AI기반 데이터 분석 및 AI Agent 개발 과정

『1과목 : 』
AI기반 데이터 분석

2025.09.22-10.02(9일, 62시간)

Prepared by DaeKyeong K

Ph.D.

삼성전자 구미사업부 G.제조팀

HS Global
HARDWARE SOFTWARE

목차

『1-10』
머신러닝 기반 데이터 분석-비지도

군집 분석

고객 세분화와 타깃 마케팅

학습 개요

학습목표

- 이 워크샵에서는 비지도 -자율학습 모델, K-means 클러스터링(K-Means Clustering)기법에 대해 알 수 있습니다.

눈높이 체크

- K-means 클러스터링(K-Means Clustering) 을 알고 계시나요?

1. 자율학습 개념

자율학습?

- **정의**
  - 비지도 학습(Unsupervised Learning) = 정답(레이블, Label)이 없는 데이터를 학습하는 방법
  - 입력 데이터만 가지고 내재된 패턴, 구조, 분포를 찾아냄
  - 지도학습(supervised learning)은 "입력 → 정답"을 맞추는 게 목표라면,
  - 비지도학습은 "데이터 안에 숨어 있는 규칙"을 발견하는 것이 목표

- **주요 특징**
  - 레이블이 없으므로 사람이 직접 정답을 주지 않아도 됨
  - 데이터의 유사성, 밀도, 차원 구조 등을 바탕으로 그룹화 또는 요약
  - 종종 탐색적 데이터 분석(EDA) 단계에서 먼저 활용

1. 자율학습 개념

대표 알고리즘

- 클러스터링 (Clustering)
  - 비슷한 데이터끼리 그룹화
  - 예: K-평균(K-Means), DBSCAN, 계층적 군집
- 차원 축소 (Dimensionality Reduction)
  - 고차원 데이터를 저차원으로 줄여서 시각화·분석 용이하게 함
  - 예: PCA(주성분분석), t-SNE, UMAP
- 연관 규칙 학습 (Association Rule Learning)
  - 데이터 내 항목들 간의 규칙 찾기
  - 예: 장바구니 분석(Market Basket Analysis) → "맥주를 산 사람은 과자를 함께 살 확률이 높다"
- 생성 모델 (Generative Models)
  - 데이터 분포를 학습해 새로운 데이터를 생성
  - 예: GAN(Generative Adversarial Network), 오토인코더(Autoencoder)

1. 자율학습 개념

활용 사례

- 마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)
  - 예를 들어, 온라인 쇼핑몰 고객 세분화의 경우에 ,
  - 온라인 쇼핑몰이 있는데, 고객 수는 많지만 모든 고객에게 같은 마케팅을 하고 있음
  - 결과: 마케팅 비용 대비 효과가 낮음 (광고 반응률↓, 재구매율↓)
  - 목표: 비슷한 성향의 고객끼리 그룹화(세분화) → 그룹별 맞춤형 마케팅 전략 수립
  - K-Means 알고리즘으로 고객을 k개의 그룹으로 나눔
  - Elbow Method, 실루엣 계수(Silhouette Score)로 최적 군집 수(k) 결정
  - 예: k=4로 결정 → 고객을 4개 그룹으로 분류

1. 자율학습 개념

활용 사례

- **마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)**

고객 그룹 해석

예시 결과:

1.가성비 추구형

1. 나이: 20대, 평균 구매 금액 적음, 할인 이벤트 반응 ↑
2. 전략: 쿠폰/프로모션 집중 제공

2.프리미엄 고객

1. 나이: 30~40대, 평균 구매 금액 높음, 브랜드 충성도 ↑
2. 전략: VIP 멤버십, 한정판 제품 추천

3.잠재 고객(이탈 위험)

1. 최근 구매 이력 없음, 방문 빈도↓
2. 전략: 리마인드 메일, 첫 구매 할인

4.충성 고객

1. 구매 빈도↑, 평균 구매액 중간, 장기 고객
2. 전략: 포인트 적립 강화, 추천 보상 프로그램

1. 자율학습 개념

활용 사례

- **마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)**

활용

•각 그룹에 맞춘 **타겟 마케팅 캠페인 실행**

•예:

- 가성비 그룹 → SNS 쿠폰 이벤트
- 프리미엄 그룹 → VIP 초대 행사
- 잠재 고객 → 재방문 유도 메일링
- 충성 고객 → 친구 추천 이벤트

1. 자율학습 개념

활용 사례

- 마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)의경우, 금융 사기 탐지 (Fraud Detection) 문제 상황에서
  - 은행·카드사에서는 하루에도 수백만 건의 결제 트랜잭션이 발생
  - 대부분 정상 거래지만, 일부는 도난 카드 사용, 계좌 해킹, 비정상 결제 같은 사기 거래
  - 문제: 사기 거래는 전체의 0.1%도 안 되는 경우가 많아, 지도학습 데이터(정답 라벨)가 부족
  - 목표: 비지도 이상 탐지 모델로 정상 패턴과 다른 의심 거래를 자동 탐지

1. 자율학습 개념

활용 사례

- **마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)의경우, 금융 사기 탐지 (Fraud Detection) 문제 상황에서**

이상 탐지 모델 적용

비지도 학습 기반 알고리즘 활용:

1.Isolation Forest

1. 의심되는 거래(Outlier)를 격리하는 방식
2. 정상 거래는 다수, 이상 거래는 소수라는 가정을 이용

2.Autoencoder (오토인코더)

1. 정상 거래 패턴을 학습
2. 재구성 오류(Reconstruction Error)가 큰 거래 → 이상치로 판단

3.DBSCAN (밀도 기반 클러스터링)

1. 정상 거래는 밀도가 높은 영역에 위치
2. 밀도에서 벗어난 거래 → 이상치

1. 자율학습 개념

활용 사례

- **마케팅 → 고객 세분화 (비슷한 성향의 고객 그룹 찾기)의경우, 금융 사기 탐지 (Fraud Detection) 문제 상황에서**

결과 해석

모델 실행 결과:

•전체 1,000,000건 중 500건을 **이상 거래 후보**로 탐지

•특징:

