INFO 284, Spring 2018, Obligatory Individual Assignment

Regression

Deadline June 1, 2018 (Inspera)

Outline

The goal of this individual project assignment is to demonstrate practical competence in using regression methods. The data set included in this assignment (Flaveria.csv) includes 48 data points. Each data point is represented using two features: “N level” and “species”. The third column in the data set called “Plant Weight (g)” is the target value for which you need to build a prediction model.

Each data point describes one plant of the species *flaveria*. Each plant is described by the nitrogen level (“N level”) used to grow it and the subspecies of Faveria (“species”) that data point belongs to. The plants were grown in a controlled laboratory environment and then plucked and individually measured (weighed). The measurement results were recorded as values for “Plant Weight (g)”. There are three possible values for the feature “N level”. These are L, M, and H indicating low, medium, and high levels of nitrogen used for growing the plant. There are six values for the feature “species". These are: brownii, pringlei, trinervia, ramosissima, robusta, and bidentis.

Task: You need to build a regression model that predicts the value for “Plant Weight (g)”. You can use any supervised learning (regression) method from the scikit-learn library. You can try as many methods as you want but you may only submit one solution.

Submit: Your python code and data set in a zip folder.

Evaluation: All models will be evaluated using a test data set (of 6 data points) that you have not been supplied with. For each student’s program, the coefficient of determination of the prediction (R^2 score which is returned by the `score()` scikit-learn method) will be calculated using the test data set. The score values will be ranked across all students. All functioning solutions will be ranked. To obtain a grade E on this assignment it is sufficient to be ranked, namely that you have submitted a solution to the regression problem that is able to do a prediction on the test data set, regardless of the `score()` that solution attains. The grades will be determined using the `score()` values - the solutions with the highest scores will get the highest grades.