주제분석 1주차 패키지

- 분석 툴은 R/Python 둘 다 가능합니다. 주제분석 1주차 패키지 문제의 조건 및 힌트는 R을 기준으로 하지만, Python을 사용해도 무방합니다.
- 제출형식은 HTML, PDF 모두 가능합니다. **.ipynb 이나 .R 등의 소스코드 파일은 불가능합니다.** 파일은 psat2009@naver.com으로 보내주세요.
- 패키지 과제 발표는 세미나 쉬는 시간 후에 하게 되며, 랜덤으로 16시 00분에 발표됩니다.
- 제출기한은 **목요일 23:59까지** 이고, 지각 시 벌금 5000원, 미제출시 10000원입니다. 패키지 무단 미제출 2회 시 퇴출이니 유의해주세요. 또한 불성실한 제출로 인한 경고를 2회 받으실 시도 퇴출이니 유의해주세요.

Chapter 0. 흐름 정리

문제1. P-SAT 카페의 PT 최종 | 주제분석, 또는 자료 공유방 게시판에 들어가주세요. 이전 분석 자료들을 살펴본 후, 가장 흥미로웠던 주제를 하나 선정하고 분석 흐름을 정리하여 하나의 파일(pdf)로 정리해주세요.

문제2. 이번 챕터 과제는 피셋 메일로 보내는 것이 아닌, 각 팀 팀장님께 제출해주세요. 팀 단톡방에 공유해서 함께 보는 것도 좋습니다.

Chapter 1. 이상치 탐지(Anomaly Detection)

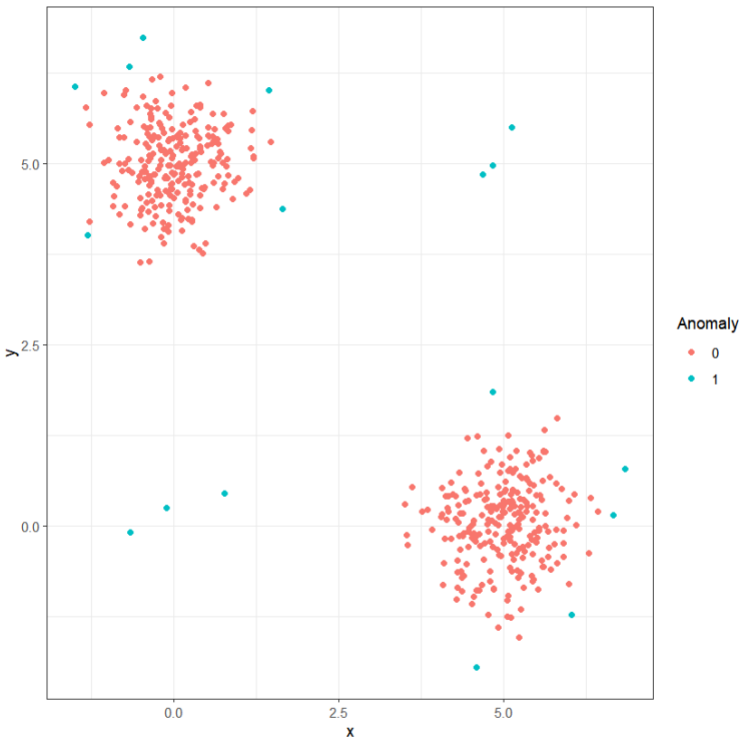
데이터 안에서 예상치 못한, 또는 일반적이지 않은 패턴을 찾는 일련의 활동을 이상치 탐지라고 합니다. 이는 이상 거래 탐지, 침입 탐지, 고객 이탈, 스팸 메일 탐지 등 클래스 불균형이 심각한 여러 Domain에서 사용이 가능합니다. 이상치 탐지는 통계적 방법, 머신러닝, 딥러닝 등 다양한 기술을 사용하여 구현될 수 있고, 이번에는 대표적인 트리 기반 이상치 탐지 모델인 isolation forest에 대해 알아보겠습니다. 우선 간단한 시뮬레이션 데이터에 대해 이상치 탐지의 원리를 확인해본 후, 실제 데이터에 대해 수행해보겠습니다. 이후 고차원 데이터를 시각화하여 수행했던 프로세스(이상치 탐지)가 타당했는지도 확인해봅시다.

