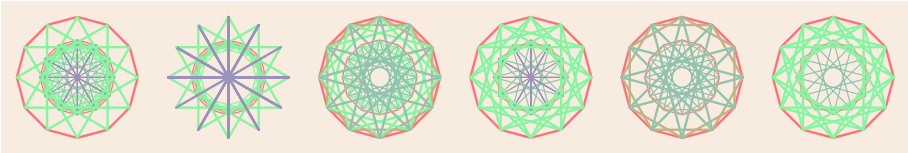
Exercise

CSC316 Machine Learning
Professor Leon Tabak

15 February 2022



Cut and paste from existing code as much as possible in this exercise.

Write a program that...

- Builds a DataFrame that contains the California Housing Prices dataset.
- Splits the dataset into a training set and test set. Do not bother with stratified sampling.
- From a single DataFrame that contains all columns except for *median_house_value*, create...
  - A DataFrame that contains only the labels.
  - A DataFrame that contains only the single categorical variable in the dataset.
  - A DataFrame that contains only the single column that contains missing values. This is the *total_bedrooms* column.
  - A DataFrame that contains new variables: *rooms_per_household*, *population_per_household*, *bedrooms_per_room*.

- Replaces all missing values in *total_bedrooms* with the median value in that column.
- Concatenates the DataFrames to create a new DataFrame that contains all numerical variables (except for *median_house_value*).
- Scales all of the numerical data in a way that makes the mean value in each column is 0.0 and the standard deviation is 1.0.
- Uses the Pandas *get_dummies()* function for one hot encoding of the catagorical variable.
- Concatenates the DataFrames to create a single DataFrame that contains all columns (except for *median_house_value*).
- Creates an instance of *LinearRegression*.
- Calls LinearRegression's *fit()* method.
- Calls LinearRegression's *predict()* method.
- Uses the emphmean_squared_error() function to produce a measure of how well the model predicts housing prices.