Data Visualization and Prediction of European Soccer

Team18: Yahsiu Hsieh Cheng-Ying Tsai
Siyuan Zhu Yuhan Wang

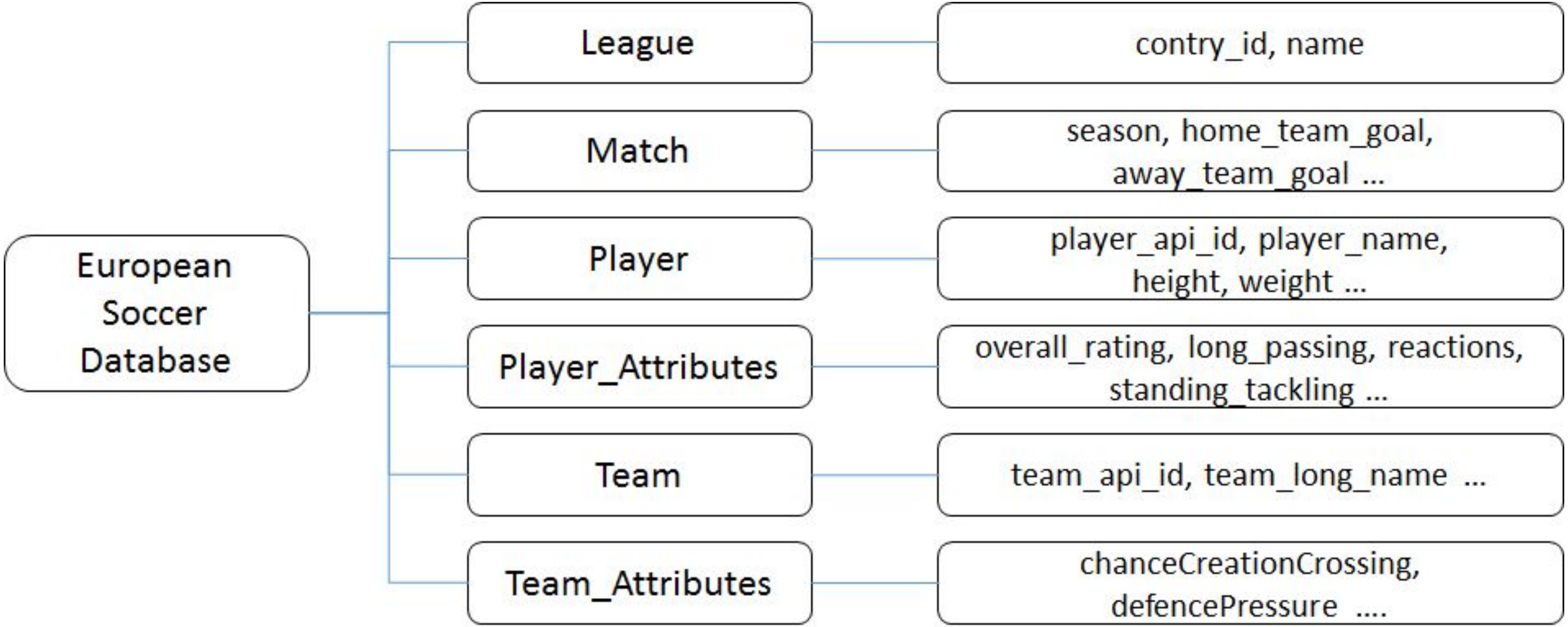
Outline

- Motivation and Objective
- Explore the data
- K-Means & PCA on players’ attributes
- Prediction

Motivation and Objective

- Over 20000 matches, more than 10000 players, in 11 European countries with their lead championship, from season 2008 to 2016.
- What are the dominant factors that determine the result of game?
- We want to see what kind of attributes of a player owns can have higher probability to win the game.
- We try to build the prediction system based on the given database.