문제0. Isolation Forest 알고리즘에 대해 알아본 후, 핵심 Concept을 1문장으로 작성해주세요.

문제1. 아래 링크에 접속해주세요. 문제1. 코드를 실행해주세요.

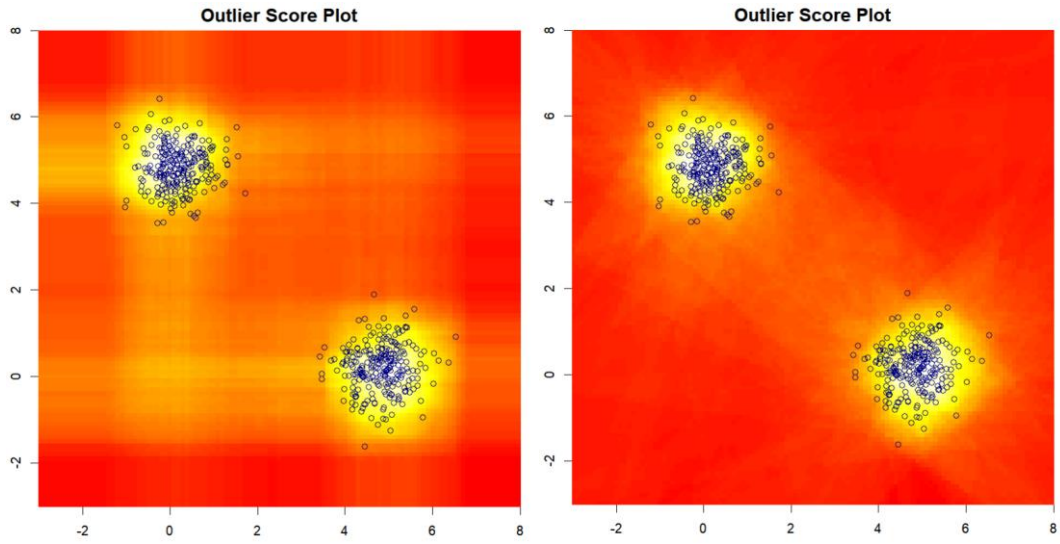
https://maize-shield-367.notion.site/1-2ef0c724839c47488cc6a63baa871b57?pvs=4

문제2. 문제2. 코드를 실행해주세요. 빈칸을 채워 아래의 plot을 그려주세요.



(HINT) theme_bw(), label과 범례 등을 제외하고는 모두 기본 옵션입니다.

문제3. 문제3. 코드의 빈칸을 채워 아래의 왼쪽 plot이 나타나게 해주세요. 이후 isolation.forest() 에서 ndim 파라미터를 2로 조정한 뒤(ndim=2), 다시 문제3 코드를 실행해주세요. 결과적으로 오른쪽 plot이 나타나야 합니다.



문제4. 문제 1과 3 코드 내에 있는 함수는 Outlier Score가 높을수록 어두운(빨간) 색으로, 낮을수록 밝은(노란) 색으로 나타나게 합니다. Outlier Score가 높을수록 우선적으로 이상치로 분류된다는 것을 고려할 때, ndim parameter를 1로 하는 것과 2로 하는 것 중 어느 것이 더 좋은 모델일지 생각해본 후 작성해주세요.

(BONUS) 위와 같은 차이가 발생한 이유를 추측해보고 의견을 작성해주세요.
(HINT) ndim parameter는 isolation forest가 데이터를 나누는 직선의 차원을 의미합니다.

문제5. creditcard.csv를 불러오세요. Class 열을 class 변수에 따로 저장한 후, 데이터에선 제거해주세요.

(HINT) class를 factor형으로 미리 지정해두면 다음에 나올 문제를 풀기 수월해집니다.

문제6. 클래스가 제거된 데이터에 대해 isolation forest 모델을 구성하고, Score가 0.6 이상이면 anomaly로 분류하도록 이상치 탐지를 수행해주세요.

