
(보안데이터분석) 연습문제_09

1. 다음 중 차원 감소(Dimensionality Reduction)에 대한 설명으로 가장 적절하지 않은 것은 무엇가?
 1. 고차원 데이터의 변수 수를 줄여 데이터 분석 및 시각화를 용이하게 한다.
 2. 데이터의 중요한 정보 손실을 최소화하면서 데이터의 복잡성을 줄이는 것을 목표로 한다.
 3. 주성분 분석(PCA)은 비선형적인 데이터의 차원 감소에 효과적인 대표적인 방법이다. ☒
 4. 과적합(Overfitting) 문제를 완화하고 모델의 일반화 성능을 향상시키는 데 도움을 줄 수 있다.
 5. 데이터 저장 공간을 절약하고 알고리즘의 계산 효율성을 높일 수 있다.
2. 다음 중 T-SNE 기법에 대한 설명으로 가장 적절하지 않은 것은 무엇가?
 1. 고차원 데이터의 각 점에 대해 주변 이웃과의 유사성을 확률적으로 표현한다.
 2. 저차원 공간에서 고차원에서의 이웃 관계를 최대한 보존하는 방식으로 데이터를 시각화한다.
 3. 주로 비선형적인 차원 감소에 사용되며, 데이터의 군집 구조를 시각적으로 잘 나타낸다.
 4. 데이터의 전역적인 구조(global structure)를 보존하는 데 초점을 맞춘 선형 차원 감소 기법이다. 비선형 차원 감소 기법이다. 또한, 데이터의 지역적인 구조(local structure) 가까운 이웃 관계를 보존하는 데 초점을 맞추며, 전역적인 구조를 항상 잘 보존하는 것은 아니다. ☒
 5. 초기 설정(Perplexity) 값에 따라 시각화 결과가 다르게 나타날 수 있다.
3. 다음 중 주성분 분석(PCA, Principal Component Analysis)에 대한 설명으로 가장 적절하지 않은 것은 무엇인가?
 1. 고차원 데이터의 분산을 최대한 보존하는 새로운 직교 기저(basis)를 찾아 데이터를 변환한다.
 2. 변수 간의 상관관계를 이용하여 데이터의 차원을 축소하고 주요한 특징을 추출한다.
 3. 비지도 학습 방법의 하나로, 데이터에 대한 사전 정보 없이 데이터 자체의 구조를 분석한다.
 4. 추출된 주성분들은 서로 독립적이며, 뒤에 오는 주성분일수록 데이터의 분산을 더 많이 설명한다. ☒

뒤에 오는 주성분은 상대적으로 앞의 주성분 보다 적은 분산을 설명하므로 정보량이 적다고 볼 수 있다.

5. 데이터 시각화, 노이즈 제거, 특징 추출 등 다양한 목적으로 활용될 수 있다.

4. 다음 중 요인분석에 대한 설명으로 가장 적절하지 않은 것은 무엇가?

1. 여러 변수들 간의 상관관계를 분석하여 잠재적인 공통 요인을 찾아내는 통계적 방법이다. 이면에 존재하는 잠재적인 요인(latent factor)을 밝히는 것을 목표
2. 변수들의 수를 줄여 데이터의 구조를 단순화하고 이해하기 쉽게 만든다.
3. 주성분 분석과 마찬가지로 변수들의 분산을 최대한 보존하는 것을 목표로 한다. ☒

주성분 분석은 변수들의 총 분산을 최대한 설명하는 주성분을 추출하는 반면, 요인 분석은 변수들 간의 공통 분산을 설명하는 요인을 추출하는 데 초점
PCA는 요인을 추출하는 방법 중 가장 일반적인 방법이며 최대우도법 등 다른 방법도 사용한다.

4. 설문조사 문항 분석이나 시장 세분화 연구 등에 활용될 수 있다.

5. 직교 회전 방법을 사용할 경우 추출된 요인들은 서로 독립적이다.

직교 회전(예: 베리맥스, 쿼티맥스)을 사용하는 경우에는 요인 간의 상관관계를 0으로 만들어 요인들을 독립적으로 해석할 수 있도록 유도





사각 회전(예: 오블리민, 프로맥스)을 사용하는 경우에는 요인들 간의 상관관계를 허용하고, 현실 세계의 복잡한 관계를 더 잘 반영하고자 함.


5. 다음 중 확인적 요인 분석(Confirmatory Factor Analysis, CFA)과 탐색적 요인 분석(Exploratory Factor Analysis, EFA)의 주요 차이점으로 가장 적절하지 않은 것은 무엇인가?

1. CFA는 사전에 설정한 요인 구조를 검증하는 반면, EFA는 데이터로부터 요인 구조를 탐색한다.
2. CFA는 분석 전에 요인의 수와 각 변수가 어떤 요인에 속할 것이라는 가설을 설정하지만, EFA는 이러한 사전 가설 없이 분석을 시작한다.
3. CFA는 모형 적합도 지수를 통해 이론적 모형과 실제 데이터의 부합 정도를 평가하지만, EFA는 주로 요인 적재치를 해석하여 요인을 추출한다.
4. CFA는 주로 새로운 척도 개발 단계에서 변수들의 잠재적인 구조를 파악하기 위해 사용되는 반면, EFA는 기존 이론을 바탕으로 개발된 척도의 타당성을 검증하는 데 사용된다. ☒
5. CFA는 변수와 요인 간의 관계를 명확하게 지정하여 분석하지만, EFA는 모든 변수가 모든 요인에 영향을 미칠 가능성을 열어두고 분석한다.

