빅데이터 분석 방법론과 데이터 과학자

07

김형준, 송완영, 안세륭
한국주택금융공사 주택금융연구원 연구위원

CONTENTS

요 약

- (연구배경) 데이터 폭증 시대 도래에 따라, 빅데이터 분석 기술에 대한 관심 증가와 함께 기술 활용을 위한 이해의 필요성 확대
  - ▶ 공사 내 빅데이터 분석에 대한 관심 및 이해 향상, 분석 방법론 활용을 위한 다양한 아이디어 수렴을 위해 본 보고서 작성
  - ▶ 빅데이터 분석 기술에서 사용되는 기계 학습과 주요 방법론을 살펴보고, 이를 수행하는 데이터 과학자의 역할과 필요 역량 소개

- (기계학습) 데이터 속에서 패턴을 찾아, 발견된 패턴을 토대로 스스로 프로그램을 수정해 나가는 인공지능의 한 분야
  - ▶ 많이 사용되는 분석 방법론을 지도학습과 비지도학습으로 분류, 각 방법론에 대하여 분석 과정을 간략히 소개
  - ▶ 보다 발전된 최신 방법론으로 딥러닝, 텍스트 마이닝, 협업 필터링을 설명하고, 응용 사례와 함께 공사 활용 방안을 제시

- (데이터 과학자) 현장에서 발생하는 대용량의 데이터 속에서 숨은 트렌드를 발견, 기업에 필요한 정보를 적시에 제공하는 분석전문가
  - ▶ 전문지식과 함께 수학·통계·프로그래밍 및 데이터베이스에 대한 이해, 효과적인 결과 전달을 위한 커뮤니케이션 및 데이터시각화 능력 요구

- (시사점) 빅데이터 분석 방법론의 활용 범위와 한계에 대한 이해를 바탕으로, 공사의 데이터 활용성을 높일 수 있는 조직문화 형성
  - ▶ 빅데이터 만능론을 경계하는 한편, 조직 내 우수한 데이터 과학자 육성을 위해 필요한 역량을 확인하고, 부족한 부분을 채워나갈 수 있는 교육 프로그램 마련 필요
  - ▶ 현장의 경험과 아이디어를 실제로 유용한 분석 결과로 이끌어낼 수 있는 통로를 마련하여, 데이터 분석을 토대로 의사결정을 뒷받침하는 빅데이터 기반의 실사구시(實事求是) 조직문화 형성 필요

* 본고의 내용은 필자의 개인 의견으로 한국주택금융공사의 공식적인 견해와 다를 수 있습니다.

I 연구배경

- 데이터 폭증 시대 도래 및 빅데이터 분석 기술에 관한 관심 증진
  - ▷ 디지털 기기 보급과 소셜네트워크 서비스 부상으로 데이터 폭증 시대 도래, 빅데이터 활용을 위한 분석 기술에 관한 관심 증가
    - – 전 세계 데이터는 매년 40%씩 증가 (맥킨지, 2011)
  - ▷ 빅데이터 처리 기술의 발달로 기존에는 불가능하던 비정형 데이터의 기계적 분석이 가능해짐에 따라 이를 활용한 서비스가 출현하는 등 해당 분야에 대한 주목도 상승

- 빅데이터 분석 방법론에 관한 이해 필요
  - ▷ 실제 현장에서 적용되는 대표적인 빅데이터 분석 방법론을 살펴보고, 공사 활용 방안에 대해 알아보고자 함

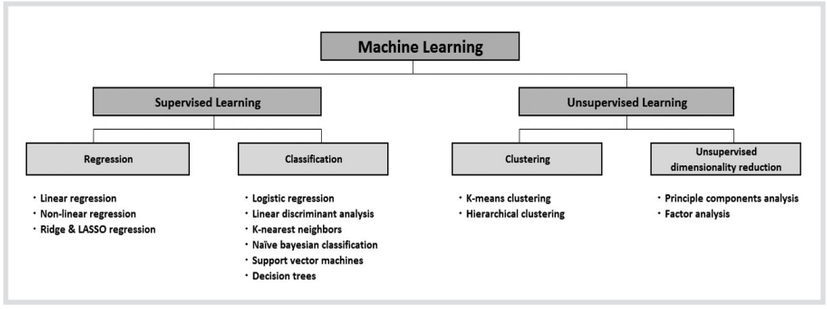
- 데이터 과학자의 역할 인식
  - ▷ 데이터 주도형 통찰력에 대한 기업들의 필요성 인식이 깊어짐에 따라, 이에 대한 솔루션을 제공하는 데이터 과학자 수요 증가
  - ▷ 향후 육성을 대비하여 데이터 과학자의 역할과 필요 역량 소개

Ⅱ 기계학습(Machine Learning)의 이해

1. 개요

- 기계학습의 정의
  - ▷ 기계학습은 인공지능의 한 분야로, 컴퓨터가 명확히 프로그램화되지 않은 상태로 학습할 수 있는 능력을 제공하는 최적화 방법론
    - – 데이터 속에서 패턴을 찾는다는 점에서 데이터마이닝과 유사하나, 발견된 패턴을 토대로 프로그램을 수정한다는 부분이 다름
- 기계학습의 분류
  - ▷ 기계학습은 학습 자료에 따라 크게 지도학습과 비지도학습으로 분류
  - ▷ 지도학습(Supervised learning)이란 정답이 주어진 학습 자료를 이용하여 예측하는 분석방법으로, Target 변수의 종류에 따라 regression과 classification으로 분류됨
    - – regression, k-nearest neighbors, naive bayesian classification, support vector machines, decision trees 등이 있음
  - ▷ 비지도학습(Unsupervised learning)이란 정답이 주어지지 않은 자료에서 그룹을 식별하기 위한 탐색적 자료 분석으로, clustering이 대표적임

[그림 7-1] 기계학습의 분류



※ 자료 : 한국주택금융공사

2. 지도학습 (Supervised Learning)

- 정답이 주어진 학습 자료를 이용하여 Target 변수를 예측하는 분석방법으로, Target 변수의 종류에 따라 크게 Regression과 Classification으로 분류됨
- Regression은 Target 변수가 연속형 자료일 때, 그 값을 예측하기 위한 모형
  - ▷ Target 변수를 Input 변수들의 선형관계를 통해 예측하는 Linear regression, 변수들의 비선형관계를 통한 Non-linear regression, Ridge, LASSO regression 등이 있음
- Linear Regression
  - ▷ 연속형인 Target 변수(Y)와 p개의 Input 변수들(X) 사이의 관계를 변수들의 선형결합(일차함수) 형태로 설명한 모형

