6/1 강의노트

머신러닝 이론

독립 변수 : 인풋 데이터

종속 변수 : 결과 데이터

Ex) 댓글이 긍정적인지 부정적인지 판별하는 모델의 경우 댓글은 독립 변수, 그것이 긍정적인지 부정적인지 판별한 것이 종속 변수

머신러닝 : 훈련 – 검증 – 테스트 3단계

머신러닝과 딥러닝의 차이

머신러닝은 어느정도 틀이 갖추어져 있음 (= 규격이 정해져 있다, 데이터를 거기에 끼워 맞춤), 딥러닝은 인공 신경망 구조를 코드로 짬 이외의 인간 개입이 없음

머신러닝은 단순 반복으로 데이터를 하나하나 해석하며, 인간의 개입이 잦음 (치와와와 머핀 특징의 차이를 직접 입력)

딥러닝은 결과만 보유하며, 데이터가 많을수록 머신러닝보다 성능이 좋음

정형 데이터 : 수치화된 데이터거나 텍스트라도 표 등으로 정리된 데이터

비정형 데이터 : 음성 데이터 등 일정한 규칙이 없는 데이터

데이터 분석이 머신러닝에서 비중이 매우 큼

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

과대적합 : 수집하는 데이터는 모든 케이스에 대한 데이터가 아님, 수집한 데이터에서만 지나칠 정도로 높은 성능을 보이면 해당 데이터가 세상의 모든 데이터로 인지하게 되는 부작용이 있어 수집 이외의 낯선 데이터에 대해서 성능이 떨어지게 됨 이를 방지하기 위하여 모델 튜닝이 가해짐

M = 1 과소적합으로 정확도가 너무 낮음

M = 3 이상적인 모델 (적합단계)

M = 9 훈련 데이터에 너무 맞추어져 있어 실제 데이터에서는 정확도가 오히려 떨어짐 (과대적합)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

이진 분류 모델 (사이킷 런) – 생선 분류하기

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

문자를 숫자로 변환하여(인코딩) 컴퓨터가 학습 가능한 형태로 가공함(LabelEncoder 클래스의 .fit\_transform)

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

데이터 개수가 많지 않고, 도미와 빙어의 개수가 유사하지 않아 실전에서 학습하기 좋은 데이터는 아님

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

.classes\_ : 인코딩한 클래스 확인 가능

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

.inverse\_transform : 인코딩한 값을 사람이 읽을 수 있는 형태로 디코딩