- 해외 결제인데 고객의 평소 패턴과 다름
- 평소 1만원 내외 쓰던 고객이 갑자기 300만원 결제
- 동일 계정으로 짧은 시간 내 다중 로그인 시도

활용

•보안팀이 **탐지된 의심 거래만 집중 조사**

•거래 실시간 차단 시스템에 연동 → 고객 피해 최소화

•장기적으로 탐지 결과를 라벨링하여 **지도학습(사기 vs 정상) 데이터셋으로 확장**

1. 자율학습 개념

활용 사례

- 이외 네트워크 침입 탐지 (Intrusion Detection)에서
  - 정상 패킷 패턴: 주기적 트래픽, 합리적 데이터 크기
  - 이상 패킷 패턴: 비정상 포트 접근, 갑작스러운 대량 트래픽 (DDoS), 새벽 시간대 다량 접속
  - 비지도 이상 탐지 모델로 네트워크 로그에서 정상/이상 행위 구분 가능

- 이미지 처리 → 차원 축소 후 이미지 압축, 데이터 시각화
- 자연어 처리 → 문서 군집화, 토픽 모델링

1. 자율학습 개념

지도학습과 비교

구분	지도학습	비지도학습
데이터	입력 + 정답(라벨)	입력만 있음
목표	정답을 예측	패턴, 구조 발견
예시	스팸메일 분류	문서 군집화

2. 클러스터링(군집)

클러스터링(군집) 분석?

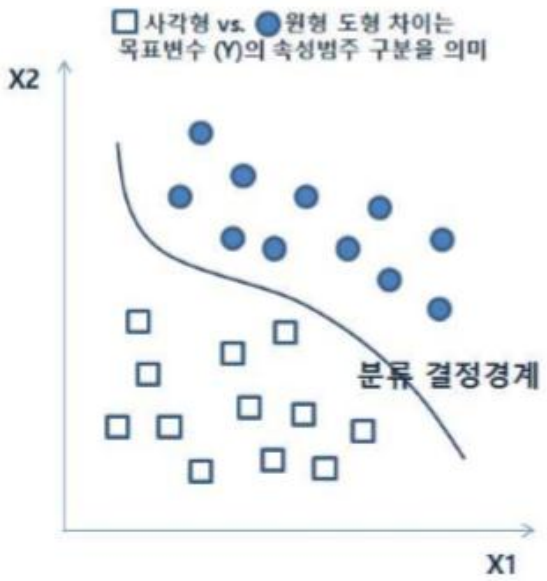
- 클러스터링(군집 분석)은 비지도 학습(unsupervised learning) 기법 중 하나
- 레이블(정답)이 없는 데이터를 유사한 특성을 가진 그룹(Cluster)으로 자동으로 묶는 과정
- 목표: 데이터의 숨겨진 구조, 패턴, 집단 분포를 찾아내는 것
- 주요 아이디어
  - 비슷한 데이터끼리는 같은 군집에, 다른 데이터끼리는 다른 군집에
  - “유사도(Similarity)” 또는 “거리(Distance)”를 기준으로 그룹화
  - 자주 쓰는 거리 척도: 유클리드 거리(Euclidean distance), 맨해튼 거리, 코사인 유사도 등

2. 클러스터링(군집)

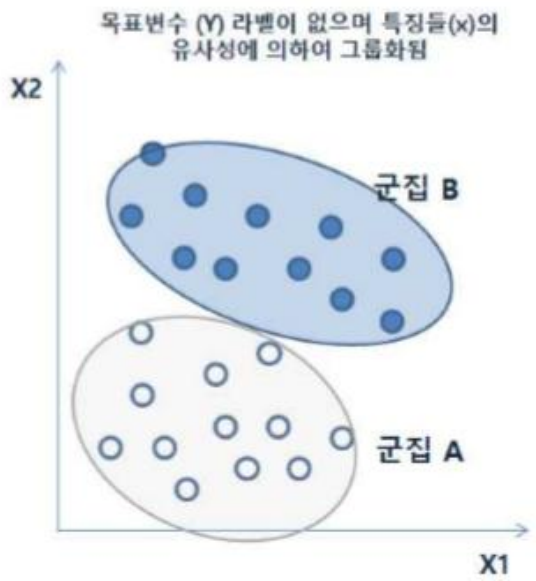
클러스터링(군집) 분석 원리

- 군집분석은 입력된 데이터가 어떤 형태로 그룹을 형성하는지가 분석의 핵심 목적이다. 따라서 군집분석에서는 입력된 데이터를 어떤 기준으로 그룹화 하는지가 첫 번째 질문이 되는데, 일반적으로는 각 데이터 간의 유사성을 기준으로 그룹화를 짓게 된다. 즉, 군집 내의 데이터는 서로 매우 유사하지만, 다른 군집과는 다르다는 원칙으로 그룹화를 진행하게 된다. 군집분석은 목적변수가 없는 상태에서 유사한 특성을 가진 개체끼리 그룹화를 한다.

(A) 분류 목적 분석기법



(B) 군집 분석



2. 클러스터링(군집)

클러스터링(군집) 분석 주요 활용 분야

- 군집분석은 데이터들에 존재하는 유사성 혹은 비유사성에 근거해서 패턴화를 하는 특성으로 인해, 매우 다양한 영역에서 활용할 수 있다. 군집분석이 주로 활용되는 분야에 대한 대표적인 예시는 다음과 같다.

1. 마케팅 등 분야에서의 고객 세분화(Segmentation)
2. 질병 및 환자 특성에 따른 유사 그룹화
3. 개체 유사성에 근거한 문서 분류
4. 디지털 이미지 인식 통한 사물 및 안면 인식
5. 금융 분야에서의 알려진 군집 이외의 사용 패턴 식별 (신용카드 사기, 보험료 과다 청구 등)
6. 공학 분야에서의 이상치 탐지 (제조 과정에서의 불량제품 자동 탐지, 통화 음질 개선을 위한 노이즈 구별 등)
7. 컴퓨터 네트워크에 비인가 된 침입 등의 비정상적 행위 탐지

2. 클러스터링(군집)

클러스터링(군집) 분석 주요 활용 분야

- 상기 몇 가지 예시에서 볼 수 있는 바와 같이, 주로 유사한 사람들을 그룹화 및 세분화하여 비즈니스에 활용하거나, 유사한 개체들을 묶어 필요한 정보를 압축하여 활용하거나, 유사성 기반의 그룹화를 응용하여 역으로 이상치(Outlier) 등의 특이점 혹은 비정상 패턴을 찾는 분야 등에 활용되고 있다. 이 외에도 매우 다양한 영역에서 응용 및 활용되고 있는 중요한 분석 기법이라고 할 수 있다.