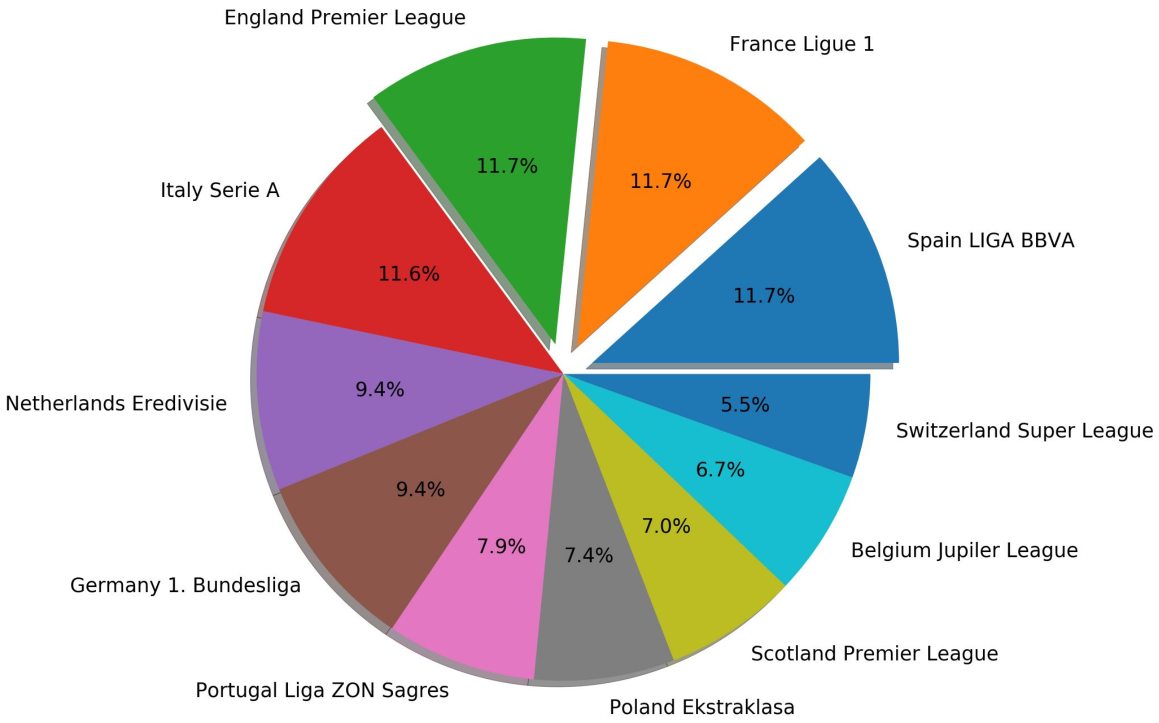
Dataset



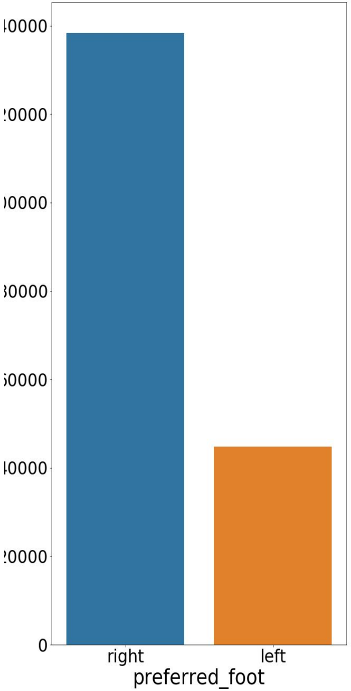
European Soccer Database

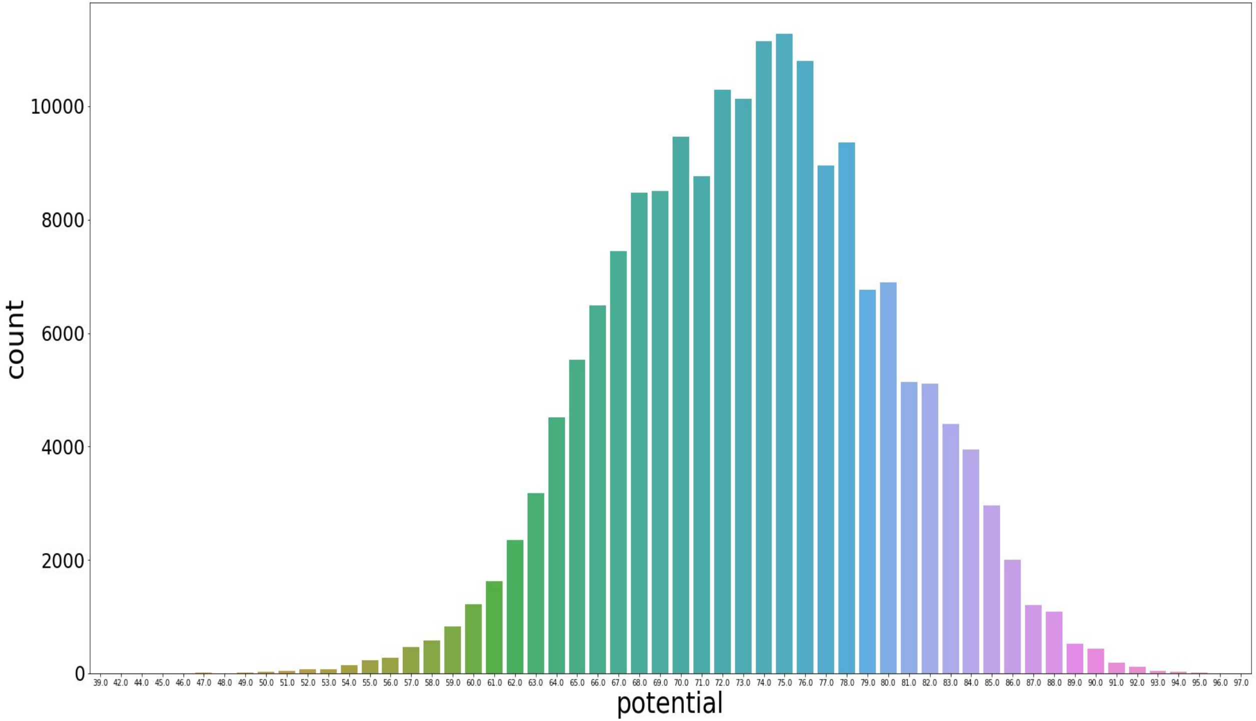
League	contry_id, name
Match	season, home_team_goal, away_team_goal ...
Player	player_api_id, player_name, height, weight ...
Player_Attributes	overall_rating, long_passing, reactions, standing_tackling ...
Team	team_api_id, team_long_name ...
Team_Attributes	chanceCreationCrossing, defencePressure ....