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\epsilon.$$

  - ▷ 회귀계수(regression coefficients) $\beta=(\beta_0,\beta_1,\cdots,\beta_p)^t$는 구하고자 하는 모수로, RSS(Residual sum of squares)를 최소화시켜 추정하는 OLS(Ordinary least squares)방식을 이용하여 구함

$$RSS=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_jx_{ij}\right)^2.$$

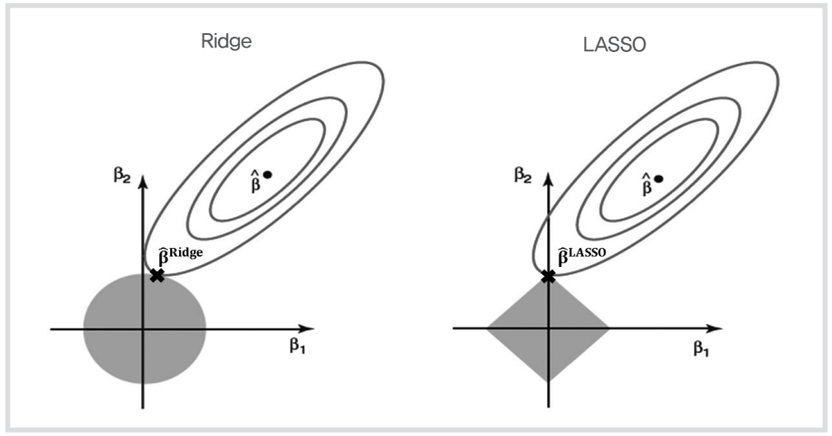
- Non-linear Regression
  - ▷ 연속형인 Target 변수와 Input 변수들 사이의 관계를 일차함수 형태가 아닌 비선형결합 형태로 설명한 모형
- Ridge & LASSO Regression
  - ▷ 관측치보다 변수의 개수가 훨씬 커 일반적인 회귀분석 적용이 불가능할 때 사용가능한 분석으로, 일반적인 회귀분석과 달리 추정하려는 모수(parameter)에 추가적인 제약을 가하여 모형을 간소화시키는 방법

▷ RSS에 특정 penalty term을 부여하여 최소화시키는 회귀계수 추정

– Ridge Regression은 모수들의 제곱합에 제약을 주어, $RSS+\lambda\sum_{j=1}^{p}\beta_j^2$을 최소화시키는 회귀계수 추정

– LASSO(Least Absolute Shrinkage and Selection Operator)는 모수들의 절대값의 합에 제약을 주어, $RSS+\lambda\sum_{j=1}^{p}|\beta_j|$을 최소화시키는 회귀계수 추정

[그림 7–2] Ridge 및 LASSO의 회귀계수



※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R", 한국주택금융공사

● Classification은 Target 변수가 두 개 이상의 집단(class)으로 나누어진 범주형 자료일 때, 어느 집단으로 분류될 것인지를 예측하는 모형

▷ 분류 방법론으로는 Logistic regression, Linear discriminant analysis, K–nearest neighbors, Naive bayesian classification, Support vector machines, Decision trees 등이 있음

● Logistic Regression

▷ Target 변수가 두 개의 집단(0 또는 1)으로 분류되어 있을 때(binary case) 가장 많이 사용하는 모델로, 로짓함수를 반응변수로 정의하고 로짓과 Input 변수들과의 관계를 선형함수로 표현한 모형

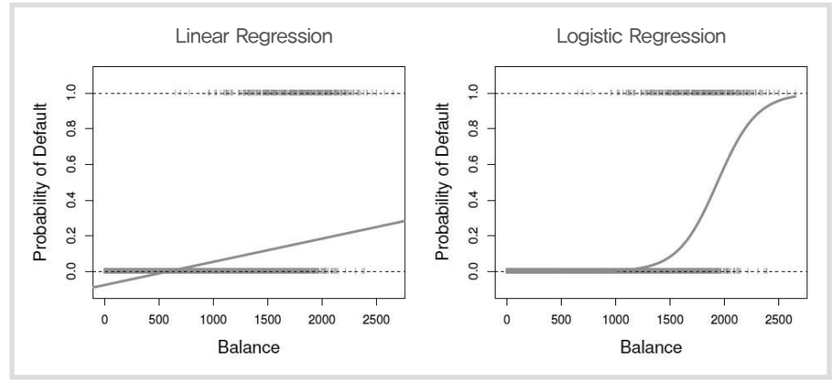
$$\log(p(X)/(1+p(X))) = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\epsilon.$$

– 집단 1에 속할 확률(prediction probability)을 구해 최종 target 변수 결정

$$p(X) = \Pr(Y=1|X) = \exp(\beta_0+\sum_{j=1}^{p}\beta_jX_j)/(1+\exp(\beta_0+\sum_{j=1}^{p}\beta_jX_j)).$$

– 확률함수 $p(X)$는 항상 0에서 1사이의 값을 가짐

[그림 7–3] Linear 및 Logistic regression의 확률함수



※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R"

● Linear Discriminant Analysis (LDA)

▷ 베이즈 정리(Bayes theorem)를 기반으로 사후확률을 계산하여, 주로 Target 변수가 세 개 이상의 집단으로 분류되어 있을 때 Input 변수들과의 관계를 설명하는 모형

▷ Input 변수 $X=(X_1,X_2,...,X_p)$가 평균 $\mu=(\mu_1,...,\mu_p)^t$, 공분산행렬 $\Sigma_{p\times p}$의 다변량 정규분포(multivariate normal distribution)를 따르고 각 집단별 분산이 같다고 가정했을 때, 선형판별식 $\delta_k(x)$가 가장 큰 집단 k를 선택하는 방식

$$\delta_k(x) = x^t\Sigma^{-1}\mu_k-\frac{1}{2}\mu_k^t\Sigma^{-1}\mu_k+\log(\pi_k).$$

▷ 집단들 사이의 차별적인 정보를 최대한 보존하면서 차원을 축소하는 기능 또한 있음

● K-Nearest Neighbors (KNN)

