REPORT

Sun Moon University

기말고사 보고서

학과명 : 글로벌 소프트웨어

교과명 : 데이터 사이언스 핵심

교수명 : 김응희 교수님

학 번 : 2017315018

이 름 : 방제호

제출일 : 2019.06.19

선문대학교

목차

데이터 소개 및 기계학습의 목표

유방암 데이터는 569개의 row와 30개의 columns를 가지고 있으며, class는 0: 악성, 1: 양성으로 분류하는 데이터 셋을 가지고 있다. 데이터의 columns는 radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension 이러한 특징의 평균, 표준 오차가 이 각 이미지에 대해 계산되어 30개의 특징이 나타난다.

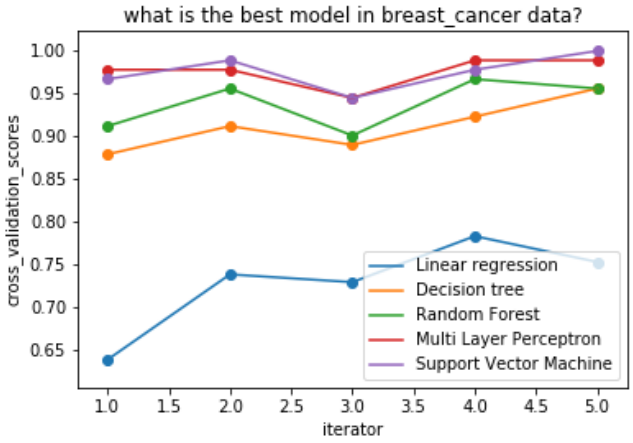
	Num_Mal_data	Num_ben_data
0	212	357

위의 데이터프레임을 보면 악성 종양인 환자는 212명이며, 양성 종양인 환자는 357명으로 약 4:6으로 unbalance하다는 것을 알 수 있다. 그래서 accuracy가 아닌 f1_score나 roc-auc 지표를 봤을 때 더 정확할 것이다.

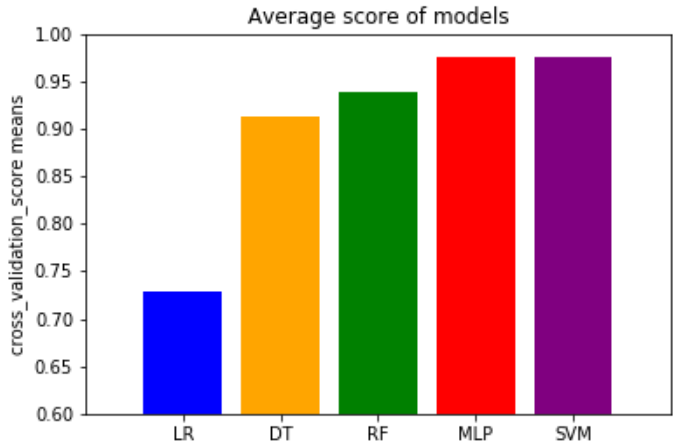
유방암은 전 세계적으로 3대 여성 암 중 하나이며, 세계적으로 심각한 공중 보건 문제이다. 조기 진단을 하면 예후와 환자의 생존 가능성을 크게 향상시킬 수 있다. 양성 종양을 더 정확하게 분류하면 불필요한 치료를 받는 환자를 예방할 수 있고, 악성 종양을 더 정확하게 분류할 시 환자를 빠르게 치료할 수 있다. 목표는 어떤 특징이 악성 또는 양성 암 예측에 가장 유용한 지 관찰하고, 유방암이 양성인지 악성인지 분류하는 것이다.

모델 구축 및 실험 과정

데이터를 가시화해서 어떤 모델이 어울릴 지 파악해보고 싶었으나 너무 어려웠고, 또한 데이터들을 가시화한 자료들을 찾아보니 아무리 봐도 어떻게 파악하는 지 이해가 안되었다. 그래서 현재 데이터를 8:2로 training data set, test data set으로 나누었다. 그리고 다섯 가지 모델들을 선정하였다. 선정한 모델들은 Linear Regression, Decision Tree, Random Forest, Multilayer perceptron, Support Vector Machine이고, 이 모델들의 hyper parameters 들은 모두 default 값으로 지정하여 모델들을 구축하였다. 그 이유는 성능을 측정할 때 hyper parameter들의 영향을 많이 받기 때문이다. default로 지정하면 모든 모델을 평등하게 판단할 수 있다. 그래서 모든 모델을 default로 구축한 후에 training date set에서 5fold로 데이터를 또 나누었다. 그 이유는 현재 training data set에서의 성능을 측정하여 어떤 모델이 이 데이터와 궁합이 좋은 지 알 수 있기 때문이다. 그래서 확인해본 결과 아래와 같은 결과가 나왔다.



