Brain Age Prediction Using Machine Learning

소프트웨어융합학과 2020105742 한지훈

목차

1. 프로젝트 소개
2. Data Preprocessing
3. Apply machine learning model
4. 성능평가/visualization
5. 결과/느낀 점

프로젝트 소개

- 목표: brain-age 정확하게 예측하는 모델 개발
- 상세 목표: brain predicted age difference (brain-predicted age–actual age) 최소화
- 의의: 1. brain age와 actual age의 비교 통해 뇌 건강 진단 가능
  2. 뇌 관련 질병의 발병 시기를 예측 가능
- 제공되는 data: 68개 regional cortical thickness, intracranial volume

Data Preprocessing - Nan value 제거

- 상세 설명: train data set의 Nan 값 포함 열 제거개발
- 이유: Nan 값이 있으면 해당 열을 preprocess하는 것 자체가 불가능하다.
- 방법: Pandas의 dropna() 함수

```
train_data = train_data.dropna(axis=0)
```

	ID	Sex	Age
478	IXI653	1.0	46.0
479	IXI661	NaN	NaN
480	IXI662	1.0	42.0

	ID	Sex	Age
477	IXI652	1.0	43.0
478	IXI653	1.0	46.0
480	IXI662	1.0	42.0

Data Preprocessing - Target 분리

- 상세 설명: train data를 X, y(target)로 분리
- 이유: X의 feature들을 통해서 target을 예측해야 하기 때문이다.
- 방법: indexing of Numpy array

	ID	Sex	Age	lbankssts	rbankssts
0	IXI002	2.0	36.0	2.314134	2.445358
1	IXI012	1.0	39.0	2.256589	2.679695
2	IXI013	1.0	47.0	2.268161	2.233568
3	IXI014	2.0	34.0	2.426588	2.517577

Train data

```
X = Train_scale
y = train_data['Age'].values
```

	ID	Sex	lbankssts	rbankssts
0	IXI002	2.0	2.314134	2.445358
1	IXI012	1.0	2.256589	2.679695
2	IXI013	1.0	2.268161	2.233568
3	IXI014	2.0	2.426588	2.517577

X

Age
36.0
39.0
47.0
34.0

Target

Data Preprocessing - Scale data

- 상세 설명: 모든 train data에 대하여 값의 크기를 조절하는 scaling 진행
- 이유: data의 크기에 따른 비중의 차이를 최소화하기 위함이다.
- 방법: Sklearn의 StandardScaler

```
[[2.0 2.314133962 2.445358474 ... 2.9754802189999996 2.9220198369999997
  1393.44]
```

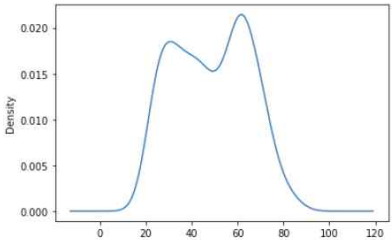
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler().fit(Train)
Train_scale = pd.DataFrame(scaler.transform(Train))
```

```
[[ 0.89921841 -0.9323089  -0.71148969 ...  0.34018861 -0.12202063
  -0.30223436]
```

Apply model - 알맞은 모델 선정

```
train_data['Age'].plot(kind = 'density')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x29476c14fc8>
```



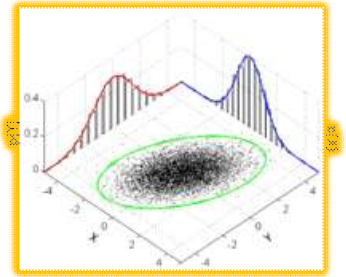
- Dataset 분석 결과: Target이 continuous하다.(나이)

continuous target의 예측에 용이한 regression model 중에서 Gaussian Process Regression 모델 사용하기로 결정

Apply model – about Gaussian Process Regression

- Gaussian Process Regression: data가 multivariate Gaussian 분포 따른다고 가정

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$



m(x): mean function

- 임의의 data x의 mean

k(x, x’): covariance function

- 임의의 두 데이터 x1, x2의 상관관계
  = Kernel

Apply model – Kernel 초기화

- 이유: Kernel을 data 분포에 맞게 설정하는 것이 model initiation에 중요하기 때문에

- 최종 kernel: RBF(Radial Basis Function) kernel + noise

```
kernel = ConstantKernel() + ConstantKernel() * RBF() + WhiteKernel()
```

Train model using defined kernel

```
model = GaussianProcessRegressor(kernel=kernel)
model.fit(X_train, y_train)
```

성능 평가

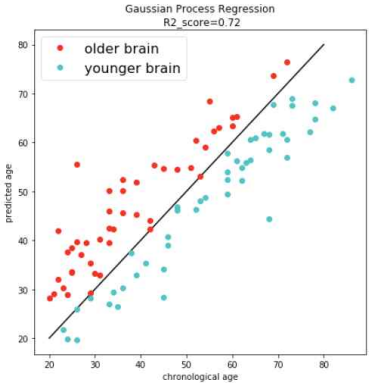
- 상세 설명: test set의 예측 값이 실제 값과 얼마나 비슷한지 확인 위한 model evaluation
- 방법: sklearn의 r2_score

r2_score

```
print('R2_score = %.2f' % r2_score(y_test, y_pred_te))
```

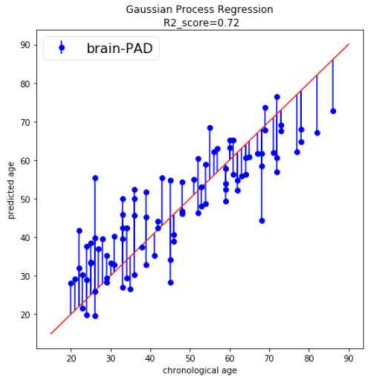
```
R2_score = 0.72
```

Visualization 1 - older/younger brain groups



- 방법: matplotlib
- 결과:
  - older/younger group 직관적으로 확인 가능
  - 두 그룹의 data point 수 거의 동일
  - data가 특정 나이대에 편중되지 않음

Visualization 2 - ‘brain-PAD’ value



- 방법: matplotlib
- 결과:
  - 전반적으로 예측, 실제 값은 큰 차이 없음
  - y = x 그래프를 따라 분포
  - 몇개 data는 실제 값과 차이 컸음

결과/느낀 점

- 프로젝트 요약

1. 주어진 feature들로 뇌 나이 예측하는 Gaussian Process Regression 모델 train
2. 두 관점의 시각화 통해 예측 결과 분석
3. train된 모델로 test data(IXI test data set, COBRE test data set)의 뇌 나이 예측

- 느낀 점

1. 수업에서 배운 machine learning 내용들을 적용할 수 있어서 뿌듯했다.
2. data science의 한 cycle을 경험해볼 수 있었던 좋은 기회

감사합니다