3장 데이터 시각화

데이터 시각화 ~ 데이터 셋의 다차원성을 이해하는데 도움.

여기서는 어떤 시각화 기법이 어떤 특징을 가진 데이터에 더 적합한 지에 대해 배움.

1. 데이터 시각화 용도.

- 데이터를 효과적으로 탐색하기에 용이.
- 주로 전처리 단계에서 사용.

① 데이터 셋에서 틀린 수치, 결측치, 중복 행, 중복 열 등을 쉽게 찾을 수 있음.

② 변수 도출과 선택에 용이. —
③ 데이터 축소 과정에서 카테고리화 하는데 도움.
- 어떤 변수를 포함시킬 것인지.
- 어떤 변수가 불필요한 변수인지.
- 데이터 구간은 적절한가
- 수치형 변수의 구간화가 필요한가

④ 데이터 수집 비용이 크다면 시각화를 통해 어떤 변수나 측정치가 더 필요한 지 알 수 있음.

2. 기본차트: 막대 차트, 선 그래프, 산점도.

동시에 하나 혹은 두 개의 데이터 열을 표시하여 데이터 탐색을 도움.

데이터의 구조, 변수의 양과 유형, 절대값의 크기와 유형 파악에 도움.

데이터 분석의 목적과 데이터에 대한 지식에 따라 어떤 차트를 선택하는가 결정.

막대차트 - 평균, 개수, 비율과 같은 단일 통계치를 그룹별로 비교하는데 유용.

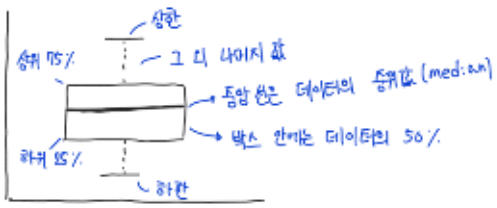
선 그래프 - 주로 시계열을 보여주기 위해 사용.
↳ 시간 프레임의 크기는 예측과제의 규모와 데이터의 속성에 따라 달라짐.

산점도 - 두 수치형 변수 간의 정보 등록이나 군집 발견과 같은 연관성을 밝히는데 도움. (비지도 학습에서)

* 분포도: 박스플롯과 히스토그램. - 수치형 변수의 전반적인 분포를 표시.

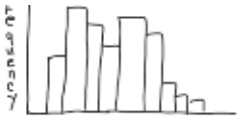
- 막대차트는 단일 변량을 표시하지만, 두 차트는 수치형 변수의 전체 분포를 보여줌.
- 데이터 마이닝 방법과 변수변환을 결정하기 위한 지도학습에 유용.
↳ 편향된 수치형 변수는 정규분포를 가정한 분석 (ex 선형 회귀, 판별분석)을 적용하려면 반드시 변환 필요. (ex 로그스케일)

· 박스플롯



분포도를 통해 예측변수의 잠재적 유효성을 판별하고 두 분포사이의 중복되지 않는 구역을 찾아낼 수 있는 데이터 마이닝 기법을 선택하는데 도움.

· 히스토그램 - 모든 X값의 출현 빈도.



x축, y축 2가지 차원뿐.

• 기본 차트와 플로팅의 주된 약점은 오직 2가지 변수만을 나타낼 수 있기 때문에 다차원적 정보를 밝혀낼 수 없다는 것.
↳ 데이터는 보통 다변량이고 분석 기법도 거기에 초점.

◦ 히트맵 : 상관관계와 결측치 시각화
↳ 수치형 데이터를 그래픽으로 나타내는 차트.
상관관계표의 시각화) 에 유용.
결측치의 시각화

색상차를 알기 어렵기 때문에.
히트맵은 큰 숫자 값을 검토하는데 유용하지만 정밀한 디스플레이는 불가능
상관관계나 결측치 처럼 한 눈에 정보를 파악하면 좋은 데이터에 유용. (큰 범위에서)
↳ 결측의 양/정도를 시각화.
(어디서, 어떤 변수가 얼마나 결측되었는가) ↦ 이를 통해 결측치 처리 방법 결정에 용이.