Linear regression 모델은 이 데이터와 적합하지 않았다. 이 데이터들의 분포가 선형 분류기로 분류하기 어려운 데이터 셋이라는 것을 알 수 있었다. 또 decision tree와 random Forest로 분류한 결과 성능이 좋게 나왔지만 그래도 Multilayer Perceptron 과 Support Vector Machine에 비하면 성능이 떨어진다. 성능이 떨어진다는 것은 MLP와 SVM에 비해 이 데이터와 궁합이 맞지 않다는 것을 알 수 있었다. 의아하게도 MLP와 SVM의 성능의 차이를 위의 그림만으로 보기엔 좀 어려운 부분들이 있었다. 처음은 MLP, 두 번째는 SVM, 세 번째는 같고, 네 번째는 MLP, 다섯 번째는 SVM이 높아서 구분하기 어려웠다. 그래서 평균 값을 구해보니 아래의 그림과 같이 나왔다.



평균값으로 봤을 때 두 모델의 평균이 같다는 것을 알 수 있었다. 그래서 SVM과 MLP 두 모델의 best hyperparameter를 찾아 비교하기로 하였다.

MLP

Hidden layer 의 개수가 1 개 일 때 2 개 일때로 분류하여 테스트를 진행하였다. 한 가지는 레이어 1 개 안의 노드의 수는 1~35 개 사이 중 하나와 sorver 중에선 lbfgs, sgd 중에서 하나와, alpha 중에서 10^{-1}~10^{-5} 중 하나와, max iter 는 1000 으로 고정시키고, random state 도 0 으로 고정시켰다. 저 많은 조합들 중에서 트레이닝 데이터 셋에서 가장 좋은 성능을 가진 조합인 alpha = 0.1, node 수 12 개, max iter 1000, random state = 0, solver = sgd 인 조합이다. 그 조합의 성능을 확인하기 위해 테스트 데이터 셋으로 예측한 후에 confusion matrix 로 표현 시 아래와 같은 표로 표현된다.

Predicted	0	1	All
True			
0	44	3	47
1	2	65	67
All	46	68	114

원래 정답이 악성인데 양성으로 분류한 것 3개 양성인데 악성으로 분류한 것 2개로 나쁘지 않은 성능이였다.

그 다음은 한 레이어 안의 노드 수는 1~35개 사이 중에서 2개의 레이어를 연결시켜서 가장 좋은 성능을 가지는 케이스를 뽑았다. 가장 좋은 조합은 alpha 1e-05, 노드 수는 각각 첫 번째 히든 레이어의 노드 개수는2개, 두 번째 히든 레이어 노드 개수는 19개로 되었다. Max iter와 random state는 위와 동일하며 solver는 lbfgs이다. 이 조합을 사용하여 만든 모델의 성능을 테스트 데이터로 테스트를 진행하여 성능 측정 후에 나오는 confusion matrix는 아래와 같은 표로 표현된다.

Predicted	0	1	All
True			
0	44	3	47
1	0	67	67
All	44	70	114

이 모델은 원래 정답이 악성인데 양성으로 분류한 것이 3개였고, 양성인데 악성으로 분류한 것은 없었다.

SVM

Hyper parameter를 찾기 위해 SVM의 parameter들은 kernel, C, gamma, degree, coef, shrinking,

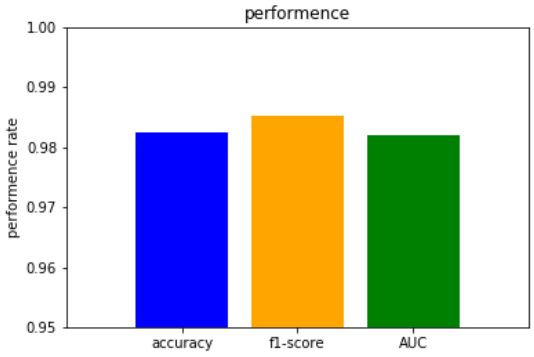
probability, tol, cache_size, class_weight, vervose, max_ider, decision_functionm_shape, random_state 이렇게 존재하는데 이 중에서 kernel, C, gamma만 조정하기로 하였다. 다른 hyper parameter들은 설명을 보니 default로 사용하는 경우가 많고 중요한 매개변수는 저 3가지라는 것을 알 수 있었기 때문에, default로 사용하고, 저 3개만 조정하기로 했다. Kernel엔 sigmoid, rbf가 있고, C는 얼마나 많은 데이터 샘플이 다른 클래스에 놓이는 것을 허용하는 지 결정한다. C가 높으면 outlier의 존재를 허용하지않겠다. 라는 마음가짐이고, 낮으면 outlier 어느 정도는 괜찮아 라는 의미가 있다. 하지만 너무 높으면 overfitting이 일어나고, 너무 낮으면 underfitting이 일어난다. 그래서 적합한 C를 찾아야한다. Gamma는 가우시안 함수의 표준편차와 관련이 있는데, 클수록 작은 표준편차를 갖는다. Gamma가 크면 데이터 포인트끼리 영향력을 행사하는 거리가 짧아지고, 작으면 커진다. Gamma는 결정 경계의 곡률을 조정한다. 따라서 이 데이터에 대한 gamma와 C의 적절한 수치를 찾고, kernel을 무엇을 사용할 지 gridsearchCV라는 기법을 이용하여 찾았다. girdsearchCV는 여러 조합들을 테스트해서 가장 좋은 성능을 내는 매개변수를 찾아냈다. SVM모델에서 가장 좋은 조합의 hyperparameter는 kernel = rbf, C = 10, gamma = 0.01, 나머진 default인 조합이고, 이 매개변수를 모델에 적용시켜 테스트 데이터로 테스트를 진행해보니 confusion matrix가 아래와 같은 표로 나타났다.