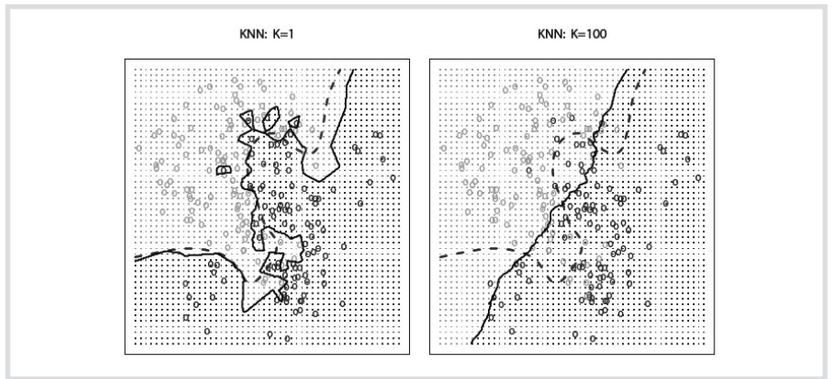
▷ 비모수적(non-parametric)인 방법으로, 사전에 K를 정의하고 가장 유사한 K개 관측치의 target 변수들로 각 집단이 나올 확률이 가장 큰 집단으로 분류하는 기법

$$\Pr(Y=j|X=x_0)=\frac{1}{K}\sum_{i\in N_0} I(y_i=j).$$

– '유사한 관측치끼리 비슷한 target 변수를 가질 것이다'는 아이디어로, x_0는 분류하기를 원하는 새로운 관측치이며 N_0는 x_0는 와 가장 유사한 K개의 관측치를 의미함

– K값이 작으면 너무 flexible하고, K값이 크면 선형에 가까워지는 특징이 있으므로 적당한 K를 정하는 게 중요

[그림 7-4] K-nearest neighbors



※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R"

● Naive Bayesian Classification (NB)

▷ 베이즈 정리(Bayes theorem)에 기반하여 많은 Input 변수들이 있을 때 단순화시켜 빠르고 쉽게 판단 내릴 때 사용하는 기법으로, 주로 문서 분류에 활용됨

$$\Pr(Y=k|X)=\frac{\Pr(X|Y=k)\cdot\Pr(Y=k)}{\Pr(X)},$$

$$\Pr(X|Y=k)=\Pr(X_1,\cdots,X_p|Y=k)=\Pr(X_1|Y=k)\cdot\cdots\cdot\Pr(X_p|Y=k).$$

– 각 변수들이 독립이라는 강한 가정 하에 시행되는 기법이지만, 실제로 변수들이 서로 독립이 아니어도 좋은 성과를 냄

● Support Vector Machines (SVM)

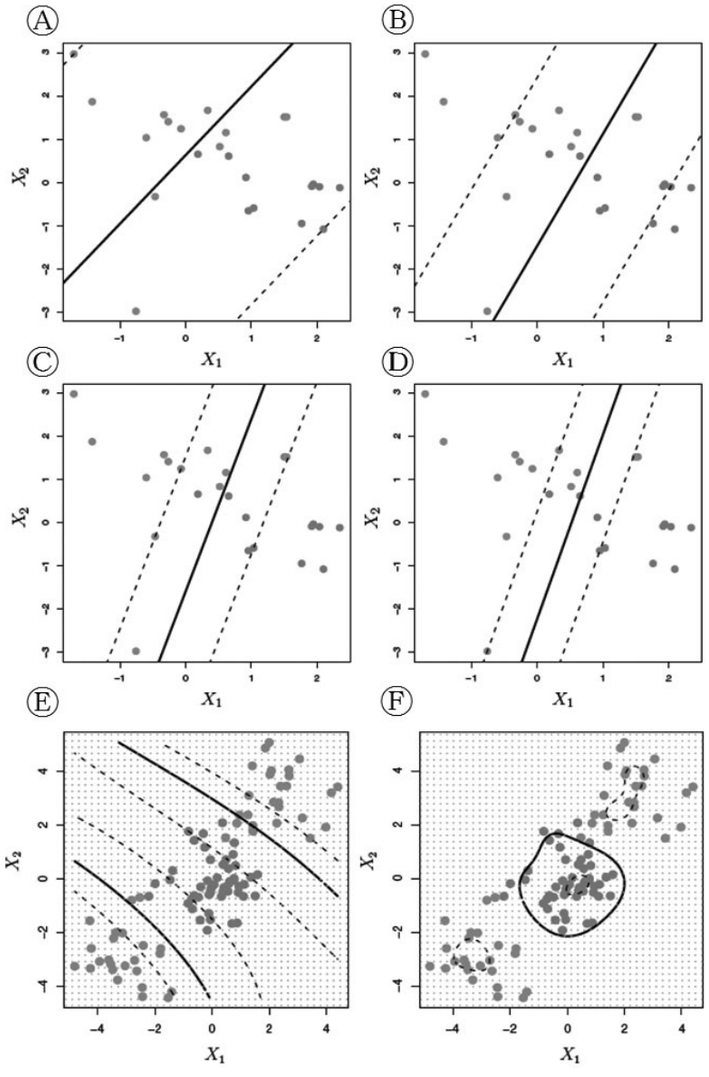
▷ 고차원의 공간에서도 데이터들을 잘 구별할 수 있도록, 약간의 오류를 허용하면서 데이터와의 직교거리인 margin을 최대화시키는 초평면(hyper-plain)을 결정하는 방법

▷ 약간의 오류를 얼마나 허락할 것인지를 설정해주는 tuning parameter 따라 [그림 7-5]처럼 두 그룹을 분류할 초평면이 다름

– Ⓐ, Ⓑ, Ⓒ, Ⓓ 순으로 잘못 분류될 오류의 한계치(tuning parameter)를 줄인 것으로, dashed line(margin) 안에 있는 support vector들에 의해서만 초평면이 결정됨

– 직선 형태의 초평면뿐만 아니라 Ⓔ, Ⓕ처럼 커널함수를 사용한 비선형의 초평면을 이용하여 간결하면서도 뛰어난 성능을 보이는 집단 분류방법임

[그림 7-5] Support vector machines

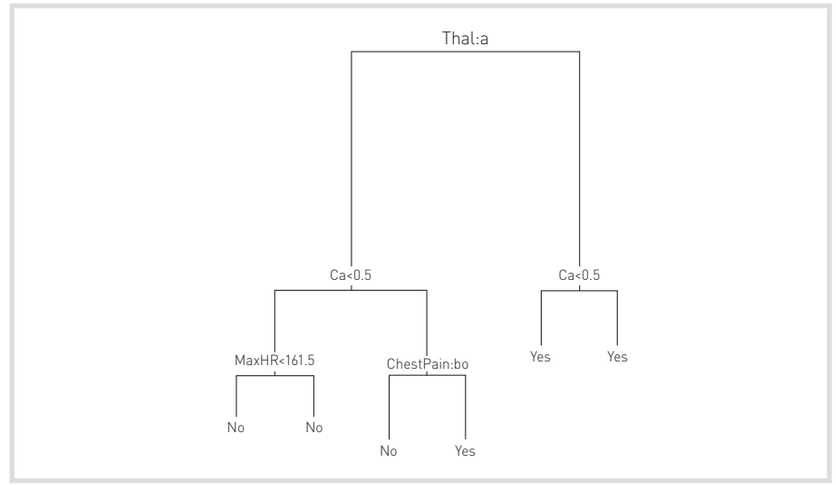


※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R", 한국주택금융공사

● Decision Trees (의사결정나무)

