2025 1학년 인공지능 일반

한국디지털미디어고등학교
KOREA DIGITAL MEDIA HIGH SCHOOL

http://dimigo.biz

목차

학습목표

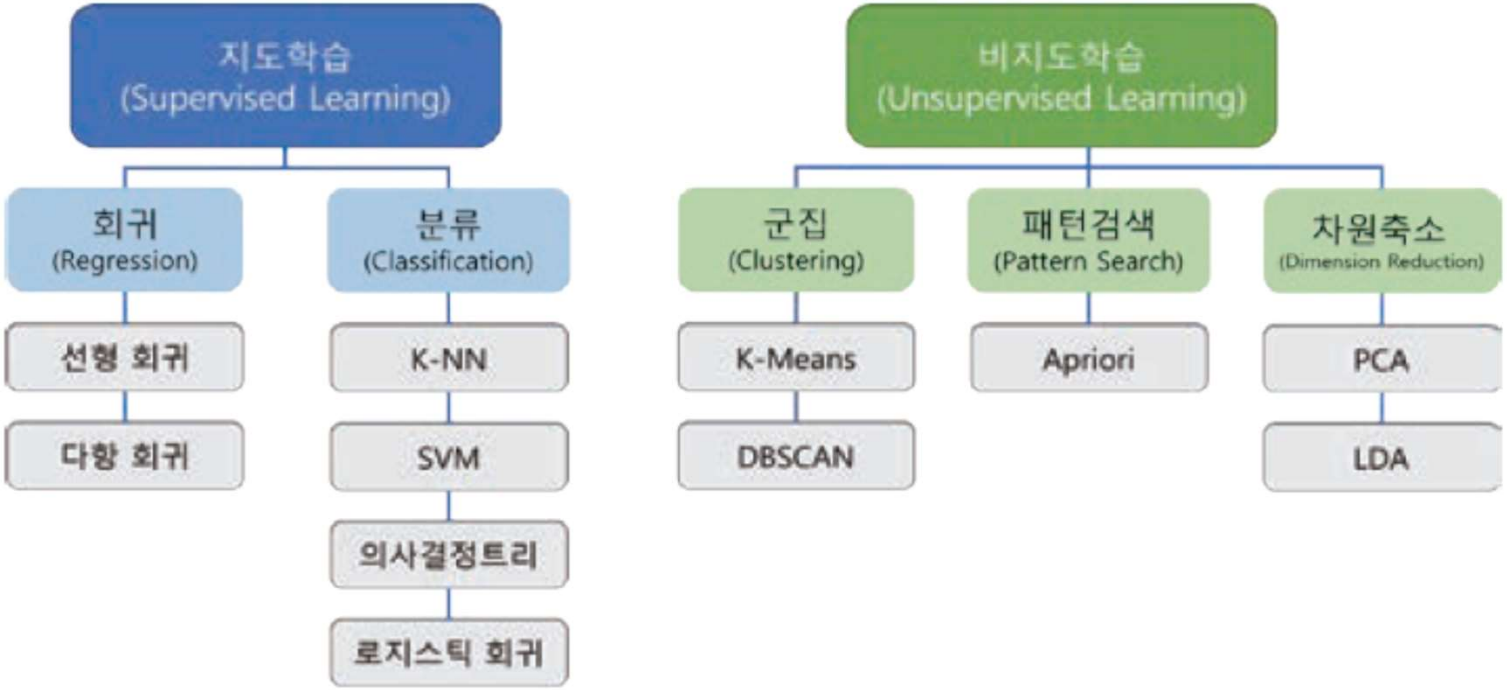
비지도 학습의 개념을 설명할 수 있다.

비지도 학습의 알고리즘의 종류를 파악할 수 있다.

목적에 맞는 비지도학습 알고리즘을 선정할 수 있다.