Explore the data - matches in leagues

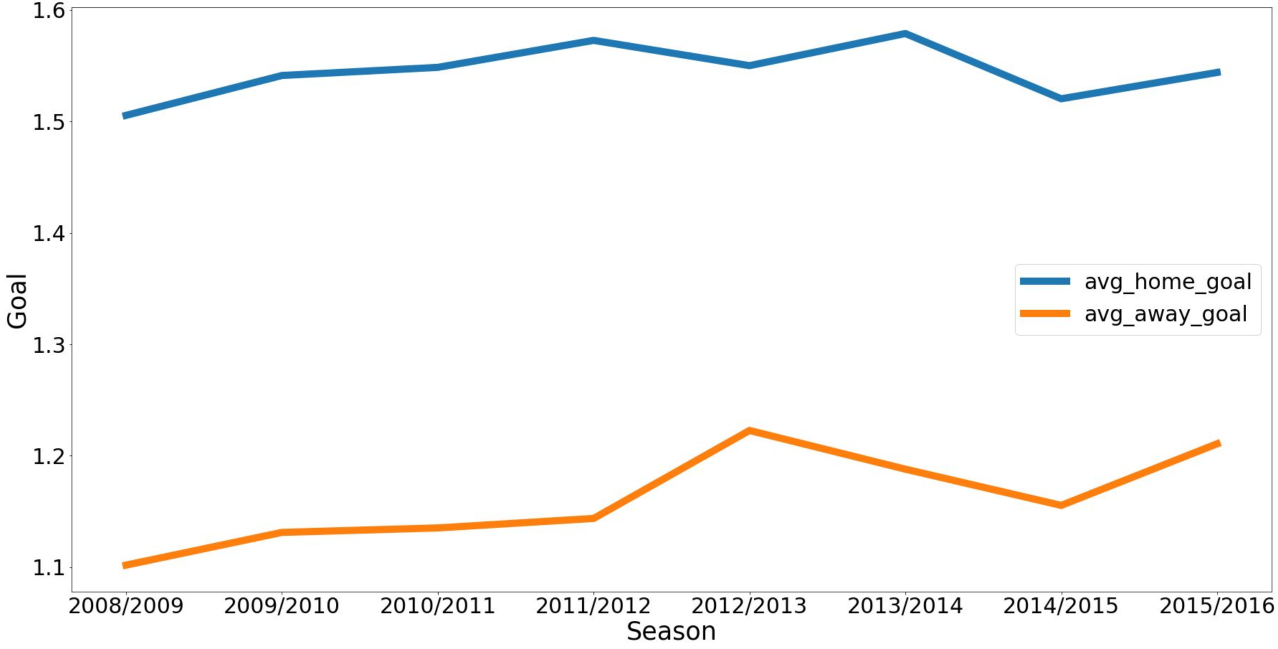


Explore the data - player attributes