▷ 의사결정규칙(decision rule)을 나무구조로 도표화하여 분류하는 분석방법

[그림 7-6] Decision Trees



※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R"

▷ 간단하고 결과를 보면 누구나 해석이 용이하다는 장점이 있지만, 예측의 정확도가 다소 떨어지는 경향이 있음

- 의사결정나무의 예측력을 높이기 위한 방안으로 다음과 같은 방법들이 사용됨
- bagging : 표본의 수를 높이면 예측의 분산이 줄어드는 아이디어를 적용해, bootstrap을 이용하여 sample을 랜덤하게 뽑아 총 B번의 의사결정나무를 실행해 다수결에 의한 방식으로 B번의 결과 중 가장 많이 나온 집단으로 분류하는 방식
- random forests : sample 뿐만 아니라, 변수들도 랜덤하게 뽑는 방식으로 Input 변수 모두를 사용하면 bagging과 동일
- boosting : B번의 bootstrap sample을 독립적으로 모델링하는 bagging과 random forests과 달리, 한 번 시행한 모델의 결과를 참조하여 순차적으로 B번 update시키는 방법

3. 비지도학습 (Unsupervised Learning)

- 비지도학습은 정답이 주어지지 않은 자료 자체만으로 특징을 발견하여 그룹을 식별하는 탐색적 자료 분석으로, 크게 Clustering과 차원축소가 있음

- Clustering이란 유사성의 개념에 기초하여 각 관측치를 몇 개의 그룹으로 분류하는 분석 방법론

  ▷ 가장 대표적인 Clustering 기법인 K-means clustering(비계층적 군집분석)과 Hierarchical clustering(계층적 군집분석)에 대하여 간략히 소개하겠음

- K-means clustering

  ▷ 비계층적 군집분석인 K-means clustering은 사전에 몇 개의 그룹으로 나눌 것인지 정하고, 관측치들 사이의 거리를 이용해 가까운 관측치들끼리 K개의 집단으로 군집을 형성하는 방법

  – 사전에 정의해야 하는 집단(초기값)에 큰 영향을 받는다는 단점이 있음

▷ K-means clustering 알고리즘

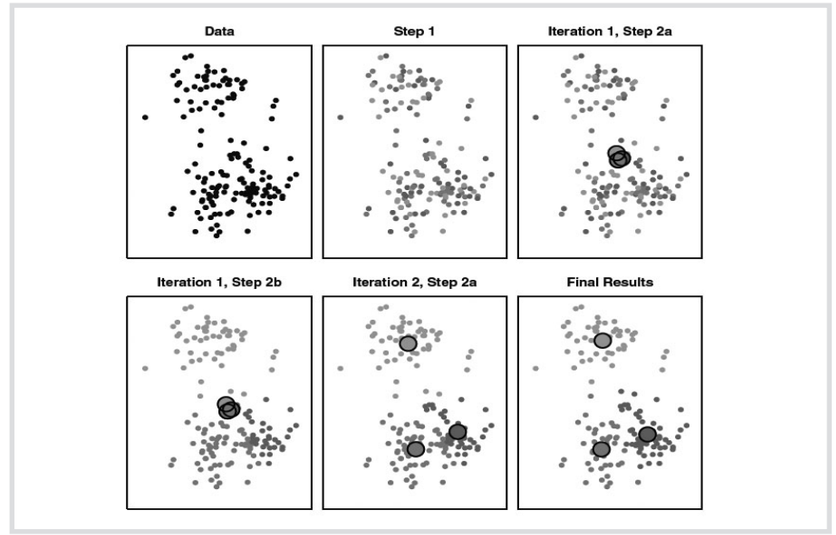
– Step1. 몇 개의 집단으로 나눌 것인지 임의로 K를 정하고, 각 관측치를 랜덤하게 1부터 K의 집단으로 분류함

– Step2. (a) K개의 집단별로 centroid(mean, median, maximum, minimum 등)를 구함

– (b) 각 관측치를 가장 가까운 centroid의 집단으로 재분류

– 각 관측치의 그룹 할당이 더 이상 변하지 않을 때까지 또는 사용자가 정의한 허용오차나 최대 반복 횟수 등에 다다를 때까지 Step2의 (a), (b) 반복

[그림 7-7] K-means clustering



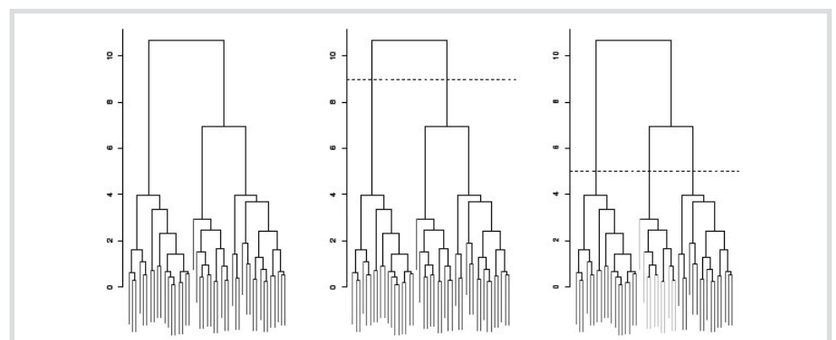
※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R"

- Hierarchical clustering

  ▷ 계층적 군집분석으로, 데이터 간의 유사성을 측정해 가장 유사한 관측치들부터 차례로 군집을 형성하는 방법

  – 초기값에 영향을 받지 않고, 의사결정나무와 유사한 나무구조의 도표(dendrogram)로 결과를 도출해 해석이 용이함

[그림 7-8] Hierarchical clustering



※ 자료 : James et al. (2013), "An Introduction to Statistical Learning with Applications in R"

