4장 차원축소

데이터의 차원 → 변수의 개수로 결정.
↳ 데이터 마이닝 알고리즘을 효율적으로 수행하기 위해선 변수 개수 축소 필요.

파일럿 / 프로토 타입 단계의 일부로 모델을 사용하기 전에 시행.

차원축소의 접근법.

① 범주를 제거하거나 결합하기 위해서 주어진 자료와 관련된 특정 분야의 지식을 도입.

② 변수 간 중복되는 정보를 검출하기 위해서 자료요약을 시행
↳ 발견 시 변수 결합 / 삭제.

③ 범주형 변수 → 수치형 변수 등 자료변환 실시.

④ 주성분 분석(PCA)과 같은 자동화된 차원축소 기술 사용.

1. 서론.

· 데이터 준비 단계 에서는 변수 변환 등으로 변수의 개수가 과도하게 증가할 수 있음.

· 이런 경우 변수 간의 과도한 상관관계가 생기거나 결과변수와 관련없는 변수가 포함될 수 있음. ⇒ 과적합유발.

· 과적합 이외에도 지도학습 알고리즘에서 별도의 계산문제가 생길 수도 있음.

· 모델에 변수가 많을 경우 데이터 수집 비용 역시 증가
↳ 모델의 차원: 모델에 의해 사용된 독립변수 / 입력변수의 개수.

2. 차원의 저주.

차원이 증가할수록 가능한 선택지는 기하급수적으로 증가.

⇒ 데이터의 패턴과 구조분석이 불가능.
↳ 정확도의 희생을 최소화하면서 차원을 축소하는 것이 매우 중요.

3. 실질적인 고려사항.

데이터 탐색의 첫 단계에서 특정 변수들이 주어진 과제에 적합한지 확인하는 과정 필요.

어떤 변수들이 가장 중요하고, 어떤 것이 쓸모없는가.
어떤 변수가 오차가 많이 생길 것인가.
분석을 계속할 때 측정이 가능한가? 비용은?
결과값이 나오기전에 측정이 가능한가?
↳ 경매에서의 입찰 횟수 같이 경매가 끝나기 전엔 알 수 없는 것들.

등을 고려

4. 데이터 요약.

데이터 요약을 통해 데이터를 더 잘 이해하고 필요없는 변수 식별 가능.

(평균(mean), 최소, 최대값, 최빈값, 중앙값(median) 등의 요약통계량.
  취합 / 피벗 테이블) → 등으로 확인가능.

5. 상관분석.

상관계수를 살펴보는 상관분석을 통해 중복되는 정보를 가진 변수들을 식별 가능.

6. 범주형 변수의 범주 개수 축소.

범주의 개수가 많은 범주형 변수가 예측변수인 경우 다수의 가변수로 변환 ⇒ 변수의 개수가 크게 늘어남.

관측치의 개수가 적은 범주들은 다른 범주와 합치기 좋은 후보. ↳ 유사하거나 가까운 범주를 합쳐서 해결.

7. 범주형 변수에서 수치형 변수로의 변환.

구간 변수 → 수치형 변수로
↳ 여러 개의 가변수가 필요없는 수치형 변수로의 변환.

8. 주성분 분석.(PCA; Principal Component Analysis)

변수들의 수가 클 때, 차원축소에 유용한 방법

데이터가 같은 스케일로 측정되고 상관관계가 높은 측정치들을 포함할 때 특히 유용.
↳ 이런 경우, 변수들을 가중선형결합 (weighted linear combination)으로 재표현 하여 원래 변수가 가지고 있는 정보를 대부분 표현. 소수의 (3개정도) 변수로 재표현.

