금융빅데이터분석학 Term paper 1

2017046271 김연준

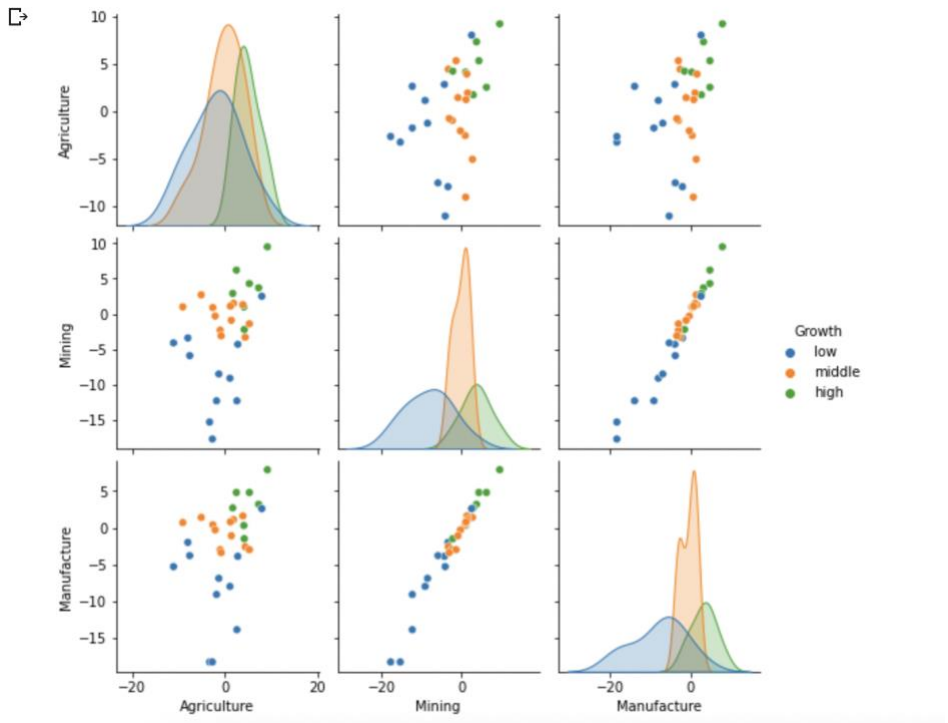
1. 주제 선정 및 이유

주제는 북한의 여러 산업 성장률(광업, 농업,제조업)으로 성장국면의 3 단계로 분류. 북한의 통계 특성상 신뢰성이 문제가 많아 성장국면이 제대로 분리될 수 있는지 의문. 따라서 직접 softmax model 로 데이터를 처리해보기로 함.

2. 개요

주제는 북한의 여러 산업의 성장률로 북한의 경제 성장률을 3 가지 category 로 분류. 출처는 한국은행경제시스템. 독립변수는 북한의 광업, 농업, 제조업 성장률 종속변수는 category 로 low ,middle, high 성장국면으로 구성. low/ middle/ high 기준은 임의의 기준 -2%~ low -2% ~ +1% middle +1%~ high 로 나눔. 이는 북한의 고난의 행군과 여러 특수한 사정을 들어 본인이 직접 설정. 본 데이터는 30 개의 행 4 개의 열로 구성. 1991 년부터 2020 년까지로 하며 주기는 1 년. 데이터 전처리 작업은 같은 성장률이며 단수차이가 크게 나지 않아 필요 없음. 데이터 분석은 soft max naïve 한 모형으로 처리하였음.

3. 본론

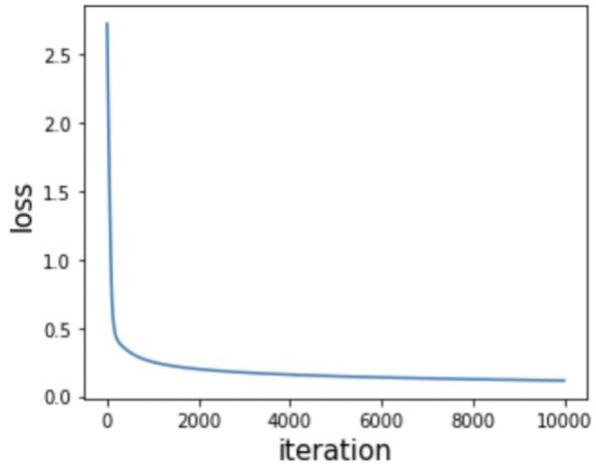


데이터의 산포도를 살펴보면 일정한 패턴 존재. 구분할 수 있는 boundary line 을 설정 가능. 그러나 어느 정도의 오차가 있을 것으로 예상.

학습률은 0.01 로 설정. Test data 와 train data 는 2:8 로 설정.