● 차원축소기법이란 기존 데이터의 정보를 유지하면서 데이터의 차원을 축소하는 분석 방법론

▷ 일부 변수들끼리 상관관계가 높거나 관심 있는 변수(Target 변수)와 관련이 없는 변수를 포함시킬 경우, 모형의 정확도와 비용 측면에서 좋지 않음

▷ 비지도학습의 대표적인 차원 축소 기법(Unsupervised dimensionality reduction)으로는 주성분분석과 요인분석이 있음

● 주성분분석 (Principle Components Analysis, PCA)

▷ 고차원의 데이터를 저차원의 데이터로 환원시키는 가장 대표적인 방법으로, 분산에 초점을 두고 변수들을 선형결합 해 변수를 축약하는 방법

– 전체 변동을 가장 잘 설명하는 축을 첫 번째 주성분으로, 이와 직교하면서 두 번째로 분산이 큰 축을 두 번째 주성분으로 두는 방식

– 전체 변수의 분산을 재생성하며 각 주성분은 변수의 공통적인 특성과 고유한 특성 모두를 반영함

● 요인분석 (Factor Analysis, FA)

▷ 변수들의 선형결합을 이용한다는 점은 주성분분석과 유사하지만, 변수들의 상호관계에 초점을 두고 상관이 높은 변수들끼리 묶어 요인으로 정하고 이에 포함되지 않거나 중요도가 낮은 변수들을 제거하는 방법

– 변수들의 공통된 분산은 반영하지만 고유의 분산은 배제하는 특성이 있음

III 기계학습의 응용

● 최근 주목받는 기계학습 방법론은 지금까지 살펴본 지도학습 및 비지도학습 방법론을 토대로 빅데이터 분석이 가능하도록 발전된 형태

▷ 강력한 계산능력(computing power)을 바탕으로 기존 방법으로는 분석이 불가능한 이미지 분석, 자연어 처리, 사용자 개개인의 마이크로 분석 등의 빅데이터 분석 영역으로 확장

▷ 대표적인 방법론과 함께, 주요 응용 분야와 공사 활용 방안을 알아보고자 함

1. 딥러닝 (Deep Learning)

● 신경세포(Neuron) vs 인공뉴런(Artificial Neuron)

▷ 신경세포(그림 7-9, A)는 두뇌와 척수를 구성하는 기본 단위

– 각 뉴런은 신호를 받아 다른 연결된 뉴런들로 전달

– 전달받은 신호가 일정 강도(역치, threshold) 이상일 때만 다른 뉴런으로 전달

▷ 인공뉴런(그림 7-9, B)은 신경세포의 역할을 모사한 수학적 모형

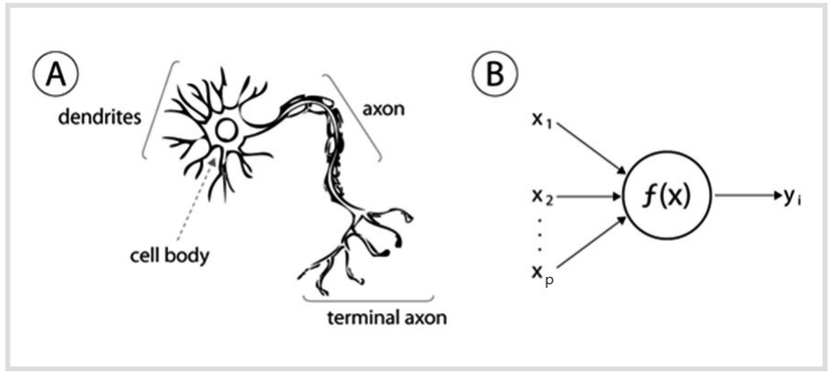
– 각 인공뉴런은 0 또는 1로 구성된 x_1, x_2, ⋯, x_p의 신호를 입력받아 각각의 가중치 w_1, w_2, ⋯, w_p에 따른 가중합계를 계산

– 과거 경험(데이터)을 바탕으로 가중치와 역치 결정

– 가중합계가 일정 강도(역치, threshold) 이상일 때 다른 뉴런으로 신호를 전달

$$y_i = output = \begin{cases} 0 & \text{if } \sum_j w_j x_j \le threshold \\ 1 & \text{if } \sum_j w_j x_j > threshold \end{cases}$$

[그림 7-9] 신경세포(Neuron) vs 인공뉴런(Artificial Neuron)



※ 자료 : Maltarollo et al. (2013), "Applications of Artificial Neural Networks in Chemical Problems"

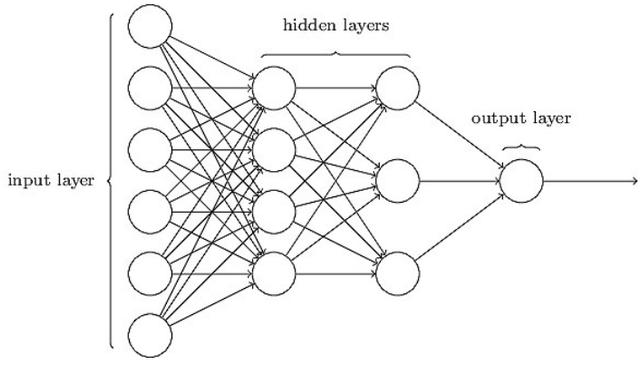
● 인공신경망(Artificial Neural Network, ANN)

▷ 단순한 구조의 인공뉴런들은 연결, 최적화를 통해 복잡한 문제 해결

▷ 입력신호와 출력신호 사이에 여러 층의 인공뉴런을 배치하고, 이를 입력계층(input layer), 은닉계층(hidden layer), 출력계층(output layer)으로 구분 [그림 7-10]

▷ 각 인공뉴런마다 가중치(w_1, w_2, ⋯, w_p)와 역치(threshold) 조절을 통한 최적화가 필요하므로 강력한 컴퓨팅 파워가 요구됨

[그림 7-10] 인공신경망 구조도



※ 자료 : Nielsen (2015), "Neural Networks and Deep Learning"

● 딥러닝(Deep Learning)

▷ 인공신경망 가운데 은닉계층이 여러 개 중첩되어 있는 것을 심층신경망(Deep Neural Network)이라 하며, 이러한 심층신경망을 사용한 기계학습을 딥러닝(Deep Learning)이라 지칭함

