|  |  |
| --- | --- |
| 교육 제목 | **머신러닝 개요 / 간단한 분류 알고리즘 훈련** |
| 교육 일시 | 2021년 11월 18일 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 내용 | 2021.11.18 19일차   1. 컴퓨터는 데이터에서 배운다.    1. 데이터를 지식으로 바꾸는 지능적인 시스템 구축       1. 수많은 데이터를 가지고 머신러닝기법을 통해 효율적으로 지식을 추출하라.    2. 머신러닝의 세가지 종류       1. 지도학습 : 레이블된 데이터 / 직접 피드백 / 출력 및 미래 예측       2. 비지도 학습 : 레이블 및 타깃 없음 / 피드백 없음 / 데이터에서 숨겨진 구조 찾기       3. 강화 학습 : 결정 과정 / 보상 시스템 / 연속된 행동에서 학습    3. 지도학습으로 미래예측       1. 지도학습의 주요 목적은 레이블된 훈련데이터에서 모델을 학습하여 본적이 없는 미래데이터에 대해 예측하는 것.       2. 지도학습의 종류로는 분류(classification)과 회귀(regression)이 있다.    4. 강화 학습으로 반응형 문제 해결       1. 강화 학습은 머신러닝의 또 다른 종류로, 환경과 상호작용하여 시스템 성능을 향상하는 것이 목적임.       2. 강화학습의 피드백은 정답 레이블이나 값이 아님. 보상함수로 얼마나 행동이 좋은지를 측정한 값이며 에이전트는 환경과 상호작용하여 보상이 최대화되는 일련의 행동을 강화학습으로 함.       3. 탐험적인 시행착오 방식이나 신중하게 세운 계획을 사용함.       4. 대표적인 예로는 체스게임이 있음.    5. 비지도 학습으로 숨겨진 구조 발견       1. 지도학습에서는 모델을 훈련할때 사전에 옳은 답을 알고 있습니다. 강화 학습에서는 에이전트의 특정 행동을 보상하는 방법을 정의합니다. 비지도 학습에서는 레이블 되지않거나 구조를 할 수 없는 데이터를 다룹니다. 비지도 학습 기법을 사용하면 알려진 출력 값이나 보상 함수의 도움을 받지않고 의미있는 정보를 추출하기 위해 데이터 구조를 탐색할 수 있습니다.       2. 군집 :서브그룹 찾기          1. 군집은 사전정보없이 쌓여 있는 그룹 정보를 의미 있는 서브그룹 또는 클러스터로 조직하는 탐색적 데이터 분석기법입니다. 분석 과정에서 만든 각 클러스터는 어느정도 유사성을 공유하고 다른 클러스터와는 비슷하지않은 샘플 그룹을 형성합니다. 이따금 군집을 비지도 분류라고 하는 이유가 여기 있습니다.          2. 클러스터링은 정보를 조직화하고 데이터에서 의미 있는 관계를 유도하는 훌륭한 도구 입니다.       3. 차원 축소 : 데이터 압축          1. 비지도학습의 또다른 하위분야는 차원 축소입니다. 고차원의 데이터를 다루어야하는 경우는 흔합니다. 즉, 하나의 관측 샘플에 많은 측정 지표가 있습니다. 비지도 차원 축소는 잡음 데이터를 제거하기위해 특성 전처리단계에서 종종 적용하는 방법입니다. 차원축소는 관련있는 정보를 대부분 유지하면서 더 작은 차원을 가진 부분 공간으로 데이터를 압축합니다.    6. 기본 용어와 표기법 소개       1. 대부분의 데이터는 행렬 및 벡터로 표기한다.       2. 머신러닝 용어          1. 훈련 샘플 : 데이터셋을 나타내는 테이블의 행          2. 훈련 : 모델 피팅, 모수 모델의 경우 파라미터 추정          3. 특성 : 데이터 테이블이나 데이터 행렬의 열, 속성, 변수 등          4. 손실함수 : 종종 비용함수와 동의어로 사용함. 일부자료에서는 손실 함수를 하나의 데이터 포인트에 대해 측정한 손실로 사용하고, 비용 함수는 전체 데이터셋에 대해 계산한 손실로 사용합니다.    7. 머신러닝시스템 구축 로드맵       1. 