|  |  |
| --- | --- |
| 교육 제목 | **머신러닝 K-means, DBSCAN** |
| 교육 일시 | 2021년 11월 24일 수요일 |
| 교육 장소 | 영우글로벌러닝 2층 |
| **교육 내용** | |
| 오전 | 1. 군집 2. <비지도 학습>    1. - 머신러닝의 한 종류로 훈련 데이터에 타깃이 따로 없다.    2. - 그러므로 스스로 유용한 무언가를 찾아 학습해야 한다.    3. - 군집, 차원 축소등이 이에 해당한다. 3. <군집>    1. - 비슷한 샘플끼리 그룹으로 묶는 작업    2. - 대표적인 비지도 학습 방법 중 하나이다.    3. - 군집 알고리즘에서 만든 그룹을 클러스터(Cluster)라고 부른다. 4. k-평균    1. - 처음에 랜덤하게 클러스터 중심을 정하고 클러스터를 만든다. 그 다음 클러스터의 중심을 이동하고, 다시 클러스터를 만드는 식으로 반복해서 최적의 클러스터를 구성하는 알고리즘    2. - 클러스터 중심 : k-평균 알고리즘이 만든 클러스터에 속한 샘플의 특성 평균값. 센트로이드라고도 부른다. 5. <엘보우 방법>    1. - 최적의 클러스터 개수를 정하는 방법 중 하나    2. - 이너셔 : 클러스터 중심과 샘플 사이 거리의 제곱 합    3. - 클러스터 개수에 따라 이너셔 감소가 꺾이는 지점이 적절한 클러스터의 개수 k가 될수 있다.    4. - LightGBM의 경우에는 최대 손실값을 가지는 노드를 중심으로 계속해서 분할하는 '리프 중심 트리 분할(leaf-wise)' 방식을 사용한다.    5. - 따라서 트리가 깊어지기위해 소요되는 시간과 메모리를 많이 절약할 수 있다.    6. - 다만 적은 데이터에 대한 과적합(overfitting)이 발생하기 쉽다 6. Clustering    1. 비지도학습 머신러닝 기법중 하나 7. 클러스터링 종류    1. Hard Clustering - 한 개체가 여러 군집에 속할수 없는 군집화 방법    2. Soft Clustering - 한 개체가 여러 군집에 속할수 있는 군집화 방법    3. Partitional Clustering - 전체 데이터의 영역을 특정 기준에 의해 동시에 구분하는 군집화 방법. 각 개체들은 사전에 정의된 개수의 군집 가운데 하나에 속하게 됨       1. 대표적으로 K-Mean 군집화    4. Hierarchical Clustering - 개체들을 가까운 집단부터 차근차근 묶어나가는 군집화 방법. \*덴드로그램을 생성함       1. 덴드로그램 → 계층적 군집에서 클러스터의 개수를 지정해주지 않아도 학습을 수행할 수 있는 것은 개체들이 결합되는 순서를 나타내는 트리 형태의 구조    5. Self-Organizing Map - 뉴럴넷 기반의 군집화 알고리즘    6. Spectral Clustering - 그래프 기반의 군집화 알고리즘 8. 클러스터링 평가 방법    1. Dunn Index 9. 클러스터링 알고리즘 종류    1. DBSCAN - Density-Based Spatial Clustering of Applications with Noise    2. KMeans 클러스터링은 개체들의 거리를 이용하여 클러스터를 나누는것에 반해, DBSCAN은 데이터 포인트가 세밀하게 몰려있는 밀도가 높은 부분을 군집화 하는 방식 10. Python Sklearn DBSCAN파라미터     1. eps → 데이터간의 최대거리     2. min\_samples → 중심객체에서 최대거리까지 포함되는 객체 개수 11. DBSCAN 작동 과정     1. 임의의 데이터를 선정하고 min\_samples에 해당되는 점을 중심객체로 선정     2. min\_samples조건을 충족하지못하면 외곽객체로 선정     3. 만약 겹치는 객체가 생기면 동일 군집으로 판단     4. 모든 객체가 조사될때까지 반복     5. 외곽, 중심객체에도 속하지않는 객체는 노이즈로 처리 12. 장점     1. 클러스터 개수 정의 불필요     2. 노이즈 처리에 강함 (노이즈객체를 따로 분류)     3. U자형, H형 모양을 띄는 데이터 분포도 잘 군집화처리가 가능 13. 단점     1. 많은 연산이 필요하여 결과처리에 속도가 느림     2. 2~3차원까지의 데이터는 연산속도가 그나마 받쳐주지만, 고차원의 데이터를 처리할때는 연산량이 급증하는 단점     3. 주변거리 하이퍼파라미터 e값을 정하기가 어려움 |
| 오후 | 1. 9장 - 웹 애플리케이션에 머신 러닝 모델 내장하기    1. 영화 리뷰 분류를 위한 모델 훈련하기    2. 학습된 사이킷런 추정기 저장    3. 데이터를 저장하기 위해 SQLite 데이터베이스 설정    4. 플라스크 웹 애플리케이션 개발    5. 첫 번째 플라스크 애플리케이션    6. 폼 검증과 화면 출력    7. 영화 리뷰 분류기를 웹 애플리케이션으로 만들기    8. 공개 서버에 웹 애플리케이션 배포    9. 영화 분류기 업데이트    10. 요약 |