    df <-  as.data.frame(rbinom(nsims,trials,theta)/trials) |> mutate(nsims = nsims,trials = trials)
    colnames(df) <- c("Ratios","Nsims","Trials")
  return(df)
```

```
}

# Create a data frame containing the draws for each number of trials
ratio_60 <- do.call(rbind, lapply(trials, binom_gen, theta, nsims)) # lapply applies elements in trial
```

> Question 7.1
>
> Suppose you are unsure whether the code to create the data frame worked. Which of the following functions should you use in order to check on the structure of the dataframe object (assuming df below stands for a generic dataframe object)?

> Question 7.2
>
> The structure checks out, but now you want to print the first 5 rows of the dataframe to check whether the values are as expected. Which of the following functions should you use?

> Question 7.3
>
> The quick peek checks also out, but you would be more at ease scrolling all data, perhaps you'll find some interesting patterns. Which of the following actions allows you to scroll through the data in a separate window (for the below, we assume that you have the code loaded in an RStudio session)? (More than one answer may be correct)

Now plot a histogram of the computed ratios for 10, 50 and 1000 trials, using the code below

```
# Plot the Distributions
subset_df <- ratio_60[ratio_60$Trials %in% c(#trial values), ] # Subset your dataframe

subset_df60 |> ggplot(aes(Ratios)) +
  geom_histogram(position = "identity" ,bins = 40) +
  facet_grid(cols = vars(Trials))  +
  ggtitle("Ratios for specific trials")
```

> Question 7.4
>
> Which histogram below is the correct one for theta = 0.6?

> Question 7.5
>
> What do these distributions refer to?

> Question 7.6
>
> Given these histograms, which number of trials gives you the most certainty about the likely red/black ratio for that table?

Question 7.7

Given the draws from the model, give an estimate about the probability p(Ratio<=0.5) for the model with 1000 trials (tolerance of 0.001).

Suppose you are now certain that theta = 0.6, plot the probability density given 1000 trials using the code below.

```
size =  # number of trials
prob =  # probability of success

binom_data <- data.frame(
  Success = 0:size,
  Probability = dbinom(0:size, size = size, prob = prob)
)

ggplot(binom_data, aes(x = Success, y = Probability)) +
  geom_point() +
  geom_line() +
  labs(title = "PMF of Binomial Distribution", x = "Number of Successes", y = "PDF")
```

Question 7.8

Which plot of the PMF is the correct one?

Question 7.9

How does the PMF plot relate to the histogram of ratios plotted earlier?

Question 7.10

Given the PMF for your model, calculate the probability for 1000 trials of observing less or equal to 500 red outcomes using theta = 0.6. Use the pbinom function in R.

markmyassignment

The following will check the functions for which `markmyassignment` has been set up:

```
mark_my_assignment()
```

```
v | F W  S  OK | Context

/ |          0 | task-1-subtask-1-tests
/ |          0 | p_red()
x | 1        3 | p_red()
--------------------------------------------------------------------------------
Failure ('test-task-1-subtask-1-tests.R:21:3'): p_red()
p_red(boxes = boxes) not equivalent to 0.5.
1/1 mismatches
[1] 0.393 - 0.5 == -0.107
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)
```

```
--------------------------------------------------------------------------------

/ |          0 | task-2-subtask-1-tests
/ |          0 | p_box()
x | 1        3 | p_box()
--------------------------------------------------------------------------------
Failure ('test-task-2-subtask-1-tests.R:19:3'): p_box()
p_box(boxes = boxes) not equivalent to c(0.4, 0.1, 0.5).
3/3 mismatches (average diff: 0.0909)
[1] 0.2909 - 0.4 == -0.1091
[2] 0.0727 - 0.1 == -0.0273
[3] 0.6364 - 0.5 ==  0.1364
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)
--------------------------------------------------------------------------------

/ |          0 | task-3-subtask-1-tests
/ |          0 | p_identical_twin()
x | 2        3 | p_identical_twin()
--------------------------------------------------------------------------------
Failure ('test-task-3-subtask-1-tests.R:16:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/500) not equivalent to 0.2857143.
1/1 mismatches
[1] 0.455 - 0.286 == 0.169
Error: Incorrect result for fraternal_prob = 1/100 and identical_prob = 1/500

Failure ('test-task-3-subtask-1-tests.R:19:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/10, identical_prob = 1/20) not equivalent to 0.5.
1/1 mismatches
[1] 0.455 - 0.5 == -0.0455
Error: Incorrect result for fraternal_prob = 1/10 and identical_prob = 1/20
--------------------------------------------------------------------------------

== Results =====================================================================
-- Failed tests ----------------------------------------------------------------
Failure ('test-task-1-subtask-1-tests.R:21:3'): p_red()
p_red(boxes = boxes) not equivalent to 0.5.
1/1 mismatches
[1] 0.393 - 0.5 == -0.107
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)

Failure ('test-task-2-subtask-1-tests.R:19:3'): p_box()
p_box(boxes = boxes) not equivalent to c(0.4, 0.1, 0.5).
3/3 mismatches (average diff: 0.0909)
[1] 0.2909 - 0.4 == -0.1091
[2] 0.0727 - 0.1 == -0.0273
[3] 0.6364 - 0.5 ==  0.1364
Error: Incorrect result for matrix(c(1,1,1,1,1,1), ncol = 2)

Failure ('test-task-3-subtask-1-tests.R:16:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/500) not equivalent to 0.2857143.
1/1 mismatches
[1] 0.455 - 0.286 == 0.169
Error: Incorrect result for fraternal_prob = 1/100 and identical_prob = 1/500
```

```
Failure ('test-task-3-subtask-1-tests.R:19:3'): p_identical_twin()
p_identical_twin(fraternal_prob = 1/10, identical_prob = 1/20) not equivalent to 0.5.
1/1 mismatches
[1] 0.455 - 0.5 == -0.0455
Error: Incorrect result for fraternal_prob = 1/10 and identical_prob = 1/20

[ FAIL 4 | WARN 0 | SKIP 0 | PASS 9 ]
```