|  |  |
| --- | --- |
| 교육 제목 | **머신러닝 주성분 분석** |
| 교육 일시 | 2021년 11월 25일 목요일 |
| 교육 장소 | 영우글로벌러닝 2층 |
| **교육 내용** | |
| 오전 | 1. # 차원의 저주 - Curse of dimension    1. 입력된 데이터의 수보다 데이터의 차원이 더 큰 경우 발생하는 문제를 차원의 저주라 합니다.    2. 데이터(벡터)가 뿌려진 벡터 공간에서 분류 또는 예측하는 가장 적합한 함수를 찾는게 기계학습, 딥러닝 학습의 목표    3. 입력한 데이터의 양은 적고, 데이터의 차원이 커지게 된다면    4. 이때 벡터 공간의 차원이 무수히 커지고 데이터는 여기저기 흩뿌려져 있는 상황입니다.    5. 이 흩어진 벡터들을 분류 예측하는 함수의 모형은 복잡해지게 됩니다.    6. 즉, 모델의 복잡도가 증가하고 예측 성능이 낮아지게 됩니다.    7. 선형대수로 표현하자면 0으로 가득한 벡터로 채워진 분산 행렬(sparse matrix)의 형태일 것입니다.    8. 이를 위해서 사용하는 방법이 Feature selection (피쳐 선택), Feature Extraction(피쳐 추출) 입니다. 2. # 피쳐 선택과 피쳐 추출    1. 피쳐 선택은 가령 500개의 차원중 N개의 차원을 골라 내는 것이고, 반면 피쳐 추출은 500개의 차원을 N차원으로 압축하는 겁니다.    2. 피쳐 선택의 경우에는 다중공산성을 고려해 상관성이 높은 피쳐들을 소거해가는 방식으로 진행할 수 있고,    3. 피쳐 추출의 경우는 PCA, SVD, MF 와 같은 차원 축소 기법들을 활용해 벡터의 차원을 줄여나갈 수 있습니다.    4. 차원이 큰 것은 차원 축소 기법 활용, 3. # PCA는 차원 축소 기법이면서, 데이터 압축 기법이기도 하고, 노이즈 제거기법이기도 합니다. 왜?    1. 주성분 분석의 기본적인 개념은 차원이 큰 벡터에서 선형 독립하는 고유 벡터만을 남겨두고 차원 축소를 하게 됩니다.    2. 이때 상관성이 높은 독립 변수들을 N개의 선형 조합으로 만들며 변수의 개수를 요약, 압축해 내는 기법입니다.    3. 그리고 이 압축된 각각의 독립 변수들은 선형 독립, 즉 직교하며 낮은 상관성을 보이게 됩니다.    4. 가령 500차원의 벡터를 주성분 분석한다는 것은 각 차원의 분산을 최대로 갖는, 분포를 설명할 수 있는 대표축을 뽑는 과정이고,    5. 주성분 분석결과 나오는 매트릭스에서 PC1 은 각 칼럼에 대한 정보 설명력이 PC5~6에 비해 높습니다.    6. 이처럼 높은 주성분들만 선택하면서 정보 설명력이 낮은, 노이즈로 구성된 칼럼들은 배제하기 때문에 노이즈 제거 기법이라고 불리기도 합니다. 4. Q. LSA, LDA, SVD 설명    1. LDA 는 토픽모델링(Topic Modeling) 기법 중 하나인 잠재디리클레할당(Latent Dirichlet Allocation, LDA)와 이니셜이 같아서 헷갈리는데 SVD (행렬분해)를 자연어처리 토픽 모델링에 적용한게 잠재디리클레할당(Latent Dirichlet Allocation, LDA)고, 또 다른 LDA (Linear Discriminant Analysis)는 PCA 에서 확장된 차원 축소 기법입니다.    2. PCA 데이터의 차원을 줄이기 위해, 공분산 행렬에서 고유 벡터/고유값을 구하고 가장 분산이 큰 방향을 가진 고유벡터에 입력데이터를 선형변환하는 컨셉이라면, LDA는 지도 학습(supervised - learning)에서 적용하는 차원 축소 기법이자,입력 데이터의 클래스(정답) 를 최대한 분리할 수 있는 축을 찾는 기법입니다.    3. SVD는 정사각행렬이 아닌 m\*n 형태의 다양한 행렬을 분해하며, 이를 특이값 분해라 말합니다.    4. 이때 분해되는 행렬은 두 개의 직교 행렬과 하나의 대각행렬이며, 두 직교행렬에 담긴 벡터가 특이벡터입니다. |
| 오후 | 1. 피쳐(Feature)를 선택/가공 하는 총 3가지 접근 법    1. 피쳐 선택 (Feature Selection) : 불필요한 피쳐는 버린다. 가령 야외활동 여부를 파악하는데 교통량은 크게 영향을 미치지 않는다고 가정한다면, 해당 피쳐는 버리는거다. 보통 각각의 피쳐가 모델이 야외 활동 여부를 파악하는데 영향을 미치는지 미치지 않는지 알고 싶다면 상관계수 값을 통해 판단한다.    2. 피쳐 추출 (Feature Extraction) : 피쳐들을 선택하는 것이 아니라, 더 작은 차원으로 피쳐들을 맵핑하는 것이다. 현재 101차원의 데이터 테이블을 50차원으로 압축하는 과정을 거치는 것인데, 여기에 속한 기법이 오늘 살펴볼 차원 축소 기법들 PCA, LDA, SVD, NMF 가 이다.    3. 피쳐 생성 (Feature Engineering)       1. 이는 데이터 테이블에서 피쳐가 부족한 상황일 때 적용하는 기법으로,       2. 해당 데이터와 만들고자하는 머신러닝 모델의 기능 활용 목적에 따라 새로운 피쳐들을 생성해 내는 것이다.       3. 가령 우리가 야외활동 판별 여부를 알려주는 머신러닝 모델을 만들고자 하는데 해당 모델은 세렝게티 지역에 특화된 모델이라고 하자 (뜬금)       4. 그렇다면, 해당 칼럼 중에 야외 동물 출현 횟수 같은 피쳐도 생성해줘야 하는 것이다... 2. 정리,    1. 만약 내가 400차원 데이터를 k 차원으로 줄이고자 한다면       1. 입력 데이터의 공분산 행렬을 구한다.       2. 공분산 행렬을 고유값 분해 해서 고유 벡터와 고유값을 구한다 (앞서 말한 최고의 화살표찾기)       3. 고유값이 가장 큰 (화살표 길이가 가장 긴) k개의 고유 벡터를 추출       4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 입력 데이터들을 선형 변환 |