PCA는 양적 변수(quantitive variable)에 사용되는 기법.
↳ 범주형 변수의 경우에는 대응분석과 같은 다른 기법이 더 적합.
(Correspondence Analysis)

(ex) 두 데이터 X, Y의 평균이 각각 106.88, 42.67이고 공분산 행렬 S가

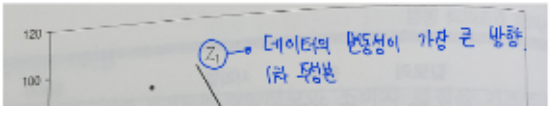
$S = \begin{bmatrix} 379.63 & -188.68 \\ -188.68 & 197.32 \end{bmatrix}$ 이면 두 변수의 상관관계는 $-0.69 = \frac{-188.68}{\sqrt{(379.63)(197.32)}}$
↳ 강한 음의 상관관계.

⇒ 두 변수의 전체 변동 중 69%는 공유하고 있는 변동이다. (한 변수의 변동이 다른 변수의 변동에서도 나타날)

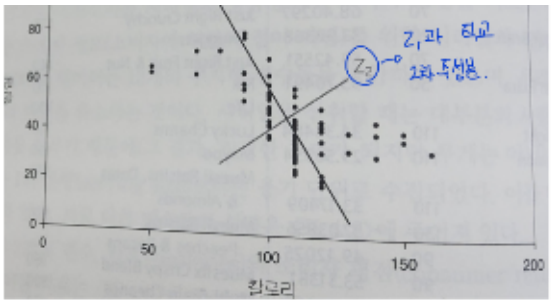
↳ 전체 변동에 각 변수가 미치는 기여정도를 최대한 활용하면서 이러한 관계를 변수의 수 축소에 사용 가능한가?
⇒ PCA는 두 변수에 있는 모든 정보를 포함하지는 못해도 대부분은 포함될 수 있는 두 변수의 선형결합을 찾는 것.
↳ 두 변수의 변동.

총 변동 = 두 변수의 분산의 합 = 379.63 + 197.32 = 577. → 변수 X가 총변동 중 66%(= 379.63/577)을 설명하고 있음.

차원 축소를 위해 한 변수를 제거하면 최소 34%의 총변동이 줄어들게 됨. ↳ 정보상실.
↳ 두 변수의 선형결합으로 생성된 새로운 변수에 두 변수 사이의 총 변동을 재분배한다면 총변동의 큰 부분을 설명할 수 있는 새로운 변수 한개만 유지 가능.



Z_1 값의 분산은 최대 ⇒ 1차 주성분 (직선과 점들의 수직거리 제곱합 최소)
↳ 주성분의



Z_2 값은 두 번째로 큰 변동성을 갖지만 Z_1과 상관관계는 0 (당연.)

Z_1, Z_2 두 직선을 이용하여 새로운 좌표를 구성. (회전행렬 이용)
↳ 가중치 행렬.

(변수 Z_1, Z_2는 원래 변수에서 평균을 뺀 값이고 평균은 0.
두 변수 Z_1, Z_2의 분산의 합은 원래 변수의 분산의 합과 동일.

$Z_i = a_{i,1}(X_1 - \bar{X}_1) + a_{i,2}(X_2 - \bar{X}_2) + \cdots +$
↳ 각 가중치 (행렬)

* 데이터의 정규화.

원래 변수들이 각 주성분에 얼마나 기여하는지를 알아보기 위해 가중치를 분석.

특정 변수들의 분산이 지나치게 크다면 주성분들이 그 변수에 큰 영향을 받게 됨.
↳ PCA이전에 데이터 정규화 필요. (원래 변수를 분산이 1인 표준화된 변수로 대체)

정규화를 통해 모든 변수들이 변동성 관점에서 동등한 중요도를 갖게 함.

(측정 단위가 공통적이고 변수의 스케일이 변수의 중요성을 나타낼 때는 정규화가 필요하지 않지만
변수들이 다른 단위로 측정되어 변수 간의 변동성을 비교하는 방법이 불분명하거나 스케일이 중요하지 않은 경우엔 정규화가 바람직.

• 분류와 예측을 위한 주성분 사용.

예측 변수로 사용될 변수의 차원축소.

- 학습데이터를 이용하여 예측 변수들에 대해 PCA 실행
- 검증 셋에서는 학습 셋의 주성분 점수 이용.

* 비선형적인 예측점수는 잃을 수 있음.