2019150445 신백록 과제1

1. SMOTE와 ADASYN에서 k-neighbors를 조절하여 변화에 대해 논의하라

일단 특성변수가 10개이기에 plot을 할 수가 없다. 따라서 pca를 통해 2차원으로 차원축소를 한 후, nonlinear svm을 통해 decision boundary를 표현하였다. 윗 줄은 original data, k=1 Adasyn, k=5 Adasyn, k=10 Adasyn, k=30 Adasyn이고, 아래줄은 k=1 Smote, k=5 Smote, k=10 Smote, k=30 Smote이다.

텍스트, 나무이(가) 표시된 사진

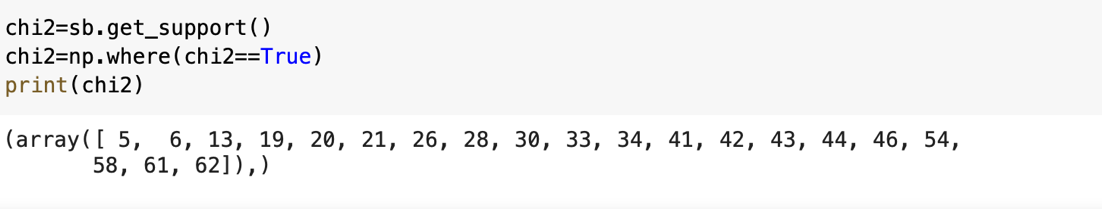
자동 생성된 설명

일단 k가 커지면 커질 수록, major class인 Red색상의 부분이 점점 더 감소하고 있는 게 눈에 띈다. Major class의 data들은 synthetic한 data를 만들어내지 않아 바뀌지 않지만 red point들이 차지하는 부분이 바뀌는 건 아마 pca때문이지 않을까 싶다.

K가 작을 땐 minor data들이 모여있는 것보다 line 형태를 이루고 있는 게 보인다. 아마 dot들을 단순히 connecting해서(특히 k=1일 때) 선 형태가 보이는 것 같다. K가 커지면 커질수록 data들이 점점 더 line형태에서 벗어나 모이는 것이 보인다.

2. 특성변수의 선택에서 교재에 있는 digit 데이터를 이용하여 목적변수가 분류일 때 chi2, f\_classif, 그리고 mutual\_info\_classif를 적용한 결과를 비교하라.

SelectKBest 함수를 통해 chi2 통계랑을 기준으로 20개의 significant한 특성변수들을 뽑아내서 나열해보면 다음과 같은 index의 특성변수들이 significant하다는 것을 알 수 있다.



이런 식으로 F 통계랑을 기준으로 20개의 significant한 특성변수들의 index와 mutual Info를 기준으로 하는 20개의 significant한 특성변수들의 index를 구하였다.

특히 F는 sample size가 무한대로 갈 때 chi2로 수렴하기에 F와 chi2의 교집합을 구해보았다. 교집합의 원소의 개수는 15개로, 완전히 같지는 않지만 F와 chi2가 비슷한 결과가 나오는 것을 알 수 있다. 아마 sample size가 1700여개로 그리 크지 않아서 20개 중 15개만 같은 결과가 나왔을 것이다. Sample size가 커지면 커질수록, significant한 feature들은 점점 더 똑같아질 것이다.

특성변수가 64개로, 이는 8\*8 픽셀의 이미지 데이터를 표현하는 것이므로 이를 그려보았다. F 통계량 혹은 chi2 혼자만 significant한 특성변수들은 검정색으로, F와 chi2 둘 다 significant한 특성변수텍스트, 낱말맞추기게임이(가) 표시된 사진

자동 생성된 설명들은 빨간색으로 색칠하였다.(column-wise로 색칠하였다.) 일단 가장 눈에 띄는 것은 오른쪽 두 줄이다. 오른쪽 두 줄은 가장자리로, 거의 숫자가 써지는 부분이 아니기에 분류에 영향을 많이 끼치지 않을 것이다. 그리고 가장자리에서 중간으로 올 수록 검정 부분과 빨간 부분이 많아지는 것을 볼 수 있고, 중간 부분의 특성변수들이 숫자의 분류에 영향을 많이 끼치는 것으로 해석할 수 있다.