2. 클러스터링(군집)

클러스터링(군집) 분석의 주요 종류

구분	기법	주요 내용
비계층 군집 (분할 기반 군집	K-평균(K-Means) 클러스터링	주어진 군집 수 k 에 대해서 군집 내 거리 제곱 합 의 합을 최소화하는 형태로 데이터 내의 개체들을 서로 다른 군집으로 그룹화하는 기법
	K-Medoids 클러스터링 혹은 (PAM : Partitioning Around Method)	K-평균 클러스터링의 보완한 기법으로서, 모든 형태 의 유사성(비유사성) 측도를 사용하며, 좌표평면상 임의의 점이 아닌 실제 데이터 세트 내의 값을 사 용하여 클러스터 중심을 정하므로 노이즈나 이상치 처리에 강건한 군집화 기법
	DBSCAN (Density Based Spatial Clustering of Application with Noise)	K-평균 기법이 K개의 평균과 각 데이터 점들 간의 거리를 계산하여 그룹화를 하는 반면, DBSCAN은 밀 도개념을 도입하여 일정한 밀도로 연결된 데이터집 합은 동일한 그룹으로 판정하여 노이즈 및 이상치 식별에 강한 군집화 기법
	자기 조직화 지도 (Self Organizing Map)	자율학습 목적의 머신러닝에 속하는 인공 신경망의 한 기법으로서 벡터 수량화 네트워크를 이용한 군 집화 기법
	Fuzzy 군집	K-평균 기법이 하나의 데이터 개체는 하나의 군집 에만 배타적으로 속하는 독점적 군집인데 반해, 퍼 지군집은 하나의 데이터 개체가 여러 개의 군집에 중복해서 속할 수 있도록 하는 중복 군집화 기법
계층적 군집	병합적(Agglomerative) 혹은 상 향식(Bottom-up) 군집화	모든 데이터 객체를 별개의 그룹으로 구성한 뒤, 단 하나의 그룹화가 될 때까지 각 그룹을 단계적으로 합쳐가는 계층적 군집기법
	분할식(Divisive) 혹은 하향식(Top-down) 군집화	모든 데이터 객체를 하나의 그룹으로 구성한 뒤, 각 데이터 점이 하나의 그룹으로 될 때까지 단계적으 로 분할에 가는 계층적 군집기법
확률 기반 군집	가우스 혼합 모형	EM (Expectation Maximization) 알고리즘 혹은 MCMC (Markov Chain Monte Carlo) 등의 알고리즘을 사용하 여 모수를 추정하는 확률 기반의 군집분석

2. 클러스터링(군집)

군집화 성능기준

- 군집화의 경우에는 분류문제와 달리 성능기준을 만들기 어렵다. 심지어는 원래 데이터가 어떻게 군집화되어 있었는지를 보여주는 정답(groundtruth)이 있는 경우도 마찬가지이다. 따라서 다양한 성능기준이 사용되고 있다. 다음의 군집화 성능기준의 예다.
- 내적 지표 (레이블 없음)
  - 실루엣 계수(Silhouette Score)
  - SSE (Sum of Squared Errors)
- 외적 지표 (레이블 있을 때)
  - Rand Index, Adjusted Rand Index (ARI)
  - NMI (Normalized Mutual Information)

2. 클러스터링(군집)

군집화 성능기준

- 일치행렬
  - 랜드지수를 구하려면 데이터가 원래 어떻게 군집화되어 있어야 하는지를 알려주는 정답(groundtruth)이 있어야 한다. N 개의 데이터 집합에서 i , j 두 개의 데이터를 선택하였을 때 그 두 데이터가 같은 군집에 속하면 1 다른 데이터에 속하면 0이라고 하자. 이 값을 N×N 행렬 T 로 나타내면 다음과 같다.

$$T_{ij} = \begin{cases} 1 & i\text{와 } j\text{가 같은 군집} \\ 0 & i\text{와 } j\text{가 다른 군집} \end{cases}$$

2. 클러스터링(군집)

군집화 성능기준

- 예를 들어 {0,1,2,3,4} 라는 5개의 데이터 집합에서 {0,1,2} 와 {3,4} 가 각각 같은 군집라면 행렬 T 는 다음과 같다.

```
import numpy as np

groundtruth = np.array([
    [1, 1, 1, 0, 0],
    [1, 1, 1, 0, 0],
    [1, 1, 1, 0, 0],
    [0, 0, 0, 1, 1],
    [0, 0, 0, 1, 1],
])
groundtruth
```

- 이제 군집화 결과를 같은 방법으로 행렬 C 로 표시하자. 만약 군집화가 정확하다면 이 행렬은 정답을 이용해서 만든 행렬과 거의 같은 값을 가져야 한다. 만약 군집화 결과 {0,1} 과 {2,3,4} 가 같은 군집라면 행렬 C 는 다음과 같다.

```
clustering = np.array([
    [1, 1, 0, 0, 0],
    [1, 1, 0, 0, 0],
    [0, 0, 1, 1, 1],
    [0, 0, 1, 1, 1],
    [0, 0, 1, 1, 1],
])
clustering
```

2. 클러스터링(군집)

군집화 성능기준

- 이 두 행렬의 모든 원소에 대해 값이 같으면 1 다르면 0으로 계산한 행렬을 일치행렬(incidence matrix)이라고 한다. 즉 데이터 집합에서 만들수 있는 모든 데이터 쌍에 대해 정답과 군집화 결과에서 동일한 값을 나타내면 1, 다르면 0이 된다.

