**산업수학 2조 제안서**

남주형, 최재원, 심현우

1. 주제

저희는 인공지능 모델을 통해 KBO 야구 승패 예측 분석을 진행할 예정입니다. 팀 승 패에 대한 중요한 지표들을 선정하여 데이터를 구축하고 모델에 따라 데이터를 처리하여 승패 예측을 진행합니다. 10개 구단 중 1개 구단을 선정하여 구단의 경기 승패 여부를 예측하고 예측 결과와 현재 진행되는 리그 결과를 비교하여 정확도를 측정해볼 예정입니다. 승부 예측 및 데이터는 이미 많은 구단에서 사용하고 있으며 이를 연구하여 더 정확한 분석 결과를 만들 수 있으면 경기 중간에도 승리를 위한 여러 전략 설정에 도움이 될 것이라고 생각합니다.

1. 데이터 종류

데이터는 승패에 큰 영향을 미칠 수 있는 것들을 사용합니다. 이닝까지 점수, 타율 등 승패에 직접적으로 영향을 끼치는 지표들과, 세부적으로 날씨, 상대 팀, 이닝별 홈런, 병살, 안타, 실책, 홈/어웨이 여부 등을 사용합니다. 데이터는 KBO 공식 홈페이지에 있는 최근 4-5년에 걸친 결과를 이용하며, 적절한 모델을 설정한 뒤에 그에 따라 데이터셋 훈련용과 검정용으로 구분하고 위에 기재된 항목별로 정리하여 사용할 예정입니다. 이후 검정용 데이터를 통해 분석 결과의 정확도를 측정하여 어떤 알고리즘이 더 정확한 결과값을 도출해낼 수 있는지 파악할 것입니다.

1. 데이터 분석 방법

KNN – parameter로 승부에 영향을 줄 만한 인자(이닝별 점수, 상대팀, 타석, 홈런갯수, 안타, 삼진, 볼넷 등등 위에 기재한 여러가지 요인들을 포함)를 여러 개 설정한 후에 피어슨 상관계수를 활용한 heatmap을 그려서 가장 영향력이 큰 인자를 3가지 선택한 후에 공식 홈페이지에 기재된 데이터를 바탕으로 2016~2018까지의 경기결과를 바탕으로 train data set을 구성하고 2019~2020년의 결과를 가져와서 test data set를 구분하여 결과를 얻어낼 예정입니다. 분석을 통해서 얻어낸 인자를 바탕으로 KNN 알고리즘을 적용하여 test data에 대해 분석을 시행하고 그 결과를 실제 결과값과 비교하여 정확도를 측정할 것입니다.

DECISION TREE – data 전처리 과정에서 얻어낸 주요 성분을 대상으로 의사결정나무 알고리즘을 적용하여 분석을 시행할 예정입니다. 분석의 목적 변수가 범주형이므로 지니지수를 이용해 가지분리를 할 것이고, CART 분석을 사용할 예정입니다. 특히 의사결정나무는 가지가 많아질 경우 과대적합의 문제가 발생할 가능성이 있기 때문에 적당한 수치를 지정하고 마디에 속한 자료가 일정 수 이하인 경우에는 분할을 정지하여 가지치기를 실시할 것입니다.

1. 분석 데이터 링크

https://www.koreabaseball.com/Schedule/Schedule.aspx