6. 다음 중 탐색적 요인 분석(EFA)의 일반적인 과정으로 가장 적절하지 않은 것을 고르시오.

1. 분석 목적 설정 및 변수 선정
2. 상관행렬 계산 및 요인 추출 방법 결정
3. 요인 수 결정 및 초기 요인 추출

4. 추출된 요인들의 해석 용이성을 높이기 위한 요인 회전(Factor Rotation) 실시
5. 이론적 배경을 바탕으로 사전에 결정된 특정 요인 구조의 적합도 평가 
7. 탐색적 요인 분석 시 고려해야 할 주요 개념 중 다음 설명에 해당하는 것은 무엇인가?
 " 각 변수가 추출된 공통 요인들에 의해 설명되는 분산의 비율을 나타내며, 1에 가까울수록 해당 변수가 요인들에 의해 잘 설명된다고 해석할 수 있습니다. "
 1. 요인의 추출법 (Factor Extraction Method)
 2. 요인의 판별 (Factor Interpretation)
 3. 공통성 (Communality) 
 4. 교차 부하량 (Cross-loading)
 5. 요인 회전 (Factor Rotation)
8. 다음 중 탐색적 요인 분석의 일반적인 절차와 관련하여 가장 적절하지 않은 설명은 무엇인가?
 1. **샘플(데이터)의 수:** 분석의 안정성을 위해 일반적으로 변수 수의 5배 이상, 최소 100개 이상의 샘플 크기가 권장된다. 일반적으로는 200 ~ 300 개 이상
 2. **정규성 검정:** 요인 분석 자체가 정규성을 엄격하게 요구하는 것은 아니지만, 일부 요인 추출 방법 및 통계적 검정의 전제 조건이 될 수 있다.
정규분포성
 3. **요인분석 적절성 지표:** KM0(Kaiser-Meyer-Olkin) 지수와 Bartlett의 구형성 검정 등을 통해 변수 간 상관관계가 요인 분석을 수행하기에 적절한지 평가한다.
KM0(Kaiser-Meyer-Olkin test) 0.7 이상이 보통, Bartlett $p < 0.05$ 이하 (상관행렬이 단위행렬이 아님)
 4. **요인의 추출법:** 주성분 분석(Principal Component Analysis)은 변수들의 총 분산을 설명하는 요인을 추출하는 데 초점을 맞춘 방법으로, 탐색적 요인 분석에서 흔히 사용된다.
 5. **요인의 판별:** 추출된 요인들의 의미를 명확하게 해석하기 위해, 각 요인에 높은 부하량을 보이는 변수들의 고유한 개념이나 특징을 중심으로 이름을 부여하는 과정은 생략해도 무방하다. 
9. 탐색적 요인 분석에서 요인 적재량(Factor Loading)에 대한 설명으로 가장 적절한 것은 무엇인가?
 1. 추출된 요인들의 총 분산을 나타내는 값으로, 고유값(Eigenvalue)과 동일한 의미를 가진다.
 2. 각 변수가 특정 요인과 얼마나 강하게 관련되어 있는지를 나타내는 상관 계수이다. 
 3. 요인 분석 모델의 적합도를 평가하는 지표 중 하나로, 0과 1 사이의 값을 가진다.

4. 요인 회전 후 요인 구조의 해석을 어렵게 만드는 변수들의 분산을 의미한다.
 5. 분석에 포함된 변수들의 수와 추출된 요인의 수를 곱한 값이다.
10. 요인 분석에서 적절한 요인 수를 결정하는 방법으로 가장 적절하지 않은 것은 무엇인가?
1. **고유값 기준 (Eigenvalue Criterion):** 고유값이 1보다 큰 요인만을 추출하는 방법으로, Kaiser의 규칙이라고도 한다.
 2. **스크리 그림 (Scree Plot):** 고유값을 크기 순서대로 나열한 그림에서 기울기가 완만해지는 지점 직전까지의 요인 수를 선택하는 방법이다.
elbow 기법
 3. **총 분산 설명량 (Percentage of Variance Explained):** 추출된 요인들이 설명하는 누적 분산 비율이 특정 기준(예: 50% 또는 60%)에 도달할 때까지 요인 수를 결정하는 방법이다.
 4. **이론적 근거 (Theoretical Basis):** 연구자가 기존 이론이나 선행 연구를 바탕으로 예상되는 요인 수를 미리 결정하는 방법으로, 탐색적 요인 분석에서 가장 중요하게 고려된다. 
 5. **평행 분석 (Parallel Analysis):** 실제 데이터의 고유값과 동일한 크기의 무작위 데이터에서 얻은 고유값을 비교하여, 실제 데이터의 고유값이 무작위 데이터의 고유값보다 큰 요인 수만 추출하는 방법이다.