$$R_{ij} = \begin{cases} 1 & \text{if } T_{ij} = C_{ij} \\ 0 & \text{if } T_{ij} \neq C_{ij} \end{cases}$$

- 즉, 원래 정답에서 1번 데이터와 2번 데이터가 같은(다른) 군집인데 군집화 결과에서도 같은(다른) 군집이라고 하면 R12=1 이다.
- 위 예제에서 일치행렬을 구하면 다음과 같다.

```
incidence = 1 * (groundtruth == clustering)  # 1*는 True/False를 숫자 1/0으로 바꾸기 위한 계산
incidence
>>
array([[1, 1, 0, 1, 1],
       [1, 1, 0, 1, 1],
       [0, 0, 1, 0, 0],
       [1, 1, 0, 1, 1],
       [1, 1, 0, 1, 1]])
```

2. 클러스터링(군집)

군집화 성능기준

◦ 이 일치 행렬은 두 데이터의 순서를 포함하므로 대칭행렬이다. 만약 데이터의 순서를 무시한다면 위 행렬에서 대각성분과 아래쪽 비대각 성분은 제외한 위쪽 비대각 성분만을 고려해야 한다. 위쪽 비대각 성분에서 1의 개수는 다음과 같아진다.

- a=T에서 같은 군집에 있고 C에서도 같은 군집에 있는 데이터 쌍의 수
- b=T에서 다른 군집에 있고 C에서도 다른 군집에 있는 데이터 쌍의 수

◦ 일치행렬 위쪽 비대각 성분에서 1의 개수=a+b

```
np.fill_diagonal(incidence, 0) # 대각성분 제외
a_plus_b = np.sum(incidence) / 2 # 대칭행렬이므로 절반만 센다.
a_plus_b

>>

6.0
```

2. 클러스터링(군집)

군집화 성능기준

- 랜드 지수(Rand Index, RI)는 클러스터링 결과의 정확도를 평가할 때 자주 사용하는 지표입니다.
- 정의
  - Rand Index는 클러스터링 결과와 실제 정답(ground truth) 간의 일치도를 측정하는 지표
  - 데이터 쌍(pair)을 기준으로 “같은 그룹에 속하는지/아닌지”를 비교
- 예시
  - 데이터 4개: A, B, C, D
  - 실제 그룹: {A,B}, {C,D}
  - 예측 그룹: {A,C}, {B,D}
  - 쌍을 나누어 비교하면…
  - 총 쌍 = 6개 (AB, AC, AD, BC, BD, CD)
  - a = 0 (같다고 맞춘 쌍 없음)
  - b = 2 (BD, AC는 다르다고도 예측했고 실제도 다름)
  - c+d = 나머지 불일치
  - 따라서 RI 값은 낮음 (≈ 0.33 정도).

2. 클러스터링(군집)

군집화 성능기준

- **계산 원리**
  - 데이터 두 개씩 짝을 지었을 때 (모든 가능한 pair),
  - a: 두 점이 같은 그룹에 속한다고 예측했고, 실제도 같음(True Positive)
  - b: 두 점이 다른 그룹에 속한다고 예측했고, 실제도 다름(True Negative)
  - c: 두 점이 같다고 예측했지만, 실제는 다름 (False Positive)
  - d: 두 점이 다르다고 예측했지만, 실제는 같음 (False Negative)

$$\text{Rand Index} = \frac{a+b}{{}_N C_2} \qquad RI = \frac{a+b}{a+b+c+d}$$

  - 정답과 예측이 일치한 비율
  - 값의 범위
  - $0 \le RI \le 1$
  - 1 → 클러스터링 결과가 완벽하게 정답과 일치
  - 0 → 완전히 불일치

2. 클러스터링(군집)

군집화 성능기준

- Rand Index는 우연에 의한 일치도를 고려하지 않음 → 그래서 Adjusted Rand Index (ARI, 조정 랜드 지수)가 더 많이 쓰임
- ARI는 -1 ~ 1 범위를 가지며, 0이면 무작위 수준, 1이면 완벽 일치

- shape[0], shape[1]를 이용하여 전체 행의 갯수와 열의 갯수를 반환

```
from scipy.special import comb
rand_index = a_plus_b / comb(incidence.shape[0], 2)
rand_index

>>

0.6
```

•**comb(n, 2)**
•n개 데이터에서 가능한 모든 쌍(pair)의 개수 = $\binom{n}{2} = \frac{n(n-1)}{2}$
•즉, 클러스터링 평가 시 비교해야 할 총 쌍의 수
•**a_plus_b**
•클러스터링 예측과 실제 라벨이 **일치하는 쌍의 개수**
•a = 같은 그룹이라고 맞춘 쌍 (True Positive)
•b = 다른 그룹이라고 맞춘 쌍 (True Negative)
•따라서 a_plus_b = a + b
•**Rand Index 계산식**

$$RI = \frac{a+b}{\binom{n}{2}}$$

•전체 쌍 중에서 "예측이 정답과 일치한 쌍의 비율"

2. 클러스터링(군집)

K-평균 군집화(K-Means Clustering)

- K-평균 군집화는 데이터를 K개의 그룹(Cluster)으로 나누는 알고리즘
- 각 그룹은 중심점(centroid)을 기준으로 형성됨
- 목적: 군집 내 데이터는 서로 가깝게, 군집 간 데이터는 멀리 떨어지게 만드는 것

2. 클러스터링(군집)

K-평균 군집화(K-Means Clustering)

- **알고리즘 동작 과정**
  - K 설정
  - 몇 개의 군집으로 나눌지(K)를 사용자가 먼저 지정
  - 초기 중심점(Centroid) 선택
  - 무작위로 K개의 중심점 선택
  - 할당 단계(Assignment Step)
  - 각 데이터 → 가장 가까운 중심점에 할당
  - 업데이트 단계(Update Step)
  - 각 군집에 속한 데이터의 평균을 계산 → 새로운 중심점으로 업데이트
  - 반복(Iteration)
  - 군집 할당과 중심 업데이트를 반복
  - 중심점이 더 이상 크게 움직이지 않거나 일정 반복 횟수에 도달하면 종료