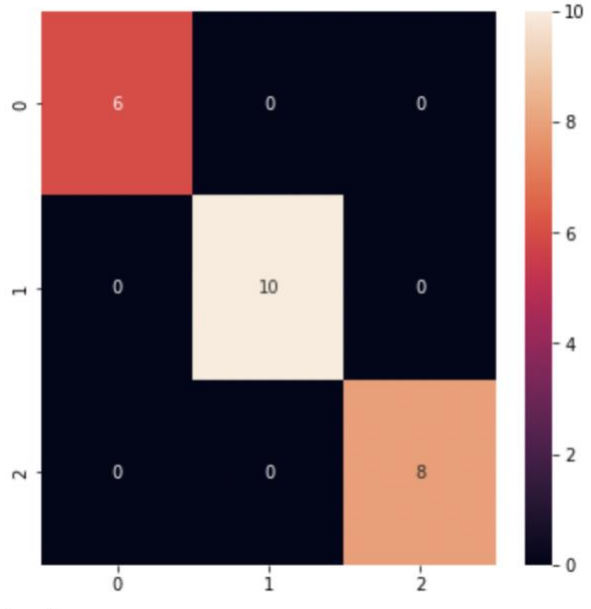
```
7500      cost: 0.12686758
8000      cost: 0.12394384
8500      cost: 0.12122622
9000      cost: 0.11868825
9500      cost: 0.11630821
10000     cost: 0.1140682
**************************************************
y = [[ 2.0289626  -0.07040556  0.25107756]
 [ 0.49482667 -1.448843    1.0554497 ]
 [ 1.1782092  -0.3305897  -1.3877922 ]]x + [-1.1045915 -2.4652555  2.7927434]
```

그렇게 10000 번의 학습을 거친 결과이와 같은 boundary line 과 loss 를 도출가능



Loss 는 2.7 부터 0.12 수준까지 감소

4. 성과 평가 및 예측



성과평과를 그림으로 표현했을 때 다음과 같음 accurancy 가 100%가 도출. 단 이는 traindata 이며 자료가 24 행 밖에 안됨을 고려하면 왜곡된 결과일 가능성

```
train
Report =
                precision    recall  f1-score   support

           0       1.00      1.00      1.00         6
           1       1.00      1.00      1.00        10
           2       1.00      1.00      1.00         8

    accuracy                           1.00        24
   macro avg       1.00      1.00      1.00        24
weighted avg       1.00      1.00      1.00        24

test
Report =
                precision    recall  f1-score   support

           0       1.00      1.00      1.00         4
           2       1.00      1.00      1.00         2

    accuracy                           1.00         6
   macro avg       1.00      1.00      1.00         6
weighted avg       1.00      1.00      1.00         6
```

Test data 와 비교해도 100% 일치 그러나 자료의 절대량이 부족함을 생각한다면 100% 유의함을 보장 불가. 이는 불가피.

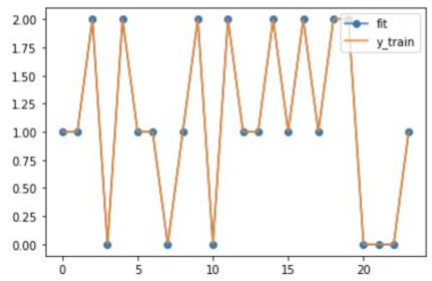
학습한 모형을 통해 각각의 경제성장률로 경제 국면이 어느 단계에 있는지 예상한다면

```
x_new_data = np.array([[0, 0, 0]], dtype=np.float32)
x_pred = predict(x_new_data).numpy()
print('Predict  in new set  = ', x_pred)
```

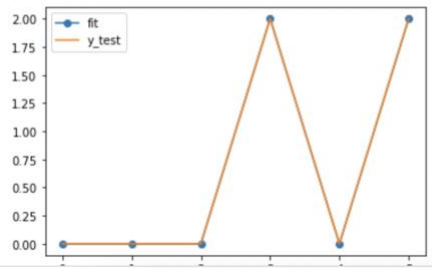
```
Predict  in new set  =  [2]
```

성장률이 모두 0% 일때 middle 국면으로 예측.

```
train
[1 1 2 0 2 1 1 0 1 2 0 2 1 1 2 1 2 1 2 2 0 0 0 1]
[1 1 2 0 2 1 1 0 1 2 0 2 1 1 2 1 2 1 2 2 0 0 0 1]
```




```
test
[0 0 0 2 0 2]
[0 0 0 2 0 2]
```



Training data 와 test data 의 predict 결과를 살펴보면 이와 같이 도출.

4. 결론 및 한계

(1) Soft max 모델로 돌린 결과, 각 산업의 성장률로 북한의 경제 성장 국면을 3 단계로 성공적으로 분류 가능. 관계식을 성공적으로 수립 가능.

(2) 데이터의 절대량이 부족함이 관계식의 유의성을 보장하지 못한다는 한계 존재