Learning Agents - Assessment 1
Spring 2024
Due 23:59 GMT on Wednesday 31st January 2024

This coursework should be submitted as a pdf version of a jupyter notebook, consisting of no more than 12 pages. To convert your notebook to a pdf, you can use 'File -> Download As' and select one of the pdf options, or 'File -> Print Preview' and then print the preview to a pdf. You should also upload your `.ipynb` file. This coursework is out of 30 and counts for 15% of the grade for the learning agents module.

Your coursework should contain your CID but it should not mention your name. You should write your own submission, including all code. You are welcome to use any sources (websites, books, etc), but you should cite them. You should not share your code or solutions with any other students. Failure to abide by these regulations will be considered misconduct.

Question 1 (20 marks)

Consider the multiple testing setting where we are determining whether there is any difference in the gene expression of a set of genes between a group of healthy patients (group A) and patients with colon cancer (group B). Data on the gene expression of of 100 genes from $n_A = 30$ healthy patients is given in the file `healthy.csv`, and the equivalent gene expressions from $n_B = 40$ patients with the cancer is given in the file `cancer.csv`.

(a) (1 mark) Write down the formal hypotheses that you wish to test.

(b) (2 marks) Are the hypotheses tests in Part a independent? Briefly discuss any implications of your answer for the multiple testing procedure.

(c) (2 marks) Conduct a multiple hypothesis test on the data in `healthy.csv` and `cancer.csv` using the Benjamini-Hochberg procedure to ensure that the False Discovery Rate (FDR) is below $q = 0.1$. Interpret your results.

(d) (3 marks) An alternative procedure for ensuring that the false discovery rate is below some $q > 0$ is outlined below.

- Calculate the p-values p_i for each of the hypothesis tests $H_{0,i}, i = 1, \dots, K$.
- Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$ be the ordered p values, and denote by $H_{0,(i)}$ the null hypothesis corresponding to $p_{(i)}$.
- Let i^* be the largest i such that $p_{(i)} \leq \frac{i}{K \sum_{n=1}^{K} \frac{1}{n}} q$.
- Reject all $H_{0,(i)}$ for $i = 1, \dots, i^*$.

Perform the multiple hypothesis test on the data in `healthy.csv` and `cancer.csv` with this correction and desired FDR $q = 0.1$. Compare your conclusions with those in Part c.

(e) (6 marks) For this part of the question, our aim is to compare the performances of the procedures in Part c and d in terms of the FDR.

(i) Explain in a few sentences why the FDR is an appropriate performance measure and why it may be difficult to calculate exactly from real data.

(ii) Instead of using real data to compare the methods, we can use simulated data. Let m be the number of 'true' hypotheses. For each $m = 0, 5, 10, 15, \ldots, K$:

- Simulate two data sets from appropriate distributions when m of the K null hypotheses are true and all hypotheses are independent, justifying your choice of distributions.
- Using the simulated data set, conduct the hypotheses tests using the Benjamini-Hochberg procedure and the procedure in Part d.
- Repeat the first two steps n times to estimate the FDR, justifying your choice of n.

Plot the FDR as a function of m for both procedures to compare their performance. Comment on your results.

(f) (3 marks) Find a set of p-values $p_1, \ldots, p_K$ such that the outputs of the Benjamini-Hochberg and procedure in Part d differ the most. What does it mean if a set of hypotheses tests have the proposed p-values? Briefly justify your answer.

(g) (3 marks) In light of your answers to the previous questions, discuss whether the Benjamini-Hochberg or the alternative procedure in Part d should be preferred for the data in `healthy.csv` and `cancer.csv`.

(Total: 20 marks)

Question 2 (10 marks)

In this question, we will consider adaptive hypothesis testing. We consider an A/B test where the designer of a website thinks that by altering the location of the adverts, users will spend more time on the website. Let version A be the original version and version B be the version with the adverts in the new location. We want to perform an adaptive hypothesis test to determine if the new advert placement does lead to an increase in the expected time spent on the website. We test the hypothesis:

$$H_0 : \mu_A \geq \mu_B \qquad \text{vs} \qquad H_1 : \mu_A < \mu_B$$

Two adaptive hypothesis tests have been performed. Test I was run with probability 0.5 of assigning a user to each group, Test II had probability 0.9 of assigning a user to group A, and probability 0.1 of assigning them to group B. Both tests were started at the same time run adaptively until the power reached 0.8, using an effect size of $\theta = 5$.

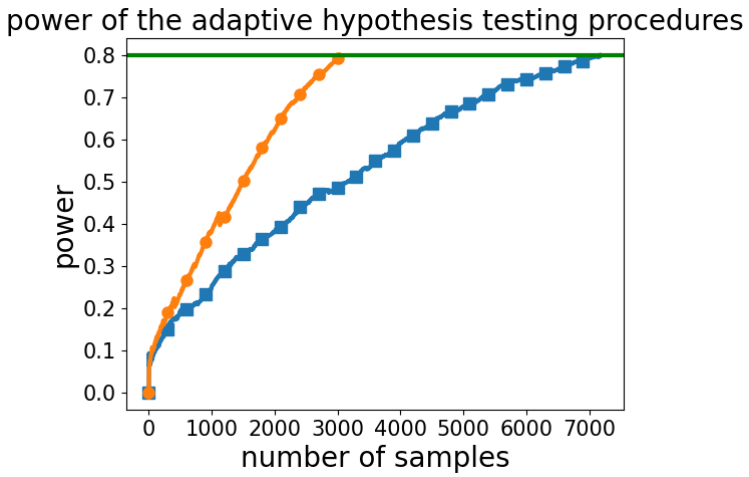

Figure 1: The power of the two procedures as a function of the total number of samples until the power exceeds 0.8.

(a) (2 marks) Figure 1 shows how the power evolves over the duration of the adaptive hypothesis tests. Which line corresponds to which allocation? Briefly justify your answer.

(b) (3 marks) What would you expect to happen to the power curves in Figure 1 if the effect size used was increased? Briefly justify your answer. How would you determine which effect size should be used in practice?

(c) (2 marks) After the data was collected, the hypothesis tests were conducted on the two data sets separately. From Test I, the p-value was $p \approx 0.0035$, whereas for Test II the p-value was $p \approx 0.0109$. How should we combine the outcomes of these two tests to come to a general conclusion and what considerations need to be taken into account when drawing conclusions from these tests?

(d) (3 marks) The stakeholder is reluctant to allow too many users to see the new version B of the website until we know it is good. Describe a procedure for adaptively increasing the proportion of users in group B. You should give pseudocode of your proposed solution, and discuss any potential issues.

(Total: 10 marks)