Machine Learning Assignment Report

2019712383 김주형

먼저 제 파이썬 프로그램의 출력에 대해서 설명 하려합니다.

출력 순서는 KNN,Logistic Regression, Decision Tree의 accuracy와 f1값이 차례대로 출력 하게끔 되어 있고 같은 파라미터와 같은 알고리즘의 accuracy와 f1는 같은 선상 그리고 각 알고리즘은 공백을 통해 구분 파라미터 변경은 바로 다음 줄에 나오도록 출력이 되도록 하였습니다.

코드의 구성요소

먼저 데이터의 값을 파악하기 위해 chart그래프를 만들어서 값을 비교 해보았습니다. 이는 평가지표의 출력 값을 가독성 좋게 하기 위해 출력 문을 빼고 알고리즘만 작성하였습니다. 데이터의 결측치를 먼저 조사 하였고 그 중 나이에 대한 결측치가 많았기 때문에 이에 대한 확인이 필요하다고 생각하였고 여러 feature중에 name으로 이 것을 유추할 수 있겠다 하여 이름에서 그 근거를 찾으려 하였습니다. 이름에서 공통적으로 등장하는 단어를 출력 하였더니 miss,mr,mrs등의 공통적인 특징을 찾을 수 있었고 capt나 mile과 같은 소수로 등장하는 또는 한 명만이 있는 값들도 있어 이 수치를 다수의 mr나 miss에 편입 시켰습니다. 이렇게 한 이유는 결국 이 단어를 활용해서 나이를 리턴 받기 때문에 공통의 나이를 가진다면 편입시켜도 무관하다는 판단이었습니다. 이렇게 나이를 판단근거로 이름을 사용한 이후 과제에서 훈련 데이터로 쓰지 말라고 하였던 ‘passengerId’, ‘Name’, ‘Ticket’, ‘Cabin’을 값들을 모두 삭제 하였습니다.

이 외에 ‘embarked’값은 결측치가 적은데다 다수가 ‘s’값이라 ‘s’로 치환해주었습니다

One hot encoding은 ‘Sex’와 ‘Embarked’ ‘Pclass’에서 진행하였는데 생존에 있어서 지대한 영향을 주는 요인들이라고 생각했기 때문에 단순 라벨링을 넘어서 원 핫 인코딩이 필요하다고 여겼고 각 ‘Sex’는 두 개 Embarked 와 pclass는 각각 세 요소로 인코딩하였습니다.

나이는 결측치 였던 것을 이름을 통해 추론 한 이후 각각의 나이대별로 라벨링을 하여 구분을 두었고 Fare 역시 큰 영향을 준 것으로 판단하기는 힘들어 등분하여 5구간으로 나누어 라벨링 하였습니다

마지막으로 ‘Parch’와’SibSp’는 같이 Family로 묶어서 단순 처리하여 전처리 과정을 마무리하였습니다.

데이터는 ‘train.csv’ 파일을 7:3으로 나누어 train 과 test용으로 분리하였고

차례대로 KNN, Logistic Regression, Decision Tree로 위에 언급한 것처럼 파라미터를 바꾸어 출력하였습니다. 평가지표로는 accuracy와 f1을 사용하였습니다.

KNN의 모델을 돌렸을 때에는 K의 개수가 5일 때

KNN accuaracy(K = 5) : 0.8432835820895522 KNN f1(K = 5) : 0.8108108108108109

값으로 가장 높았습니다.

Logistic regression 모델을 썼을 때에는 여러가지 파라미터를 바꾸어보았으나

Logistic Regression accuracy(iter==20) : 0.8134328358208955

Logistic Regression f1 : 0.7727272727272728

모든 값이 동일하게 나오는 문제점을 보였습니다. 해당 문제에 관하여 분석을 진행하고 있으나 아직 뚜렷한 문제점을 찾지 못하고 있어 추가적인 작업을 해야할 것입니다.

Decision Tree에서는 Decision Tree accuracy : 0.8134328358208955

Decision Tree f1 : 0.766355140186916

값을 보였고 이에 더하여

Bagging을 해주었을 때에

Decision Tree with Bagging accuracy(n=5) : 0.835820895522388 Decision Tree with Bagging f1(n=5) : 0.808695652173913

가장 높은 값을 얻었음을 확인 하였습니다.