전처리 : 특성 추출 및 스케일 조정 / 특성 선택 / 차원 축소 / 샘플링       2. 학습 : 학습데이터셋을 가지고 적절한 모델을 선택하여 학습을 시작한다.       3. 평가 : 테스트 데이터셋으로 예측을 해보고 평가해본다. 만약 결과가 좋지않을 경우, 이전 단계로 돌아가서 다시 학습을 시작한다.       4. 예측 : 검증 데이터셋을 가지고 예측함.    8. 전처리 : 데이터 형태 갖추기       1. 주어진 원본 데이터의 형태와 모습이 학습 알고리즘이 최적의 성능을 내기에 적합한 경우는 거의 없다.       2. 많은 머신러닝 알고리즘에서 최적의 성능을 내려면 선택된 특성이 같은 스케일을 가져야함.       3. 일반적으로 특성을 0 ~ 1 범위로 변환하거나, 평균이 0이고 단위분산을 가진 표준 정규분포로 변환하는 경우가 많음.       4. 일부 선택된 특성은 매우 상관관계가 높아 어느정도 중복된 정보를 가질 수 있음. 이때는 차원 축소 기법을 사용하여 특성을 저차원 부분 공간으로 압축함. 특성 공간의 차원을 축소하면 저장 공간이 덜 필요하고 학습 알고리즘을 더 빨리 실행이 가능함.       5. 데이터셋은 훈련 데이터셋과 테스트 데이터셋으로 나누어 별도로 보관한다.    9. 예측 모델 훈련과 선택       1. 현실적으로 가장 좋은 모델을 훈련하고 선택하기 위해 최소한 몇가지 알고리즘을 비교해야함.       2. 여러 모델을 비교하기전 성능을 측정할 지표를 결정해야하며, 분류에서 널리 사용되는 지표는 정확도임.    10. 모델을 평가하고 본적없는 샘플로 예측        1. 모델을 학습한 뒤에는 테스트 데이터셋을 이용하여 이번에 본적이 없는 데이터에서 얼마나 성능을 내는지 예측하여 일반화 오차를 예상합니다.        2. 다만 이는 특성 스케일 조정과 차원 축소 같은 단계에서 사용한 파라미터는 훈련 데이터셋만 사용하여 얻은 것임을 주목해야함. 나중에 동일한 파라미터를 테스트 데이터셋은 물론 새로운 모든 샘플을 변환하는데 사용해야합니다. 그렇지않으면 테스트 데이터셋에서 측정한 성능은 과도하게 낙관적인 결과가 됩니다. 2. 간단한 분류 알고리즘 훈련    1. 인공뉴런 : 초기 머신러닝의 간단한 역사       1. 인공뉴런 : 두 개의 클래스가 있는 이진 분류 작업          1. 두 클래스는 간단하게 1(양성클래스) 와 -1(음성 클래스)로 나타낸다.          2. 이후 입력값 x와 이에 상응하는 가중치 벡터 w의 선형 조합으로 결정함수를 정의합니다.          3. 최종 입력(net\_input) = wx, wx, wx, … wx          4. 이 특정 샘플의 최종입력이 사전에 정의된 결정함수안의 임계값보다 크면 클래스 1로 예측하고, 그렇지않으면 클래스 -1로 예측합니다. 퍼셉트론 알고리즘에서 결정함수는 단위 계단 함수를 변형한 것입니다.          5. 머신러닝 분야에서 음수 임계 값을 절편이라고 합니다.       2. 퍼셉트론 학습 규칙          1. 로젠블라트의 초기 퍼셉트론 학습 규칙             1. 가중치를 0 또는 랜덤한 값으로 초기화합니다.             2. 각 훈련 샘플 x에서 다음 작업을 합니다.   출력 값 y를 계산합니다.  가중치를 업데이트합니다.   * + - 1. 여기서 출력값은 앞서 정의한 단위계단함수로 예측한 클래스 레이블입니다.       2. 가중치의 변화량은 학습률 \* 타깃과 예측값 일치여부 \* 입력값으로 구성되어있고, 학습률은 보통 0 ~ 1사이의 실수입니다.       3. 만약 퍼셉트론이 예측한 값이 일치하면, 가중치는 업데이트되지않습니다.       4. 만약 퍼센트론이 예측란 값이 불일치하면, 가중치는 타깃 클래스 방향으로 이동됩니다.  1. 퍼셉트론 구현 3. 모델이 분류 그래프 그리기 |