Uppsala University

Introduction to Machine Learning, Big Data, and AI

Assignment 1

General information

- The recommended tool in this course is R (with the IDE R-Studio). You can download R **here** and R-Studio **here**. There are tons of tutorials, videos and introductions to R and R-Studio online. You can find some initial hints from **RStudio Education pages**.
- When working with R, we recommend writing the report using R markdown and the provided **R markdown template**. The remplate includes the formatting instructions and how to include code and figures.
- Instead of R markdown, you can use other software to make the PDF report, but the the same instructions for formatting should be used. These instructions are available also in **the PDF produced from the R markdown template**.
- Report all results in a single and *anonymous* pdf.
- The course has its own R package `bsda` with data and functionality to simplify coding. To install the package just run the following (upgrade="never" skips question about updating other packages):

  1. `install.packages("remotes")`
  2. `remotes::install_github("MansMeg/BSDA", subdir = "rpackage", upgrade="never")`

- Many of the exercises can be checked automatically using the R package `markmyassignment`. Information on how to install and use the package can be found **here**. There is no need to include `markmyassignment` results in the report.
- Common questions and answers regarding installation and technical problems can be found in Frequently Asked Questions (FAQ).
- Deadlines and information on how to turn in the assignments can be found in Studium.
- You are allowed to discuss assignments with your friends, but it is not allowed to copy solutions directly from other students or from internet. Try to solve the actual assignment problems with your own code and explanations. Do not share your answers publicly. Do not copy answers from the internet or from previous years. We compare the answers with urkund. All suspected plagiarism will be reported and investigated.
- If you have any suggestions or improvements to the course material, please post in the course chat feedback channel, create an issue, or submit a pull request to the public repository **here**

Information on this assignment

The exercises of this assignment are not necessarily related to chapter 1, but rather as an introduction to the course. The second exercise refreshes your basic computer skills and guides you to some basic R functions. In the last three ones you will first solve the problems using pen and paper (you can, for example, write the equations in markdown or scan and include hand written answers), and then implement the final equations in R (and then you can use `markmyassignment` to check your results).

Reading instructions: Chapter 1 in BDA3, O'Hagan (2004) "Dicing with the unknown".

To use markmyassignment for this assignment, run the following code in R:

```
library(markmyassignment)
assignment_path <-
  paste("https://github.com/MansMeg/BSDA/",
        "blob/main/assignments/tests/assignment1.yml", sep="")
set_assignment(assignment_path)
# To check your code/functions, just run
mark_my_assignment()
```

Don't include `markmyassignment` results in the report.

---

1. **(General questions)** See formating instructions in the template **here**.

2. **(Basic probability theory notation and terms)**. This can be trivial or you may need to refresh your memory on these concepts. Note that some terms may be different names for the same concept. Explain each of the following terms with one sentence:

   - probability
   - probability mass
   - probability density
   - probability mass function (pmf)
   - probability density function (pdf)
   - probability distribution
   - discrete probability distribution
   - continuous probability distribution
   - cumulative distribution function (cdf)
   - likelihood
   - aleatoric uncertainty
   - epistemic uncertainty

3. **(Basic computer skills)** This task deals with elementary plotting and computing skills needed during the rest of the course. You can use either R or Python, although R is the recommended language and we will only guarantee support in R. For documentation in R, just type `?{function name here}`.

   a) Plot the density function of Beta-distribution, with mean $\mu = 0.2$ and variance $\sigma^2 = 0.01$. The parameters α and β of the Beta-distribution are related to the mean and variance according to the following equations

   $$\alpha = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad \beta = \frac{\alpha(1-\mu)}{\mu}.$$

   Hint! Useful R functions: `seq()`, `plot()` and `dbeta()`. Later on we will also use the more flexible `ggplot2` for plotting.

   b) Take a sample of 1000 random numbers from the above distribution and plot a histogram of the results. Compare visually to the density function.
   Hint! Useful R functions: `rbeta()` and `hist()`

   c) Compute the sample mean and variance from the drawn sample. Verify that they match (roughly) to the true mean and variance of the distribution.
   Hint! Useful R functions: `mean()` and `var()`

   d) Estimate the central 95% probability interval of the distribution from the drawn samples.
   Hint! Useful R functions: `quantile()`

4. **(Bayes' theorem)** 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.

   The researchers are happy with these preliminary results (about 97% success rate), and wish to get the test to market as soon as possible. How would you advise them? Base your answer on Bayes' rule computations.

   Hint : Relatively high false negative (cancer doesn't get detected) or high false positive (un-necessarily administer medication) rates are typically bad and undesirable in tests.

   Hint : 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*.

5. **(Bayes' theorem)** 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. $P(A) = 0.4$).

   a) What is the probability of picking a red ball?
   b) If a red ball was picked, from which box it most probably came from?

   Implement two functions in R that computes the probabilities. Below is an example of how the functions should be named and work if you want to check them with `markmyassignment`.

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

```
##   red white
## A   2     5
## B   2     5
## C   1     1
```

```
p_red(boxes = boxes)
```

```
## [1] 0.3928571
```

```
p_box(boxes = boxes)
```

```
## [1] 0.29090909 0.07272727 0.63636364
```

Note! This is a test case, you will need to change the numbers in the matrix to the numbers in the exercise.

6. **(Bayes' theorem)** 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 values, see Exercise 1.6, page 28, in BDA3). American male singer-actor Elvis Presley (1935 – 1977) had a twin brother who died in birth. Assume that an equal number of boys and girls are born on average. What is the probability that Elvis was an identical twin? Show the steps how you derived the equations to compute that probability.

   Implement this as a function in R that computes the probability.

   Below is an example of how the functions should be named and work if you want to check your result with `markmyassignment`.

```
p_identical_twin(fraternal_prob = 1/125, identical_prob = 1/300)
```

```
## [1] 0.4545455
```

```
p_identical_twin(fraternal_prob = 1/100, identical_prob = 1/500)
```

```
## [1] 0.2857143
```