01. 지도 학습

■ 비지도 학습의 활용



01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석(clustering)

- 비지도학습의 일종으로 주어진 각 개체들의 유사성을 분석해서 높은 대상끼리 일반화된 그룹으로 분류하는 기법
- 규칙, 결과 없이 주어진 데이터들을 가장 잘 설명하는 그룹 또는 클러스터를 찾을 수 있는 방법으로 복잡하고 다양한 대상들을 이해하기 쉽게 구분
- 이상치에 민감하여 신뢰성과 타당성 검증이 어려우나 사전 정보없이 특정패턴, 속성을 파악하기 위한 효과적인 그룹 분류기법
- 유통, 서비스 등 업종 분야에서 VIP핵심 고객들을 군집화 하거나 마케팅 조서에서 실제 앱 이용자들을 더 잘 이해하기 위해 이용자 정보와 이용 패턴 데이터를 수집하여 고객 세그멘테이션을 군집 분석 알고리즘을 통해 진행될 수 있다.

■ 군집분석의 기본적인 가정

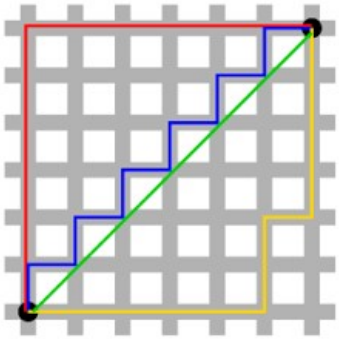
- 하나의 군집 내에 속한 개체들의 특성은 동일하다.
- 군집의 개수 또는 구조와 관계없이 개체 간의 거리를 기준으로 분류한다.
- 개별 군집의 특성은 군집에 속한 개체들의 평균값으로 나타낸다.

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석의 척도

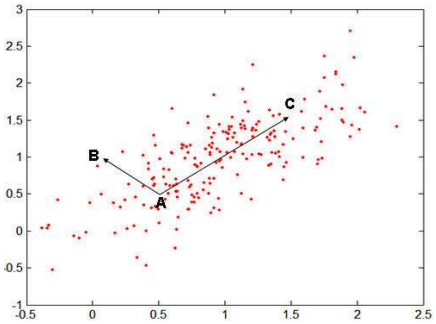
- 유클리드 거리(Euclidean Distance) : 두 관측치 간의 직선거리

- 맨하탄 거리(Manhattan Distance) : 절대값을 합산하는 방식. 가로지르지 않고 도달하는 최단거리



- 민코프스키 거리(Minkowski Distance) : 유클리드 거리와 맨하튼 거리 일반화. 민코우스키 공간:3차원 유클리드 공간에 시간이 결합한 4차원적 다양체

- 마할라노비스 거리(Mahalanobis Distance) : 변수의 분산과 상관성을 고려한 거리 측정 방식. 정규분포에서 특정값이 얼마나 평균에서 멀리 있는지.



- 자카드 거리 (Jaccard distance): 범주형 데이터 비유사성 측정 지표. 비교 대상 두 객체를 특징들의 집합으로 간주.

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석의 종류

- 병합방식

  – N군집에서 시작, 하나의 군집이 남을 때까지 순차적으로 유사한 군집들을 병합

- 분할방식

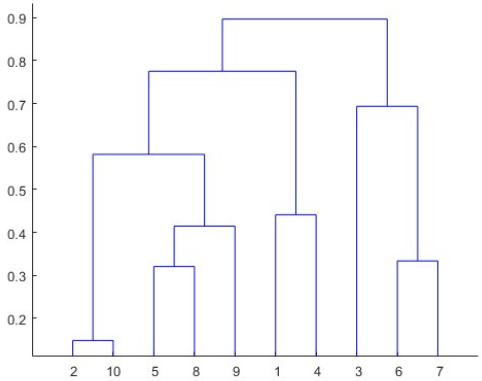
  – 전체 하나의 군집에서 시작, N군집으로 분할

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석의 종류

- 계층적 군집분석 – 군집수 명시가 필요하지 않고 덴드로그램을 통해 결과 표현을 시각화
  - 계층적병합 군집화 : N개의 군집으로 시작하고 가장 근접하고 유사한 두 개의 군집들이 1개 군집으로 병합, 가장 거리가 짧은 두 개의 군집들이 순차적으로 병합
  - 최단연결법 : 군집과 군집/데이터 간의 거리 중 최단거리 값을 거리로 산정.
  - 최장연결법 : 군집과 군집/데이터 간의 거리 중 최장거리 값을 거리로 산정.
  - 평균연결법 : 군집과 군집/데이터 간의 거리의 평균거리 값을 거리로 산정
  - Ward연결법 : 군집 내 편차들의 제곱합을 고려한 군집 내 거리를 기준으로 한다.