3. 다차원 시각화.

기본차트에 색상, 크기, 여러 패널을 적용하거나 다양한 기능을 구현하여 더 풍부한 정보 표현 가능 → 여러 변수들을 함께 관찰 가능. → 복잡한 정보를 효율적으로 표현. | 데이터를 더 높은 차원으로 표현하는 게 아니라 정보를 더 이해하기 쉽게 해줌
↳ 기본차트는 최대 2차원인데 반해

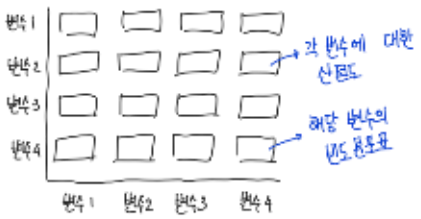
• 속성 변수추가.

차트에 더 많은 변수를 포함하기 위해선 변수의 형태 고려 필요.

- 범주형 정보 - 색조, 모양, 다중 패널이 효과적
- 수치정보 - 색상강도 조절, 크기 변화 등.
- 시간정보 - 애니메이션을 사용하면 효과적

• 산점도 매트릭스 - 다중 패널 산점도를 이용하는 특별한 차트.

↳ 수치형 변수들 간의 연관성 분석, 아웃라이어 탐지, 군집식별과 같은 비지도학습에 유용.
(지도 학습에서는 변수선택과 변수변환을 도움.



애니메이션 등은 시간흐름에 따른 변화를 보여주기는 좋으나 데이터 탐색에는 효율적이지 못함.

◦ 차트조절 - 스케일 조절, 팁 처리 계층구조, 확대축소, 필터링.

전처리 과정 ~ 모델이 효과적으로 학습할 수 있도록 변수 선택, 차원 변수 생성 등을 수행.

(수치스케일 변경
수치형 변수 범주화
범주형 변수의 변수 재조정 등)

- 스케일 조절 : 변수 간의 관계 부각 가능.
  (현재 스케일에서는 파악하기 힘든 관계들을 파악 가능.
  어느 한 특에 밀집한 데이터를 효율적으로 분석 가능.

- 집계와 계층.
  시간 단위에서 년,월,일 등으로 변경하거나 계층에 따라 집계.

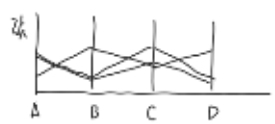
- 확대축소와 패닝.
  패턴이나 아웃라이어를 발견하려면 필요.
  다른 특정 영역을 탐지하여 새로운 조건, 변수, 함수 모델 등을 구성 가능.

- 필터링.
  특정 관측들을 제거하고 데이터 보기 or 다른 데이터에 의해 만들어진 노이즈 제거 가능.

• 주석선과 데이터 레이블.
  패턴과 아웃라이어 발견에 도움.
  (주석선 : 참고사항 제공, 패턴의 형태를 쉽게 파악.
  인-플롯 레이블 : 아웃라이어와 관심 탐색에 효과적.

• 대용량 데이터 셋으로 스케일 업.
  관측 수가 많을 때, 산점도 같이 개별 관측치를 나타내는 차트는 비효율적.
  박스플롯과 같이 집계차트 이외에 다른의 대안 존재.
  ① 샘플링 : 무작위 샘플 추출 후 차트 생성
  ② 표시 사이즈 축소.
  ③ 표시 색의 투명도 사용.
  ④ 데이터를 구분하여 서브셋 생성 (다중패널 구성)
  ⑤ 집계를 사용
  ⑥ 지터링을 사용 (적은 양의 노이즈를 추가하여 개별 표시를 쉽게 이동)

• 다변량 플롯 : 평행좌표 플랏.
  평행좌표플랏 : 각 변수를 위한 수직축 생성. 개별 수치들은 이 수직축들을 서로 연결.



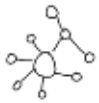
  관측에 유용한 변수 탐색 가능. 어떤 변수들을 그룹화할 지 분석 가능
  비지도학습에 유용. 군집, 아웃라이어, 변수들 간의 특성정보를 가져다줌.

