|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 학과 | 컴퓨터 전자시스템공학 | 학번 | 201904458 | 이름 | 이준용 |
| 구분 | 내용 | | | | |
| 학습 범위 | 기계학습 6장 비지도 학습  6.1 절 비지도 학습을 지도/준지도 학습과 비교  6.2 절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환  6.3 절 k-평균과 친밀도 전파 알고리즘  6.4 절 커널 밀도 추정과 가우시안 혼합 | | | | |
| 학습 내용 | **기계학습 6장 비지도 학습**  **6.1 절 비지도 학습을 지도/준지도 학습과 비교**   * 지도 학습 = 모든 훈련 샘플이 레이블 정보를 가짐   비지도 학습 = 모든 훈련 샘플이 레이블 정보를 가지지 않음  준지도 학습 = 지도+비지도   * 중요한 두 가지 사전 지식   1) 매니폴드 가정 = 모든 샘플은 매니폴드와 가까운 곳에 있다. 고차원 데이터로부터 저차원의 Locally Euclidian 를 구한다. 고차원 공간에 내재한 저차원 공간이므로 학습 데이터를 저차원 매니폴드 공간에 표현한다.  2) 매끄러움 가정 = 샘플은 어떤 요인에 의해 변화한다. 카메라의 위치, 조명 등 획득한 특징 공간에서 위치가 조금씩 바뀜.  **6.2 절 비지도 학습의 일반 연산으로 군집화, 밀도 추정, 공간 변환**  군집화 = 유사한 샘플을 모아 같은 그룹으로 묶는 일. Ex) 영상 분할, 맞춤 광고  밀도 추정 = 데이터로부터 확률분포를 추정하는 일. Ex) 분류, 생성 모델 구축  공간 변환 = 원래 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일(매니폴드) Ex) 데이터 가시화, 데이터 압축, 특징 추출.     * 1. **절 k-평균과 친밀도 전파 알고리즘** * k-평균 = x 개의 데이터 셋에 k 개의 군집단을 찾아내는 작업. 군집의 개수 k 는 주어지는 경우와 자동으로 찾아야 하는 경우가 있음. 군집화를 부류 발견 직업이라 부르기도 한다.      * K-평균은 샘플의 평균으로 군집 중심을 갱신   k-medoids 는 대표를 뽑아 뽑힌 대표로 군집 중심을 갱신한다(k-평균에 비해 잡음에 둔감).   * k-평균 알고리즘에서 초기 군집 중심이 달라지면 최종 결과가 달라진다. 다중 시작은 서로 다른 초기 군집 중심을 가지고 여러 번 수행한 다음, 가장 좋은 품질의 해를 취함(오차가 적은). * K-평균은 군집 크기 불균형, 군집 밀도 차이, 특이한 분포로 인해 한계가 발생한다.   EM 기초 = k-평균에서 훈련집합과 군집집합은 각각 입력단과 출력단에서 관찰 가능하다. K-평균은 z의 추정과 a 의추정을 번갈아 가면서 수행하는 EM 알고리즘이다.  가우시안으로 샘플의 소속 정보를 개선(E 단계) -> 샘플의 소속 정보로 가우시안 개선(M 단계) -> 가우시안으로 샘플의 소속 정보 개선(E 단계) -> 샘플의 소속 정보로 가우시안 개선(M 단계)… 이런 과정 반복.    친밀도 전파 알고리즘 책임 행렬 R과 가용 행렬 A라는 두 종류의 친밀도 행렬을 이용하여 군집화 군집 개수 k를 자동으로 알아냄 ⎝ 가까울수록 유사도가 증가하고 멀수록 유사도가 감소한다.  예) i와 k, k’라는 parameter가 있을 때 k가 k’보다 i에게 가까울 경우 k는 i에게 대표 데이터로서의 가산점을 부여하지만 i가 k’로부터 대표 데이터로서의 가산점을 받고 있을 경우 i는 굳이 k에게 대표 데이터 가산점을 부여하지 않는다. 받는 대표 가산점이 많을수록 다른 데이터에게 영향력을 행사하기 수월해진다.   * 자기 유사도 – 유사도의 최솟값(적은 군집), 중앙값(중간 군집), 최댓값(많은 군집) 중에서 선택 * 자가 친밀도 는 친밀도 전파 알고리즘에서 사용하는 식을 그대로 사용한다. 는 새로운 식   **6.4 절 커널 밀도 추정과 가우시안 혼합**  밀도 추정 문제 = 어떤 점 x에서 데이터가 발생할 확률, 확률분포 P(x)를 구하는 문제.  커널 밀도 추정에서 히스토그램 방법이 있는데 특정 공간을 칸의 집합으로 분할한 다음, 칸에 있는 샘플의 빈도를 세어 추정한다. 그러나 칸의 크기와 위치에 민감하고 매끄럽지 못하며 게단 모양을 띠는 확률밀도함수가 되는 단점이 존재. 커널 밀도 추정의 근본적인 문제는 샘플을 모두 저장하고 있어야 하는 메모리 기반 방법이고 데이터의 희소성이 존재할 뿐만 아니라 데이터가 낮은 차원의 경우로 국한하여 활용해야 한다.  그걸 보완시킨 방법이 가우시안 혼합이다. 데이터가 가우시안 분포를 따른다고 가정하고 평균 벡터와 공분산 행렬을 추정한다. 대부분 데이터가 하나의 가우시안으로 불충분하다.    EM 알고리즘을 이용할 경우  𝜃를 모르므로 난수로 넣고 풀이하게 됨  가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계) -> 가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계) -> 가우시안으로 샘플의 소속 정보 개선(E단계) -> 샘플의 소속정보로 가우시안 개선(M단계)… | | | | |