2. 클러스터링(군집)

K-평균 군집화(K-Means Clustering)

목적 함수
K-평균은 다음 목적 함수(J)를 최소화하는 것이 목표

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2$$

- C_i: i번째 군집
- μ_i: i번째 군집의 중심(centroid)
- $|| x - \mu_i ||^2$: 데이터와 군집 중심 간의 거리 제곱

2. 클러스터링(군집)

K-평균 군집화(K-Means Clustering)

- **장단점**
  - 장점
    - 구현이 간단하고 빠름
    - 대규모 데이터셋에도 적용 가능
  - 단점
    - K 값을 미리 알아야 함
    - 이상치(outlier)에 민감
    - 구형(spherical) 형태의 군집에 잘 맞음 (비구형 데이터에는 성능 저하)

- **K값 결정 방법**
  - Elbow Method (엘보우 기법)
    - K에 따른 SSE(오차 제곱합) 변화를 확인
    - 감소 폭이 꺾이는 지점(elbow)을 최적 K로 선택
  - 실루엣 계수 (Silhouette Score)
    - 군집 응집도(cohesion)와 분리도(separation)를 종합 평가

2. 클러스터링(군집)

K-평균 군집화(K-Means Clustering)

- **활용 사례**
  - 마케팅: 고객 세분화 (구매 성향별 그룹)
  - 이미지 처리: 색상 압축 (이미지 색상 팔레트 단순화)
  - 문서 분석: 뉴스 기사 클러스터링 (주제별 분류)
  - 추천 시스템: 유사 취향 사용자 그룹화

2. 클러스터링(군집)

K-평균++ 알고리즘

- K-평균++(K-Means++) 알고리즘은 K-평균(K-Means) 군집화의 초기 중심 선택 방법을 개선한 알고리즘.
- K-평균의 한계
  - 기존 K-평균 알고리즘은
    - 무작위로 K개의 중심(centroid)을 선택
    - 군집 할당 → 중심 업데이트 반복
  - 문제점:
    - 초기 중심 선택에 따라 결과가 크게 달라짐
    - 중심이 잘못 잡히면 지역 최적해(local optimum)에 빠져 품질이 떨어짐

- K-평균++의 아이디어
  - 무작위로 선택하되, 멀리 떨어진 점들이 중심으로 선택될 확률을 높여 초기 중심을 더 "좋게" 선택

2. 클러스터링(군집)

K-평균++ 알고리즘

K-평균++ 알고리즘 절차

1.첫 번째 중심을 데이터 포인트 중 무작위로 선택
2.각 데이터 포인트 x에 대해, **가장 가까운 이미 선택된 중심까지의 거리** $D(x)$계산
3.$D\left(x\right)^2$을 확률 분포로 사용하여, 멀리 떨어진 점이 **중심으로 선택될 확률** ↑

$$P(x) = \frac{D\left(x\right)^2}{\sum_i D\left(x_i\right)^2}$$

1.이렇게 선택을 반복하여 총 K개의 초기 중심을 정함
2.이후는 일반 K-평균 알고리즘과 동일 (할당 → 중심 업데이트 반복)

2. 클러스터링(군집)

K-평균++ 알고리즘

- **장점**
  - 더 좋은 초기 중심 선택 → 군집 품질 향상
  - 반복 횟수와 계산 시간이 줄어듦
  - 지역 최적해 문제 완화

- **예시 비교**
  - K-평균 (랜덤 초기화)
    - 초기 중심이 몰리면 → 비효율적 클러스터링 발생
  - K-평균++
    - 중심이 데이터 공간 전체에 잘 퍼져 있어 → 안정적이고 좋은 성능

『1-10』
머신러닝 기반 데이터 분석-비지도

군집 분석

고객 세분화와 타깃 마케팅

학습 개요

학습목표

- 이 워크샵에서는 시장 내 다양한 고객군을 체계적으로 분류하고 분석할 수 있다
- 데이터 기반의 과학적 세분화 방법론을 실무에 적용할 수 있다
- ROI를 고려한 효과적인 타깃 선정과 마케팅 전략을 수립할 수 있다
- 브랜드 포지셔닝을 통해 경쟁우위를 확보하는 전략을 설계할 수 있다

눈높이 체크

- 실제 기업 사례 분석을 통한 전략 도출 (5개 이상)
- 개인 프로젝트: 특정 브랜드의 STP 전략 완성
- 포지셔닝 맵 작성 및 전략적 시사점 도출

1. 고객 세분화의 전략적 이해

세분화의 본질과 비즈니스 임팩트

- 세분화란 무엇인가?
- 전체 시장을 동질적인 특성을 가진 소그룹으로 나누어, 각 그룹에 최적화된 마케팅 전략을 수립하는 과정입니다.
- 왜 세분화가 필요한가? - 비즈니스 관점

◦ 마케팅 효율성 극대화

– 평균 마케팅 비용 30-50% 절감 효과
– 전환율 2-3배 향상 (개인화 메시지 효과)

◦ 고객 만족도 및 충성도 증대

– 맞춤형 서비스 제공으로 고객 이탈률 감소
– 고객생애가치(CLV) 증가 ※ 고객생애가치(CLV, Customer Lifetime Value)는 한 명의 고객이 기업과 관계를 맺는 전체 기간 동안 기업에 가져다줄 순이익의 총합

◦ 경쟁우위 확보

– 틈새시장 발굴을 통한 블루오션 창출
– 브랜드 차별화 포인트 명확화

1. 고객 세분화의 전략적 이해

세분화의 4가지 핵심 기준

- **인구통계적 세분화 (Demographic)**
  - 주요 변수들:
    - 연령: 베이비부머, 밀레니얼, Z세대 등
    - 성별: 젠더 마케팅의 진화
    - 소득: 고소득층, 중산층, 서민층
    - 교육수준: 대졸 이상, 고졸, 전문대 등
    - 직업: 전문직, 사무직, 서비스직 등
    - 가족구성: 1인 가구, 신혼부부, 육아세대, 실버세대

