|  |  |
| --- | --- |
| 교육 제목 | ML Classification 과 Regression |
| 교육 일시 | 10월 15일(금) |
| 교육 장소 | 강의실 |
| **교육 내용** | |
|  | * 이진분류   머신러닝에서 여러 개의 종류(혹은 클래스(class)라고 부릅니다) 중 하나를 구별해 내는 문제를 분류(classification)라고 부릅니다. 2개의 클래스 중 하나를 고르는 문제를 이진 분류(binary classifiacation)라고 합니다. 여기서 클래스는 파이썬 프래그램의 클래스와는 다릅니다. 용어를 혼동하지 마세요.   * 사이킷런   사이킷런 은 각 특성의 리스트를 세로 방향으로 늘어뜨린 2차원 리스트가 기본이기 때문에 여기에 맞게 데이터를 정제 해줘야 한다.   * K –최근접 이웃 알고리즘   어떤 데이터에 대한 답을 구할 때 주위의 다른 데이터를 보고 다수를 차지하는 것을 정답으로 사용.   * 기본개념   특성 : feature, 열 등 다양한 이름으로 불리운다. 데이터를 표현하는 하나의 성질  머신러닝 알고리즘이 데이터에서 규칙을 찾는 과정을 '훈련’이라고 함. 사이킷런에서는 fit() 메서드가 역할  사이킷런에서는 정확도를 출력 0~1 사이의 값으로 출력함  정확도 =(정확히 맞힌 개수) / (전체 데이터 개수)   * 지도 학습   지도학습을 하기 위해서는 훈련 세트와 테스트 세트로 나누어야 함.  주의할 점은 훈련 세트와 테스트 세트를 나눌 때 샘플링 편향이 일어날 수 있으며 이렇게 되면 제대로 훈련이 되지 않을 수 있다. 그러면 정확도가 낮아지게 된다.   * 비지도 학습   타깃 데이터가 없음. 무엇을 예측하느냐 보단 데이터가 어떤 특징을 갖갖추었느냐 볼 때 유용함   * 훈련 세트   훈련세트는 모델을 훈련할 때 사용하는 데이터. 보통 훈련 세트가 클수록 좋음. 따라서 테스트 세트를 제외한 모든 데이터를 사용   * 테스트 세트   전체 데이터에서 20~30% 를 테스트 세트로 사용하는 경우가 많음. 전체 데이터가 아주 크다면 1%만 덜어내도 충분할 수 있다.   * 데이터 전처리   데이터 스케일링 : 데이터 특성이 각각 다르다면 정확도가 낮아질수 있기 때문에 각 데이터의 특성을 확인하고 스케일링이 필요하다 싶으면 스케일링 을 진행해 주어야 한다. (예: 부동산데이터에서 어떤 데이터는 35평으로 표기되어있고 어떤 데이터는 84^2M로 표기되어 있을경우)  데이터 전처리는 머신러닝 훈련 데이터를 주입하기 전에 가공하는 단계로 때로는 데이터 전처리에 많은 시간이 소모되기도 함. |
| 전체 내용 |