Project 1

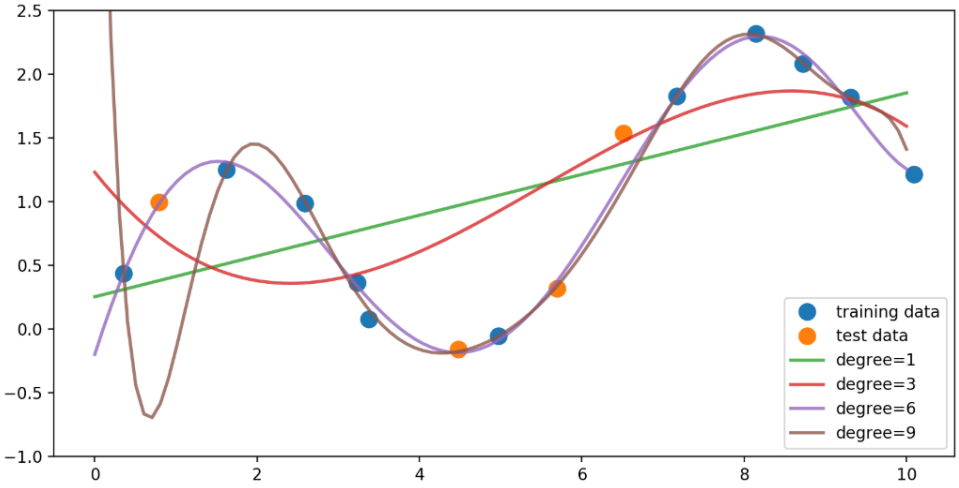
Upload data file "Project 1-Data.csv" and perform following studies:

(a) Split data into training data and test data by

X_train, X_test, y_train, y_test = train_test_split(x, y, random_state=0)

Scatter plot training dataset and test dataset on one figure

(b) Write a program that fits a polynomial LinearRegression model on the *training data* X_train for degrees 1, 3, 6, and 9. (Use PolynomialFeatures in sklearn.preprocessing to create the polynomial features and then fit a linear regression model) For each model, find 100 predicted values over the interval x = 0 to 10 (e.g. np.linspace(0,10,100)) and store this in a numpy array. The first row of this array should correspond to the output from the model trained on degree 1, the second row degree 3, the third row degree 6, and the fourth row degree 9. Plot fitted value lines on the same figure as (a). You program should print out the fitted data as an array of 4x100 and plot should look like



(c) Write a program that fits a polynomial LinearRegression model on the training data X_train for degrees 0 through 9. For each model compute the R-square (coefficient of determination) regression score on the training data as well as the test data, and return both of these arrays in a tuple. Based on the R-square scores (degree levels 0 through 9), what degree level corresponds to a model that is underfitting? What degree level corresponds to a model that is overfitting? What

choice of degree level would provide a model with good generalization performance on this dataset?

(d) Training models on high degree polynomial features can result in overly complex models that overfit, so we often use regularized versions of the model to constrain model complexity, as we saw with Ridge and Lasso linear regression. For this question, train two models: a non-regularized LinearRegression model (default parameters) and a regularized Lasso Regression model (with parameters alpha=0.01, max_iter=10000) both on polynomial features of degree 12. Return the R-square score for both the LinearRegression and Lasso model's test sets.