Capstone :
Date A Scientist project.

—

Machine Learning Fundamentals

Sing Er Liu 2018.10.09

Outline

1. Data Description
2. Regression Model
3. Classification Model
4. Conclusion

Data Description

- I use data cleaning technique such as dropna, and delete the outlier by using logic symbol.
- The augment data what I create are “education_code”, “avg_word_length”, “text_word_counts” and “i_me_frequency” .

New Data Frame after augmenting

drinks_code	smokes_code	drugs_code	orientation_code	sex_code	education_code	essay_len	avg_word_length	i_me_frequency	text_word_counts
2.0	1.0	0.0	0	0	0.0	2644	4.453608	13	485
3.0	0.0	1.0	0	0	5.0	1453	4.215827	17	278
2.0	0.0	NaN	0	0	1.0	5517	5.212838	21	888
2.0	0.0	NaN	0	0	0.0	477	5.012658	1	79
2.0	0.0	0.0	0	0	0.0	725	5.269565	2	115
2.0	0.0	NaN	0	0	0.0	2469	4.613636	22	440
2.0	NaN	0.0	0	1	0.0	1917	4.369748	6	357
2.0	0.0	0.0	0	1	0.0	1240	6.034091	3	176
2.0	2.0	NaN	0	1	0.0	2247	5.034946	9	372
0.0	0.0	0.0	0	0	0.0	2420	4.591224	24	433

Regression

Questions

- **Q1: Using ‘essay_len’ to predict ‘age’**
- **Q2: Using ‘essay_len’, ‘avg_word_length’, ‘age to predict ‘income’**
- **Q3 : Using ‘i_me_frequency’, ‘avg_word_length’ to predict ‘age’**

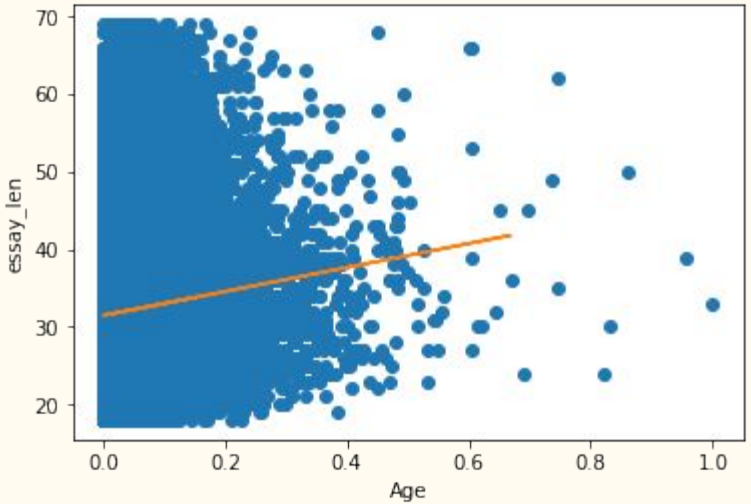
Linear Regression vs KNN Regression

- **Using ‘essay_len’ to predict ‘age’**

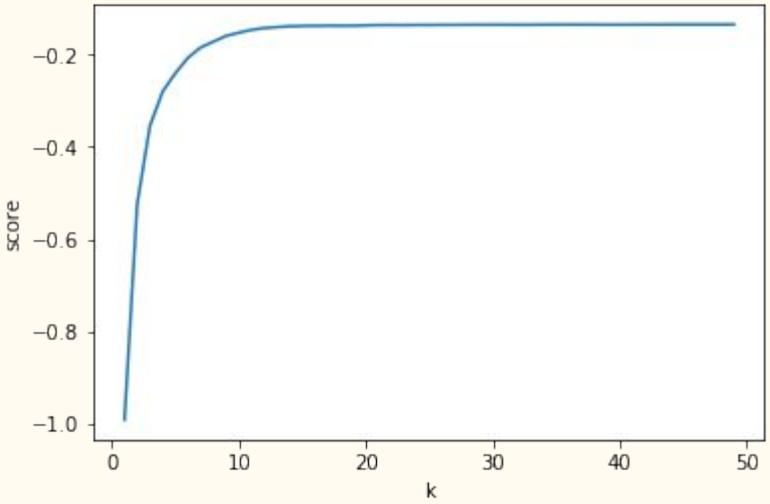
Train score: 0.007
Test score: 0.004

Train score: 0.154
Test score: -0.138

Running Time: 1.99ms



32.7ms



Linear Regression vs KNN Regression

- **Using ‘essay_len’ to predict ‘age’**

1. Both model’s performance are not good, the best accuracy score is under 0.2.
2. KNN is much better than Linear Regression on training score, but Linear Regression is better than KNN on test score
3. KNN’s training score is much higher than its training score, it might be the overfitting problem.
4. The running time for Linear Regression is better.