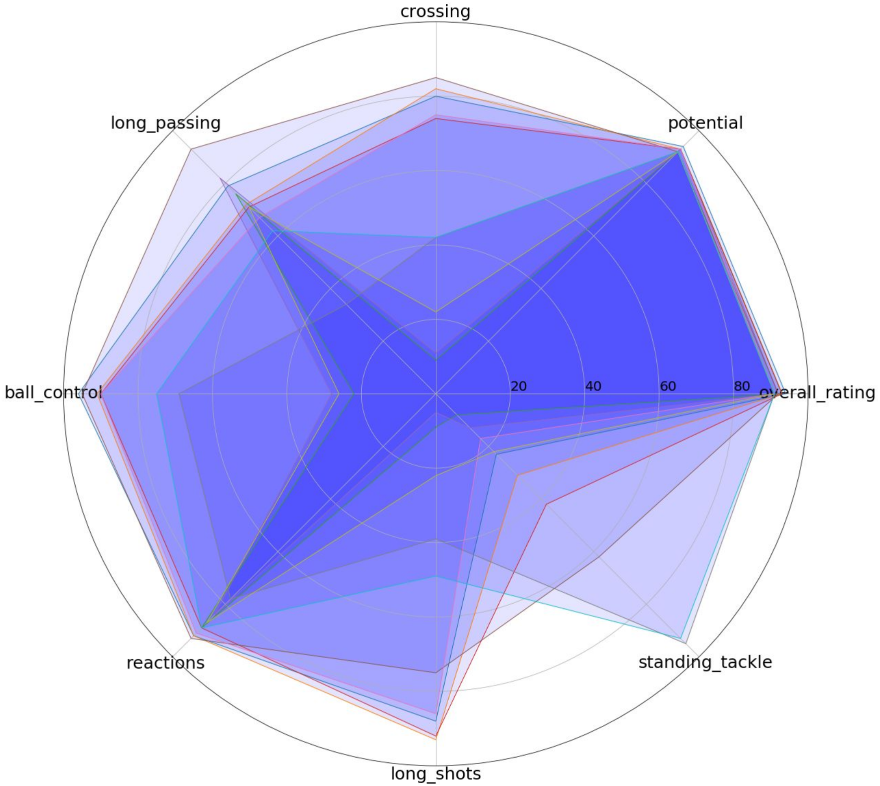




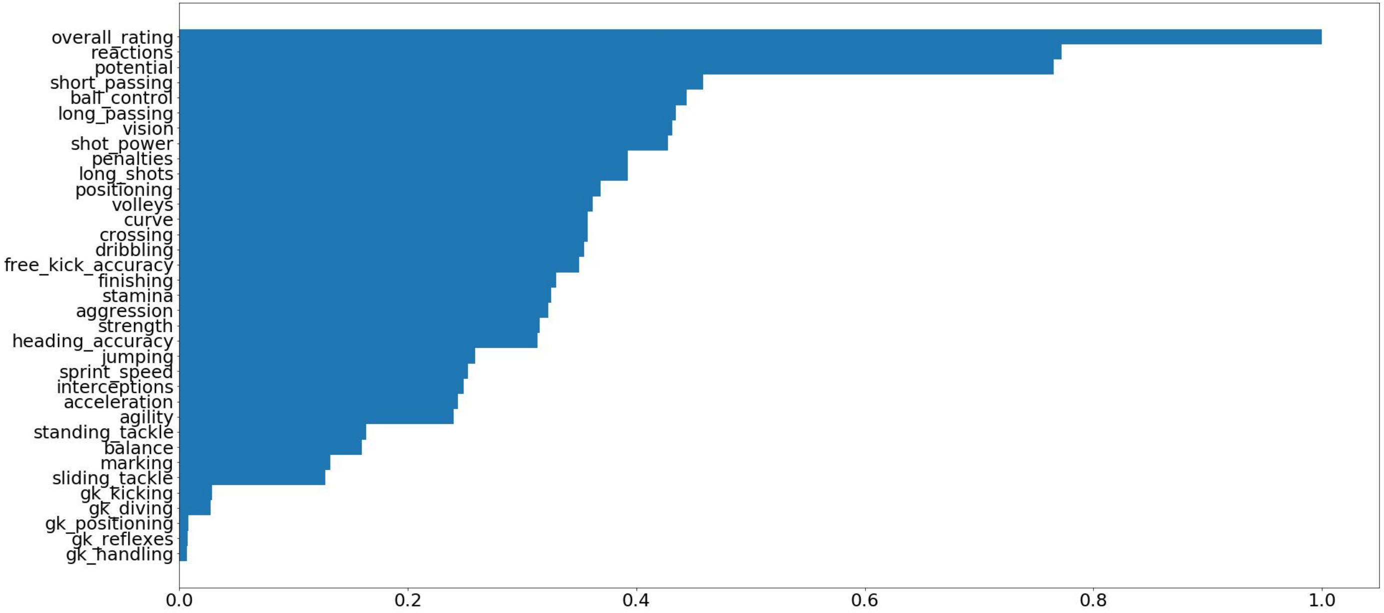
Explore the data - does home advantage exist?



