<군집화, 의사결정 나무 분석>

정보융합학부 2018204002 박정재

#1. 군집화 분석

1)데이터 수집 및 전처리

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

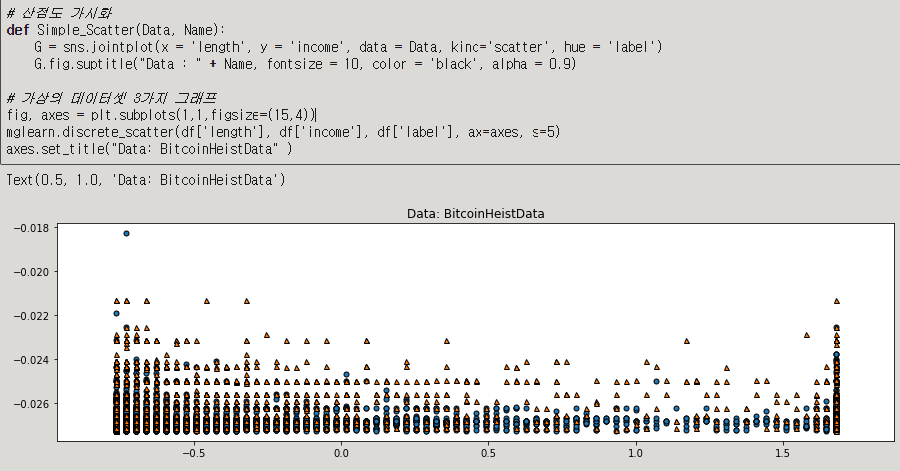
자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

먼저 군집화 분석을 할 환경을 위해 라이브러리를 설치해준다. 그리고, 데이터는 비트코인 랜섬웨어 주소 데이터를 사용하였다. 비트코인 랜섬웨어 주소에 따른 길이, 루프, 수입 등의 데이터를 나타낸다. 이 때, 랜섬웨어 주소에 따라 길이와 수입의 값들을 군집화 분석 해보기로 했다. 먼저 label 데이터를 범주형에서 수치형으로 바꾼 후 데이터를 정규화 해준다. 그 후, 이상치를 제거하고 두 분류로 제거하기위해 label데이터를 수정해주었다.

2) 데이터 분석



데이터의 분포정도를 알기 위한 Simple\_Scatter함수를 만들어주고, 길이와 수입에 따라 산점도 분포를 나타내었다. 분포를 보면 그룹별로 그렇게 잘 모이지 않는다는 것을 알 수 있다.

3) K-Means

텍스트이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

테이블이(가) 표시된 사진

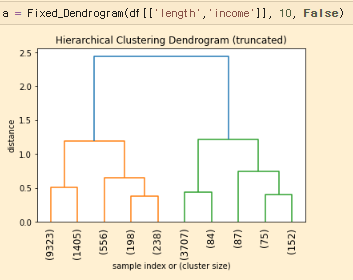
자동 생성된 설명

우선, K-Means함수를 만들어 K-Means분석을 수행할 수 있게 했다. 첫번째로, 초기n값을 1로 설정한 후 k를 2~5까지 증가시켜 Simple\_Scatter와 같이 비교했고 두번째는 n값을10, 마지막은K-Means++를 통해 분석해 보았다. 세가지 분석모두 분석결과가 k를2로 설정했을 때가 가장 좋은 score값을 가지고 분포도 두 그룹으로 나누어져 좋은 모습을 보여주었다.

4) Hierarchical clustering

텍스트이(가) 표시된 사진

자동 생성된 설명



텍스트이(가) 표시된 사진

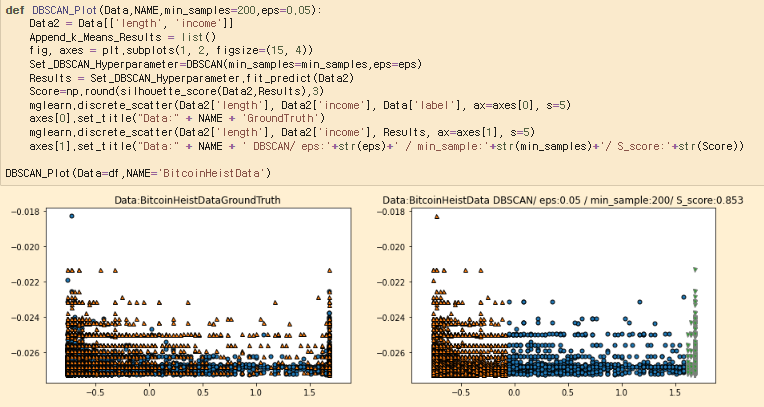
자동 생성된 설명

테이블이(가) 표시된 사진

자동 생성된 설명

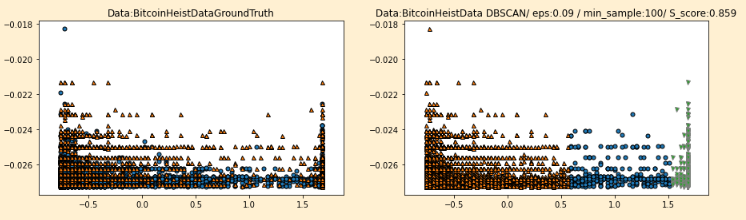
Hierarchical clustering 마찬가지로 우선 Simple\_Scatter를 통해 데이터의 분포를 확인하고 Fixed\_Dendrogram 함수를 구현하여 트리형태로 데이터를 구분하여 군집별로 어느 정도의 데이터를 가지는지를 확인했다. 그 후 Hierarchical clustering를 구현하기 위한 Hclust\_Plot을 만들어 산점도와 같이 분석해보았다. 분석결과는 K-Means와 거의 유사한 결과로 k값이 2일 때 score값이 가장 좋고 그룹이 두개로 나누어지는 좋은 결과를 보였다.

