Bayesian Statistics: Assignment 2

Shota Gugushvili

11 May 2020

1 Summary

Use *either* **R** *or* **Python** and write a script performing Bayesian logistic regression on the Titanic dataset. Document the script.

2 Description

The accompanying file *titanic.csv* contains data on a part (about 2/3) of the Titanic passengers. It gives information on the class of travel (first, second, third), survival (1 stands for survival, 0 for death), names (with titles), gender, age, number of siblings and children, ticket numbers, fares paid, cabin numbers, and ports of embarkment (Southampton, Cherbourg, Queenstown). Data are incomplete and messy; for instance, cabin numbers are largely missing.

The task is to build a Bayesian logistic regression classifier for this dataset. The response is survival (1 or 0), the features used for prediction can be any of the attributes named above, or their combinations. You choose. Missing values can be imputed at will.

Randomly split the data into two parts: training set (600 passengers) and test set (291 passengers).

Use the training data to fit the model. Results of exploratory data analysis can be provided to motivate your choices. Any decisions you make must be based on the training data only. Test data is untouchable, holy and invisible at this stage.

Use the test data to assess accuracy of the resulting classifier. Accuracy must be measured with 0-1 loss (raw accuracy) and with sensitivity/specificity. ROC curve and AUC must be provided.

The following is a naive classifier that must be employed as a baseline:

(i) All women survived.

(ii) All men died.

Your logistic regression classifier is worthless, unless it performs better than this naive classifier.

Software options are **Python** and **R**, with **PyStan**, **PyMC3**, **RStan** and **rstanarm** inclusive. Each has documented examples on Bayesian logistic regression. Choose the tool yourself.

You must produce a report containing your analyses and findings. In your script, it is not required that you wrap up everything under one command. The script may consist of individual parts. It is not necessary that the script implements everything from scratch.

Important points are the following:

- The report must provide background information and read like a good, self-contained story. Also, a brief description of Bayesian logistic regression (with suitable scientific references) must be provided. Use, e.g. the **R** of **Python** help style for inspiration.
- The script must be well-documented and explain each of the steps that you undertook.
- Language and grammar should be correct, *typso* should be filtered out.

3 Hints

In **Python**, **scikit-learn** can compute ROC curve and AUC. There are many options to that end in **R**.

4 Guidelines

The assignment must be submitted by mailing your solution to me no later than May 31, 2020. The address is `shota.gugushvili@wur.nl`. Retain a copy in your Sent folder, in case something goes wrong.

If **R** is used, ideally the submission must be prepared using **R Markdown**. In that case submit both the pdf file and the source Markdown file. Less desirable is a pdf file made with LaTeX and a separate .R file. Either way, zip the two files and name the archive LastnameFirstnameStudentnumberAssignment2.

If **Python** is used, then ideally submit the assignment as a Jupyter Notebook. Less desirable is a pdf file made with LaTeX and a separate .py file. In that case, zip the two files and name the archive LastnameFirstnameStudentnumberAssignment2.

Word files will not be accepted and will result in an automatic failure.

5 Grading

Grading is on the $1 - 10$ scale. For a score 6, the script must be (essentially) complete and functional, and the accompanying text must contain enough details to reconstruct what you did.

Good luck!