Assignment 3

STAT30270 – Statistical Machine Learning

Deadline - Wednesday 26 April at 17:00

Data: DeepSolar database

The data is a subset of the *DeepSolar* database, a solar installation database for the United States, 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, which have been collected from readily available data sources (NASA Surface Meteorology and Solar Energy and the American Community Survey). As such, the database can be employed to relate key environmental, weather and socioeconomic factors with the adoption of solar photovoltaic energy production.

More information about this database is at the link:

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

The dataset `data` in the file `data_hw3_deepsolar.RData` contains a subset of the *DeepSolar* database. Each row of the dataset is a "tile" of interest, that is an area related to a detected solar power system. A *solar power system* consists of a set of solar panels on top of a building, or at a single location such as a solar farm, or a similar system to convert solar irradiance to energy. For each system, a collection of features record social, economic, housing, geographical, and meteorological aspects of the tile (area) in which the system has been detected. Information about these features are reported below:

- `solar_system_coverage`: a binary indicator variable indicating the coverage of solar power systems in a tile. `low` for low-to-medium coverage, `high` for high coverage (more than 10 systems).
- `average_household_income`: average annual houeshold income (US dollars).
- `employ_rate`: employment rate.
- `population_density`: population density in a tile (mile^2).
- `housing_unit_count`: total number of housing units.
- `housing_unit_median_value`: median housing unit value (US dollars).
- `occupancy_vacant_rate`: ratio of vacant housing units.
- `heating_fuel_gas_rate`: ratio of house units using gas as heating fuel.
- `heating_fuel_electricity_rate`: ratio of house units using electricity as heating fuel.
- `heating_fuel_oil_rate`: ratio of house units using oil as heating fuel.
- `land_area`: total land area (mile^2).
- `water_area`: total water area (mile^2).
- `air_temperature`: air temperature (Celsius).
- `earth_temperature`: earth temperature (Celsius).
- `daily_solar_radiation`: daily solar radiation (kWh/m^2).

Task: Predict solar power system coverage

The target variable is `solar_system_coverage`. 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-to-medium number of solar power systems, while it takes outcome `high` if the tile has a larger number of solar power systems (more than 10).

Detection and estimation of the coverage level of a tile is an expensive process. In fact, labeling of the tiles in relation to their solar power system coverage involves processing and analysis of satellite image data, and implementation of complex and computationally intensive advanced machine learning methods. Researchers have interest in building a reliable and scalable classifier, which, using the available data on environmental, weather, and socioeconomic measurements, can predict the classification of new tiles associated with their solar power system coverage. Predictions from this "simpler" classifier could subsequently be used to aid the tile labeling process, making it less computationally expensive in some instances.

The task is to build a supervised learning model to predict if a tile can be considered as containing low-to-medium or high solar power system coverage, using the available social, economic, housing, geographical, and meteorological features.

1. Implement at least 3 different supervised learning methods to predict if a tile has high (`high`) solar power system coverage or not, on the basis of the input features. Employ an appropriate framework to compare and tune the different methods considered, evaluating and discussing their relative merits. *(70 marks)*
2. Select the best model a predicting if a tile has high solar power system coverage from the available numerical features data. *(10 marks)*
3. Use appropriately some test data in order to evaluate the generalized predictive performance of the best selected classifier. Provide a discussion about the ability of the selected model at detecting correctly `high` and `low` coverage. *(20 marks)*

Instructions and guidelines

- **Discuss and motivate the various decisions taken in all stages of the analysis**.
- If you wish, you can use only a subset of the features of the data in the model. However, for full marks you **must clearly motivate your choice** and why some features are discarded.
- You will not be evaluated on the basis the predictive performance of your classifiers, but you would need to show that attempts have been considered to build a classifier with reasonable performance.
- Submitting only code does not provide any marks.

Submission rules

- Write a short and tidy report and submit it as a **single pdf file** (approximately max 8-10 pages, code excluded).
- Include the R code used for analysis in the report. The report can be produced using R Markdown, with the code included in the main text or as an appendix. **The code must be working and the analysis must be reproducible in all parts**.
- Multiple submissions before deadline are allowed and only the latest one will be considered for marking.
- Submission after deadline will incur in penalization as UCD rules (see "Module details" document).
- **Plagiarism is strictly prohibited and will incur in serious penalization** (see "Module details" document and "Information materials" tab).