Linear Regression vs KNN Regression

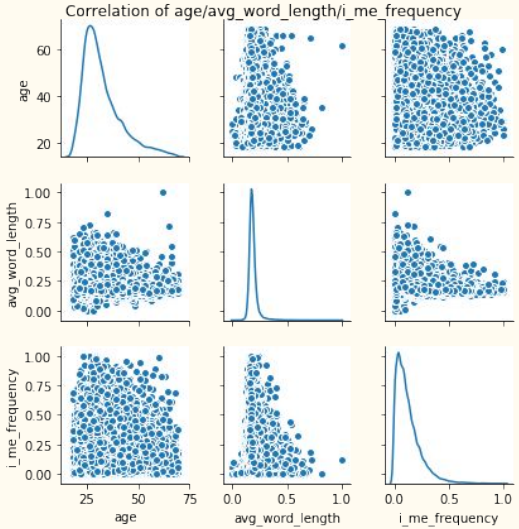
- **Using ‘essay_len’, ‘avg_word_length’, ‘age to predict ‘income’**

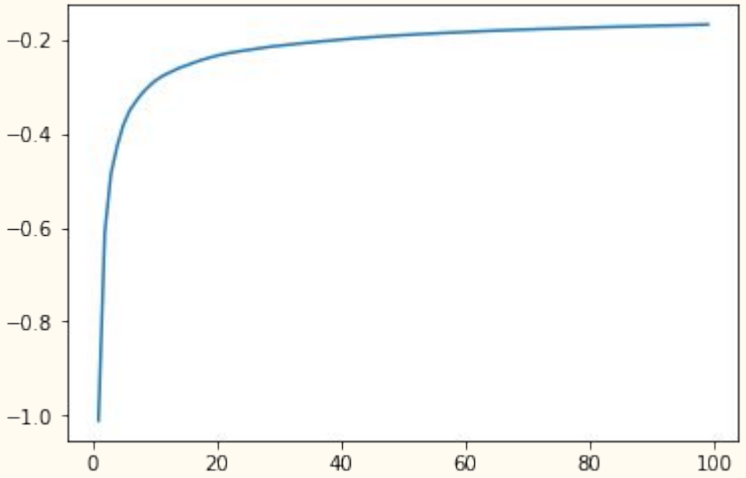
Train score: 0.0041
Test score: 0.0037

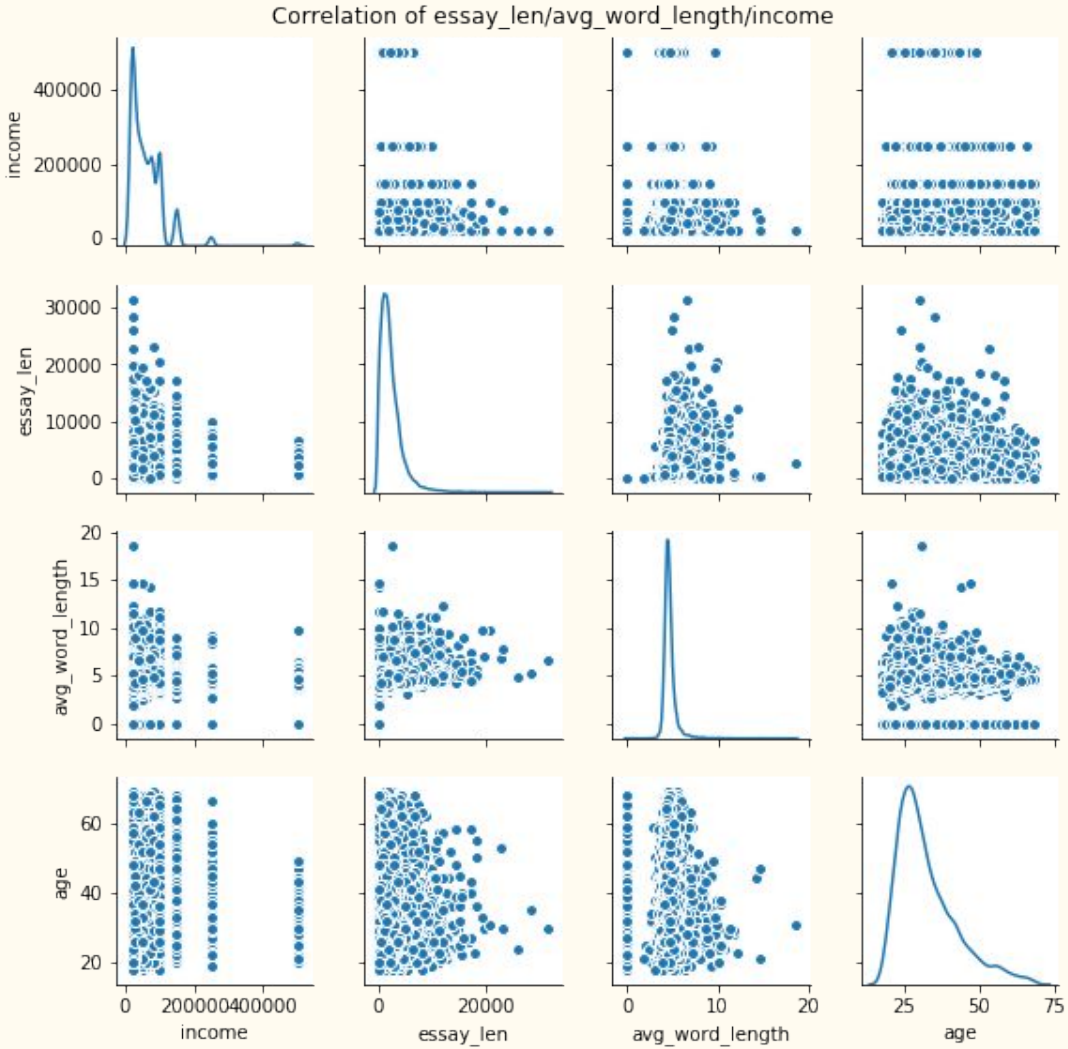
Train score: 0.9569
Test score: -0.2879

Running Time: 1ms

6.99ms







The graph shows that these variables do not have much correlation and the distribution all are Skewed distribution

Linear Regression vs KNN Regression

- **Using ‘essay_len’, ‘avg_word_length’, ‘age to predict ‘income’**

