01 부

기본 개념

제 1장

서론

정보기술 시대의 진화는 데이터의 시대(Data Age), 정보의 시대(Information Age)를 거쳐서 현재에는 지식의 시대(Knowledge Age)에 이르렀다. 하지만 오늘날 우리는 다시 데이터의 시대에 살고 있다는 표현이 더 적절할 것 같다. 매일 엄청난 양의 데이터가 생성된다. 데이터를 분석해 정보 또는 지식을 도출해야 하는데, 이 과정이 데이터의 생성 속도를 따라가지 못해서 '데이터는 풍부하지만 정보나 지식은 빈약한 시대'가 되었다. 그에 따라 중요한 의사결정이 데이터의 정확한 분석으로부터 획득한 정보나 지식에 기반하기보다는 의사결정자의 직관에 의존하는 경우가 종종 발생한다. 이러한 현대의 데이터 홍수 속에서 데이터 분석의 중요성은 더욱더 강조되고 있다.

이 책은 데이터 애널리틱스(Data Analytics) 기법을 사용해 모델을 구축하는 과정을 다룬다. 이 책에서 주로 대상으로 하는 데이터는 정형 데이터다. 비정형 데이터의 중요성이 점차 증가하고 있지만, 이 책에서는 다양한 데이터 애널리틱스 기법을 효과적으로 설명하기 위해서 정형 데이터에 집중했다. 제16장 딥러닝(Deep Learning)에서 비정형 데이터인 텍스트(Text) 데이터를 일부 다루지만, 나머지 장들에서는 모두 정형 데이터를 다룬다.

데이터 애널리틱스와 관련된 여러 용어가 있는데, 이 장에서는 이 용어들의 간단한 정의를 들어서 각 용어 간의 관계를 알아보고 이 책의 구성에 관해 기술한다.

1. 인공지능, 머신러닝, 딥러닝

인공지능(Artificial Intelligence, AI)은 인간의 지능적 행동의 모사(Simulation)와 관련된 연구를 수행하는 학문 분야다. AI 정의에 '컴퓨터 과학(Computer Science)의 한 분야'라는 말이 포함되는 경우가 있는데, 이 책에서는 그 말을 넣지 않았다. AI에는 기본적으로 인간에 대한 연구가 있어야

한다. 우리가 접하는 많은 학문이 인간에 대한 연구와 관련돼 있다. 예를 들어 경영학, 경제학, 인문학, 정치학, 언어학, 사회학, 심리학, 철학 등 거의 모든 학문이 인간에 대해 연구하는 학문이다. AI는 바로 이런 학문에 체계적인 연구 방법론을 제공한다. 1950년대 중반 AI가 태동할 당시, 컴퓨터 과학을 전공한 연구자들이 많은 기여를 한 것은 사실이지만, 이제는 컴퓨터 과학 이외의 전공 분야에서도 AI 연구가 활발히 진행되고 있다. 그러므로 AI를 컴퓨터 과학의 한 분야로 제한할 필요는 없다.

머신러닝(Machine Learning)은 컴퓨터가 실제 세계의 관찰과 상호작용을 통해 데이터나 정보를 수집해 자신의 지식수준을 향상시킴으로써 인간처럼 학습하고 행동하도록 하는 AI의 연구 분야다. 딥러닝은 머신러닝의 한 분야로서 은닉층이 세 개 이상인 신경망 구조를 사용한다.

인공지능, 머신러닝, 딥러닝 간의 관계는 그림 1.1과 같이 표현할 수 있다.

