금융빅데이터분석학 Term paper 2

2017046271 김연준

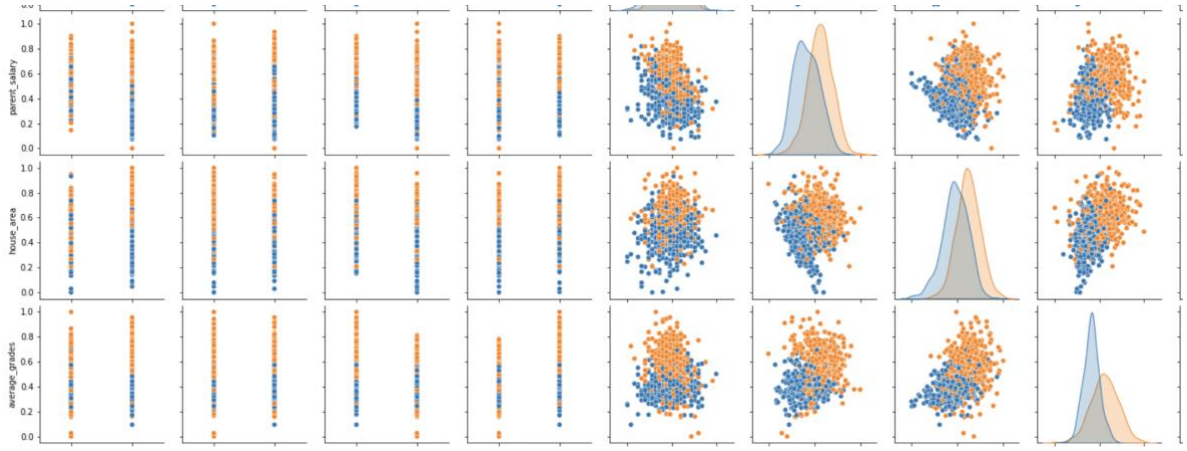
1. 주제 선정 및 이유

주제는 직업학교/인문계, 남/녀, 도시/촌락, 부모님 나이, 부모님의 임금, 부모님의 대학진학 여부, 부모님의 집과의 거리 7 개의 요인으로 대학 진학 미진학을 구분.
Deep Neural Network 로 이진분류. 주제 선정의 이유는 대학교육의 여부가 다양한 요인들에 따라서 구분 되는지 실증적으로 검증하고자 함.
출처 : https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset

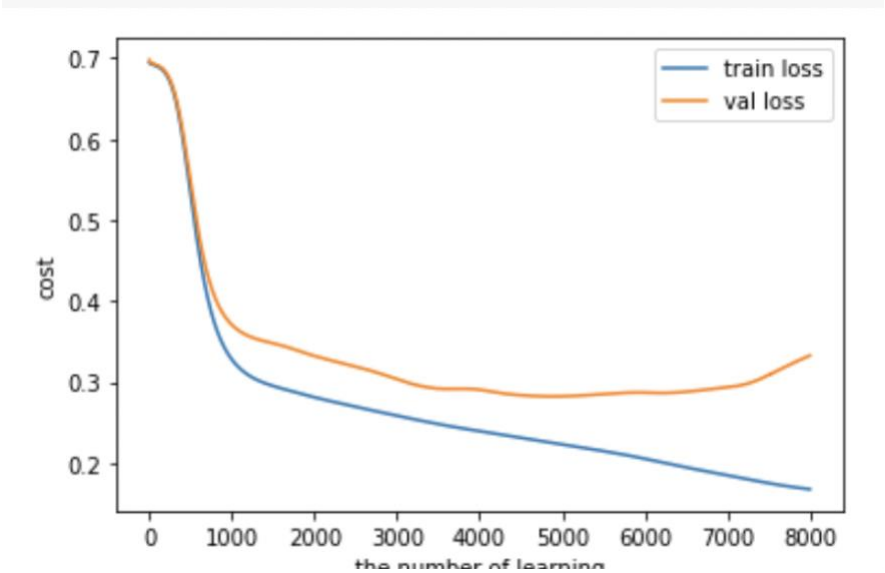
2. 개요 및 데이터 전처리

주제는 여러가지 요인에 따른 대학 진학 여부를 구분.
본 데이터는 1000 개의 행 8 개의 열로 구성. 데이터 전처리 작업은 자료들간의 단수 차이 때문에 Min-Max 로 처리해 0~1 의 값으로 표현. Train data 는 700 행 Test data 가 300 행으로 7:3 수준으로 진행.
또한 더미 데이터는 one – hot 인코딩으로 나타낼 수도 있지만 엑셀상에서 0,1 로 구분. 직업학교 0 인문계 1 여 0 남 1 촌락 0 도시 1 부모님 대졸 1 부모님 대학 미진학이 0 으로 나타냄.
Keras 라이브러리를 사용. 학습횟수는 5000 번. 학습률은 0.001 batch size 는 128. Hidden layer 는 2 개 각각의 노드는 7 개. 모형의 깊이가 깊지 않기 때문에 모두 sigmoid 사용. 비용함수는 binary cross -entropy 사용.

