



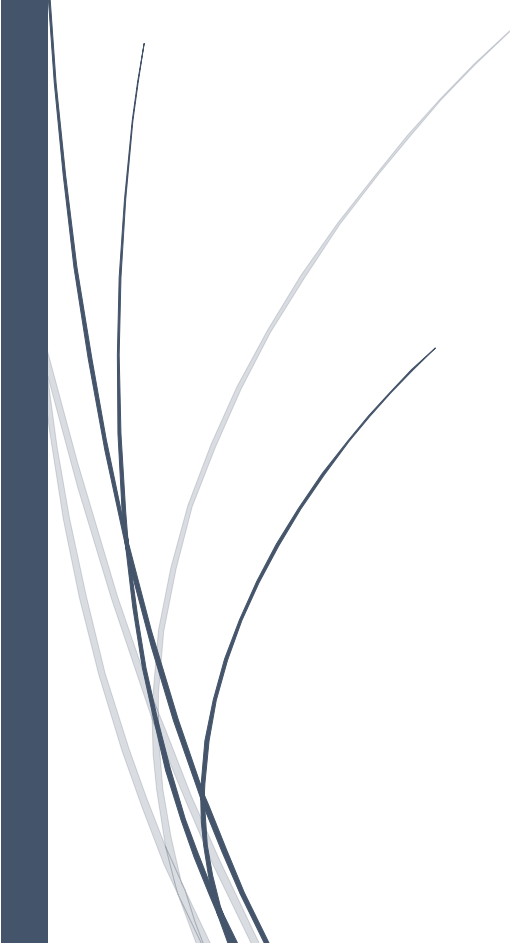
2202-CSE-5334-202- DATA MINING

Term Project Report

Student ID: 1001778514

Name: Haibo Wang

05/09/2020



The task of the term project is given the review, predict the rating. And the data is game geek review data, which comes from <https://www.kaggle.com/jvanelteren/boardgamegeek-reviews>

what is the contribution

The paper gives the experimental method of linear regression to predict continuous values. Based on the linear regression model, I have completed more experiments.

1. Linear regression model

The purpose of linear regression is to obtain the linear relationship between the output vector Y and the input feature X , and find the linear regression coefficient θ .

For the model performance, using least squares and gradient descent to optimize the model.

2. Ridge regression model

Add regularization term, use ridge regression.

3. Lasso model

Lasso regression can make the coefficients of some features smaller, and even make some coefficients with smaller absolute values directly become 0, thereby enhancing the generalization ability of the model.

4. ElasticNet CV model

Use cross-validation on hyperparameters to help us choose the appropriate hyperparameters

5. Different depth of decision tree regressor (depth = 1, 3, 6)

The decision regression tree is a regression tree, and the average value observed by the leaf nodes can be used as the predicted value.

By setting parameters such as the depth of the tree, construct different models and conduct different experiments

how to solved it

Adjust the data distribution to match the uniform distribution

```
def uniform_norm(X):  
    X_max = X.max(axis=0)  
    X_min = X.min(axis=0)  
    return (X - X_min) / (X_max - X_min), X_max, X_min
```

Adjust the model parameters and vector dimensions multiple times, add regularization terms, and train the model multiple times. Not only that, I also generate word vectors through word2vec, and use the angle of the vectors to calculate the similarity.

```
In [17]: train_X  
  
Out[17]: array([[ 0.          ,  0.          ,  0.          , ...,  0.          ,  
                  0.          ,  0.          ],  
                [ 0.          ,  0.          ,  0.          , ...,  0.          ,  
                  0.          ,  0.          ],  
                [ 0.01431885,  0.12670898, -0.07094727, ..., -0.04813232,  
                  0.01752624,  0.01533203],  
                ...,  
                [-0.06306966, -0.00065104,  0.04785156, ..., -0.07063802,  
                  0.08671061, -0.03645833],  
                [ 0.          ,  0.          ,  0.          , ...,  0.          ,  
                  0.          ,  0.          ],  
                [ 0.          ,  0.          ,  0.          , ...,  0.          ,  
                  0.          ,  0.          ]])
```

Use multiple evaluation criteria (mean square error, mean absolute error, root mean square error) to evaluate different models, and then adjust the parameters to select the model with the best effect.

experiments and findings

Through multiple experiments and multiple adjustments of parameters, the Ridge regression model was found to be the best, and the evaluation results are shown in the figure.

```

LinearRegression-----
MSE1 2.8854243471035645
RMSE1 1.6986536866305517
MAE1 1.3083312347715774
*****
Ridge-----
MSE2 2.885415844166543
RMSE2 1.6986511837827514
MAE2 1.3083306964122348
*****
Lasso-----
MSE3 2.915484791934709
RMSE3 1.7074790751088895
MAE3 1.3161503162305628
*****
ElasticNetCV-----
MSE4 2.8857893026017605
RMSE4 1.6987611081614038
MAE4 1.308493856250108
*****
DecisionTreeRegressor(max_depth = 1)-----
MSE5 2.9129757355782995
RMSE5 1.7067441916052621
MAE5 1.3161520659962913
*****
DecisionTreeRegressor(max_depth = 3)-----
MSE6 2.907673685531794
RMSE6 1.7051902197502171
MAE6 1.3151368087476636
*****
DecisionTreeRegressor(max_depth = 6)-----
MSE7 2.9016749312409535
RMSE7 1.7034303423506796
MAE7 1.3126936170451564

```

Reference

Notes on linear regression analysis[J], Robert Nau Fuqua School of Business, Duke University, 2014