1. **초록**

선정한 데이터는 Kaggle에서 가져온 데이터로, 카드 사기에 대한 문제를 담고 있는 데이터이다.   
<https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud>

본 실험에서는 이미 Labelling 되어 있는 Y값이 존재한다. (정상: 0, 사기: 1)   
이를 활용하여 이상탐지 알고리즘을 통해 해당 학습 모델을 구축한다.   
그 후 해당 데이터의 정상과 비정상을 탐지하고, 이를 실제 값과 비교하여 이들의 효율성을 측정하려 한다.

실험에서는 SVM과 SVDD 알고리즘을 적용하여 정확도 측면의 비교 실험을 수행했으며, 실험 결과 SVM이 FRR 0.002%, FAR 0.94%, 그리고 F1 0.11%로 SVDD보다 더 좋은 성능을 보였다. 본 실험을 통해서 이상탐지가 현장에 적용될 수 있음을 느꼈고, 본 데이터를 기준으로 봤을 때 현장에 어떻게 적용될 수 있을 것 같다는 생각이 들었다

1. **문제 설명**

이 데이터는 카드 관련

1. **이상탐지를 적용하려는 데이터의 EDA**

이 데이터는 7개의 독립변수와 (=x) 1개의 종속변수(= y)가 존재한다.   
이는 다음과 같다:

Distance\_from\_home: 거래가 일어난 장소와 집 사이의 거리를 나타낸다.

Distance\_from\_last\_transaction: 마지막 거래 장소와 집 사이의 거리를 나타낸다

Ratio\_to\_median\_purchase-price: 구매 값/ 중앙값 의 비율을 나타낸다.

Repeat\_retailer: 같은 소매인에게 반복적으로 거래가 일어났는지를 암시한다.

Used\_chip: 카드를 긁어서 계산했는지 확인한다.

Used\_pin\_number: 핀 번호를 입력하여 계산했는지 확인한다.

Online\_order: 거래가 온라인으로 이루어졌는지 확인한다.

Fraud: 거래가 사기였는 지 확인한다.

전체 샘플의 개수는 1,000,000개가 존재하였고, 그중 정상 데이터는 912597개, 그리고 비정상 데이터는 87043개였다. (약 9.57%)  
대부분의 correlation은 낮은 편이였고, 유일하게 ratio\_to\_median\_purchase\_price와 fraud만이 유의미한 0.46 상관성을 지니고 있었다.

(, 샘플 몇개, 각 변수는 어떠한 분포를 따르고, 결측치/이상치가 얼마나 포함되며,, 등등)데이터의 샘플은 10000개가 있지만, RAM의 부족으로 인해 02%

결측치는 존재하지 않고 / 이상치는

1. **본인이 선정한 새 이상탐지 알고리즘 간략 소개**

이번에 선정된 새로운 이상탐지 알고리즘은 SVDD(Support Vector Data Description)로,  
 Boundary-based Novelty Detection의 일종이다.   
전 처리된 데이터를 가지고 라그랑지 듀얼 함수를 구한 후, 도출된 알파 값에 따라 데이터를 분류한다.   
그 후 결정된 support vector를 이용해 radius를 계산, 이를 사용하여 Classification을 구한다.   
SVM이 d-1차원의 hyperplane boundary를 찾아 이상치를 찾는 것과는 달리   
SVDD는 최소 체적의 구를 찾는 것을 목적함수로 두고, 비선형 SVM을 응용하여 이상치 탐지를 위한 boundary를 제공한다. (참고로, 만일 모든 데이터들이 unit vector로 normalize되어 있다면 SVDD와 1 SVM은 동일하다.)

이 알고리즘은 또한 Patch-SVDD, Deep SVDD 등 다양한 변칙적 하위 알고리즘이 존재한다.

1. **실험결과 및 결론**
2. SVDD(Support Vector Data Description)는 Boundary-based Novelty Detection의 일종으로, 2004년 Tax와 Duin에 의해 개발된 알고리즘입니다. Binary classification을 수행하기 위한 boundary를 제공한다는 점에 있어 1-SVM과 그 용도는 비슷하나, 1-SVM이 원점에서 가장 멀리 떨어진 hyperplane boundary을 찾는 것에 그 목적을 두었다면 SVDD는 정상 데이터를 감싸안는 최소 체적의 구를 찾는 것을 목적함수로 두 이상치 탐지를 위한 boundary를 제공하는 알고리즘이라 할 수 있습니다.

(모든 데이터들이 unit vector로 normalize 되었다는 가정 하에 SVDD와 1-SVM은 동일한 알고리즘이 됩니다.)