  - 실무 적용 예시: 현대카드의 연령별 세분화 전략
    - 20대: M카드 (혜택 중심, 간편결제)
    - 30-40대: THE BLACK (프리미엄, 라이프스타일)
    - 50대 이상: 골드카드 (안정성, 전통적 혜택)

1. 고객 세분화의 전략적 이해

지리적 세분화 (Geographic)

- **세분화 단위:**
  - 거시적: 국가, 대륙, 권역
  - 중범위: 시/도, 광역시, 도시 규모
  - 미시적: 구/군, 동/읍면, 상권

- **현대적 지리 세분화:**
  - O2O 마케팅: 위치 기반 실시간 프로모션 ※ O2O 마케팅은 Online to Offline의 약자
  - 날씨 연동: 기상 조건에 따른 상품 추천
  - 교통 패턴: 출퇴근 루트 기반 광고 노출

1. 고객 세분화의 전략적 이해

심리적 세분화 (Psychographic)

- **핵심 요소:**
  - 가치관: 환경 의식, 건강 지향, 효율성 추구
  - 라이프스타일: 워라밸 중시, 소확행, 미니멀 라이프
  - 성격: 외향적/내향적, 모험적/안정적
  - 관심사: 취미, 여가활동, 문화생활

- **VALS (Values and Lifestyles) 프레임워크 활용:**
  - 혁신가형 (Innovators) → 프리미엄 신제품 얼리어답터
  - 성취자형 (Achievers) → 명품, 지위재 소비
  - 체험자형 (Experiencers) → 트렌디한 제품, 감성 소비

1. 고객 세분화의 전략적 이해

행동적 세분화 (Behavioral)

- **구매 행동 기반:**
  - 구매 시점: 정기 구매, 계절성, 이벤트 연동
  - 구매 빈도: 헤비유저, 라이트유저, 비사용자
  - 브랜드 충성도: 충성고객, 브랜드 스위처, 가격 민감층
  - 혜택 추구: 품질 중시, 가격 중시, 편의성 중시

- **디지털 행동 패턴:**
  - 온라인 구매 여정: 검색 → 비교 → 구매 → 리뷰
  - 채널 선호도: 온라인 쇼핑몰, 오프라인 매장, 앱
  - 콘텐츠 소비: SNS 활동, 영상 시청, 리뷰 작성

2. 타깃 마케팅 전략 수립

타깃팅의 정의

- 세분화된 고객군 중에서 기업의 자원과 역량을 집중할 가장 매력적이고 접근 가능한 세그먼트를 선택하는 의사결정 과정

2. 타깃 마케팅 전략 수립

타깃팅 전략의 3가지 유형

- **무차별 마케팅 (Undifferentiated Marketing)**

◦ 특징:
– 전체 시장을 하나의 큰 세그먼트로 간주
– 표준화된 제품과 마케팅 믹스 사용
– 대량생산을 통한 비용 효율성 추구

◦ 성공 사례:코카콜라 초기 전략 (1950-1980년대)
– "Real Thing" 슬로건으로 전 세계 통일
– 표준화된 맛과 브랜딩
– 규모의 경제를 통한 비용 우위 확보

◦ 적용 조건:
– 시장 초기 단계이거나 고객 니즈가 동질적인 경우
– 자원이 제한적인 중소기업의 초기 진입 전략

2. 타깃 마케팅 전략 수립

타깃팅 전략의 3가지 유형

- **차별화 마케팅 (Differentiated Marketing)**

◦ 전략적 접근:
– 복수의 세그먼트를 타깃으로 설정
– 각 세그먼트별 차별화된 마케팅 믹스 개발
– 시장 점유율 확대와 위험 분산 효과

◦ 성공 사례 심화 분석: 현대자동차그룹의 다중 세그먼트 전략
1. 제네시스: 럭셔리 세그먼트
– 타깃: 고소득 전문직, 기업 임원
– 포지셔닝: "한국형 럭셔리의 새로운 정의"
2. 현대: 대중 실용 세그먼트
– 타깃: 실용성 중시하는 일반 소비자
– 포지셔닝: "합리적 선택의 완성"
3. 기아: 스타일리시 감성 세그먼트
– 타깃: 개성을 중시하는 젊은층
– 포지셔닝: "Movement that inspires"

2. 타깃 마케팅 전략 수립

타깃팅 전략의 3가지 유형

- **집중 마케팅 (Concentrated Marketing)**

◦ **전략 특징:**
– **하나의 세그먼트에 모든 자원 집중**
– **틈새시장에서의 전문성과 지배력 확보**
– **높은 수익성과 브랜드 충성도 달성**
◦ **니치 마케팅 성공 사례:**
◦ ※ 니치 마케팅(Niche Marketing) 은 전체 시장이 아닌, 특정하고 좁은 틈새시장(niche market)을 공략하는 마케팅 전략
– **오리온 초코파이 vs 롯데 가나초콜릿**
– **오리온 초코파이:**
• **타깃: 추억과 정서를 중시하는 전 연령층**
• **차별점: "정"이라는 한국적 감성 어필**
• **전략: 꾸준한 브랜드 스토리 구축**
– **롯데 가나초콜릿:**
• **타깃: 프리미엄 초콜릿을 추구하는 소비자**
• **차별점: 가나 카카오의 고급스러운 맛**
• **전략: 품질과 전통에 기반한 브랜드 구축**

2. 타깃 마케팅 전략 수립

타깃 세그먼트 평가 기준

● 시장 매력도 분석

◦ 규모와 성장성 (Market Size & Growth)

– 평가 매트릭스:

- 현재 시장 규모: 억 원 단위
- 연평균 성장률(CAGR): 3년 기준
- 시장 성숙도: 도입기/성장기/성숙기/쇠퇴기
- 고객 확장 가능성: 인접 세그먼트로의 확산 여부

◦ 수익성 (Profitability)

– 수익성 지표:

- 고객 평균 구매액 (Average Order Value)
- 고객 생애 가치 (Customer Lifetime Value)
- 단위당 마진율
- 고객 획득 비용 대비 수익 (CAC/LTV Ratio)

