1. 데이터 분석이란?

데이터 분석이란 수집된 데이터에 숨어 있는 정보를 찾아 가치 있게 만드는 일, 즉 데이터에서 새로운 의미와 가치를 지닌 정보를 생산하는 일을 말한다. 이 데이터 분석을 누가 어떻게 하느냐에 따라서 중요한 정보가 발견되거나 아무 의미 없는 정보만 생선되는 등 데이터의 가치가 바뀐다.

1.1 데이터 분석 프로세스

1. 문제 정의 (목적, 목표)
2. 데이터 수집(데이터 정의, 데이터 수집)
3. 데이터 분석(데이터 전처리, 기초 통계 분석, 모델 구축 및 평가)
4. 검증 및 고찰

1.2 데이터 분석 알고리즘의 특징과 분류

데이터 분석 알고리즘은 **통계적 방법**과 **인공지능 방법**으로 나눌 수 있다.

통계적 방법은 수집된 데이터에 대해서 어떤 규칙을 갖고 있는지 분석하고, 발견된 규칙을 알고리즘과 같이 만들어서 활용하는 방법이다.

인공지능 방법은 대량의 데이터로부터 데이터에 대한 규칙을 알고리즘이 찾아내게 만드는 방법이다.

이렇게 찾은 규칙은 사람이 해석할 수 있는 알고리즘인 **화이트박스 알고리즘**과 해석할 수 없는 **블랙박스 알고리즘**으로 나눌 수 있다.

화이트 박스 알고리즘으로는 결정 트리가 있어, 트리 구조를 그래프로 그려 보면 알고리즘이 어떤 규칙을 만들어 냈는지 이해 할 수 있다.

블랙박스 알고리즘으로는 요즘 유행하는 딥러닝의 기본구조인 신경망이 있다. 이 신경망의 학습 결과는 가중치 벡터형태로 나타나지만, 사람이 해석하기에는 무리가 있다.

이런 알고리즘을 이용해서 데이터 분석을 해야한다. 데이터 분석에서는 주로 변수라는 용어가 많이 사용된다. 이 변수는 크게 **종속변수(dependent vairable)** 와 **독립변수(independent variable)** 두 가지로 나눌 수 있다. 종속변수는 결과를 나타내는 변수로서, 일반적으로 Y로 표시한다. 그리고 독립변수는 종속변수의 원인에 해당하는 변수로서, 일반적으로 X로 표시한다.

종속변수의 개수	독립변수의 개수
1개 일변량 분석	1개 단변수 분석
2개 이변량 분석	2개 이상 다변수 분석
3개 이상 다변량 분석	

독립변수와 종속변수를 이용해 분석하는 데이터 분석 알고리즘 중에서

통계적 알고리즘으로는 분산분석, 회귀분석, 주성분분석, 요인분석, 판별분석과 같은 알고리즘이 있으며,

인공지능 알고리즘에는 결정 트리, 신경망, 유전 알고리즘, 서포트 벡터 머신 등 수많은 알고리즘이 있다.

여기서 회귀분석, 주성분분석은 인공지능의 한 가지인 데이터 마이닝 분야에서도 사용되며 요즘에는 통계적 알고리즘이나 인공지능 알고리즘 중 어디에 속해 있는지에 대한 경계가 애매모호해지고 있다.

이런 데이터 분석 알고리즘들은 크게 예측, 압축, 분류를 목적으로 이용할 수 있으며 특징은 다음과 같다.

1.

목적		종속변수	독립변수	주요 알고리즘
예측	종속변수 예측	O	O	**linear regression**, support vector regression
압축	차원 축소	X	O	**principal component analysis**, factor analysis
분류	닮은 데이터의 그룹화	O	O	decision tree(supervised learning), **Mahalanobis-Taguchi system(supervised learning)**, self-organizing map(unsupervised learning)

1)**예측 알고리즘**은 종속변수와 독립변수 사이의 인과관계를 이용해 모델을 만들어 종속변수의 값을 예측한다. 주요 알고리즘으로는 선형 회귀분석(linear regression analysis)이 있으며, 5장 에서 세 가지 회귀분석 방법에 대해 설명한다.

2)**압축 알고리즘**은 데이터의 차원을 축소하기 위해 사용되는 알고리즘으로 독립변수들 간의 관계를 분석해 정보를 압축하는 알고리즘이다. 3장에서 대표적인 알고리즘인 주성분분석을 설명한다.

3)**분류 알고리즘**은 분류와 군집화로 나눌 수 있다. **분류**는 종속변수를 기분으로 독립변수의 특징을 학습시켜 분류하는 알고리즘이다. 이렇게 정답(종속변수)을 알려주고 학습시키는 방법을 교사학습(supervised learning, 지도학습)이라고 하며 대표적인 알고리즘으로 결정 트리가 있으며, 4장에서는 마할라노비스-다구치 시스템이 있다.

종속변수가 없는 **군집화**는 독립변수의 속성을 파악해 비슷한 속성을 가진 데이터끼리 군집화하는 알고리즘이다. 교사학습과는 반대로 정답(종속변수)이 없는 것이 특징이며, 이런 학습방법을 비교사 학습(unsupervised learning, 비지도 학습 또는 자율 학습)이라고 한다. 대표적인 알고리즘으로는 신경망의 한 종류인 자기 조직화 지도(self-organizing map)가 있다.

위 표에서 굵은 색이 본 책에서 다루는 내용이다.

회귀분석은 수익 예측이나 생산량 예측, 그리고 종속변수에 영향을 미치는 변수를 찾기 위한 가장 기본적인 알고리즘으로 알려져 있다. 회귀분석을 이용해서 1차 분석을 진행한 후에, 좀 더 고도화된 알고리즘은 PCR, PLS, 서포트 벡터 회귀 등을 사용하게 된다.

주성분분석은 많은 데이터를 압축할 때 쓰는 알고리즘으로 알려져 있다. 또, 데이터를 다른 방향으로 바라볼 수 있게 변환하여 데이터를 분석할 수 있는 도구도로 많이 사용된다. 특히, 여론조사나 앙케이트 결과를 종합 분석 할 때 많이 사용된다.

마할라노비스-다구찌 시스템은 정상/ 비정상을 분류하는 최적의 알고리즘이다. 우리가 일반적으로 분류하는 대상은 개, 고양이, 자동차와 같이 특정 물체를 분류하는 것도 있지만, 정상과 비정상을 분류하는 경우가 대부분이다. 특히, 비정상의 종류가 너무 많아 모든 비정상의 상태를 정의 할 수 없을 때, 정상 상태만을 정의해서 정상/비정상을 분류하는 대표적인 시스템이기 때문이다.