San José State University
Computer Science Department
CS156, Introduction to Artificial Intelligence, Fall 2021

Homework #3

Objective:

This homework's objective is to implement a regression model to predict quality of concrete based on its composition.

Details:

For this assignment you will be using the dataset containing concrete quality as well as information about composition . This dataset is publicly available from kaggle dataset repository and you can find more information about it: https://www.kaggle.com/maajdl/yeh-concret-data

Use input file homework3_input_data.csv for this assignment. The rows of this csv-formatted file are different countries surveyed for this report. The columns are the variables assessed in this report. You will not use all the columns of this file for your prediction model, so read the next paragraph carefully.

Implement a regression model to predict the concrete compressive strength in megapascal (csMPa) based on the other independent variables: *cement, slag, flyash, water, superplasticizer, coarseaggregate, fineaggregate, age*. Notice that the predictor variable *csMPa* is a continuous variable. Your model will predict a real number rather than a categorical label.

You can use the code from my notebook examples as a reference to help you get started:
- Regression.Boston.ipynb

Your submission should include the following:

1. Load the dataset.
2. Plot all independent variables vs. the dependent variable (similar to how I demonstrated in the Regression.Boston.ipynb Jupyter notebook). This will show you the relationship between each independent variable and the dependent variable.
3. Compute and plot a correlation matrix between the independent variables (similar to how I demonstrated in the Regression.Boston.ipynb Jupyter notebook). This will show you how independent all the input variables are between each other.
4. Break the data into the training and test datasets.
5. Train a linear regression model to predict the output/dependent variable (*Happiness Score*) based on the input variables I specified in the description of this assignment.
6. Report (print out) the *mean squared error* and *coefficient of determination* for the test data as your model performance indicators. Remember to use the set aside test data for this.

Submission:

Email your assignment submission to me at Yulia.Newton@sjsu.edu and the grader (Gursimran Singh) at gursimransingh@sjsu.edu. Make sure to email this submission by 11:59pm on the due date listed in Canvas. Your sent email is the proof of submission. The subject of the email should say "CS156 Assignment 3". In the body of the email list your name as it appears on the class roster and your student ID. Attach to this email both the pdf of your Jupyter notebook, which contains the solution for this homework assignment, as well as the notebook itself (the notebook file with .ipynb extension). Make sure to submit both files, otherwise the submission will not be considered complete.

Grading:

I will return the grades as fast as we can grade this homework. Normally it should not take more than a few weeks.

A total of 10 points are possible for this homework assignment.