2. 타깃 마케팅 전략 수립

타깃 세그먼트 평가 기준

- **접근 가능성 (Accessibility)**

◦ **유통 채널 접근성**
– 기존 유통망 활용 가능성
– 새로운 채널 구축 비용과 시간
– 온라인/오프라인 채널 선호도

◦ **커뮤니케이션 접근성**
– 타깃 고객의 미디어 소비 패턴
– 효과적인 메시지 전달 채널 보유 여부
– 인플루언서/KOL 네트워크 활용 가능성

– ※ KOL 네트워크는 Key Opinion Leader Network의 약자로, 영향력 있는 전문가나 오피니언 리더(KOL, Key Opinion Leader)들을 조직적으로 연결하여 마케팅이나 브랜딩 활동에 활용하는 네트워크

2. 타깃 마케팅 전략 수립

타깃 세그먼트 평가 기준

- **경쟁 강도 분석**

◦ **직접 경쟁자 분석**
– 시장 내 주요 플레이어 수와 점유율
– 기존 브랜드들의 포지셔닝 현황
– 진입 장벽의 높낮이

◦ **대체재 위협**
– 고객 니즈를 충족하는 대안적 솔루션
– 기술 발전에 따른 시장 변화 가능성

3. 포지셔닝 전략의 실무 적용

포지셔닝의 전략적 의미

- **포지셔닝 정의**
  - 타깃 고객의 마음속에서 경쟁 브랜드와 구별되는 고유한 가치와 이미지를 구축하여, 구매 의사결정 시 최우선적으로 선택받을 수 있도록 하는 마케팅 전략

- **포지셔닝의 3가지 핵심 요소**
  - 차별화 (Differentiation): 경쟁자와 다른 고유한 특징
  - 관련성 (Relevance): 타깃 고객에게 의미 있는 가치
  - 신뢰성 (Credibility): 약속한 가치를 실제로 제공할 수 있는 능력

3. 포지셔닝 전략의 실무 적용

포지셔닝 전략 유형

- **제품 속성 기반 포지셔닝**
  - 볼보자동차: "안전성"
    - 핵심 메시지: "가족의 안전을 지키는 자동차"
    - 지원 증거: 충돌 테스트 최고 등급, 안전 기술 특허
    - 일관된 커뮤니케이션: 광고, 제품 개발, 서비스 모든 영역

- **편익/결과 기반 포지셔닝**
  - 나이키: "Just Do It"
    - 기능적 편익: 운동 성능 향상
    - 감정적 편익: 도전 정신, 성취감
    - 자아 표현적 편익: 능동적이고 역동적인 이미지

3. 포지셔닝 전략의 실무 적용

포지셔닝 전략 유형

- **사용 상황 기반 포지셔닝**
  - 레드불: "에너지 드링크"
  - 특정 상황: 집중력이 필요한 순간, 피로감을 느낄 때
  - 타깃 확장: 학생 → 직장인 → 운동선수 → 게이머

- **사용자 기반 포지셔닝**
  - 애플: "Think Different"
  - 타깃 사용자: 창의적이고 혁신적인 사고를 하는 사람들
  - 브랜드 페르소나: 도전적, 미니멀, 프리미엄

3. 포지셔닝 전략의 실무 적용

포지셔닝 맵 작성 및 활용

- **2차원 포지셔닝 맵 구성 요소**
  - 예시: 커피 브랜드 포지셔닝 맵
    - Y축: 프리미엄 정도 (High ↑)
    - X축: 접근성/편의성 (High →)

  - 사분면 분석:
    - 1사분면 (고프리미엄 + 고편의성): 스타벅스
    - 2사분면 (고프리미엄 + 저편의성): 블루보틀
    - 3사분면 (저프리미엄 + 저편의성): 동네 커피숍
    - 4사분면 (저프리미엄 + 고편의성): 맥도날드 맥카페

- **포지셔닝 맵 활용 방법**
  - 화이트 스페이스 발굴: 경쟁이 없는 빈 공간 식별
  - 경쟁자 분석: 직접/간접 경쟁자의 위치 파악
  - 이동 전략 수립: 현재 위치에서 목표 위치로의 이동 경로
  - 차별화 포인트 도출: 독특한 위치 선점 가능성 검토

3. 포지셔닝 전략의 실무 적용

포지셔닝 문장 작성법

- **기본 템플릿**
  - For [구체적인 타깃 고객]
  - Our [브랜드명] is a [제품 카테고리]
  - That provides [핵심 편익/가치]
  - Unlike [주요 경쟁자/대안],
  - We [고유한 차별점/증거].
- **실제 적용 예시**
  - 삼성 갤럭시 노트 시리즈:
  - For 생산성을 중시하는 비즈니스 프로페셔널들,
  - Our 갤럭시 노트 is a 프리미엄 스마트폰
  - That provides S펜을 활용한 창의적 작업 환경
  - Unlike 일반적인 터치폰들,
  - We 필기와 디지털이 완벽하게 결합된 새로운 경험을 제공합니다.

4. 마케팅 시나리오

트리거 기반 자동화 워크플로우

- **신규 가입 고객 온보딩:**
  - Day 1: 웰컴 메시지 + 브랜드 소개
  - Day 3: 첫 구매 할인 쿠폰 (15% 할인)
  - Day 7: 인기 상품 추천 + 고객 후기
  - Day 14: 미구매시 → 추가 할인 (20% 할인)
  - Day 30: 설문조사 + 맞춤 상품 추천

- **휴면 고객 활성화:**
  - 1단계: 그리움 메시지 + 컴백 할인
  - 2단계: 신상품 소식 + 무료배송
  - 3단계: VIP 혜택 제안 + 개인화 상품
  - 4단계: 최종 할인 + 한정 기간 설정

4. 마케팅 시나리오

행동 기반 실시간 트리거

- **장바구니 이탈 시나리오:**
  - 30분 후: 푸시 알림 "장바구니에 상품이 기다리고 있어요"
  - 2시간 후: 이메일 "놓치면 아까운 상품들"
  - 1일 후: SMS "마지막 기회! 5% 추가 할인"
  - 3일 후: 개인화된 대안 상품 추천

