대회 회고

대회 소개

신용카드 사기 거래 탐지 AI 경진대회

- 각각의 다른 id를 갖는 거래 내역마다 `V1 ~ V30` 까지의 거래 Feature을 기록하고 있는 특징 정보가 주어짐. 대부분은 정상 거래이지만 un-supervised learning을 통해서 각각의 거래 내역을 구분할 수 있어야 한다.
- `train.csv` 파일만을 모델 학습에 사용을 할 수 있고, `val.csv` 파일은 단지 모델 평가와 데이터 통계 정보 확보에 사용을 할 수 있다.
- Normal / Fraud 거래인지의 여부를 확인 할 수 없는 신용 카드 데이터
  - `train.csv`: (11384, 31)
  - `val.csv` : (28462, 32)
  - `test.csv` : (142503, 31)

해야 하는 것

- 정상/사기 거래 여부가 기록이 되어 있지 않은 `train` 데이터를 사용해서 latent space로 임베딩을 하고 ground truth label이 주어져 있는 val 데이터를 앞서 train 데이터를 사용해 구한 <임베딩을 위한 네트워크>에 통과시켜서 나온 결과물로 classification

생각해본 방법

- 30개의 feature들을 Embedding을 한 후에 그 벡터를
  1. Latent Space를 학습 시킬 때에 군집화(`DEC : Deep Embedded Clustering` , K-means)
  2. NN에 입력해서 classification
  3. unsupervised learning을 target의 분포를 추정함으로서 supervised learning인 것 처럼 네트워크에 적용

시도

1. 오토인코더 baseline(encoder: 30->64->128으로 feature 증가시킴+decoder: encoder 대칭되게 구성) : `0.9305289388`(val score 0.90312028782757557에 수렴
2. AE+nosie(DAE) : `0.9284014343`
3. z-score → validation set에서 의미있는 feature 17개(신뢰구간 95%)만 남겨서 autoencoder 학습 : `0.5312460906`
4. EllipticEnvelope로 fake label 생성 → up sampling → labeled classification AutoEncoder(baseline): `0.3983979445`
5. auto encoder로 validation data와 유사한 데이터를 생성해서 학습을 시킨 뒤에 인코더 단에서 나온 feature vector로 군집화를 해보고 앙상블로 서로 다른 threshold로 군집 결과 판단: invalid data error -> 제출 x
6. Swap noise를 추가하여 input data에 noise를 추가하려고 했지만 test data를 training에 포함해야 하는 방법이기 때문에 대회 규칙에 어긋나서 시도 못해봄
7. 1d gan : validation 스코어 자체가 너무 낮아서 제출 x

한계점 & 후기

1. 기본적으로 비지도 학습이라는 대회의 주제 특성상 <학습 가능> 데이터가 너무 제한적
   - `train.csv` 파일 만에는 정답 라벨이 없었고 정답이 있는 데이터로는 학습을 시키는 어떠한 행위도 불가했기 때문에 어쩔 수 없이 정답 라벨이 주어진 데이터를 기반으로 **data generating**을 하는 것이 제일 해볼 수 있을 법한 시도
2. 데이터의 속성의 개수가 너무 적었다. 사실 column의 개수가 많다면 차원을 줄였다 늘리는 encoder-decoder의 구조를 고안해 볼 수 있을 것이고, 차원 축소를 한다는 것의 의미가 있었을 테지만 이 경우에는 column의 개수가 고작 31개밖에 없었기 때문에 특별히 EDA를 하는 것의 목적이나 효과를 많이 보지 못했던 것 같다.
   - 그러나 지금 생각해 보면 오히려 변수의 개수가 많다고 해서 좋은 결과가 나오는 것은 아니기 때문에 결과가 제공이 되는 `val.csv` 를 사용해서 변수를 선택하는 시도를 해도 좋았을 것이라고 생각한다.

     → 예를 들면 4분위로 나누어서 실제 fraud data인데 전체 분포중에 2/4 ~ 3/4 사이에 있는 변수라면 절대로 이상치를 감지에 도움이 안될것이므로 제외 가능 변수라고 판단할 수 있었을 것이다.

- 실제로 1등한 수상자의 결과물을 보면 `FRUFS` : **Feature Relevance based Unsupervised Feature Selection** 이라는 패키지를 사용해서 변수의 중요도를 얻어 변수를 선택적으로 사용하였다.

3. 비지도 학습의 방법론이나 알고리즘에 대해 좀더 알아볼 수 있는 충분한 시간이 있었으면 좋았을 것 같다. 이미지 데이터의 경우에는 cut paste, noise2void등과 같은 자기 주도 학습을 많이 하기 때문에 익숙하지 않았던 정형 데이터 + 비지도 학습의 조합은 상당히 어려웠던 것 같다.