

Assignment 02: Using Scikit-Learn and statsmodel for Regression and Classification

CSci 574: Machine Learning

Description

In this assignment you will be performing a regression and classification task using the `scikit-learn` machine learning framework, and the Python `statsmodel` library. You should work through the tutorial on using `scikit-learn` and `statsmodel` before doing this assignment, as well as work on the materials from our units on regression and classification tasks.

For the first part of this assignment, I recommend looking through the following tutorials on using Scikit Learn and the statsmodel library for linear regression:

[A Beginners guide to Linear Regression in Python with Scikit-Learn](#)

[Simple Linear Regression with statsmodel Library](#)

I am using this material as a reference when developing the work and questions for Task 1 below.

Instructions:

- You need to use the class development environment and make sure that you are pushing your assignments to your GitHub classroom and they are successfully passing the autograder.
- Avoid using for-loops and while loops, unless you are explicitly told to do so.
- Do not modify the `### TESTED FUNCTION [function name]` cells. These cells call unit tests on the functions you are asked to write for these assignments.
- All functions you need to write should be placed into the `src/assignment_tasks.py` file. Functions that are tested and graded are imported from there into this notebook. You have again been given the function declaration and appropriate Docstrings for the functions in this assignment.
- After coding your function, run the `### TESTED FUNCTION` cell to determine if it is passing the assignment unit tests and that your result is correct.

Objectives

After this assignment you will:

- Have experience using the basics of the `scikit-learn` library and API.
- Be introduced to the `statsmodel` library for doing statistical analysis.
- Be familiar with some concepts of linear and logistic regression, such as the model parameters and the mse and R^2 score that you can obtain from fitting a linear or logistic regression model.
- Have some practice in plotting and visualization of data.
- Also get a little practice on encoding data and imputing missing data for a dataset.
- Get an introduction to the concept of a decision boundary, that we will go into much more detail later on in this class.

Let's get started!

Overview and Setup

Assignment Tasks

Assignment Submission

Additional Information