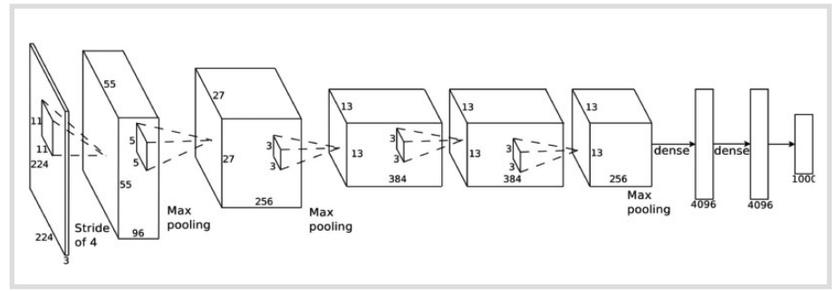
– 기존에 사용된 인공신경망의 국소 최적치(local minimum) 문제를 극복하기 위해 모든 작동을 세세하게 프로그래밍(rule-base)하지 않고 주어진 데이터로부터 지속적으로 학습하여 특정 판단에 적용하는 기계학습의 한 부분

● 딥러닝은 기존 기계학습 방법론에서 발생하는 과적합(over-fitting) 문제를 해결하기 위하여 Convolutional Neural Network(CNN) 방법론과 Recurrent Neural Network(RNN) 방법론을 사용

▷ convolutional neural network는 주어진 정보를 여러 영역으로 나누어 제한된 정보를 통해 관계성 높은 영역끼리 분석하는 방법으로, 이미지 인식 분야에서 특히 높은 성과를 나타냄

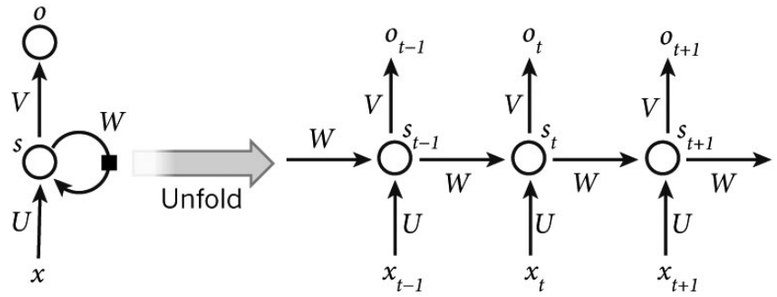
▷ recurrent neural network는 모든 입력과 출력이 각각 독립적이라고 가정하지 않고 순차적으로 처리하는 방법으로, 시계열 자료나 문맥 인식 등에 사용됨

[그림 7-11] Convolutional Neural Network 개요



※ 자료 : Smirnov et al. (2014), "Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks"

[그림 7-12] Recurrent Neural Network 개요



※ 자료 : LeCun, et al. (2015), "Deep learning"

● 응용 분야

▷ 이미지 및 영상 인식

– 딥러닝은 이미지 및 영상 인식에서 두각을 드러내며, 인식된 정보를 바탕으로 자동주행, 영상분류, 자연어 처리 등 다양한 분야에서 활용

▷ IBM 왓슨, 구글 알파고 등

– 정보의 맥락(context)에 대한 이해를 바탕으로, 기존 기계학습에서는 불가능했던 인간의 창의성이나 직관력을 흉내 내는 수준에 도달

– 알파고의 CNN(Convolutional Neural Network)은 다음 돌을 놓을 위치를 선택하는 정책망(policy network)과, 해당 위치에 돌을 놓았을 때 승리 확률을 예측하는 가치망(value network)이라는 2개의 심층신경망을 활용

– 몬테카를로 트리 검색[1]을 통해 정책망으로 다음 착점을 고려하고, 가치망으로 승률을 계산하여 가장 승률이 높은 곳에 착수

● 공사 활용 방안

▷ 공사 주택담보대출 연체율 분석 및 예측·관리

– 신용평가사에서는 이미 딥러닝을 활용한 신용평가모형을 개발 중에 있으며, 설명력과 예측력도 우수한 것으로 나타남

– 공사 보유 주택담보대출의 정보와 함께 대출상환현금흐름 정보를 분석하여 대출조건·차입자·담보주택 특징에 따른 연체율 예측 모형 개발을 통해 연체율 관리는 물론, 취약계층 대상의 상품 개발로 확대 가능

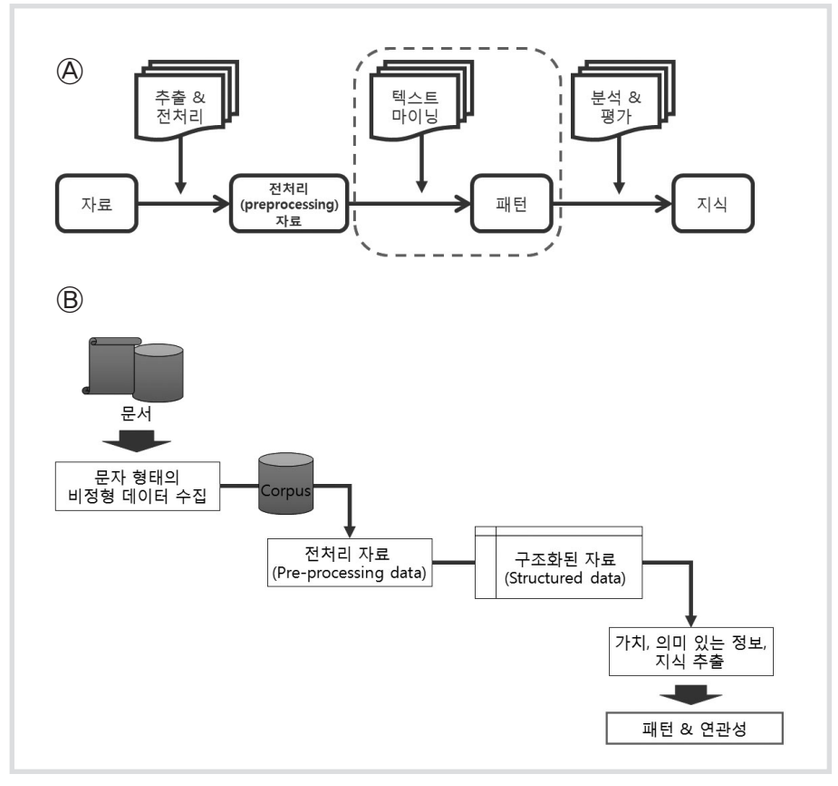
1) 무작위 대입을 통해 예상 확률을 알아낸 뒤 가장 가능성이 높은 수를 선택하는 기법

