1장 서론, 2장 데이터마이닝 프로세스 개요

1장 : 서론

비즈니스 애널리틱스
↳ 이 3가지능력이 필수.

① 해당 비즈니스에 대한 근본적인 이해를 바탕으로 유효한 질문을 던지는 능력
② 다양한 소스로 부터 정형, 비정형 데이터를 찾고 저장하고, 최적으로 처리하는 능력
③ 다양한 머신러닝, 통계 기법을 사용하여 의사결정을 위한 새로운 인사이트를 발견하는 능력.

- 비즈니스 애널리틱스 (Business Analytics; BA) - 의사결정에 필요한 정량적인 데이터를 산출하는 업무/기술.
  ↳ 기본적으로 데이터를 탐색하고, 변수관계를 측정하고, 변수의 값을 예측.

  - 효과적으로 사용하기 위해선 가치가 숨겨져있는 비즈니스 환경에 대한 이해와 데이터 마이닝 기법에 대한 정확한 이해가 필요.

- 데이터 마이닝 - 통계 + 기계학습.

여기서는 (어떤 종류의 데이터 및 문제를 풀기 위해서는 어떤 방법을 사용하여야 하는가?
분석 방법들은 어떻게 작동하는가?
각 분석방법들의 요구사항, 장·단점은 무엇인가?
각 분석 방법의 성과를 어떻게 평가하여야 하는가?) 에 대해 초점.

↳ 데이터의 크기, 패턴의 유형, 분석방법의 가정을 충족시키는지, 노이즈는 심한지, 목표는 무엇인지에 따라 방법 결정.

〈용어 정리〉

알고리즘 - 특정 데이터 마이닝 기술을 실행하기 위한 자세한 과정

신뢰 - "만약 A와 B를 샀다면 C도 구매되었다" 라는 형태의 연관규칙에 관한 측정치. A, B가 구매되었다면 C도 구매된 것이라는 조건부 확률.

홀드아웃 데이터 (홀드아웃 표본) - 모델을 구할 때 쓰는 데이터 샘플은 아니지만 그 모델의 수행을 평가할 때 쓰인다.

모델 - 데이터 세트에 적용되는 알고리즘.

관측 - 측정된 것들에 대한 분석 단위. (인스턴스, 샘플, 예제, 케이스, 레코드, 패턴, 행 으로도 불림)

예측 - 연속 결과 변수의 예측값 : 추정.

예측변수 - 예측 모델에서 입력 변수로 사용. X로 표기. 특성, 입력변수, 독립변수, 필드 라고도 함.

프로파일 - 관측 결과 측정된 값들.

응답 - 지도학습에서 예측되는 변수. Y로 표기. 종속변수, 출력변수, 목표변수, 결과변수라고도 함.

원수 – 예측된 값이나 계층. 새로운 데이터를 분석한다는 것은 트레이닝 데이터로 개발된 모형을 사용해서 새로운 데이터의 클라값을 예측.

성공 클래스 – 2진법 나의 관심클래스.

지도 학습 – 결과 변수를 아는 레코드를 알고리즘에 제공하는 과정. 알고리즘은 결과 변수를 모르는 상태에서 새로운 레코드들을 이용하여 이 값을 예측하는 것을 배움.

평가 데이터 – 모델 구축과 선택 과정 과정에서 모델이 얼마나 잘 예측하는지를 알아보기 위한 데이터.

학습 데이터 – 모델을 구축하는데 사용되는 데이터.

비지도학습 – 분석 대상의 결과값 이외에 특징을 더 알아보려는 분석.

