Homework 2 Report (Zillow Challenge, Kaggle)

Anjul Kumar Tyagi
111482304

Linear Regression

Given n features and k samples of each of these n features, if we want to predict a linear model that would satisfy the equation: $(X^TX)^{-1}$ W = X^TY, where X is the matrix of n-features with k entries, one on each row, W is the weight matrix that is assigned to each of these features and Y is the predicted outcome. The maximum absolute weight signifies the most important feature from the data set.

While we train the data, we know X and Y and calculate W. To test the data, we use this W to calculate Y for a given test set X. The formula for linear regression comes from the **Least Square** form where $(X^TX)^{-1}$ is a square matrix transformation of X with least square error.

Polynomial Regression

Polynomial regression tries to fit the data with a non linear polynomial unlike Linear regression. For a degree 3 polynomial, the equation would be like A x_1^2 + B x_2 + C where A, B are weights for features x_1 and x_2 and C is a constant.

The main idea behind polynomial regression is that it tries to fit a model in a similar fashion as the linear regression assuming that the higher powers of features are linearly independent. For example, say feature **x** is nonlinear, but there is a possibility that x^2 will be linear and hence, the **OLC (Ordinary Least Square)** method as described in Linear regression can be applied for polynomial regression as well.

Evaluation and Results Discussion

Submission results for Polynomial regression:

- 2 degree, Zestimate = 0.114
- 3 degree, Zestimate = 0.089

It is clear that from the above results that higher degree polynomial will fit the data more precisely than a lower degree polynomial. As we go higher in the degree of the polynomial, the error on training data will decrease but after a certain point of time, due to overfitting, the error on testing data may increase.

Submission results for Linear Regression:

- Zestimate = 0.065

Interesting thing that happened in my results were that the Linear regression model performed better than the higher degree polynomials. The reasons can be the following:

- Lack of features: Because only 8 features out of the total 52 were used for prediction, it may be the case that the feature that was actually distributed in relevance to the higher degree polynomial was missed. And because the feature set was small, linear regression was better as it had lesser number of points to take care of in one degree.
- Overfitting: Overfitting actually depends on the feature selection. Because there are low number of features, a higher degree polynomial might easily fit all of them and hence result in overfitting, which in turn leads to more errors in the test data.

R^2 value analysis

R^2 value for linear regression: **0.0039**
R^2 value for polynomial regression: **0.0098**

The R^2 value clearly shows which model is a better predictor of the training data. R^2 value is proportional to the degree of predictive power that a model has for a data. As polynomial regression has higher R^2 value that the Linear Regression, it shows that a polynomial curve fits the training data better than a linear regression.

Interesting things learnt

- Normalization: The improvement on the results of Linear regression after normalizing the data was a great lesson on how important normalization is. Also, it was interesting to know that normalization is generally helpful in cases where the data is continuous. That's why, I didn't use normalization for features like: bedroom count, year built etc.
- Outliers: How important can removing outliers be can be seen from the improvement of the results of Linear Regression. Linear regression is greatly affected by outliers.
- One of the most interesting part I learnt from the data is about the number of houses sold during different months a year. I expected them to be high at around Christmas but then I understood it's not that way.
- From part-1 of the homework, **log-error** is highly correlated with the **area of the apartment** i.e. bigger the house, difficult it is to predict the actual value.
- Replacing Nan: I understood how drastic it could be replacing Nan with the mean of the column values in some cases. It can be realized from the initial results of linear regression which gave an R^2 value to be around **-4**.
- Higher degree isn't always good: In case we're working on less features, linear regression can outperform polynomial regression.