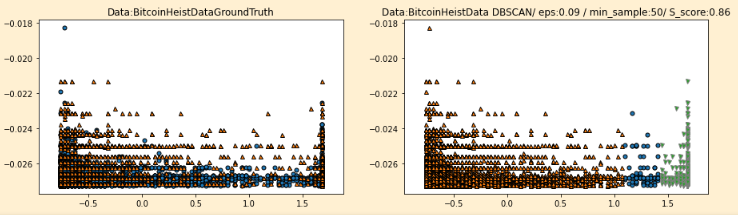
5) DBSCAN



텍스트이(가) 표시된 사진

자동 생성된 설명





위의 두가지 분석들과 마찬가지로 DBSCAN을 하기위한 DBSCAN\_Plot 함수를 만들어 주었다. 먼저, 파라미터들을 min\_samples=200, eps=0.05으로 설정하고 산점도와 같이 비교를 해보았다. 임의로 값을 설정하여 구분이 잘 되지 않은 모습이다. 이 때, min\_samples와 eps가 각각 (100, 0.03), (50, 0.03) 정도의 값을 가질 때 적절한 모습을 보인다는 것을 확인했다. 따라서, 파라미터의 값들을 늘리고 줄여가며 분석해본 결과 (50, 0.09)의 파라미터 값을 가질 때 가장 산점도와 유사하게 거의 두분류로 분류되고 score값도 괜찮게 나왔음을 알 수 있다.

#2. 의사결정나무

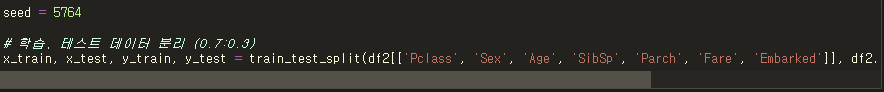
1. 데이터 수집 및 전처리

텍스트이(가) 표시된 사진

자동 생성된 설명

의사결정나무 분석을 하기위한 기본 환경 구축 라이브러리를 설치하고 데이터를 가져왔다. 데이터는 타이타닉 관련 데이터로 사고당시 사망유무에 대해 성별, 나이 등의 많은 데이터분포를 나타낸다. 우선 데이터의 결측치를 제거해주었고, 성별과 승선과 같은 범주형 변수들을 수치형으로 변환해 주었다.

2)의사결정나무 구축 및 가시화



텍스트, 스크린샷, 모니터, 화면이(가) 표시된 사진

자동 생성된 설명

텍스트, 표지판, 다른, 묶음이(가) 표시된 사진

자동 생성된 설명

먼저, 학습데이터와 테스트데이터를 분리하기위해 seed를 설정해주고 테스트를 70% 수행하는 것으로 설정한다. 이를 바탕으로 의사결정나무 모델을 구축하고 tree.dot 파일을 생성하여 먼저 max\_depth에 제약이 없을 때의 의사결정나무 모델을 시각화 해준다.

3) 의사결정나무 모델 테스트

텍스트이(가) 표시된 사진

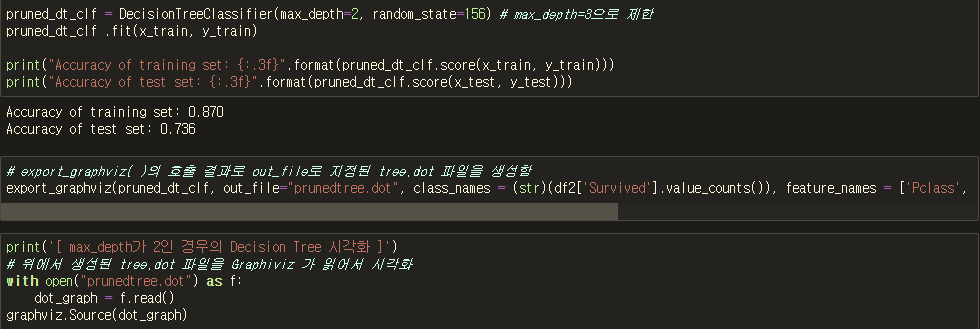
자동 생성된 설명

X의 테스트데이터를 이용하여 예측된 y를 만들고 정확도, recall, matrix로 학습결과를 보면 훈련정확성이 1.0임을보아 너무 과적합 했고 아직 부족함을 알 수 있다. 정확도는 84%정도, recall은78%, F1의 값은 0.81이 나왔다.

4) 교차검증 및 결과비교분석

텍스트이(가) 표시된 사진

자동 생성된 설명



텍스트, 표지판이(가) 표시된 사진

자동 생성된 설명

앞선 의사결정나무 모델보다 정확도를 높이기 위해 교차검증을 사용하였고, score의 평균값은 0.75정도가 나왔고, test셋에 대한 score값도 0.75정도가 나왔다. 마지막으로 더욱 정확도를 높혀주기위해 가지치기를 수행하였고, 정확도는 84%에서 87%까지 올랐음을 알 수 있다. 그 후 max\_depth의 값을 2로 제한을 둔 결과 가시화의 결과가 더욱 간결하고 정확성까지 높이게 되었음을 알 수 있다.