Explore the data - top 10 players attributes

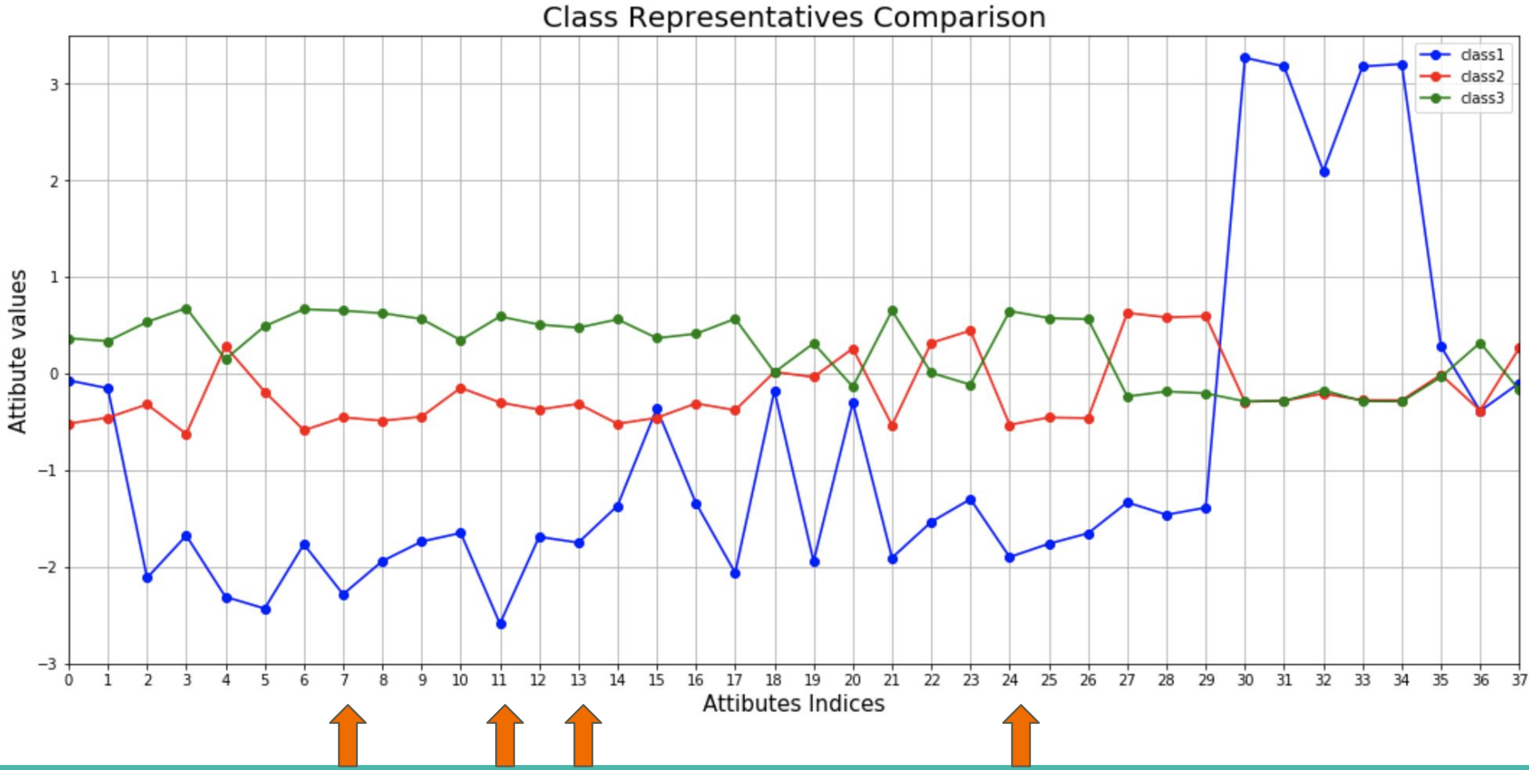


Explore the data - correlation between overall rating and other attributes



K-means Analysis on players dataset

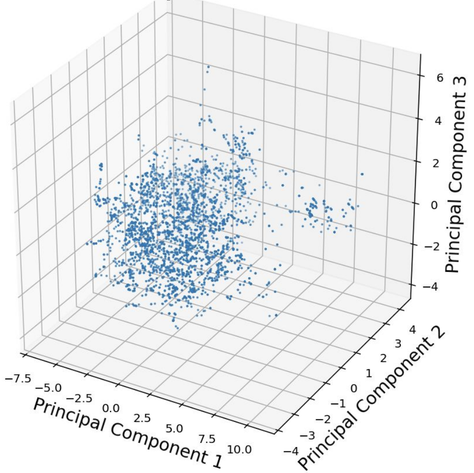
K=3 best classifies the players dataset according to their 38 attributes.



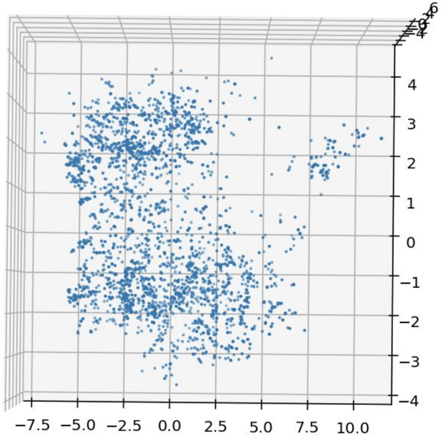
