STAT30270/STAT40590 – Statistical Machine Learning

Data analysis project

Deadline - Friday 24 April at 18:00

Data: DeepSolar database

The data is a subset of the *DeepSolar* database, a solar installation database for the US, built by extracting information from satellite images. Photovoltaic panel installations are identified from over one billion image tiles covering all urban areas as well as locations in the US by means of an advanced machine learning framework. Each image tile records the amount of solar panel systems (in terms of panel surface and number of solar panels) and is complemented with features describing social, economic, environmental, geographical, and meteorological aspects. As such, the database can be employed to relate key environmental, weather and socioeconomic factors with the adoption of solar photovoltaics energy production.

More information about this database is at the link:

`http://web.stanford.edu/group/deepsolar/home`

The dataset `data_project_deepsolar.csv` contains a subset of the *DeepSolar* database. Each row of the dataset is a "tile" of interest, that is an area corresponding to a detected solar power system, constituted by a set of solar panels on top of a building or at a single location such as a solar farm. For each system, a collection of features record social, economic, environmental, geographical, and meteorological aspects of the tile (area) in which the system has been detected. Information about the features are in the file `data_project_deepsolar_info.csv`.

Task: Predict solar power system coverage

The target variable is `solar_system_count`. This variable is a binary variable indicating the coverage of solar power systems in a given tile. The variable takes outcome `low` if the tile has a low number of solar power systems (less than or equal to 10), while it takes outcome `high` if the tile has a large number of solar power systems (more than 10).

Use the supervised classification methods described in this course to predict the solar power system coverage of a tile given the collection of predictor features. Use a range of methods and evaluate/discuss their relative merits. In the end, you should base your conclusions and interpretations on the best model you find. Write a report summarising your analysis.

Notes:

- Some preliminary exploratory analyses and pre-processing could be useful.
- Some features are related to households, other to entire areas. Moreover, some features are derived from others.
- It is not necessary to include all the predictor variables in the predictive model. Some variables can be discarded upon providing valid motivation and explanation.

Only STAT40590 students - Discard the target variable `solar_system_count` and the variable `state` (recording the US state of a tile). Perform a cluster analysis on a subset of the predictor variables of your choice and compare the obtained clustering with the classification into states. Motivate the choice of the subset of predictor variables in relation to the purpose of the cluster analysis and provide an interpretation of the detected clusters.

Guidelines

Indicatively, the report should include the following sections:

1. **Title** - Project title.
2. **Abstract** - The abstract should concisely summarise the information in the report.
3. **Introduction** - The introduction should provide relevant background information. The introduction should also be used to introduce the data analyzed as well as the main research question(s) and why the question(s) is/are worth answering.
4. **Methods** - The methods section should describe the entire analysis approach and contain the discussion about all modelling stages. Any data preprocessing step should be briefly described here. This section should also refer to the methods used to assess the performance.
5. **Results and discussion** - This section should present the results of the analysis and an interpretation of these.
6. **Conclusion** - The conclusion should summarise the important points made in the discussion.
7. **References** - The references section should contain citations for any source that was referenced throughout the report. R packages that were not part of any of the lab sessions or lectures for this course should be clearly referenced also.

Please, use a concise writing style and integrate all relevant tables and figures into the text. Comments/discussions must be added to tables and figures included in the text.

The body of the report should not exceed 10 pages (12 pages for STAT40590). Please, submit the report document as a single PDF file. Note that few additional pages could be used if needed (because of figures or output), but the report would still need to be concise.

Include the R code used for analysis in the appendix. The report can also be produced using R Markdown, with the code included in the main text. The code must be working and the analysis must be reproducible in all parts. Please, ensure that this code is tidy and appropriately commented.