〈 2장. 데이터마이닝 프로세스 개요 〉

목적 결정 → 데이터 수집 → 데이터 탐색 및 정제 → 데이터마이닝 방법 결정 → 최종 모델 결정 → 성능평가 → 적용

〈 데이터 모델링 과정 〉

1 데이터 마이닝의 핵심 아이디어

- 분류 – 범주형 변수
- 예측 – 연속형 변수
- 연관 규칙 및 추천시스템 ← 연관(판타적) 분석, 협업 필터링
  - 연관(판타적) 분석: 일반적인 패턴
  - 협업 필터링: 개개인의 패턴
- 예측 애널리틱스 – 데이터에 내재되어있는 패턴 탐색.
- 데이터 축소 및 차원축소 – 변수의 수가 적당하고, 비슷한 속성의 관측치들을 묶어서 분석할 때 효과적. ⇒ 많은 수의 데이터를 적은 수의 그룹으로 요약하는 과정을 데이터 축소.
  변수의 개수를 줄이는 과정을 차원축소 ↳ 예측성능을 높이고 해석을 쉽게 함.
- 데이터 탐색 및 시각화
  - 데이터의 전반적인 이해/ 이상치 탐지
  - 시각화를 통한 효율적인 탐색 가능.
- 지도학습과 비지도학습.

· 지도학습. – 분류 / 예측이 필요한 변수가 있는 경우 이를 반응변수로 놓고, 예측변수와의 관계를 통해 모델링하는 방법. → 모델링 시 필요한 데이터: 학습데이터.

학습데이터 → 모델 → 검증데이터 → 예측 – 미래데이터 예측.
- 모델: 반응변수 – 예측변수 사이의 관계를 학습.
- 검증데이터: 모델의 성능 검증

ex : 단순 선형 회귀분석)
반응변수 Y와 예측변수 X 사이의 관계를 선형식으로 표현한 것.
↳ 실제 Y값과 예측 Y값의 차이의 최소합.

· 비지도학습. – 예측/분류 하고자 하는 변수가 없을 때 사용.
– 데이터 내 연관규칙을 찾고, 비슷한 관측치끼리 군집, 차원축소.

⇒ 지도/비지도가 동시에 쓰이기도 함.
↳ 비지도 학습으로 군집화하고 지도학습으로 군집별 예측.

[2] 데이터 마이닝의 단계.

1. 데이터 마이닝 프로젝트의 목적을 명확히 설정.
   ↳ 결과의 사용처, 영향을 받는 대상, 일회용인지 지속적인지.
2. 분석에 필요한 데이터의 획득.

3. 데이터의 탐색, 정제, 전처리. - 데이터 분석을 위한 준비가 되어 있는가?
   ↳ 결측치의 처리, 데이터의 범위는 적절한가, 이상치는 없는가.
4. 필요시 데이터의 축소. - 불필요 변수의 제거, 변수 값의 변환 (연속 → 이산)
   ↳ 각 변수들의 의미를 명확히 보고, 무엇을 모델에 포함시킬건지 정함.
5. 데이터 마이닝 문제 정의 - 분류 / 예측 / 군집 등 어떤 문제로 볼 것인가?
   ↳ 1단계의 문제를 모델로 재해석.
6. 데이터 분할. (지도학습의 경우)
   ↳ 학습 / 검증 / 평가 데이터로 분할.
7. 사용할 데이터 마이닝 기법 선택 - 회귀분석, 계층군집 등

8. 알고리즘을 수행하여 과제 수행
   ↳ 가장 좋은 모델을 찾는 과정
9. 알고리즘 결과의 해석.
   ↳ 효율적인 알고리즘을 찾고, 검증 데이터를 이용해 성능평가.
10. 모델적용.

3. 데이터 분석 사전단계

◦ 데이터 베이스로부터 샘플링 – 모든 데이터를 다 사용하는 것보다 일부만 쓰는 것이 효율적이면 이 과정이 필요.

◦ 오버샘플링 – 클래스별 오분류의 중요도를 가정하여 가중치를 다르게 설정.
↳ 희소사건일 경우, 희소사건의 가중치를 크게 줌으로써 균형적인 균형을 맞춤.
↳ 이 희소사건이 유의미할 때.

◦ 데이터 전처리와 정제 – 어떤 변수를 어떻게 처리해서 무엇을 선택할 것인가.

- 변수의 종류 : 범주형 – 범주에 순위가 없을 경우 범주를 가변수로 바꿔서 사용하기도 함.
  연속형 – 연속형을 범주형으로 바꾸는 것이 적절할 때도 있다.
- 변수의 선택 : 무작정 많은 변수가 좋은 결과를 보장하지는 않는다.
  → 신뢰성높은 모델을 구축하기 위해선 꼭 필요한 변수만 사용할 필요가 있음.
  변수가 늘어날수록 더 많은 관측치가 필요하고, 전처리 작업도 증가하게 됨.

