EEM 409 Makine Öğrenmesi’ne Giriş
EEE 6113 Machine Learning

Regression Assignment

1. Load house sales data (home_data.csv) and explore.
   a. Print a frequency plot of the data with respect to features like zip code and house price.
   b. What is the highest and lowest house sale prices in the data? What is the average house sale price?
   c. Determine the neighborhood (zip code) that has the highest average house sale price. Select only the houses with this zip code, and compute the average price.

2. Select the houses that have ‘sqft_living’ higher than 2000 sqft but not larger than 4000 sqft. What fraction of the all houses have ‘sqft_living’ in this range?
   Note: If you are using SFrame, you can use logical filters to select its rows (see Logical Filter section of this documentation)

3. Build a simple (linear) regression model that predicts price from square feet.
   Note: You can choose to use whichever you like: scikit-learn or turicreate for the regression model.

4. Randomly split the data to train-and-test sets with 80%-20% proportions, respectively.
   Note: When doing the train-test split, make sure you use seed=0, so you get the same training and test sets.

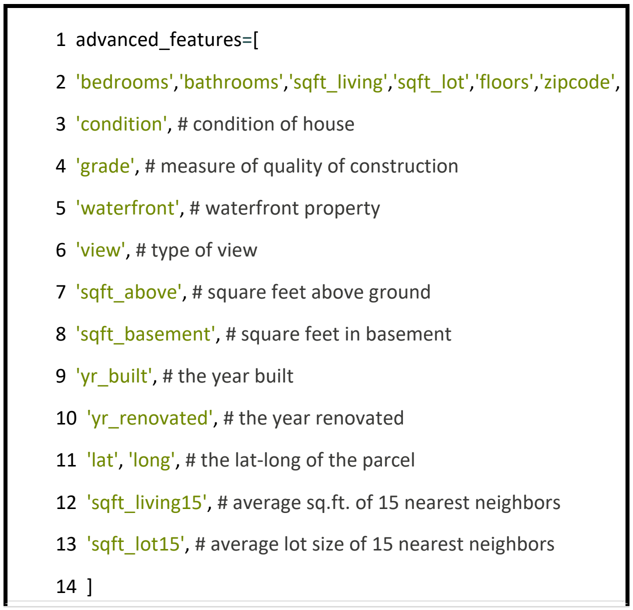
5. Train your simple regression model on your training set, and print the progress of the training.

6. Evaluate the performance of your trained model. What is the RMSE (root mean square error)?

7. Explore your trained model. Print the coefficients of your trained model.

8. Now, obtain the predictions of your trained model on the test set. Plot the test set (square feet vs. price) and draw the linear regression line of your trained model on the plot.

9. Build a simple linear regression model using the following features (*advanced_features*):



```
advanced_features=[
'bedrooms','bathrooms','sqft_living','sqft_lot','floors','zipcode',
'condition', # condition of house
'grade', # measure of quality of construction
'waterfront', # waterfront property
'view', # type of view
'sqft_above', # square feet above ground
'sqft_basement', # square feet in basement
'yr_built', # the year built
'yr_renovated', # the year renovated
'lat', 'long', # the lat-long of the parcel
'sqft_living15', # average sq.ft. of 15 nearest neighbors
'sqft_lot15', # average lot size of 15 nearest neighbors
]
```

Compute the RMSE (root mean squared error) on the test_data for the model using *advanced_features*. What is the difference in RMSE between the model trained with *square feet* and the one trained with *advanced_features*?

Note: Both models must be trained on the original sales train dataset, not the one filtered on `sqft_living`.

10. In the above, we used a linear regression (order=1) model. But instead you may use a polynomial (higher order) regression model as well.
    Now, build two polynomial regression models, one with order=2 and another with order=5. Compute the RMSE on the test data for the models using the *'sqft_living' feature*, and *advanced_features*.

11. Plot the test set (square feet vs. price) and draw the regression lines of your trained models (order=2 and order=5) on the plot.

Note: Check if turicreate toolkit contains a polynomial regression model. If not, you may check scikit-learn toolkit based solutions helpful for this task (see the following documentation for example).

12. Apply learned models to make predictions for houses with IDs '5309101200', and '1925069082'. Explore the details of these houses: how many square feet they have, how many bedrooms they have, etc? What are the predictions of your models for these houses? How much difference there is between their actual and predicted prices? Which model would you prefer, why?