Topics in Artificial Intelligence (2017 Spring) : Book Report

비즈니스를 위한 데이터 과학

(빅데이터를 바라보는 데이터 마이닝과 분석적 사고)

저자 : 포스터 프로보스트 , 톰 포셋 / 번역 : 강권학 / 출간 : 2014-06-24

소속 : 컴퓨터공학과

ID : 1632036004

성 명 : 이 지 운

비즈니스를 위한 데이터 과학

(빅데이터를 바라보는 데이터 마이닝과 분석적 사고)

저자 : 포스터 프로보스트 , 톰 포셋 / 번역 : 강권학 / 출간 : 2014-06-24

1. 개 요

데이터 과학을 직접 응용할 일이 없더라도 데이터 과학을 이해하는 일은 매우 중요하다. 데이터 분석적 사고 방식에 익숙해지면 프로젝트를 평가하는 데 도움이 된다. 예를 들어 어떤 컨설턴트나 잠재적인 투자자가 데이터에서 지식을 추출하는 업무를 개선하고자 제안할 경우, 제안서를 체계적으로 평가함으로써 제안이 과연 타당한지, 아니면 문제가 있는지를 판단할 수 있다. 그렇다고 해서 프로젝트가 성공한다고는 확신할 수 없지만 적어도 제안서에 있는 결함이나 비현실적 가정, 빠진(놓치는) 부분은 알아낼 수 있다. (Data-Driven 과 같은 사고를 할 수 있다.)

대상 독자

- 데이터 과학자와 함께 일을 하거나 데이터 과학 중심의 프로젝트를 관리하는 사람들
- 데이터 과학 벤처 기업에 투자하려는 기업가
- 데이터 과학 프로젝트를 구현하려는 개발자
- 데이터 과학자를 지망하는 사람

2. 데이터 과학에 대한 이 책의 개념적 접근 방법

이 책에서는 데이터 과학에서 가장 중요한 기본 개념을 설명한다. 이 개념의 일부는 각 장의 '제목'이 되기도 하고 다른 일부는 설명을 통해 자연스럽게 소개된다 (설명에 들어 있는 개념은 기본 개념이라고 표시되어 있지 않다). 이 개념들은 문제에 대한 계획을 세우는 일부터 데이터 과학 기법을 적용하고 더 나은 의사 결정을 하기 위해 결과를 배치하는 과정까지 폭 넓게 적용될 뿐만 아니라 다양한 비즈니스 분석 방법론 및 기법의 기반이 되기도 한다. 세부내용은 다음과 같다.

가. 데이터 과학을 기업 조직에 결합하는 방법에 대한 개념

- 데이터 과학팀을 모집, 조직, 육성하는 방법이 포함된다.
- 데이터 과학이 경쟁력을 향상시키도록 생각하는 방법이 포함된다.
- 데이터 과학 프로젝트를 성공적으로 수행하기 위한 전략적 개념이 포함된다.

나. 데이터 분석적으로 사고하는 일반적인 방법

- 이 개념을 갖고 있으면 적절한 데이터를 찾아내 적절한 방법을 적용하는 데 도움이 된다.
- 여러 상위 수준의 데이터 작업과 '데이터 마이닝 프로세스'가 이에 포함된다.

다. 실제로 데이터에서 지식을 추출하는 일반적인 개념

- 방대한 데이터 과학 작업 및 작업에 사용하는 알고리즘의 기반이 된다.

3. 독후 : 데이터 과학과 분석적 사고

데이터 과학에 입문하는 사람들에게 도움이 될 것이며, 비즈니스 문제에 데이터 과학 문제를 적용하는데 중점을 두었다. 고객 이탈, 타겟 마켓팅, 위스키 분석처럼 실제 비즈니스에서 발생하는 익숙한 문제를 여러 곳에서 예제로 다뤘다. 알고리즘을 나열하기보다는 데이터 과학에 깔려 있는 개념을 이해할 수 있으며, 문제 해결 방법을 알려준다.

데이터를 현명하게 사용하면 비즈니스 경쟁력을 새로운 차원으로 끌어올릴 수 있으며, 데이터가 주도하는 환경에서 성공하려면 엔지니어, 분석가, 관리자 모두 자신 앞에 놓여 있는 선택사항, 설계 결정 사항, 장단점을 반드시 이해하고 있어야 한다. 흥미로운 예제, 명확한 설명, '방법' 뿐만 아니라 '이유'도 자세하고 폭넓게 설명하고 있으므로, 데이터 주도 시스템을 개발하고 응용하는 업무를 수행하려는 사람에게 완벽한 입문서이다.

