|  |  |
| --- | --- |
| 교육 제목 | **머신러닝 KNN 모델** |
| 교육 일시 | 2021년 11월 22일 월요일 |
| 교육 장소 | 영우글로벌러닝 2층 |
| **교육 내용** | |
| 오전 | 1. K-최근접이웃(K-Nearest Neighbor, KNN) 2. 모델 개요 3. KNN은 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 k개 이웃의 정보로 새로운 데이터를 예측하는 방법론. 4. 만약 회귀(regression) 문제라면 이웃들 종속변수(y)의 평균이 예측값. 5. KNN을 모델을 별도로 구축하지 않는다는 뜻으로 게으른 모델(Lazy model) 6. Instance-based Learning 7. 데이터로부터 모델을 생성해 과업을 수행하는 Model-based learning과 대비되는 개념으로, 별도 모델 생성과정 없이 각각의 관측치(instance)만을 이용하여 분류/회귀 등 과업을 수행한다는 취지 8. KNN의 하이퍼파라메터(Hyper parameter)는 탐색할 이웃 수(k), 거리 측정 방법 두 가지입니다. k가 작을 경우 데이터의 지역적 특성을 지나치게 반영하게 됩니다(overfitting). 반대로 매우 클 경우 모델이 과하게 정규화되는 경향이 있습니다(underfitting). 9. 다만 KNN 알고리즘이 이러한 경계면을 직접 만드는 것은 절대 아니고, 새로운 데이터가 주어졌을 때 어느 범주로 분류되는지를 보기 좋게 시각화했다는 점에 주의하셔야 합니다. 10. 거리 지표 11. KNN은 거리 측정 방법에 따라 그 결과가 크게 달라진다 12. Euclidean Distance 13. 가장 흔히 사용하는 거리 척도입니다. 두 관측치 사이의 직선 최단거리를 의미합니다. 14. Manhattan Distance 15. A에서 B로 이동할 때 각 좌표축 방향으로만 이동할 경우에 계산되는 거리입니다. 16. Taxi cab Distance라고도 불립니다. 17. Correlation Distance 18. 데이터의 pearson correlation을 거리 척도로 직접 사용합니다. 개별 관측치 하나하나가 아니라 데이터 전체의 경향성을 비교하기 위한 척도입니다. 다시 말해 두 개 데이터 패턴의 유사도를 반영할 수 있습니다. |
| 오후 | 1. KNN 수행시 주의점    1. KNN 수행 전 반드시 변수를 정규화(Normalization)해 주어야 합니다.    2. 변수별로 평균과 분산을 일치시키는 정규화 작업을 반드시 KNN 적용 전 수행해 주어야 합니다. 2. KNN의 장단점    1. KNN은 학습데이터 내에 끼어있는 노이즈의 영향을 크게 받지 않으며 학습데이터 수가 많다면 꽤 효과적인 알고리즘이라고 합니다. 특히 마할라노비스 거리와 같이 데이터의 분산을 고려할 경우 매우 강건(robust)한 방법론으로 알려져 있습니다. 네이버, 카카오 등 현업에서도 KNN을 두루 사용하고 있는 것으로 전해집니다.    2. 아울러 k가 1인 1-NN의 오차 범위는 다음과 같다는 사실이 증명되어 있습니다. (Idealerr=주어진 데이터에 적합 가능한 가장 이상적인 모형의 오차) 바꿔 말해 1-NN에 한해서는 모델 성능을 어느 정도 보장할 수 있다는 이야기입니다.    3. 그러나 최적 이웃의 수(k)와 어떤 거리척도가 분석에 적합한지 불분명해 데이터 각각의 특성에 맞게 연구자가 임의로 선정해야 하는 단점이 있습니다.    4. 또 새로운 관측치와 각각의 학습 데이터 사이의 거리를 전부 측정해야 하므로 계산 시간이 오래 걸리는 한계점이 존재합니다. 3. KNN의 장단점 4. 장점    1. 단순하고 효율적이다.    2. 훈련 단계가 빠르다.    3. 기저 데이터 분포에 대한 가정을 하지 않는다.    4. 수치 기반 데이터 분류 작업에서 성능이 우수하다.    5. 특별한 훈련이 없기에 lazy learning(게으른 학습)이라고도 불린다. 5. 단점    1. 모델을 생성하지 않아 특징과 클래스간 관계를 이해하는데 제한적이다.    2. 적절한 k의 선택이 중요하다.    3. 데이터가 많아지면 분류 단계가 느리다.    4. 명목 특징 및 누락 데이터를 위한 추가 처리가 필요하다.    5. 더미코딩: 명목 특징을 수치 형식으로 변환하는 코딩    6. 수치 형식으로 변환하여 유클리디안 거리 공식에 대입한다. 6. KNN 자체가 데이터 분석에서 중요한 변수를 파악하는 것이 아니라 데이터 분석가가 직접 중요하다고 여기는 변수를 선정하고 알고리즘을 적용하기에 사전 분석 작업이 많이 요구된다. 7. 훈력 단계는 짧지만 예측 단계는 상대적으로 느리다. 8. KNN 사용방법 9. 각 column의 데이터마다 scale이 다를 것이기에 수치 형식의 데이터들의 값을 모두 0~1 사이로 변환시키는 정규화(Normalization) 과정을 거쳐야 한다. 10. 예를 들어 class A는 데이터 분포가 0~10, class B는 데이터 분포가 0~100000일 수도 있기에 이런 경우 각 변수들의 차이를 해석하기 어렵다. |