- 덴드로그램(dendrogram) : 개체들이 결합되는 순서를 나타내는 트리형태 구조



01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석의 종류

- 비계층적 군집분석(분할적 군집)
  - K-Means 군집분석 : 군집들 내부의 분산을 최소화하여 각각의 사례를 군집들 중 하나에 할당, 개별 유형의 특징 파악 및 좌표기반 군집 분석으로 대용량 데이터 처리/분산 처리에 용이
    - » K개(중심점 임의지정)의 초기 군집으로 시작
    - » 가장 가까운 중심을 가진 군집에 할당
    - » 2-3번 과정을 허용오차 이내 반복
    - » 군집 중심 재설정, 관찰치 변동시 군집 중심 재계산
  - 밀도 기반 클러스터링(DBSCAN) : 개체들의 밀도 계산을 기반으로 밀접하게 분포된 개체들끼리 그룹핑. 파라미터로 밀도계산 범위(epsilon)와 하나의 그룹으로 묶는 최소 개체수(minPts)가 필요
    - » 임의의 점p에서 epsilon 범위내에 포함된 점들의 개수가 minPts이상이면 p중심 군집 형성
    - » P와 동일한 군집에 있는 다른 q에 대해서 1번 단계 진행
    - » 더 이상 그룹에 포함할 개체가 없으면 해당군집이 아닌 다른 점 중심으로 1,2번 진행
    - » 만약 r 중심 epsilon 범위 내에 minPts이상 개체가 존재하지 않을 시 r은 아웃라이어로 처리

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 군집분석의 종류

- 비계층적 군집분석(분할적 군집)
  - 확률분포 기반 클러스터링(Gaussian Mixure Model) : 전체데이터의 확률분포가 가우시안 분포조합으로 이루어졌음을 가정하고 각 분포에 속할 확률이 높은 데이터들 간 군집을 형성하는 방법. 개별 데이터가 정규분포상에서 어떤 분포에 속할지 더 높은 확률로 배정된 부문으로 군집화

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 핵심 개념 이해

▪ K-평균 알고리즘

- k개의 중심점을 임의 위치로 잡고 중심점을 기준으로 가까이 있는 데이터를 확인한 뒤 그들과의 거리(유클리디안 거리의 제곱을 사용하여 계산)의 평균 지점으로 중심점을 이동하는 방식
- 가장 많이 활용하는 군집화 알고리즘이지만, 클러스터의 수를 나타내는 k를 직접 지정해야 하는 문제가 있음

▪ 엘보 방법

- 왜곡: 클러스터의 중심점과 클러스터 내의 데이터 거리 차이의 제곱값의합
- 클러스터의 개수 k의 변화에 따른 왜곡의 변화를 그래프로 그려보면 그래프가 꺾이는 지점인 엘보가 나타나는데, 그 지점의 k를 최적의 k로 선택