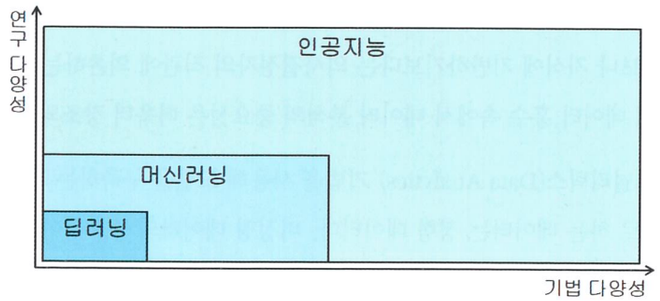

그림 1.1 인공지능, 머신러닝, 딥러닝

2. 데이터 사이언스와 데이터 애널리틱스

데이터 사이언스(Data Science)는 과학적 방법, 프로세스, 알고리즘, 시스템을 사용해 다양한 형식의 데이터로부터 지식과 통찰력을 추출하는 융합 분야다. 데이터 사이언티스트(Data Scientist)가 되기 위해서는 다음과 같이 4가지 분야의 기량이 필요하다.

- **데이터 애널리틱스**: 통계학, 머신러닝 등의 기법을 활용하는 능력.
- **프로그래밍과 데이터베이스**: 컴퓨터 기초지식, 선형대수, 프로그래밍 언어(파이썬, R 등), 데이터베이스(SQL, NoSQL 등), 빅데이터(MapReduce, Hadoop, Hive, Pig 등), 시각화 도구(ggplot, Tableau 등).

- **해당 영역의 지식과 정서적 지능**: 기업 운영에 대한 지식, 데이터에 대한 관심, 문제를 해결하려는 성격, 전략적·주도적·창의적·혁신적·협력적 품성, 새로운 기술에 도전하려는 마음가짐.
- **의사소통 능력**: 상관과 협업하는 능력, 능숙한 발표 능력, 데이터로부터 도출한 지식과 통찰력을 의사결정으로 변환하는 능력, 시각적 표현 능력.

데이터 애널리틱스는 데이터로부터 유용한 정보와 지식을 도출해내는 기법과 프로세스다. 데이터 애널리틱스와 데이터 어낼리시스(Data Analysis)는 서로 혼용되기도 하는 용어다. 하지만 엄밀하게 말하면 데이터 어낼리시스는 데이터를 분석하는 과정 자체만을 의미하지만, 데이터 애널리틱스는 데이터로부터 유용한 정보와 지식을 도출하기 위한 모델링 기법들, 그 기법들로 데이터를 분석하는 과정, 그리고 신뢰할 수 있는 방법과 원칙에 입각하여 모델을 구축하는 과정 전반을 일컫는 용어다. 그러므로 데이터 애널리틱스가 데이터 어낼리시스를 포함하는 개념이라고 할 수 있다. 데이터 애널리틱스에는 다양한 기법이 사용되는데, 가장 기본으로 사용되는 기법은 그림 1.2와 같이 통계학 기반 기법과 머신러닝 기반 기법이다.

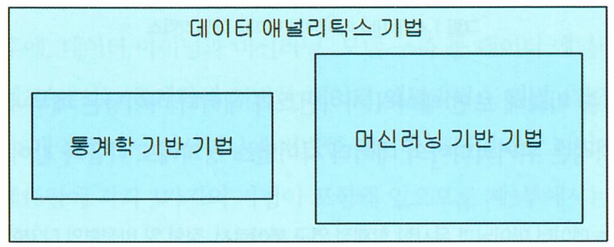

그림 1.2 데이터 애널리틱스 기법의 구성

데이터 사이언스와 데이터 애널리틱스 간의 관계는 그림 1.3과 같이 표현할 수 있다.