2. 텍스트 마이닝 (Text Mining)

● 방대한 양의 비정형화 데이터(텍스트)를 자연어 처리, 문서 처리, 통계방법론 기술 등을 적용하여 가치 있는 정보를 추출하는 방법

▷ 보고서, 논문, 신문, SNS 등 정형화 되지 않은 자료이고 그 양이 방대하여 사람이 분석하기 힘든 문서형태의 자료를 대상으로 함

[그림 7-13] 텍스트 마이닝 개요



※ 자료 : 한국주택금융공사

● 텍스트 마이닝의 핵심은 정성적 텍스트를 벡터화(vectorization)하여 정량적 분석이 가능한 자료형태로 가공하는 기술

▷ 수많은 문서 형태의 자료로부터 전처리과정(preprocessing)을 거쳐 정량적 분석이 가능한 형태의 자료로 만들고 Naive bayesian classification, Support vector machines, 의사결정나무 등의 통계적 분석 방법론과 결합하여 사용됨

▷ 단어 및 문맥의 출현 횟수를 세는 count based method, 단어에서 문맥 또는 문맥에서 단어를 예측하는 predictive method 등을 적용

▷ 다음 [그림 7-14]과 [표 7-1]은 한국주택금융공사 홈페이지의 주택보증 관련 상담 문의 게시글을 추출하여 단어별 빈도수를 워드 클라우드로 시각화한 결과

[그림 7-14] 워드 클라우드 예시

※ 자료 : 한국주택금융공사

▷ 또한 [표 7-1]은 [그림 7-14]과 동일한 자료를 이용하여 기준 단어와 같이 사용될 가능성이 높은 연관어를 분석한 결과

– [표 7-1]의 연관어 괄호 안의 숫자는 기준단어와 같이 사용될 확률 (단위: %)

[표 7-1] 연관어 분석 결과 예시

기준단어	연관어
한국주택금융공사	방문(51), 상품(51), 신용보증서(27) 등
전세자금보증	전환대출(33), 필요서류(33) 등
대출	전세자금(36), 불가(23), 신혼부부(23), 수령(21) 등
한도	3억(28), 개인별(28), 어려움(28), 한시적(28), 획일적(28) 등

※ 자료 : 한국주택금융공사

● 응용 분야

▷ 고객 관리 서비스

– 소셜 미디어 데이터 분석을 통하여 시장 및 고객 정보를 파악하고 해당 브랜드나 제품에 대한 다양한 의견과 감성반응 모니터링

– 설문조사나 게시판 분석 등을 통해 문제해결의 효용성 및 속도 개선

– 보험사는 텍스트 분석을 통해 사기를 방지하고 클레임을 빠르게 처리

▷ 기업 내부 Smart Search

– 기업 내부에서 생산된 문서와 외부의 전문가 의견을 분석하여 컨텐츠 자동분류, 관심정보 추천, 용어 분석 등을 실시

– 검색 소요시간 80% 감소, 업무 리드타임 40% 단축

▷ 연구 논문의 주요 내용과 추세 분석

– 논문 초록 정보에 텍스트 마이닝 기술을 적용하여 최신 연구 트렌드와 관심사항을 요약 및 분석

● 공사 활용 방안

▷ 정책보도자료 분석을 통한 국정 방향 탐색

– 주택시장 및 주택금융 관련 정책보도자료 분석을 통해 정책 관심사의 변화를 진단하고 향후 관심사 예측에 보조 자료로 사용

▷ 수요조사 등에서 정성적 평가를 정량적 평가지표로 치환

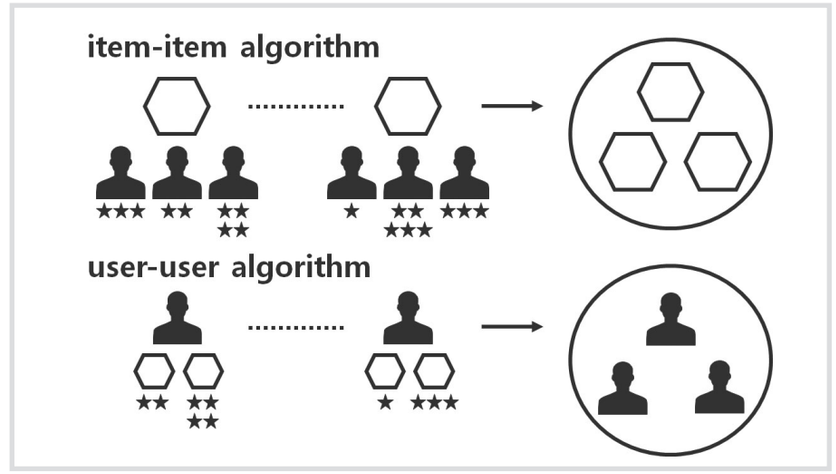
– 고객이 주택금융공사에 바라는 점 등 정성적 조사내용을 정량적으로 분석하여 상품에 대한 인식과 변화를 확인

3. 협업 필터링 (Collaborative Filtering)

- 다수의 사람들의 행동 정보를 이용하여 마케팅의 타겟인 사용자와 비슷한 성향의 사용자들이 좋게 평가한 상품을 추천해주는 맞춤형 서비스
  - ▷ 사용자의 선택과 평가 정보를 분석함으로써 취향과 기호를 추론하여 관심 가질만한 새로운 것을 추천해줌
  - ▷ 시공간 제약이 약한 온라인 시장에서는 개인화된 서비스 제공을 통해 다양한 사용자들이 구매할 확률이 높은 상품을 찾아내는 것이 중요
    - – 오프라인 시장에서는 전통적인 베스트셀러 추천 방식이 적합

- 협업 필터링의 핵심은 평가 자료를 바탕으로 사용자 간 혹은 상품 간의 유사도(Similarity)를 파악하여 추천에 의한 선택 가능성을 증대시키는 것
  - ▷ 사용자기반이든 상품기반이든 상품에 대한 선호도를 기반으로 유사성 계산
    - – 사용자가 상품을 얼마나 자주 구입하는지, 선호도 평가 값, 상품 클릭 횟수 등을 모두 고려함
  - ▷ 비슷한 사용자가 선호하는 상품을 추천해주는 user-based recommendation, 구매했던 상품과 연관성 있는 상품을 추천하는 item-based recommendation 등이 있음
  - ▷ 사용자의 평가가 필요하지 않고 함께 소비된 상품을 분석할 때에는 관련성 분석(affinity analysis) 방법론 사용

