Lab 1 + Project 1

Requirement

- You are encouraged to cooperate with your friends to finish these exercises (maximum group size = 4), but individual submission is also allowed.
- You can finish the lab in either Chinese or English.
- You can print out your report, including the code, results, figures and observations.
- Hand in to the instructor when you attend the lecture, on or before 23 Mar, 2023 (Thursday). Late submissions will be penalized by a 20% deduction in score.

Note

In this lab, Exercise 1 is compulsary. For Exercise 2, 3 and 4, you are allowed to choose one of them to finish.

You will need data sets from `R` packages `bayesrules`. Descriptions of the data sets can be found in the documentation of the packages (you can find them in the lab1 folder). There are three approaches to use the data

- If you use `R` (recommended), you can install the corresponding package and directly load the data sets from the library.
- If you use `R` but do not want to install the package, you can find the data sets in the lab1 folder ("`.rda`" files), and `load` them into your `R console`.
- If you prefer to use `Python` (note that there may be more coding effects needed for some tasks here when `Python` is used), you can use the package `pyreadr` to read in the data sets in the lab1 folder ("`.rda`" files).

compulsary Exercise

Exercise 1:

Hypothesis Testing

In this exercise, you will need either of the following data set from the `bayesrules` package:

- `fake_news`: A dataset containing data behind the study "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media" (`https://arxiv.org/abs/1809.01286`). The news articles in this dataset were posted to Facebook in September 2016.

- `pulse_of_the_nation`: Cards Against Humanity's "Pulse of the Nation" project (`https://thepulseofthenation.com/`) conducted monthly polls into people's social and political views, as well as some silly things. This data includes responses to a subset of questions included in the poll conducted in September 2017.

Take a look at the documentation to see what are the available features in the data sets. **Before you browse through the data**, brainstorm with your friends to come up with a testable hypothesis using two-sample test, based on the available features and your wisdom. Some examples: "fake news tend to use more words in capital letters in the text", "the people who prefer to be wise but unhappy (rather than unwise but happy) have higher income".

Problems:

1. Pick a data set and write down your question of interest. What is the corresponding null hypothesis?

2. Plot some graphs to visualize the features related to your question above.

3. Test the hypothesis based on the data set (using two-sample test). Compute the p-value. At the significance level $\alpha = 0.05$, do you reject the hypothesis or not?

   Remark: if you want to use t-test, be aware that it may not be appropriate if the sample deviate from normal distribution a lot (small deviation is OK), or the sample size is too small. In those cases, you can use the permutation test or the Wilcoxon rank sum test, which can be done in the `coin` package in `R` (see Sec. 12.2 of the reference book "R in action 2nd Ed" for more details; the book is available in our QQ group).

4. Compute the effect size for the problem of interest. For two-sample test, a straightforward measure is Cohen's d

   $$d = \frac{\bar{Y} - \bar{X}}{s_p}, \tag{1}$$

   where s_p is the standard deviation of the combined sample.

   A rule of thumb interpretation for Cohen's d is:

   $|d| < 0.2$: negligible effect;

   $0.2 \leq |d| < 0.5$: small effect;

   $0.5 \leq |d| < 0.8$: medium effect;

   otherwise: large effect.

5. Alex works in a newspaper office. He wanted to write headline news based on the data, so he generated many hypotheses as above, and decided to pick the tests with p-values below level $\alpha = 0.05$ and claimed one of them as a significant discovery and reported the corresponding p-value. Explain to him why this is problematic.

Choose One of the Following Exercises to Finish

Exercise 2:

GMM for Weather Data

In this exercise, you will need the `weather_australia` data from the `R` package called `bayesrules` (`https://github.com/bayes-rules/bayesrules/`). It is a sub-sample of daily weather information from a large weather data set for Australian cities.

Format: it is a data frame with 300 daily observations and 22 variables, including records for various weather conditions, as well as the location and the date of that observation.

As the observations spanned the whole year, we will consider those recorded in summer (in Australia, take it to be Dec, Jan and Feb), and focus on a few weather features in the afternoon: `windspeed3pm`, `humidity3pm`, `pressure3pm` and `temp3pm`. The target of this lab exercise is to fit a probabilistic model to these observations.

Problems:

1. Load the `weather_australia` data in to `R`. Read the documentation on the descriptions of the available features.

   Take a subset of the data, keeping the observations corresponding to summer (Dec, Jan and Feb).

   Further create a data frame, keeping the following "afternoon" features: `windspeed3pm`, `humidity3pm`, `pressure3pm` and `temp3pm`, and use some graphical methods to visualize the data.

   For above tasks, you need to keep only some of the rows and columns of the original data frame. If you don't know how to do that in `R`, take a look at Sec. 4. 10 of the reference book "R in action 2nd Ed" (available in our QQ group).

2. Use Gaussian mixture models (GMM) to fit the above-obtained data with 4 continuous variables. Use BIC as a criteria to select the model.

   Describe your observations and interpret the model parameters of the fitted GMM.

3. How many clusters are identified, and what is the shape of the GMM model being selected? Does it meet your expectation?

   As there are additional features available in the original data set (in particular, the location of the weather station), think of a way to evaluate the outcome of the GMM clustering.

4. From the fitted GMM, which features can differentiate the clusters most effectively?

5. Compute the 95% confidence intervals of mean parameters $\boldsymbol{\mu}_k$ of each cluster k. You can use the bootstrap method for this purpose.

Exercise 3:

Probability Integral Transform and The Null Distribution of p-values

Suppose that X is a continuous random variable, with density $f_X(x)$ and cumulative distribution function (CDF)

$$F_X(x) = \int_{\infty}^{x} f_X(x')\mathrm{d}x'. \tag{2}$$

Consider transformation of random variable $Y = T(X)$, with a transformation map $T(\cdot)$ (essentially a “1D normalizing flow”).

The probability integral transform (PIT) states that, using $F_X(\cdot)$ as the transformation map, the random variable $Y = F_X(X)$ follows a standard uniform distribution with density

$$f_Y(y) = 1, \quad 0 \leq y \leq 1. \tag{3}$$

(Remark: in the spirit of normalizing flow, you can reverse the PIT to generate a random variable X following a complex distribution f_X, by transformation $X = F_X^{-1}(Y), \ \ Y \sim \text{Unif}[0, 1]$.)

Problems:

1. Prove the probability integral transform.
2. Consider a one-sided hypothesis testing task. Under the null hypothesis H_0, suppose the test statistic T has density $f(t|H_0)$, and the p-value is computed as $p = \int_T^\infty f(t'|H_0)\mathrm{d}t'$.

   Show that under H_0, the p-value follows a standard uniform distribution.
3. Perform computer simulations to explore whether the conclusion in Step 2 also holds for two-sided tests or not.

   For simplicity, consider a one-sample t-test ($H_0 : \mu_X = \mu_0$). Generate a random sample of size n, drawn from $\mathcal{N}(\mu_0, 1)$. Test H_0 and extract the p-value (by using `t.test(...)$p.value` in `R`). Replicate the simulation for $M = 10000$ times, and plot the distribution of the obtained p-values.

   Instead of writing `for` loop, you may find the function `replicate()` useful.

(Remark: the property that the p-value under the null follows a standard uniform is very useful in multiple testing.)

Exercise 4:

If you are already an `R` expert and find the above exercises too straightforward, you are encouraged to suggest tasks of your interests to maximize what you can learn from the course, e.g.,

- Use methods of density estimation and hypothesis testing on data sets of your interests.
- Explore some interesting methods and/or applications of generative models (such as optimal transport methods, GAN, normalizing flow, diffusion models, etc).