• 대화형 시각화.
  대화형 시각화 (차트를 변경하는 것이 쉽고, 빠르며, 이전 상태로 복귀가능 / 관련있는 멀티차트들이 쉽게 결합되고, 한 화면에 디스플레이 가능. / 어떤 두 개의 차트가 서로 연결되어있으면 한 쪽의 변경이 다른 쪽에도 반영.) 이런 성질을 가져야함.
  정적인 차트생성기 (ex 엑셀) 에 비해 데이터 탐색을 쉽고 빠르게 할 수 있음.
  ↳ 이런 변경을 하려면 차트를 여러 개 생성하거나 새로운 변수 생성 필요.

5. 특화된 시각화

특정 정보 (ex 계층, 네트워크, 공간)은 기존 차트로 묘현이 힘들기 때문에 특화된 차트 필요.

- 네트워크 그래프 - 노드, 링크로 구성



연관성 규칙을 탐색하는데 효과적.

- 계층데이터의 시각화 : 트리맵.

  계층구조, 대규모 데이터 탐색 - 트리맵이 효과적.

- 공간정보의 시각화 : 지도 차트.

  지도 위에 데이터를 표시 - 공간과 데이터 상의 패턴 파악에 효율적.

6. 요약 : 주요 시각화 및 작업.

예측

- 박스플롯, 막대차트, 산점도의 y축에 결과변수를 배치
- 병렬 박스플롯, 막대차트, 그리고 멀티패널을 사용하여 결과변수와 범주형 예측변수 간의 관계를 탐색
- 산점도를 사용하여 결과변수와 수치형 예측변수 간의 관계를 탐색
- 분포도(박스플롯이나 히스토그램)를 사용하여 결과변수(그리고/또는 수치형 예측변수들)의 변환 필요성을 결정
- 상호작용 조건의 필요성을 찾기 위해 색상/패널/크기를 추가하여 산점도를 분석
- 여러 가지 집계 수준과 확대축소를 사용하여 다른 행동양식을 보이는 데이터 영역을 찾고, 글로벌 패턴과 로컬 패턴의 수준을 평가

분류

- y축에 있는 결과변수를 막대 차트를 사용하여 범주형 예측변수와 결과변수의 관계를 탐색
- 컬러코드화된 산점도(색상은 결과변수를 표시)를 통해 결과변수와 쌍별 수치형 예측변수 간의 관계를 연구
- 병렬 박스플롯을 통해 결과변수와 수치형 예측변수 간의 관계를 연구. 결과변수에 관한 수치형변수의 박스플롯 시각화. 다른 수치형 예측변수도 동일한 시각화작업 수행. 가장 분리된 박스가 유용한 예측변수임
- 평행좌표 차트에서 결과변수를 표시하기 위해 색상을 사용
- 분포도(박스플롯이나 히스토그램)를 사용하여 결과변수(그리고/또는 수치형 예측변수들)의 변환 필요성을 결정
- 상호작용 조건의 필요성을 찾기 위해 색상/패널/크기를 추가하여 산점도를 분석
- 여러 가지 집계 수준과 확대축소를 사용하여 다른 행동양식을 보이는 데이터 영역을 찾고, 글로벌 패턴과 로컬 패턴의 수준을 평가

시계열 예측

- 패턴의 종류를 결정하기 위해 다양한 시간으로 집계한 선 그래프들을 생성
- 확대축소와 패닝을 사용해서 여러 가지 단기 시계열을 찾고, 서로 다른 행동양식을 보이는 데이터 영역을 결정
- 글로벌 패턴과 로컬 패턴을 찾기 위해 다양한 집계 수준을 사용
- 시계열 데이터의 결측치 식별
- 적절한 모델을 선택하기 위해 여러 유형의 추세선을 겹쳐보기

비지도학습

- 관측의 쌍별 관계와 군집을 식별하기 위해 산점도 매트릭스를 생성
- 상관관계표를 검토하기 위해 히트맵을 사용
- 다른 행동양식을 보이는 데이터의 영역을 찾기 위해 여러 집계수준과 확대축소를 사용
- 데이터 군집을 찾기 위해 평행좌표 차트를 생성