1. KNN’s performance is very good at training data since it’s score is more than 0.9.
2. KNN is much better than Linear Regression on training score, but Linear Regression is better than KNN on test score
3. KNN’s training score is much higher than its training score, it might be the overfitting problem.
4. The running time for Linear Regression is better.

Linear Regression vs KNN Regression

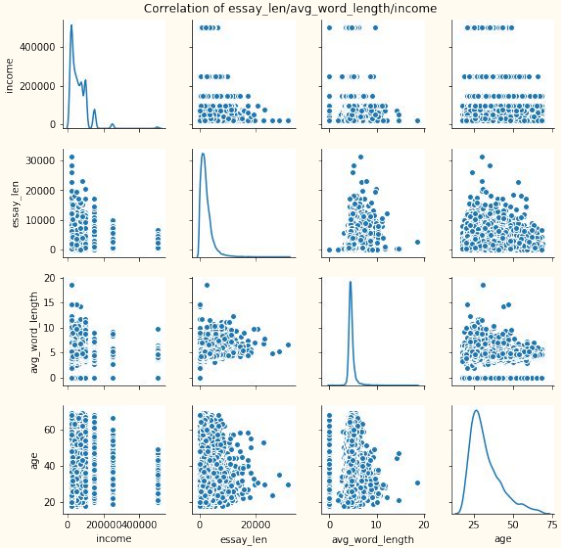
- **Using ‘essay_len’, ‘avg_word_length’, ‘age to predict ‘income’**

