2022 k-ium

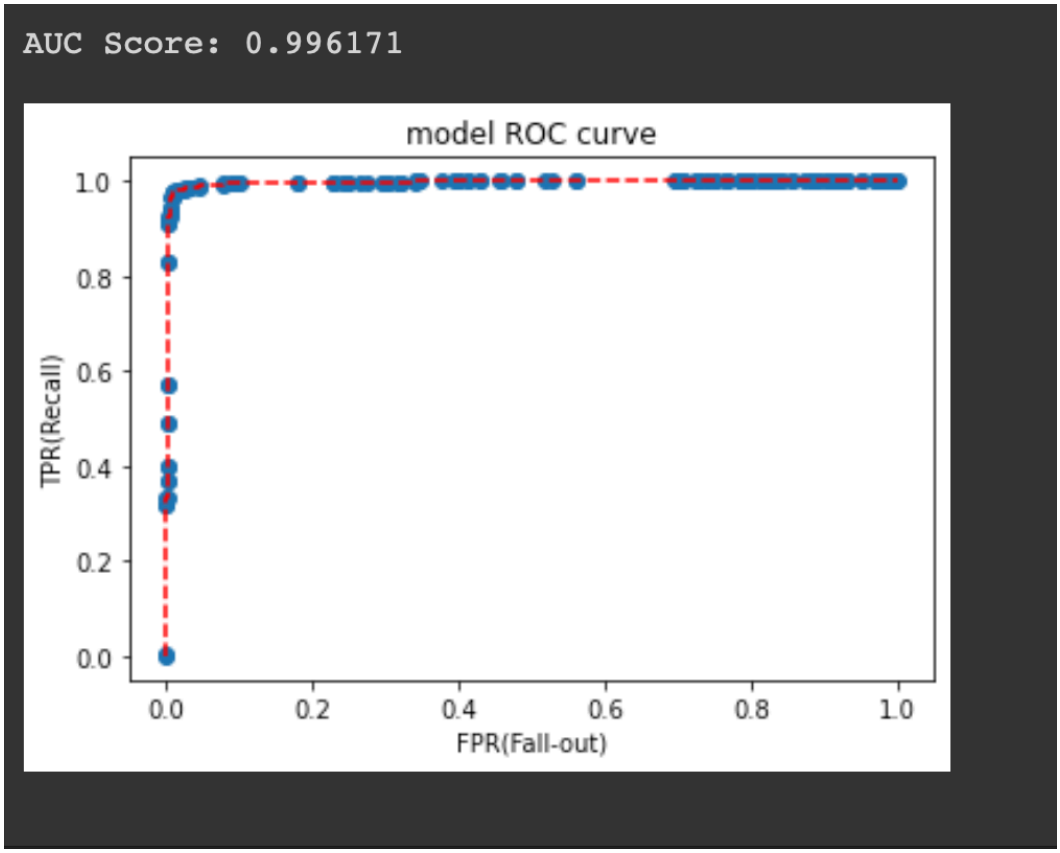
의료 인공지능 경진대회

급성 허혈성 뇌졸중 판독문 분류를 위한 인공지능 모형을 개발하라!

팀명: SignAI

단국대학교 컴퓨터공학과 정민준

본문



2차 데이터 셋(ValidationSet_2차.csv)을 제출한 모델을 사용해서 출력된 모델 값을 output(제출용).txt에 기록했습니다. output(제출용).txt에 기록된 모델 값을 바탕으로 연산한 C-statistics 값은 0.996171로 정확도가 매우 높은 것을 확인할 수 있습니다.

제출된 모델은 Apache 2.0 라이선스를 가진 "bert-base-multilingual-cased"를 기반으로 만들어졌습니다. 해당 사전학습 모델을 주최 측을 통해서 공개 된 "TrainSet _1st.csv" 파일을 통해서 학습시켰습니다. Transformer를 이용했고 토크나이저 역시 "bert-base-multilingual-cased"를 사용했습니다. Batch 사이즈는 16, optimizer는 AdamW에 학습률과 eps는 각각 2e-5, 1e-8로 설정했습니다. epoch는 4번으로 수행했습니다.

데이터 전처리는 'Findings' 와 'Conclusion' 열을 'Findings'라는 이름의 열로 합쳐서 입력 차원을 축소시켰습니다. 실제 사람이 판단을 할 때도 두 개의 데이터를 모두 참고하고 연관관계도 고려해서 판단한다는 것에서 착안했습니다. 실제로 두 개의 열 병합 후 학습 및 판별로 'Findings' 와 'Conclusion' 열 사이의 관계 등 잠재적인 요소 추출에 도움을 줘서 유의미한 성능 향상이 있다고 판단됩니다.

```
cnt_true_0 = 0 #실제 정답이 0일때, 모델이 1로 분류(False Positive)
cnt_true_1 = 0 #실제 정답이 1일때, 모델이 0으로 분류(False Negaitive)

for i in range(len(test_ans)):
  if test_ans[i] != labels[i] and labels[i] == 0:
    cnt_true_0 += 1 #(FP)

  elif test_ans[i] != labels[i] and labels[i] == 1:
    cnt_true_1 += 1 #(NP)

print("FP의 개수: ",cnt_true_0)
print("NP의 개수: ",cnt_true_1)
```

```
FP의 개수:  12
NP의 개수:  13
```

실제로 2654건의 정답 데이터(labels)와 모델의 확률 값을 numpy의 argmax 함수로 판단한 결과(test_ans) 잘못 분류된 데이터는 총 25개였습니다. 이중 FP의 개수는 12개, NP의 개수는 13개로 성능이 우수한 모델임을 확인할 수 있습니다.