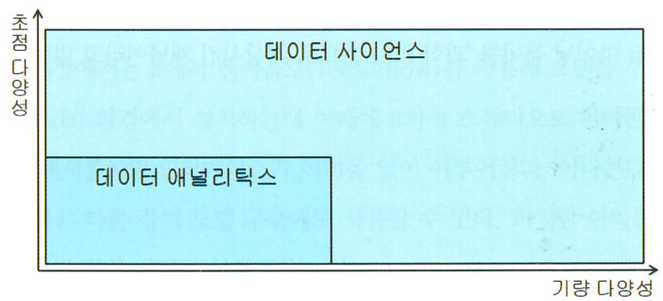

그림 1.3 데이터 사이언스와 데이터 애널리틱스

3. 데이터 마이닝과 데이터 애널리틱스

데이터 마이닝(Data Mining)은 대량의 데이터로부터 의미 있는 패턴과 규칙을 발견하기 위해 탐색(Exploration)과 분석(Analysis)을 하는 비즈니스 프로세스다. 데이터 마이닝은 위에서 언급한 다른 분야와 비교해 특히 프로세스가 강조되는 분야로서, 여러 단계로 구성된 방법론들이 제시돼 있다. 이 단계에 데이터 애널리틱스 과정이 포함된다. 그러므로 데이터 마이닝과 데이터 애널리틱스 간의 관계는 그림 1.4와 같이 표현할 수 있다.

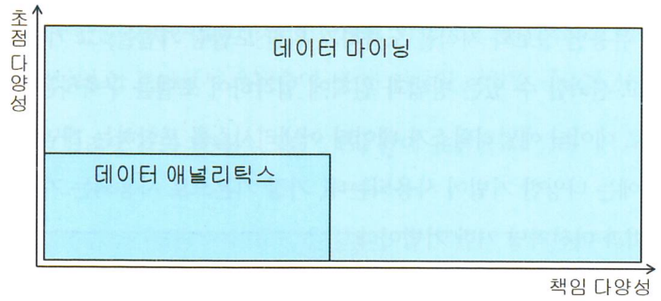

그림 1.4 데이터 마이닝과 데이터 애널리틱스

그림 1.3과 그림 1.4를 비교해 보면 데이터 사이언스와 데이터 마이닝은 매우 유사한 분야임을 알 수 있다. 이러한 유사성은 위키피디아의 데이터 사이언스 정의에도 다음과 같이 잘 나타나 있다.

> 데이터 사이언스는 데이터 마이닝과 유사한 학제적 연구 분야로서, 정형 및 비정형의 다양한 형식의 데이터에 과학적 방법, 프로세스, 알고리즘, 시스템 등을 적용하여 지식과 통찰력을 추출하는 융합 분야다.

데이터 사이언스와 데이터 마이닝은 사실 같은 분야를 지칭하는 것으로 볼 수 있는데, 데이터 마이닝은 1990년 중반부터 발전해온 오래된 연구 분야고, 데이터 사이언스는 2000년대 초반에 태동한 용어로서, 데이터 마이닝 분야를 '과학'의 수준으로 격상시킨 개념이라고 말할 수 있다.

4. 이 책의 구성

이 책은 그림 1.5와 같이 총 3부 16개의 장으로 구성되어 있다.

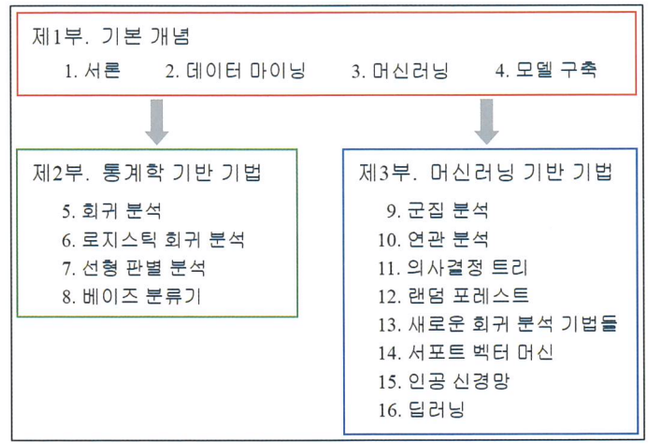

그림 1.5 이 책의 구성

제1부에서는 서론 후에, 데이터 마이닝과 머신러닝, 모델 구축 등 데이터 애널리틱스에 대한 기본 개념을 설명한다. 제2부에서는 통계학에 기반한 데이터 애널리틱스 기법 4가지에 대해서 설명한다. 제3부에서는 총 8개의 장에서 머신러닝에 기반한 데이터 애널리틱스 기법에 대해서 설명하고 있는데, 제13장과 제16장에 각각 3가지의 기법이 포함돼 있으므로 제3부에서는 총 12가지 기법에 대해서 설명한다.

