DATA SCIENCE FINAL PROJECT

Footbal Data set

November 12, 2019

Please use these data sets and complete following tasks. You need to write a report where you explain the preprocessing/data exploration steps for the data, the method(s) you used, and the result you received (plots and/or numbers). Remember, you need to return the report as a single PDF file (5-8 pages) and you do not need to include your code. You are allow to use different method but you need to be able to explain the method in a way that your peer students could understand. You may use all the material provided during the course, build-in function in different packages, and online sources. Just remember to cite the sources outside the course sources. The outline of the report is given in the slide "Final project and Peergrade".

1 Description of the data set

The data was collected and distributed by Football-Data.co.uk. For this project we are using the results of the Premier league in seasons 2016, 2017 and 2018. From the original dataframe only following columns were used:

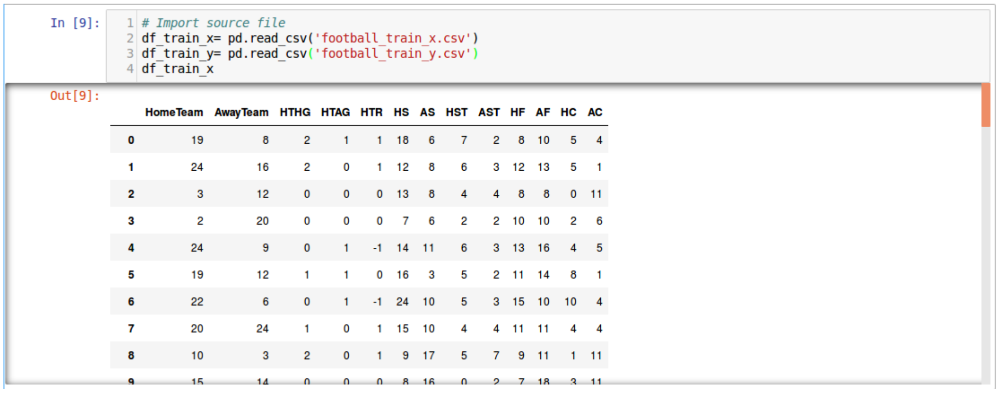
- "HomeTeam", "AwayTeam": names of the home and away teams, mapped between 0 and 19.
- "HTHG", "HTAG": half time goals of home and away teams.
- "HTR": half time results, 1, if home team won, -1 if it lost, 0 for a draw.
- "HS", "AS": number of shots by home and away teams.
- "HST", "AST": number of shots on target by home and away teams.
- "HF", "AF": number of fouls by home and away teams.
- "HC", "AC": number of corners of home and away teams.

The football match is considered interesting, if the final goal difference for the match is more or equal to 3 goals.

The data have been split in train and test (approximately 70% of the data are the training set, and 30% are the test set), and are divided in 4 .csv files:

- football_train_x: contains 798 observations and 13 columns as features.
- football_train_y: contains the labels "FTG" (full time goal which is the total number of goals in the whole match) that you will need for the regression task and "Interest" for classification for these 798 observations.
- football_test_x and football_test_y sets have the same respective structure and contain 342 observations.

Here an example showing how to read the files in pandas:



In [9]:

```python
# Import source file
df_train_x= pd.read_csv('football_train_x.csv')
df_train_y= pd.read_csv('football_train_y.csv')
df_train_x
```

Out[9]:

	HomeTeam	AwayTeam	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC
0	19	8	2	1	1	18	6	7	2	8	10	5	4
1	24	16	2	0	1	12	8	6	3	12	13	5	1
2	3	12	0	0	0	13	8	4	4	8	8	0	11
3	2	20	0	0	0	7	6	2	2	10	10	2	6
4	24	9	0	1	-1	14	11	6	3	13	16	4	5
5	19	12	1	1	0	16	3	5	2	11	14	8	1
6	22	6	0	1	-1	24	10	5	3	15	10	10	4
7	20	24	1	0	1	15	10	4	4	11	11	4	4
8	10	3	2	0	1	9	17	5	7	9	11	1	11
9	15	14	0	0	0	8	16	0	2	7	18	3	11

Please complete the following tasks and include your results and analysis in your report:

2 Data exploration

Before starting, explore your data. You are required to complete the followings;

- Correlation: Plot a correlation matrix of all features.
- Pair plots: Plot several pair plots for a few interesting features. You might get some inspiration from the correlation matrix.
- PCA: Prepare a PCA projection using the first 2 components. Plot cumulative explained variance.

3 Regression

The goal of this task is to predict the number of goals of the match based on all or some of the features. You should complete the following; You can A

- Try to understand and explain which features are the strongest predictor for total number of goals in a match and why.
- Analyze your data, see if there is a linear trend or not. Try different regression models. If you want, you can also try a quadratic model. Explain your result(s). If you try different models, remember to compare them and explain all of them.
- Adopt PCA to reduce the dimension of your data set. Select the right number of components by checking the variance explained by components. Use the projected train and test set to predict the total number of the goals for matches in the test set. compare your result obtained after adopting the PCA with without PCA and explain your results/observation.

4 Classification

The task is to classify the football match as interesting or not based on its proximity to other classes. For this purpose you should;

- Apply one or two different method(s) for classification. If you don't know where to start, you might want to try KNN-classification first. You can also try logistic regression. Describe, how you were choosing the optimal number of neighbors, and plot the confusion matrix for your results. Try to explain the results you get.
- Perform PCA on the data, since you have rather many different features, and then use first two components for plotting.
- Try to select the correct number of components for projecting your data to lower dimension by looking at the cumulative explained variance plot. Use the your projected data and the method(s) you used previously for classification task. Compare your result and explain your observation.

- If you are using more that one method, explain all and compare their results.

Tips: You may find the following commands and methods useful. From “sklearn.linear_model" use “LinearRegression()" and “LogisticRegression", and from class “sklearn.neighbors" the function “KNeighborsClassifier" See their documentations and online sources for more details information.

After finishing all the tasks, write a report based on the result you obtained in different part and upload it to the Peergrade platform.