**데이터베이스 응용 보고서**

12131588 장수환

**+ training 데이터 생성 방법 :**

- 데이터를 random하게 300개 정도 생성한다. 생성된 데이터의 class는 대부분 2와 3인 중앙값으로 몰리게 된다. 따라서 양 극단 값을 생성 되도록 임의로 범위를 조절해서 0, 1, 4의 클래스가 나오는 샘플들도 어느정도 생성한 후 해당 데이터를 shuffle 하였다.

**+ network 구조 :**

- DNNClassfier를 사용.

**+ 각 설정값 별 accuracy 표**

**- batchsize = 250 & train\_step = 2500으로 고정일 때 엑셀 데이터 개수 별 정확도**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **테스트**  **데이터갯수** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 500 | 0.7 | 0.45 | 0.8 | 0.8 | 0.9 | 0.9 | 0.95 |
| 1000 | 0.99 | 0.96 | 0.97 | 0.22 | 0.97 | 0.96 | 0.68 |
| 1500 | 0.8 | 0.99 | 0.95 | 1 | 1 | 0.94 | 0.7 |
| 2000 | 1 | 0.8 | 0.89 | 1 | 1 | 0.98 | 1 |
| 2500 | 0.7 | 0.99 | 0.95 | 1 | 1 | 1 | 1 |

\* 각 데이터는 training 데이터 생성 방법에 기술한 방식대로 생성

고찰 : 높은 데이터의 개수를 가진 샘플의 경우 정확도가 완전히 일치하는 경우가 드물게 나온다. 반면에 낮은 데이터의 개수를 가진 샘플의 경우 정확도가 1이 나오는 경우가 표본에선 없었으며 많은 실험을 하더라도 거의 안나올거라 생각된다. 하지만 데이터의 개수가 정확도와 꼭 비례되서 나오는 것 같진 않다. 2500개의 데이터로 실험 했을 때 처음 정확도는 0.7로 저조한 정확도를 보였다. 반면에 낮은 데이터 500개로 실험 했을 때 높은 정확도가 다소 나오는 현상을 볼 수 있었다. 따라서 데이터의 개수와 정확도를 타협했을 때 위 실험 결과로는 1500개의 데이터 개수를 택하는 것이 바람직하다.

**- 데이터의 개수가 1500으로 고정일 때 batchsize와 train\_step 변화별 정확도**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **테스트**  **Size&step** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 50/500 | 0.23 | 0.24 | 0.3 | 0.2 | 0.4 | 0.23 | 0.5 |
| 200/2000 | 0.62 | 0.91 | 0.57 | 0.76 | 0.32 | 0.69 | 0.45 |
| 250/2500 | 0.8 | 0.99 | 0.95 | 1 | 1 | 0.94 | 0.7 |
| 300/3000 | 0.68 | 0.2 | 0.45 | 0.91 | 0.7 | 0.87 | 0.8 |

고찰 : batchsize와 train step도 역시 분포가 일정하지 않게 나온다. 확실한건 지극히 작은 사이즈일 때는 저조한 정확률을 보이고 너무 높은 사이즈로 설정시에도 저조한 정확률을 보이므로 적절하게 중간정도의 사이즈로 설정하는 것이 좋아보인다.

따라서 본인은 batchsize 와 train step을 250/2500, 샘플 데이터 개수를 1500개로 설정하였다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 0.8 | 0.99 | 0.95 | 1 | 1 | 0.94 | 0.7 |

**평균 0.911**