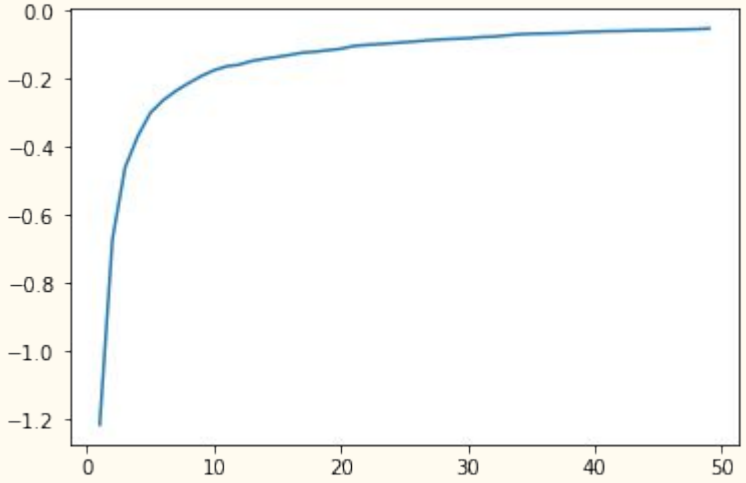
Train score: 0.0011
Test score: 0.0003

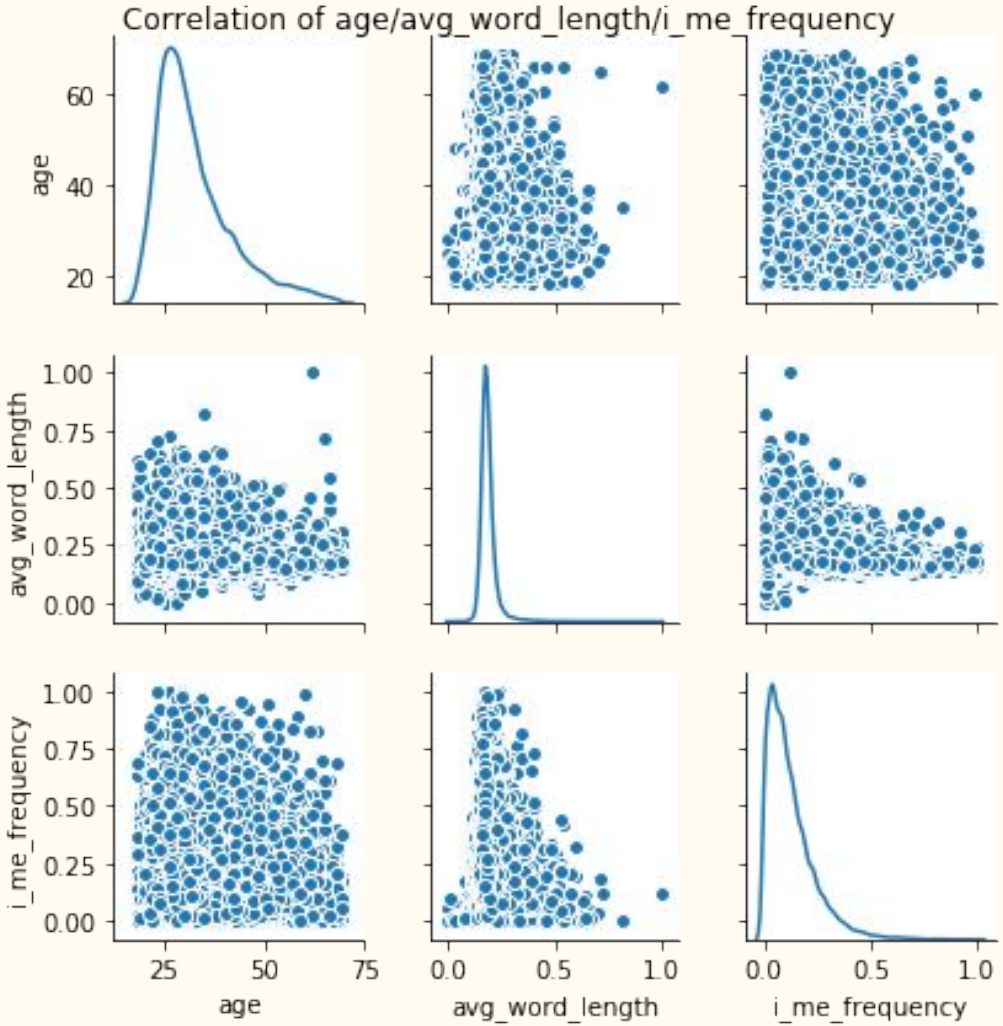
Train score: 0.9454
Test score: -0.0930

Running Time: 2ms

31.2ms







The graph shows that these variables do not have much correlation and the distribution all are Skewed distribution

Linear Regression vs KNN Regression

- **Using ‘essay_len’, ‘avg_word_length’, ‘age to predict ‘income’**

1. KNN’s performance is very good at training data since it’s score is more than 0.9.
2. KNN is much better than Linear Regression on training score, but Linear Regression is better than KNN on test score
3. KNN’s training score is much higher than its training score, it might be the overfitting problem.
4. The running time for Linear Regression is better.