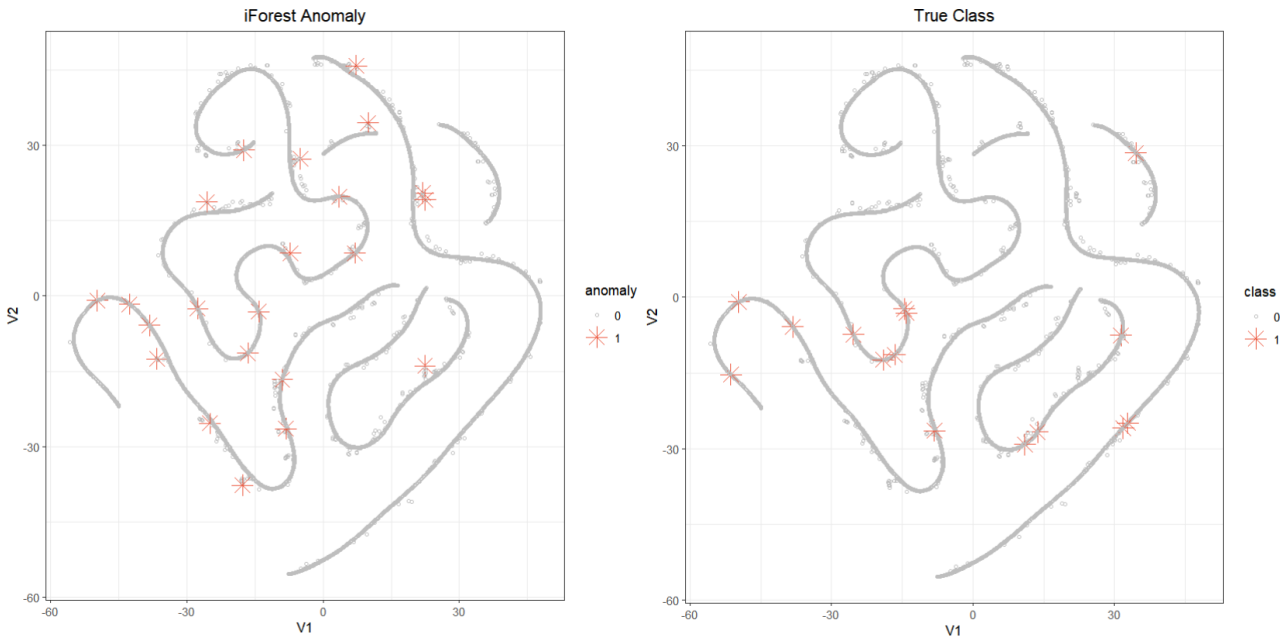
문제7. 혼동행렬을 생성해, 실제 class와 anomaly 여부의 결과를 비교해보세요. Recall을 구해주세요.

(HINT) 혼동행렬은 table()로 간단하게 확인할 수 있습니다.

문제8. Rtsne 패키지를 불러와주세요. tsne를 통해 Class가 제거된 데이터를 2차원으로 축소해주세요.

(HINT) set.seed(3233)를 실행 후, tsne 함수 내에서도 (seed = 3233)를 선언해줘야 결과가 항상 같습니다. perplexity=50으로 설정해주세요.

문제9. 차원 축소된 데이터에 대해 Anomaly, Class에 따라 각각 아래와 같이 시각화해주세요.



(HINT1) class에 따라 색뿐만 아니라 크기, 모양도 조절할 수 있습니다. 사용한 옵션은 다음과 같습니다.

scale_color_manual(values = c("grey", "#ED553B"))

scale_size_manual(values = c(1, 5))

scale_shape_manual(values = c(1, 8))

(HINT2) random seed나 설정에 따라 다른 plot이 나올 수 있습니다. 이번 문제는 자유롭게 시각화해주셔도 좋지만, 색깔과 모양은 클래스마다 다르게 해주세요.

(BONUS) tsne로 차원을 축소한 후 anomaly detection을 적용하는 것과, 차원 축소 없이 원본 데이터에 anomaly detection을 적용하는 것을 비교해보세요. (각각에 대해 문제7, 문제9를 수행해 결과를 비교해보세요.)