이 책의 초반부는 개념 및 예제, 비즈니스 측면에서 풀어나갔으며, 중반부는 모델, 과적합, 유사도, 군집, 이웃, 시각화, 증거, 확률, 마이닝 등 데이터 분석 기법 설명한다. 후반부에는 비즈니스 전략 및 결론으로 끝을 맺는다. 챕터별 새부 내용을 나름대로 정리한 내용은 아래와 같다.

1. 개요: 데어터 분석적 사고 방식

- 데이터의 제공하는 기회를 데이터 주도 의사 결정을 축으로 분석 활용하면 어떤 이익이 있는지 설명한다. (데이터 마이닝과 데이터 과학의 차이점 포함)

2. 비즈니스 문제와 데이터 과학 해결책

- 데이터 마이닝의 프로세스 (비즈니스이해->데이터이해->데이터준비->모델링->평가->배치)와 분석 기법 및 기술 (통계학, 데이터베이스 쿼리, 데이터 웨어하우스, 회귀분석, 기계학습과 데이터마이닝)

3. 예측 모델 개요 : 연관성에서 감독 세분화

- 모델 유도(예측)을 기반으로 감독 세분화 및 시각화 과정 설명

4. 데이터에 대한 모델 적합화

- 회귀분석 및 선형/비선형 함수, 벡터 기계 설명

5. 과적합화 문제 해결

- 일반화(Generalization)와 과적합화(Overfitting) : 데이터 마이닝은 모델 복잡도와 과적합화 문제간의 싸움 (균일화(Regularization)를 통해 모델 복잡도 통제)

6. 유사도, 이웃, 군집

- 비슷한 항목을 찾아내는 일, 예측 모델링, 개체 군집화을 통해 유사도 사용 방법 설명, 유사도 및 거리 계산 방법, 최근접 이웃법 계열의 방법 설명

7. 결정 분석적 사고 1: 좋은 모델은?

- 모델의 적절한 평가 척도 고안이 필요하며, 기댓값은 좋은 틀을 제공한다.

8. 모델 성능 시각화

- 모델의 평가 결과를 시각화하는 것은 평가 업무에서 매우 중요한 부분이며, 훈련 데이터 및 표본을 이용해 결과를 예측해야 한다. (수익 곡선 및 수신자 운용 특성 곡선 등은 중요한 시각화 도구이다.)

9. 증거와 확률

- '각 타겟이 특정값을 어떻게 생성하는가?'라는 질문을 통해 새로운 기법 설명 (생성기법(Generative Method), 베이지안 기법(Bayesian Method)

10. 텍스트 표현 및 마이닝

- 텍스트를 특정 벡터로 변환하는 방법 설명 (각 문서를 개별적인 단어로 분할하거나 TFIDF 공식을 이용해 각 단어에 값을 할당하는 방법)

11. 결정 분석적 사고 2 : 분석 공학

- 문제에 대해 데이터를 분석적으로 장려함으로써 데이터 마이닝의 역할을 명확히 하고 비즈니스 제약, 비용, 효과를 고려ㅏ며, 문제를 단순화하기 위한 가정을 명확히 표현하는 것

12. 기타 데이터 과학 작업 과 기법

- 항목드의 동시 발생 또는 연관성 찾아내기 : 상품 구매
- 전형적인 행위의 프로파일링 : 신용카드 사용량이나 고객 대기 시간
- 데이터 항목 간의 연결 예측 : 사람들 간의 소셜 네트워크에서의 연결
- 데이터를 관리하기 쉽게 만들거나 데이터에서 숨은 정보를 찾아내기 위해 축소 : 잠재적인 영화 선호도
- 모델을 하나의 전문가로 생각하고 모델을 조합하기 : 영화 추천 모델 개선
- 데이터 간의 인과 관계 도출 : 소셜 네트워크에 연결된 사람들이 동일한 상품을 구입하는 이유

13. 데이터 과학과 비즈니스 전략

- 데이터 과학은 운용도 아니고 공학도 아니다.

4. 독후 : 결론

'명확히 설명할 수 없다면 그것을 제대로 알고 있는 것이 아니다.'

(알버트 아인슈타인)