PCA Analysis

For visualization purposes, we performed a 3-component PCA operation and shows all players data on the projected 3 principal components space.

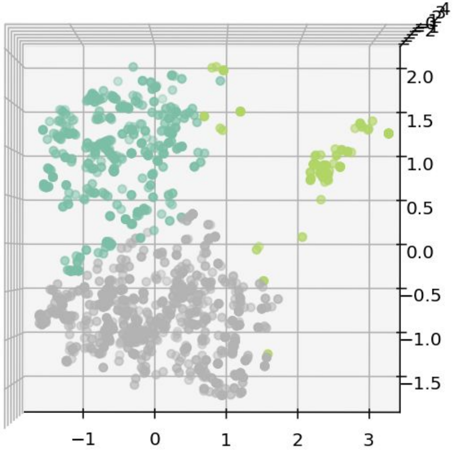
3 component PCA



PCA Top View



K-means on PCA result



Prediction

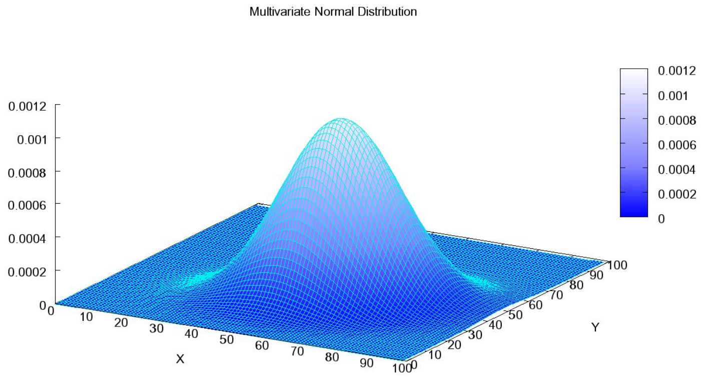
Input: player attributes of all team members in two teams.

