|  |  |
| --- | --- |
| 교육 제목 | Normalization, data Split, OnehotEncording |
| 교육 일시 | 10월 26일(화) |
| 교육 장소 | 강의실 |
| **교육 내용** | |
|  | * Normalization(정규화)   데이터에 여러 특성이 들어있을 경우 서로 규격이 맞지 않는 경우가 있다. 이럴 경우 그 데이터를 그대로 학습시키게 되면 결과에 오차가 커지는 경우가 있다. 이럴때 정규화를 해준다. 정규화는 표준화와도 어쩔땐 비슷한 의미로 사용되기도 한다. 표준화는 결국 z-score로 값을 변환하는 것으로 전체 보유한 데이터를평균이 0이고 분산이 1이 되도록 만드는 것이다. 정규화의 경우에는 정의가 좀 복잡한데, 표준화를 모두 포함하여 전체적으로 두리뭉실하게 말할 때에는 정규화가 전체를 포함하기도 하는데 여기서는 Mean Noramalization만 설명하여 -1과 1사이에서 값을 가지고 샘플 평균이 0이 되는 샘플로 만드는 과정이 대표적이다. 정규화를 위한 함수를 간단히 만들어보자면  Def min\_max\_norm(x):  Return(x – 데이터 최소값) / (데이터 최대값 – 데이터 최소값)  Def standard\_norm(x) :  Return (x – 데이터 평균) / 데이터 표준편차   * Data Split   데이터 분할은 중요하다. 컴퓨터가 학습 할 수 있게 적절한 샘플을 취해 골고루 섞어 트레인 데이터셋과 테스트 데이터셋을 나눠야 학습망의 성취율이 증가한다. 파이썬의 내장 함수들을 이용해 자신이 직접 데이터 셋과 테스트 셋을 나눌수도 있지만 사이킷런 패키지는 이런 데이터 분할을 쉽게 해주는 역할을 하므로 데이터가 커지면 사이킷런 패키지의 train\_test\_split 패키지를 잘 활용하자.   * One hot Encording   모델링을 할 때 숫자가 아닌 독립변수 값을 쓸 수 없기 때문에, 범주형 독립변수는 적절히 처리해야 한다. 범주형 독립변수를 처리하는 가장 일반적인 방법은 더미변수로 변환하는것이다. 해당하는 값에 대하여 맞으면 1, 아니면 0을 부여한다. 즉 고유값에 해당되는 특징에는 1, 나머지는 0을 부여한다. Data split과 마찬가지로 사이킷런에서는 LabelEncoder 패키지를 지원하므로 잘 활용하자 LabelEncoder는 독립 변수가 아닌 종속 변수(라벨)에 대해 사용한다. 문자열이나 정수로된 라벨 값을 0~K-1 까지의 정수로 변환   * Callback   Callback 함수의 early Stoping을 활용하면 적절한 학습 중단점을 찾을수 있다. H5는 우리가 학습한 가중치까지 들어가 있는 저장 모델이다 . |
| 전체 내용 |