이 책의 제5장부터 제16장까지에는 R을 사용해 모델을 구축하는 예제가 기술돼 있고 관련 소프트웨어의 설치 방법이 부록 A, B, C에 기술돼 있다. 부록 A와 B를 따라 R과 R 스튜디오(Studio)를 설치하면, 제5장부터 제13장 4.1절까지 그리고 제14장과 제15장의 모델 구축에 사용할 수 있다.

제13장 4.2절과 제16장에서는 R에서 텐서플로(Tensorflow)를 사용해 모델을 구축한다. 텐서플로는 아나콘다(Anaconda) 환경에서 설치하는데 이때에 R과 R 스튜디오도 아나콘다 환경에서 동시에 설치해줘야 R과 텐서플로의 인터페이스가 어려움 없이 이루어진다. 아나콘다 환경에서 설치된 R과 R 스튜디오를 물론 다른 장의 모델 구축에도 사용할 수 있다. 하지만 아나콘다 환경에서 R과 R 스튜디오를 설치하면 최신 버전이 설치되지는 않는다.

본 저자는 다음과 같이 권장한다. 부록 A와 B를 따라 설치한 R과 R 스튜디오를 제13장 4.1절까지의 모델 구축에 사용한 후에 R과 R 스튜디오를 제거한다. 그 다음에 부록 C에 기술된 방법에 따라 아나콘다, R, R 스튜디오, 텐서플로를 모두 동시에 설치한 후에 제13장 4.2절부터 제16장까지의 모델 구축에 사용한다.

5. 참고문헌

1) 6 Vital Data Science Skills Every Data Scientist Must Possess!
https://www.proschoolonline.com/blog/data-science-skills/

2) Data Analysis
https://en.wikipedia.org/wiki/Data_analysis

3) Data Analytics
https://searchdatamanagement.techtarget.com/definition/data-analytics

4) Data Analytics vs Data Analysis
https://www.educba.com/data-analytics-vs-data-analysis/

5) Data Science
https://en.wikipedia.org/wiki/Data_science

6) Data Science
https://www.datarobot.com/wiki/data-science/

7) Data Scientist: The Sexiest Job of the 21st Century
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

8) Definition – What does Artificial Intelligence (AI) mean?
https://www.techopedia.com/definition/190/artificial-intelligence-ai

9) Definition – What does Data Analytics mean?
https://www.techopedia.com/definition/26418/data-analytics

10) Definition – What does Data Science mean?
https://www.techopedia.com/definition/30202/data-science

11) Definition – What does Deep Learning mean?
https://www.techopedia.com/definition/30325/deep-learning

12) Definition – What does Machine Learning mean?
https://www.techopedia.com/definition/8181/machine-learning

13) Definition of Artificial Intelligence
https://www.merriam-webster.com/dictionary/artificial%20intelligence

14) Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann Pub., 2012.

15) Linoff, G. and M. J. A. Berry, *Data Mining Techniques: For Marketing, Sales, and CRM*, 3rd ed., Wiley Pub., 2011.

16) Modern Data Scientist Skill Set
https://mywebvault.wordpress.com/2017/05/18/modern-data-scientist-skill-set-marketing-distillery/

17) What is Machine Learning?
https://www.techemergence.com/what-is-machine-learning/

18) What's The Difference Between Data Analytics and Data Analysis?
https://www.inteliment.com/blog/our-thinking/whats-the-difference-between-data-analytics-and-data-analysis/