◦ 얼마나 많은 변수와 관측치가 필요한가?

– 모집단을 설명하는 통계적 추론보다는 많은 관측치가 필요 ex) 변수 당 10개의 관측치

↳ 미래를 예측해야하기 때문에. | 6x m x p (m: 클래스의 수, P: 변수)

- 변수의 개수는 작을수록 효과적.
  ↳ 어떤 변수를 포함시킬 것인가에 대한 고민필요.

◦ 이상치

일반적인 데이터의 범위를 벗어난 데이터. → 포함될 경우 모델에 큰 영향을 끼칠 수 있음.
↳ 적절한 제거가 필요하지만 일반적인 방법이 없으므로 해당 데이터에 대해 잘 아는 사람이나 추가적인 분석필요

◦ 결측치

- 어떤 변수의 데이터의 일부분이 측정이 되지 않은 것.
- 결측치가 적은 경우 해당 변수를 삭제하면 되지만 많을 경우에는 삭제할 데이터가 너무 많아짐.
  ⇒ 결측치의 대체가 필요.

- 평균, 중앙값 등으로 대체 → 다른 데이터들을 온전하게 살릴 수 있음.
  ↳ 분산 등이 달라질 수 있으나 어차피 검증데이터로 검증

- 예측에 미치는 변수의 중요도를 고려하여 삭제 / 다른 변수의 정보를 이용하여 대체
  ↳ 제일 좋은 건 또 데이터를 확보하는 것.

- 결측치인지 실제 데이터값이 0인지도 구분 필요.
  ↳ 해당 데이터에 관한 깊은 이해가 요구됨

◦ 데이터 정규화 및 리스케일링.

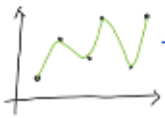
- 특정 알고리즘들은 신뢰성있는 결과를 얻기위해 정규화 필요.
  ↳ 정규분포화 / 표준 정규분포화.
- 데이터의 단위가 다를 경우에 효과적인 분석을 위해서.

4. 예측력과 과적합.

지도학습 시에 중요한 점 - 미래 데이터를 얼마나 잘 예측하는가.
⇒ 데이터 분할과 과적합의 개념필수.

◦ 과적합.

- 모델의 가장 기본적인 목적은 미래의 예측.
  ↳ 현재 데이터를 설명하는 것도 중요하지만 주 목적은 아님.


→ 현재 데이터만 딱 맞춘건 몰라도 미래 예측은 무리.

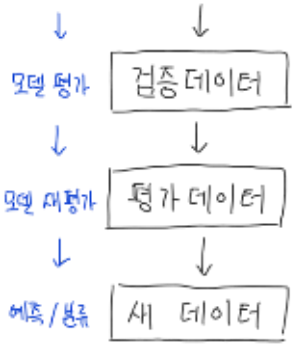
- 모델이 현재 데이터에 과하게 적합된 경우 ⇒ 과적합.

- 불필요하게 포함된 변수가 있는 경우 과적합의 위험.
- 관측치가 변수의 개수보다 충분하지 않은 경우 엉뚱한 상관관계가 포함될 수 있음.

◦ 데이터 분할의 사용과 방법.

- 학습과 검증을 같은 데이터로 한다면 우연히 성능이 좋은 모델이 될수도 있음.
- 데이터의 분할 : 학습 / 검증 / 평가.
  ↳ 랜덤하게 분류 / 시계열에 따른 분류.

모델 구축 | 학습 데이터



모델 평가에 있어 가장 객관적인 검증은 모델구축에 전혀 사용되지 않은 데이터를 이용한 검증.

- 교차검증. 데이터의 양이 적은 경우 분할이 불가능 ⇒ 교차검증 사용.
  전체 데이터를 중첩되지 않도록 k개의 폴드(fold)화.
  k-1 부분을 이용한 모델 구축 / 나머지로 검증.
  k번 시행 후 예측값을 평균하여 모델의 성능을 검증.