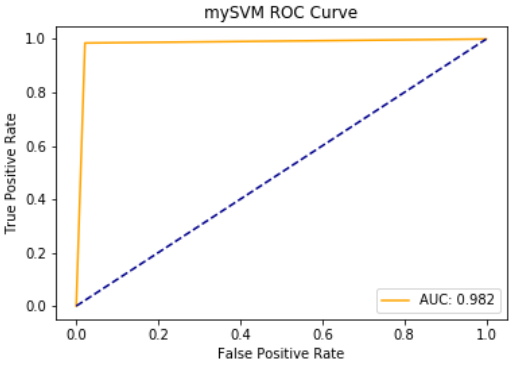
Predicted	0	1	All
True			
0	46	1	47
1	1	66	67
All	47	67	114

악성인데 양성으로 분류한 것 하나 양성인데 악성으로 분류한 것 하나씩 존재하였다.

최종 모델 및 성능

위의 실험 결과를 바탕으로 하여 최종모델을 선택하였다. MLP 2 번째 케이스와 SVM 모델 중 고민을 하였다. MLP는 양성을 다 맞추는 대신 악성을 3개 틀렸고, SVM은 양성과 악성 1개씩 틀렸다. 우리가 암으로 판단을 할 때 MLP는 암인데 암이 아니라고 판단한 사람이 3명인 것이다. 하지만 SVM은 암인데 암 아니라고 한 사람 1명, 암이 아닌데 암이라고 한 사람 1명이라고 하였다. MLP를 사용할 경우 암인데 아니라고 판정을 해서 3명은 사망할 가능성이 크다. 하지만 SVM 모델을 사용할 경우 암인 환자 1명만 사망하고 나머지 1명은 2차 검사를 더 진행하여 비용을 추가적으로 지불을 하는 경우가 발생하지만 목숨엔 영향이 없다. 이러한 암인 지 아닌 지 판별하는 이유는 사람의 목숨을 잃지 않게 하는 것이 목표이기 때문에 최종모델을 SVM 모델로 선정하게 되었다. SVM을 새롭게 나만의 인공지능으로 만들었다. 인공지능의 hyperparameter는 kernel = rbf, C = 10, gamma = 0.01, 나머진 default인 모델이다.





위의 측정 방법들인 accuracy, f1-score, ROC-AUC를 이용하여 성능을 확인해보았을 때 모두 0.98을 넘었다. 데이터가 6:4로 밸런스가 맞지 않기 때문에 accuracy뿐만 아니라 f1-score, roc-auc를 이용하여 성능을 측정하였다.

결론

Default 데이터인 breast cancer 데이터와 잘 맞는 모델은 선형 모델이 아닌 비선형 모델과 잘 어울린다는 것을 알 수 있었으며, 그 중에서도 SVM과 MLP 모델이 이 데이터 셋에 가장 잘 어울린다는 지표가 나왔다. 그래서 두 모델을 내가 지정한 범위 내에서 hyperparameter를 미세 조정하여 가장 좋은 모델들을 1가지씩 가져왔으며, 그 모델들의 confusion matrix를 보고서 이 데이터를 잘 예측하고, 분류하고, 이 인공지능의 목표에 알맞은 SVM 모델을 선택하였다. 선택한 SVM 모델은 양성을 악성이라고 판 별하는 것 1개와 악성을 양성이라고 판단하는 것 1개로 총 2개 예측에 실패하였다. 하지만 SVM 모델의 매개변수 C가 10임에도 불구하고 각각 1 개씩 틀린 것을 보아하니 틀린 데이터 2개가 outlier일 가능성이 크다는 것을 알 수 있었다. 이 1개씩 틀린 데이터의 outlier가 어떤 feature로 인해 outlier인 지 판별한다면 더 정확한 인공지능이 될 수 있다고 생각한다.