Classification

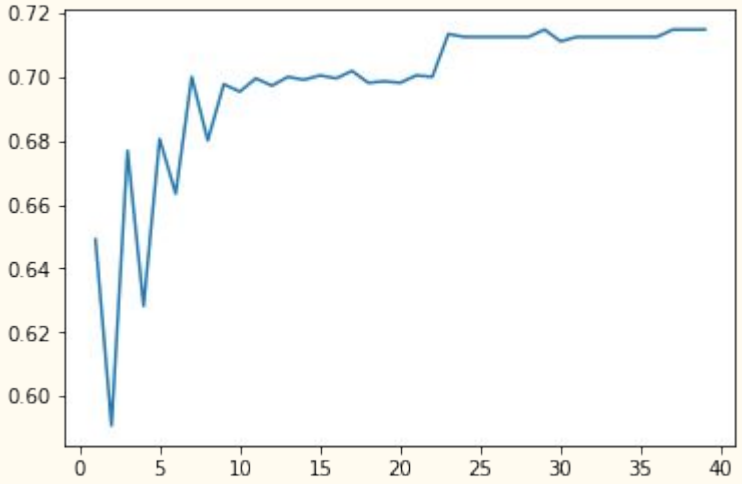
Questions

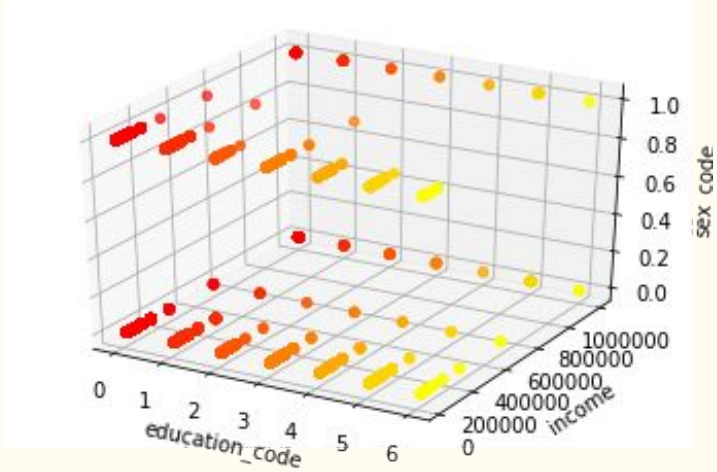
- **Using ‘education_code’, ‘income’ to predict ‘age’**

KNN vs SVM

- **Using ‘education_code’, ‘income’ to predict ‘age’**

KNN	SVM
Train score: 0.7126	Train score: 0.728
Test score: 0.7306	Test score: 0.713
Running Time: 24ms	629ms





Linear Regression vs KNN Regression

- **Using ‘education_code’, ‘income’ to predict ‘age’**

1. Both models’ performance are good at training/test data since the scores are all more than 0.7.

2. SVM is slightly better than KNN on training score, but KNN is slightly better than SVM on test score

3. Both models’ training score are similar with the training score, it might has no overfitting or underfitting problem.

4. The running time for KNN is better, but when the data is linear, the SVM does not need to choose the parameter, which KNN need to choose ‘k’.

Conclusion

Conclusion

For Regression problem, I failed to find the best variable to fit the data I want, the data seems not has correlation between them, and some of the data, such as ‘income’, has the problem of too much useless information. The other problem is the continuous data is rare, I need to create some from essay.

For Classification problem, I can predict the ‘age’ form ‘education’ and ‘income’, and bothe model performance are good. Considering of the efficiency, the KNN is faster, but need to choose the k value, so for this topic, I’ll choose SVM to train the model.

The End

—

Machine Learning Fundamentals
Sing Er Liu 2018.10.09