MA0218 Exercise 4 : Linear Regression

Workflow

1. Download the .ipynb files and data files posted corresponding to this exercise and store them in a single folder.
2. Open and explore the .ipynb files (notebooks) that you just downloaded and go through "Preparation" as follows.
3. The walk-through videos posted on NTU Learn (under Course Content) may help you with this "Preparation" too.
4. Create a new Jupyter Notebook, name it MatID_Exercise4_solution.ipynb, where "MatID" is your Matric Number.
5. Solve the "Problems" posted below by writing code and comments in MatID_Exercise4_solution.ipynb notebook.
6. Submit the Notebook MatID_Exercise4_solution.ipynb to your respective Lab Group's Course Site on NTU Learn.
7. Talk to your TA at the Lab Session regarding submission portal and/or procedure before you submit your solution.

Try to solve the problems on your own. Take help and hints from the "Preparation" codes and the walk-through videos. If you are still stuck, talk to your TA in the Lab Session to get help/hints. Try *not to discuss this problem* with your classmates.

Note : Don't forget to import the Essential Python Libraries required for solving the Exercise. Write code in the usual "Code" cells, and notes/comments in "Markdown" cells of the Notebook. Check the preparation notebooks for guidance.

Preparation

M3 LinearRegression.ipynb Check how to perform Linear Regression on the Pokemon data (pokemonData.csv)

Problems

Download the dataset train.csv and the associated text file data_description.txt posted with this Exercise.

Problem 1 : Predicting SalePrice using GrLivArea

Import the complete dataset "train.csv" in Jupyter : houseData = pd.read_csv('train.csv')
Use the following Numeric variables from the dataset in this problem : GrLivArea and SalePrice

a) Plot SalePrice against GrLivArea using any appropriate bivariate plot to note the strong linear relationship.

b) Print the correlation coefficient between these two variables to get a numerical evidence of the relationship.

c) Import Linear Regression model from Scikit-Learn : from sklearn.linear_model import LinearRegression

d) Partition the dataset houseData into two "random" portions : Train Data (1100 rows) and Test Data (360 rows).

e) Training : Fit a Linear Regression model on the Train Dataset to predict or estimate SalePrice using GrLivArea.

f) Print the coefficients of the Linear Regression model you just fit, and plot the regression line on a scatterplot.

g) Print Explained Variance (R^2) and Mean Squared Error (MSE) on Train Data to check Goodness of Fit of model.

h) Predict SalePrice in case of Test Data using the Linear Regression model and the predictor variable GrLivArea.

i) Plot the predictions on a Scatterplot of GrLivArea and SalePrice in the Test Data to visualize model accuracy.

j) Print the Mean Squared Error (MSE) on Test Data to check Goodness of Fit of model, compared to the Training.

Problem 2 : Predicting SalePrice using Other Variables

Perform all the above steps on “SalePrice” against each of the variables “LotArea”, “TotalBsmtSF”, “GarageArea” oneby-one to perform individual Linear Regressions and obtain individual univariate Linear Regression Models in each case.

Problem 3 : Best Uni-Variate Model to Predict SalePrice

Compare and contrast the four models in terms of Explained Variance (R^2) and Mean Squared Error (MSE) on Train Data, the accuracy of prediction on Test Data, and comment on which model you think is the best to predict “SalePrice”.

Feel free to comment throughout the notebook (using markdown) to explain and justify your solution and conclusion.

Extra Resources

You may read more about the LinearRegression model you use in this exercise in the following references.

LinearRegression : https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Other Linear Models (Scikit Learn) : https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model