3. 본론



변수가 8 개라서 모든 자료의 산포도를 표현하기엔 너무 방대해 몇몇 자료만 표기. 어느 정도 규칙성이 존재. 이진 분류가 가능할 것으로 예상.



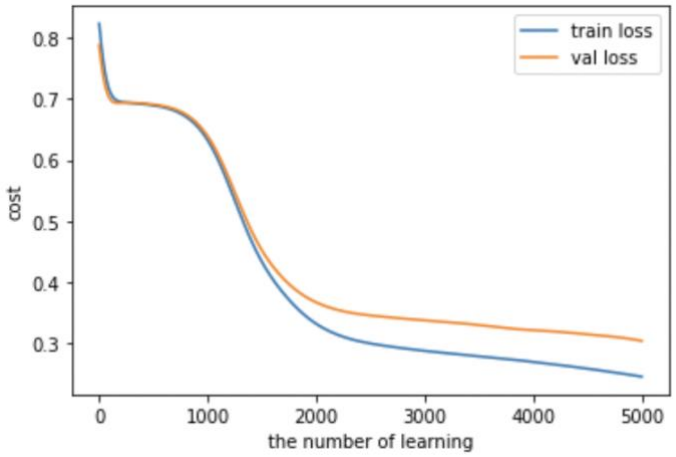
5000 번부터 성과 평가의 loss 가 늘어나기 시작 그러므로 5000 번에서 early stop 이 바람직함.

```
Epoch 4995/5000
1/1 - 0s - loss: 0.2447 - accuracy: 0.8943 - val_loss: 0.3034 - val_accuracy: 0.8800
Epoch 4996/5000
1/1 - 0s - loss: 0.2447 - accuracy: 0.8943 - val_loss: 0.3034 - val_accuracy: 0.8800
Epoch 4997/5000
1/1 - 0s - loss: 0.2447 - accuracy: 0.8943 - val_loss: 0.3034 - val_accuracy: 0.8800
Epoch 4998/5000
1/1 - 0s - loss: 0.2446 - accuracy: 0.8943 - val_loss: 0.3033 - val_accuracy: 0.8800
Epoch 4999/5000
1/1 - 0s - loss: 0.2446 - accuracy: 0.8943 - val_loss: 0.3033 - val_accuracy: 0.8800
Epoch 5000/5000
1/1 - 0s - loss: 0.2446 - accuracy: 0.8943 - val_loss: 0.3033 - val_accuracy: 0.8800
```

5000 번의 학습 결과 89.43% 의 정확도와 성과 평가에서는 88%의 정확도가 도출.
성공적으로 boundary – line 이 도출.

```
weights =
  [[ 0.7495071]
 [-2.286268 ]
 [-3.323321 ]
 [ 2.1856759]
 [ 2.2564797]
 [ 1.6934271]
 [ 2.5630593]]
biases  =
  [-0.6402032]
```

Boundary line 은 이와 같이 도출.



5000 번의 학습 동안의 loss 를 그래프로 나타내면 이와 같음

4. 성과 평가

```
[[319  38]
 [ 36 307]]
              precision    recall  f1-score   support

         0.0       0.90      0.89      0.90       357
         1.0       0.89      0.90      0.89       343

    accuracy                           0.89       700
   macro avg       0.89      0.89      0.89       700
weighted avg       0.89      0.89      0.89       700

**************************************************
Accuracy  =  0.8942857142857142
Precision =  0.8898550724637682
Recall    =  0.8950437317784257
F1        =  0.8924418604651163
[[132  11]
 [ 25 132]]
              precision    recall  f1-score   support

         0.0       0.84      0.92      0.88       143
         1.0       0.92      0.84      0.88       157

    accuracy                           0.88       300
   macro avg       0.88      0.88      0.88       300
weighted avg       0.88      0.88      0.88       300

**************************************************
```

Train, test data 가 높은 성과 지표들을 나타내며 수치 또한 유사. 이는 boundary line 이 성공적으로 설정되었음을 의미.

5. 결론 및 한계

(1) Deep Neural Network 방법론을 통해서 여러가지 요인에 의한 대학 진학 여부를 효과적으로 분류.

(2) 따라서, 부모의 경제력, 학벌, 임금 등 다양한 요인들로 대학 진학의 여부를 구분 가능.

(3) 학습에서의 hidden layer 의 개수와 node 의 개수 그리고 사용 모형 등을 바꾼다면 더 좋은 학습 결과가 나올 가능성 존재한다는 한계. 그러나 이는 많은 시간이 필요하므로 생략.