▪ 실루엣 분석

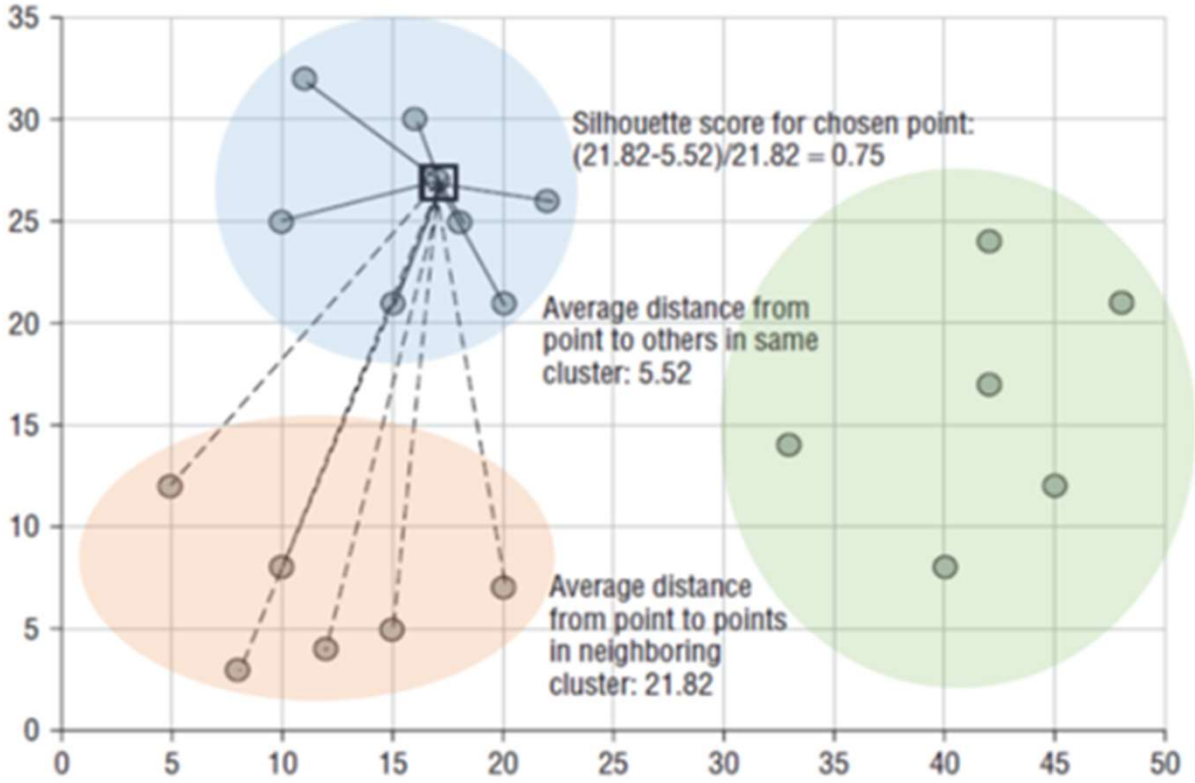
- 클러스터 내에 있는 데이터가 얼마나 조밀하게 모여있는지를 측정하는 그래프 도구
- 데이터 i가 해당 클러스터 내의 데이터와 얼마나 가까운가를 나타내는 클러스터 응집력 a(i)
- 가장 가까운 다른 클러스터 내의 데이터 와 얼마나 떨어져있는가를 나타내는 클러스터 분리도b(i)를 이용
- 실루엣 계수 s(i)를 계산
- -1에서 1 사이의 값을 가지며 1에 가까울수록 좋은 군집화를 의미

- $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$

01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 실루엣분석

- 각 군집간의 거리가 얼마나 효율적으로 분리 되어 있는지 나타냄.
- 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝다는 의미.



01. [K-평균 군집화 분석 + 그래프] 타깃 마케팅을 위한 소비자 군집 분석하기

■ 실루엣분석

- **실루엣 계수**
  - 전체 실루엣 계수의 평균값
  - 사이킷런의 silhouette_score() 값은 0~1사이의 값. 1에 가까울 수록 좋음.
  - 전체 실루엣 계수의 평균값과 더불어 개별 군집의 평균값의 편차가 작아야 함.
  - 개별군집의 실루엣 계수 평균값이 전체 실루엣 계수의 평균값에서 크게 벗어나지 않아야 함.
  - 전체 실루엣 계수의 평균값은 높지만 특정 군집의 실루엣 계수 평균값만 유난히 높고 다른 군집들의 실루엣 계수 평균값은 낮으면 좋은 군집화가 아님.
- sklearn.metrics.silhouette_samples(X, labels, metric='euclidean', **kwds) : 인자로 X feature 데이터 셋과 각 피처 데이터 셋이 속한 군집 레이블 값인 labels 데이터를 입력해주면 각 데이터 포인트의 실루엣 계수를 계산해 반환.
- sklearn.metrics.silhouette_score(X, labels, metric='euclidean', smaple_szie = None, **kwds) : 인자로 X feature 데이터 셋과 각 피처 데이터 셋이 속한 군집 레이블 값인 labels 데이터를 입력해 주면 전체 데이터의 실루엣 계수 값을 평균해 반환. 즉 np.mead(silhouette_samples()) 이다. 이 값이 높을수록 군집화가 어느정도 잘 되었다고 판단할 수 있다. (절대적인 기준이 될 수는 없다.)

THANK YOU

한국디지털미디어고등학교
KOREA DIGITAL MEDIA HIGH SCHOOL