EPA 1315. Data Analysis and Visualization

Project Description

Due Date: 8:00 am, Monday 13th November 2017

The learning objective of this assignment is to demonstrate your capability to independently design and implement a Bayesian model. Other learning objectives involve executing and reporting on a data analysis problem, and the ability to use R to manipulate and visualize data.

Choose a set of variables from the World Development Indicators database. Other data sources may be permissible; please discuss with the instructor. I expect that your examples will be comparable in complexity to those discussed by Krushke. Design a suitable model using the examples in the book by Krushke.

Be modest in introducing complexity to your model. If you are ambitious, please consider performing your modelling in two steps. Begin with a simple mode and lock-in your results. Only then add a second phase to your modelling exercise where you expand the model. Document your results in a short policy brief of 2400 to 4000 words. Any additional words beyond 5200 will not be graded.

The project is worth 35% of the grade, and is due as noted above. No late assignments are accepted, but project retakes in subsequent quarters may be permitted in the case of incompletes or course failures. You may work alone, or have up to five partners on this project.

You may optionally ask for formative assessment of your preliminary work. This is not graded but may be to your benefit in improving the work. The following table can be used to guide the structure of your policy brief. The table shows eight questions. These questions stem from the CRISP-DM data mining process, as well as the Bayesian modelling approach discussed in Krushke. You may organize your brief using the CRISP-DM headings, but it is not required. Treat each of the subsections with equal weight or importance. That is, each section should be roughly 300 to 500 words in length.

Table 1. Sections and Subsections of the Work

|  |  |  |  |
| --- | --- | --- | --- |
| Subsection | CRISP-DM | CRISP-DM Questions | Bayesian Modelling Questions |
| 1 | Problem Understanding. | Describe a research or policy problem. Who is interested in the problem or why? Is there any available theory which can guide your choices of variables, relationships or priors? |  |
| 2 | Data Understanding. | Which variables and relationships are relevant to the problem? Document your data, including the source of the data, the variables in the data, the measurement scales in the data. | Identify the data relevant to the research questions. |
| 3 | Data Preparation. | Describe how you prepare the data, including transformation or recoding of data, and handling of missing values. |  |
| 4 | Modelling. | Describe the modelling process. | Define a descriptive model for the relevant data. |
| 5 |  |  | Specify a prior distribution on the parameters. |
| 6 |  |  | Use Bayesian inference to re-allocate credibility across parameter values. |
| 7 | Evaluation. | Evaluate the quality of the model.  Consider the next steps in improving or extending your model. | Check that the posterior predictions mimic the data with reasonable accuracy. |
| 8 | Deployment. | Given theory and data can you predict future behaviour or guide additional policy?  Discuss how you will communicate or report your findings to decision-makers. Or, communicate how you will disseminate the findings on the cloud or social media. Or describe how your model might be put into a production system or algorithm. |  |

The following table provides some ideas about additional files, features or graphics which you might considering adding to your policy brief.

Table 2. Features for Possible Inclusion in Your Policy Brief

|  |  |  |  |
| --- | --- | --- | --- |
| Subsection | CRISP-DM | Bayesian Modelling Questions | Files, Features or Graphics |
| 1 | Problem Understanding. |  | A system diagram might be suitable. |
| 2 | Data Understanding. | Identify the data relevant to the research questions. | A data dictionary might be suitable. |
| 3 | Data Preparation. |  | Excerpts from the R-code might be presented. The whole of the R code might be included as an embedded file. |
| 4 | Modelling. | Define a descriptive model for the relevant data. | A bar chart, line chart, or scatter plot might be suitable for showing the predictor variables. |
| 5 |  | Specify a prior distribution on the parameters. | A hierarchical influence diagram might be suitable. |
| 6 |  | Use Bayesian inference to re-allocate credibility across parameter values. | Excerpts from the R-code might be presented. The whole of the R code might be included as an embedded file. You might include one or more of the diagnostic plots. |
| 7 | Evaluation. | Check that the posterior predictions mimic the data with reasonable accuracy. | You might include the posterior predictive plot. |
| 8 | Deployment. |  | A visualization which easily communicates your finding might be suitable. |

The following table provides a rubric which provides some guidance into how your contribution may be assessed.

Table 3. A Rubric for Grading

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 5 (or less) | 6 | 7 | 8 | 9 (or more) |
| 1 | A compelling problem is not described | A problem is poorly described | A problem is compellingly described | A problem is compellingly described but the decision-maker is not well-discussed | A compelling problem is described; an interested decision-maker is well-discussed or relevant theories are described |
| 2 | Not presented | A relevant predicted variable is presented | A relevant predicted and predictor variable is presented | A data dictionary is presented | A data dictionary is presented and additional exploratory insight is offered |
| 3 | No cleaning | No discussion | Appropriate cleaning was performed but not discussed | Appropriate cleaning was performed and discussed | Appropriate additional efforts beyond the lab were used in cleaning |
| 4 | Not given | The descriptive model is poorly discussed | The descriptive model is discussed | The descriptive model is well-discussed | Additional descriptive insight into the model is offered in words or figures |
| 5 | Not given | The model selected is not discussed | The model selected is discussed | The model selected is strongly motivated; the priors are well-discussed or parameterized | The model selected is strongly motivated and additional insight is provided |
| 6 | The model appears to be in error | A model discussed from Krushke was presented, but the diagnostics are not discussed | A model discussed from Krushke was presented and the diagnostics are used | The model was expanded after initial estimation | The model involved additional self-study from Krushke |
| 7 | The posterior predictive distribution appears to be in error | The posterior predictive distribution is not considered | The posterior distribution does not match with reasonable accuracy | The posterior distribution does not match with reasonable accuracy, but a good discussion is presented why | The posterior distribution matches with reasonable accuracy, and a good discussion is presented why |
| 8 | Not given | The deployment of the model is poorly described | The deployment of the model is described | The deployment of the model is well-described | The deployment is well-described and additional insight is given into its use and dissemination |