**2018-1 IT 집중교육 1**

**homework 2**

1조

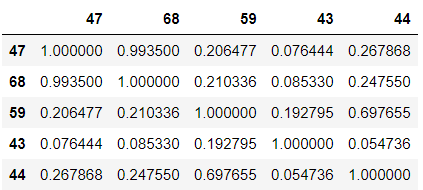
민태홍

201320991

제출일자 :03/19

1. 데이터 분석(2)-A에서 feature selection에 사용한 알고리즘을 간단히 설명하고 몇 개의 feature를 왜 사용했는지 기술

일단 결론은 최종 feature를 '47','68','59','43','4'로 잡았다. 그 이유는 각 feature의 coefficient matrix를 구해 예측해야 되는 86 column과의 가장 관계가 높은 5가지를 뽑았을 때 '47','68','59','43','44'가 나왔다. 또 뽑힌 각각의 feature의 관계를 봤을 때

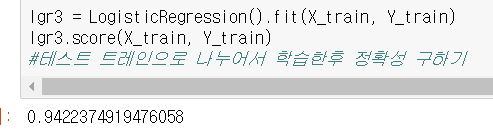


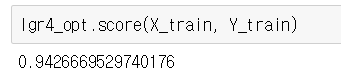
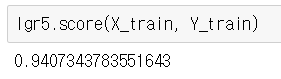
47(1) 68(2) 관계가 높음 44(5)랑 59(3)관계가 높았다 따라서 5개중 가장 낮은 44를 빼고 데이터를 봤을 때 나이를 넣기로 결정하였다.

1. 데이터 분석(2)에서 regularization term을 찾는 과정에 대해 간단히 설명하고 training accuracy를 기술하고 5-cross validation을 수행 결과를 제시한다.

1번에서 coefficient가 높은 5개의 feature 만을 뽑아서 logistic regression을 사용하였을 때 train의 score는 0.9395가 나왔다. 이것을 기준으로 더 정확하게 하기위해서 일단

train데이터와 test데이터를 합쳐서 정규화를 시킨 후 다시 원래크기만큼 train 데이터와 test 데이터로 나누었다. 나눈 후 train데이터를 5-cross validation을 사용하여 logistic regression을 학습시키고 accuracy를 구해 보았을 때 아래의 그림과 같이 나왔고



Lasso를 이용하여 5-cross validation을 사용하고 C값을 최적화를 하여 학습시키고 accuracy를 구해 보았을 때  이와 같이 나오게 되었다. 또 ridge를 5-cross validation을 사용하고 c값을 최적화하여 하였을 때  이와 같이 accuracy가 나오게 되었다.

따라서 최종적으로 lasso를 이용한 regression을 사용하였다.

1. 위에서 분석한 내용을 토대로 데이터 분석 결과의 해석 진행

최종 결과값을 보았을 때 다 0으로 나오게 된 것을 확인하였다.

왜 0인가 하며 다른 모델들을 다 사용하여 해보았을 때 가장 정확성이 낮은 아무 feature만을 뽑았을 때 1이 섞여 있었다. 다른 정확성이 높은 모델들 예를들어 coefficient가 높은 5개의 feature를 뽑았을 때, laso모델을 사용하였을 때 ridge모델을 사용하였을 때 등등 여러 train accuracy가 높은 모델들은 결과값이 다 0으로 나온 것을 확인하였다. 따라서 데이터분석을 해보았을 때 처음에 학습시킨 데이터가 제대로 된 특징이 없는 것 같았다. 우리가 예측해야하는 결과값과 모든 feature에서의 coefficient가 가장 높은 값이 0.11이 나온 것을 보았을 때 처음 데이터가 좋은 데이터가 아닌 것 같다고 결론이 지어졌다.

**[2] 최적화 과제**

1. 최종 결과는 0.6685가 나왔다.

우선 확률을 높이기 위해 train데이터와 test데이터를 합쳐서 결측 값을 처리하고 구하려는 class의 feature와의 coefficient를 구해보았다. 구해서 상위 4개의 feature를 구해보았는데 'V91','V16','V85','V8'가 나왔다 이 feature를 뽑아서 다시 train데이터와 test데이터의 크기로 나눈 후 1차적으로 5-nn분류기를 사용하여 test의 accuracy를 구해 보았을 때 0.506이 나왔다. 이 accuracy는 float형식의 feature만 사용하여 coefficient를 구한 것이 기 때문에 낮게 나온 것 같다고 생각이 들어 다시 data를 합친 후 one-shot을 사용하기위해 dummy를 이용하여 class와의 상관관계가 높은 4개의 feature를 뽑아 보았는데 'V51','V8','V3','V69'가 나왔다. 4개의 feature를 사용하여 5nn분류기를 사용하여 test의 accuracy를 구하면 0.608이 나왔다.

앞에서 구한 4개의 feature와 방금 구한 4개의 feature의 class와의 상관관계를 확인해 보았을 때 V51','V8','V3이 다른 feature보다 coefficient가 월등히 높아 이 세개의 feature를 pix하였다.

V51','V8','V3세개의 feature를 이용하여 정규화를 시키고 5-NN분류기를 사용하였을 때 test accuracy = 0.648 Decision tree 분류기를 사용하였을 때 test accuracy = 0.658 svm분류기를 사용하였을 때 test accuracy = 0.628이 나왔다. 더 최적화하기 위해 K-NN의 K값을 구하도록 최적화를 시행해 보았을 때 k가 9일 때 test accuracy = 0.662 가 나왔고 svm을 최적화 하여 gamma=2이고 c값이 3이상이면 최대 test accuracy = 0.6658이 나왔다 . 마지막으로 random forest모델을 이용하였을 때 test accuracy = 0.652가 나왔다.

자세한 내용은 제출한 코드에 주석을 달아 놓았습니다