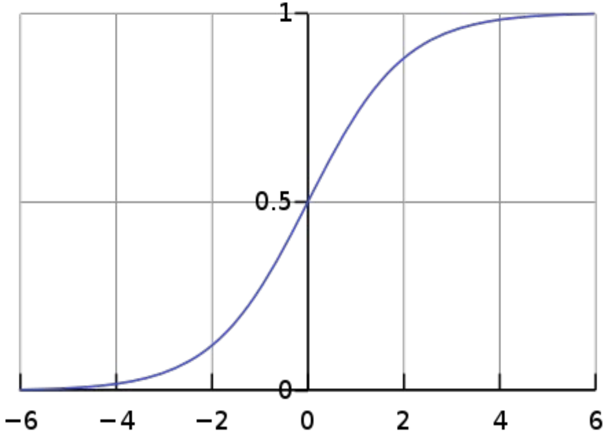
Output: the result of the game (home team wins, home team loss or a tie)

method	Accuracy (%)
K-Nearest Neighbors	43.96
Gaussian Classifier	46.62
Gaussian Classifier + PCA	50.06
Fully connected neural network	53.38

Prediction analysis

Since there are only three possibilities for the result of a game, the accuracy of our model is not satisfactory.

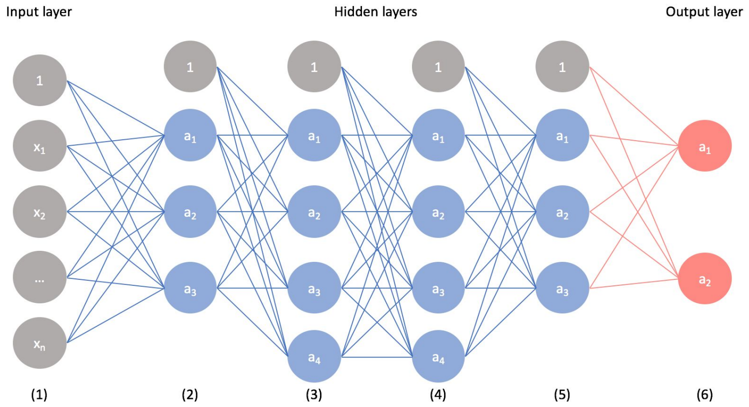


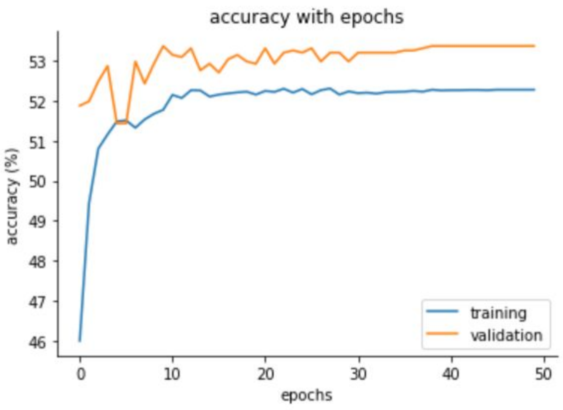


It is reasonable that the accuracy of gaussian distribution model and k-nearest neighbors is not ideal because the underlying data distribution doesn’t meet the requirements of the above two models.

Why neural network doesn’t work on this dataset?

However, it seems mysterious why fully connected neural network doesn’t works on this dataset.





Our hypothesis is that the cost function should be highly non-convex so that gradient descent can not easily find the global minimum.

Thank you!

Citations:

Gaussian plot:

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

5-layer neural netwrok diagram:

https://www.jeremyjordan.me/convolutional-neural-networks/