|  |  |
| --- | --- |
| 교육 제목 | 머신러닝 |
| 교육 일시 | 2021년 10월 19일 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 오전 | \* 어제 교육한거 반복 복습  \* 로지스틱 회귀 , 먼저 데이터 프레임에서 한가지 열을 타깃을로 만들고 나머지 5개 열은 입력 데이터로 사용합니다. 나머지 5개열을 리스트로 나열하고 to.numpy()를 이용하여 넘파이 배열로 바꿉니다. 머신러닝에서 기본적으로 데이터 세트 2개가 필요합니다.. 그다음 사이킷런의 StandardScaler 클래스를 사용해 훈련 세트와 테스트 세트를 표준화 전처리하고, 로지스틱 회귀를 사용하여 0.5를 기준으로 0또는 1에 가까워지는 그래프를 그려봅니다. z = np.arange(-5, 5, 0.1) 간격으로 배열 z를 만듭니다. 그다음 z위치마다 시그모이드 함수를 계산합니다. 지수 함수 계산은 np.exp() 함수를 사용합니다. 그다음 로지스틱 회귀로 이진 분류를 위해 블리언 인덱싱을 사용하여 두가지만 골라내여 | (or) 을 사용하여 True False 인 배열을 반환합니다. 이 배열은 | (or) 안에 입력된 값은 Ture를 그외 는 False를 나타내게됩니다. 이데이터로 linear\_model 에 있는LogisticRegression 클래스를 사용하여 샘플 몇개를 사용하여 예측해봅니다. 사이킷런의 타깃값은 알파벳순으로 정렬하여 사용합니다. classes\_ 속성에서 확인해 보죠 |
| 오후 | \* 확률적 경사 하강법 : 경사 하강법은 바로 이런 방식입니다. 가장 가파른 경사를 따라 원하는 지점에 도달하는것이 목표입니다. 딱 하나의 샘플을 훈련세트에서 랜덤하게 골라 가장 가파른 길을 찾습니다. 이처럼 훈련 세트에서 랜덤하게 하나의 샘플을 고르는것이 확률적 경사 하강법입니다. 대표적인 분류용 클래스는 SGDClassifier 입니다. 사이킷런의 리니어 모델 패키지 아래에서 임포트하면 됩니다. 여기서 loss='log'로 지정하여 전체 훈련 세트를 max\_iter=10 을 지정해 10회 반복한후 스코어 값을 추출하면 정확도가 낮게 나옵니다. partial\_fit() 메서드를 호출하고 다시 훈련세트와 테스트 세트의 점수를 확인해 보겠습니다. 아직 점수가 낮지만 에포크를 한 번 실행하니 정확도가 향상되었습니다. 에포크가 횟수가 적으면 모델이 훈련 세트를 덜학습하게 되고 에포크호시수가 충분히 많으면 훈련세트를 환전히 학습할 것입니다. 바꾸어 말하면 적은 에포크 횟수 동안에 훈현한 모델은 훈련세트와 테스트 세트에 잘 맞지 않는 과소적합된 모델이 가능성이높고 반대로 많은 에포크 횟수 동안에 훈현한 모델은 훈련 세트에 너무잘 맞아 테스트 세트에는 오히려 점수가 나쁜 과대 적합된 모델일 가능성이 높습니다.  partial\_fit() 메서드만 사용해보곘습니다. 그러러면 전체클래스 레이븡을 메서드에 전달해주어야 합니다 이를 위해 np.unique() 함수로 train\_target에 있는 7개의 생선 목록을 만듭니다. 또 에포크마다 훈현 세트와 테스트 세트에 대한 점수를 기옥하기 위해 2개의 리스트를 준비합니다. 300번의 에포크 동안 훈현을 반복하여 진행합니다. 그래프로 그려보게 되면 데이터가 작기 때문에 아주 잘 드러나지는 않지만 백 번째 에포크 이후에는 훈현 세트와 테스트의 점수가 조금씩 벌어지고 있습니다. 이모델의 경우 백번째 에포크가 적절한 반복 횟수로 보입니다. |