- **브라우징 패턴 분석:**
  - 특정 카테고리 3회 이상 방문 → 해당 카테고리 할인 정보
  - 고가 상품 반복 조회 → 분할 결제 옵션 안내
  - 리뷰 집중 읽기 → 구매 확신 강화 콘텐츠

5. 성공 사례 심화 분석

넷플릭스: 데이터 기반 초개인화 세분화

- **전통적 미디어와의 차별점**
  - 기존 TV 방송 vs 넷플릭스 전략

  - 기존 방송:
    - 시간대별 시청률 기반 편성
    - 연령/성별 등 거시적 세분화
    - Push 방식의 일방향 콘텐츠 제공

  - 넷플릭스:
    - 개인별 시청 패턴 분석
    - 76,000개 이상의 마이크로 세그먼트
    - AI 알고리즘 기반 개인화 추천

5. 성공 사례 심화 분석

넷플릭스: 데이터 기반 초개인화 세분화

- **세분화 방법론**

◦ 행동 데이터 수집
– 시청 완료율, 중단 지점, 재시청 횟수
– 검색 키워드, 클릭 패턴, 평점 데이터
– 시청 시간대, 기기별 사용 패턴

◦ 컨텐츠 DNA 분석
– 장르, 배우, 감독, 제작 국가
– 스토리 아크, 감정 곡선, 테마
– 시각적 요소, 음악, 분위기

◦ 개인화 추천 엔진
– 협업 필터링 + 콘텐츠 기반 필터링
– 딥러닝 모델을 통한 취향 예측
– A/B 테스트를 통한 지속적 최적화

5. 성공 사례 심화 분석

넷플릭스: 데이터 기반 초개인화 세분화

- **성과 지표**

  - 개인화 추천의 시청률: 80% 이상
  - 고객 유지율: 연간 93% (업계 평균 대비 20%p 높음)
  - 신규 콘텐츠 발굴 성공률: 70% 이상

5. 성공 사례 심화 분석

스타벅스: 위치 기반 맥락적 마케팅

- O2O 마케팅의 진화
  - 1세대: 단순 위치 알림
    - "근처 스타벅스에서 할인 받으세요"
  - 2세대: 상황 인식 마케팅
    - "출근길 7:30, 평소 아메리카노 주문 → 사이즈업 할인 제안"
  - 3세대: 예측적 개인화
    - "내일 비 예보 + 평소 실내 선호 → 따뜻한 음료와 실내 좌석 추천"

5. 성공 사례 심화 분석

스타벅스: 위치 기반 맥락적 마케팅

- **세분화 전략의 다층 구조**

◦ 지리적 세분화
- 반경 500m 이내 고객 인식
- 매장별 고객 유형 분석 (비즈니스/주거/쇼핑 지역)
- 교통 패턴과 연동한 타이밍 최적화

◦ 시간적 세분화
- 출근 시간대: 테이크아웃 + 빠른 서비스
- 점심 시간: 식사 메뉴 + 휴식 공간
- 저녁/주말: 디저트 + 사교 공간

◦ 행동적 세분화
- 주문 패턴: 단골 메뉴 vs 신메뉴 시도형
- 결제 방식: 앱 결제 vs 카드 결제
- 매장 이용: 테이크아웃 vs 매장 이용

5. 성공 사례 심화 분석

무신사: 패션 커머스의 세분화 혁신

- **기존 패션 쇼핑몰과의 차별점**

◦ **전통적 패션 쇼핑몰:**
– 브랜드별, 카테고리별 상품 진열
– 할인/이벤트 중심 프로모션
– 범용적 스타일 제안

◦ **무신사의 접근법:**
– 개인별 스타일 DNA 분석
– 코디 콘텐츠 기반 상품 발견
– 또래 집단의 트렌드 데이터 활용

5. 성공 사례 심화 분석

무신사: 패션 커머스의 세분화 혁신

- **세분화 방법론**

◦ 스타일 기반 세분화
– 스트릿/캐주얼/미니멀/빈티지 등 12개 스타일군
– 착용 사진 AI 분석을 통한 자동 분류
– 개인별 스타일 선호도 점수 산출

◦ 소셜 세분화
– 또래 집단별 인기 아이템 분석
– 인플루언서/스트릿 스냅 연계
– 지역별 패션 트렌드 차이 반영

◦ 구매 여정 기반 세분화
– 브라우징형 vs 목적 구매형
– 가격 민감도별 상품 추천
– 브랜드 충성도별 마케팅 메시지

5. 성공 사례 심화 분석

무신사: 패션 커머스의 세분화 혁신

- **성과 측정**

  - 개인화 추천 클릭률: 일반 상품 대비 3.2배
  - 평균 구매 단가: 연간 15% 증가
  - 고객 재방문율: 월 65% (패션 이커머스 평균 40%)

6. 실습

실습32 : 군집 분석

문제
[Mall_Customers.csv] 업로드
koreanize_matplotlib-0.1.1-py3-none-any.whl, NanumBarunGothic.ttf 업로드한 라이브러리를 설치하고 Matplotlib 한글 사용 환경을 설정 한 다음 나눔체로 한글을 표현해 줘. 이 데이터를 탐색해줘

6. 실습

실습33 : 군집 분석과 LDA 알고리즘

문제
[data.xlsx]파일 첨부
'사용 중인 언어','학습 희망하는 언어','사용 중인 데이터베이스', '학습 희망하는 데이터베이스', '사용 중인 클라우드 프랫폼','사용 중인 웹 프레임워크','사용 중인 IDE','사용 중인 OS'라는 말이 포함된 칼럼만 선택해서 데이터를 추출해 주세요.

THANK YOU.

앞으로의 엔지니어는 단순한 '코더'나 '기계 조작자'가 아니라 뇌-기계 인터페이스를 통해 지식과 능력을 즉각 확장하는 존재(뉴로-인터페이스: Neuro Interface)가 될 수 있습니다.

🎯 목표 달성을 위한 여정이 시작됩니다.
✨ 궁금한 점이 있으시면 언제든 문의해주세요!
🚀 함께 코더와 프롬프트 전문가로 성장해 나갑시다!