QBUS3830
Advanced Analytics

Semester 2, 2018

Homework Task 4: Additive Models

1 Task description

In this task you will work with the California Housing Dataset. Our objective to model the average house price in a geographic area as a function of characteristics of the housing stock, demographics, and location. An interesting feature of this dataset is the presence of spatial information: we have the latitude and longitude for each census tract.

The accompanying Jupyter Notebook provides a standard analysis. We find that the gradient boosting method has substantially better generalisation performance than a linear regression, which is not surprising since the latter cannot account for spatial effects.

Your task is to implement a Python class that estimates the following model:

$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ij} + g(\text{latitude}_i, \text{longitude}_i) + \varepsilon_i,$$

where g is an unknown function. Fit the model and compare the test performance to that of the benchmark models.

The estimation will probably be very slow to run, maybe hours. Test your code with a small sample, and use the print function to keep track of the progress.

A generalised additive model can outperform gradient boosting for this problem. However, to obtain this result you would need to estimate a model of the type

$$Y_i = \beta_0 + \sum_{j=1}^{6} f_j(x_{ij}) + g(\text{latitude}_i, \text{longitude}_i) + \varepsilon_i.$$

You can also try a neural network for your own knowledge and practice.

2 Rules

The code for fitting the generalised additive model must be your own work.

3 Rubric

You will get the full marks if you follow the instructions and fit the model correctly, as measured by performance on the test data.