[그림 7-15] 협업 필터링 개요



※ 자료 : 한국주택금융공사

- 응용 분야
  - ▷ 넷플릭스, 구글 뉴스, 페이스북 등
    - – 사용자의 서비스 사용 패턴을 분석하여 적절한 상품, 기사, 친구 등을 추천함으로써 사용자의 선택을 유도
    - – 실질적인 상업적 가치가 있는 것으로 보고되어 많은 연구 개발 진행
  - ▷ 아마존은 협업 필터링을 활용하여 예측배송 실행
    - – 과거 구매 기록, 상품 검색 이력, 위시 리스트, 상품 카트, 상품 조회 시간 등 수집된 모든 데이터 반영
    - – 사용자의 구매 패턴을 분석하여 구매가 예상되는 물품을 미리 물류창고로 이동시켜 배송시간 단축

- 공사 활용 방안
  - ▷ 연구보고서 수요계층 분석
    - – 주택금융연구원 홈페이지에 접속한 IP 주소를 기준으로, 함께 조회되는 연구보고서를 파악하여 주제에 따라 관심 수요계층 분류
    - – 함께 조회되는 분야에 대한 복합 연구 주제 검토

Ⅳ 데이터 과학자 (Data Scientist)

1. 정의 및 역할

- 데이터 과학자란, “빅데이터 세상에서 새로운 것을 발견하는데 도움이 되는 전문 지식과 풍부한 호기심을 갖고 있는 중요 전문가” (Thomas et al., 2012)

  ▷ 현장에서 발생하는 대용량 데이터를 분석하여 숨은 트렌드를 발견하고 기업에 필요한 정보를 제시하는 역할

- 최근 기업들이 데이터 주도적 통찰력(data-driven insight)에 대한 이해가 깊어짐에 따라 해당 역할에 대한 수요 증가

2. 필요 역량

- 데이터 과학자는 전문분야의 지식뿐만 아니라 수학·통계학·컴퓨터 프로그래밍·데이터베이스·커뮤니케이션 등 많은 능력이 요구됨

- 수학 및 통계학 지식

  ▷ 데이터 분석을 위한 통상의 통계학 지식은 물론, 기계학습 등의 최신 분석 트렌드를 이해하기 위해서는 선형대수 등의 수학적 지식 또한 필요

- 프로그래밍 능력 및 데이터베이스에 관한 이해

  ▷ 수학 및 통계학 지식을 바탕으로, 이를 구현할 수 있는 통계 패키지 활용 능력과 프로그래밍 능력 필요

  ▷ 데이터의 규모가 커질수록 데이터베이스와 분산처리에 관한 이해 필요

- 도메인 지식 및 소프트 스킬

  ▷ 전문 분야에 대한 지식과 함께 데이터가 발생하는 현장에서 통용되는 암묵지(暗默知)에 관한 이해 필요

- 커뮤니케이션 및 시각화

  ▷ 데이터 주도적 통찰을 고위 경영진과 실무진에 효율적으로 전달하기 위하여 데이터 시각화 및 커뮤니케이션 능력 요구

[표 7-2] 데이터 과학자의 필요 역량

수학 및 통계	프로그래밍 및 데이터베이스
- 기계 학습 - 통계 모형 - 실험 설계 - 베이즈 추론 (Bayesian inference) - 최적화 방법론 - 지도학습: 의사 결정 나무, 로지스틱 회귀 등 - 비지도학습: 클러스터링, 차원 축소 등	- 기초 컴퓨터 과학 기초 - 스크립팅 언어 (예: 파이썬) - 통계 패키지 (예 : R) - 데이터베이스 (SQL 및 NoSQL) - 관계대수 (relational algebra) - 병렬 데이터베이스 및 병렬 쿼리 처리 - MapReduce 개념
도메인 지식 및 소프트 스킬	**커뮤니케이션 및 시각화**
- 전문 분야에 관한 이해 - 데이터에 대한 호기심 - 문제 해결 능력 - 전략적, 적극적, 창의적, 혁신적, 협업적	- 고위 경영진 및 실무진과의 의사소통 - 스토리텔링 기술 - 데이터에 기반한 직관으로 의사결정 조언 - 데이터 시각화 및 디자인

※ 자료 : 한국주택금융공사

V 결론 및 시사점

- 빅데이터 분석은 최근 다양한 분야에서 관심을 받고 있으나, 각각의 방법론에 따라 활용 분야와 범위가 다름
  - ▷ 앞에서 살펴본 빅데이터 분석 방법론에 대한 이해를 바탕으로, 방법론의 활용 범위와 한계를 명확하게 인지하여 빅데이터 만능론에 대한 경계 필요
  - ▷ 조직 내 데이터 주도적 통찰력을 높이기 위해서는 데이터 분석 결과와 현장의 암묵지를 연결하는 원활한 커뮤니케이션 필요

- 데이터 분석과 활용을 높이기 위하여 데이터 중심적 의사결정과 조직구조를 갖춘 '데이터 거버넌스'에 주목
  - ▷ 맥킨지 조사에 따르면 데이터 기반 조직은 23배 더 많은 고객을 확보하고, 6배 더 유지하며, 19배 더 많은 수익률을 얻는 것으로 나타남
  - ▷ 데이터 중심적 조직문화 양성을 위해서는 ① 보유한 데이터의 목록 정리와 품질 관리, ② 조직 구성원의 데이터 활용 능력 육성, ③ 필요 즉시 적절한 인사이트를 줄 수 있는 통계치 작성 능력 등이 필요

- 데이터 과학자 육성을 통한 실사구시(實事求是) 조직문화
  - ▷ 조직 내 데이터 과학자 육성을 위하여 필요한 역량을 확인하고, 개개인의 부족한 부분을 채워나갈 수 있는 교육 프로그램 마련이 요구됨
  - ▷ 현장의 경험과 아이디어를 실제로 유용한 분석 결과로 이끌어낼 수 있는 통로를 마련하여, 데이터 분석을 토대로 의사결정을 뒷받침하는 빅데이터 기반의 실사구시(實事求是) 조직문화 형성

참고문헌

1. James et al. (2013), "An Introduction to Statistical Learning with Applications in R".
2. Maltarollo et al. (2013), "Applications of Artificial Neural Networks in Chemical Problems".
3. Nielsen (2015), "Neural Networks and Deep Learning".
4. Smirnov et al. (2014), "Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks", AASRI Procedia, (6), 89–94.
5. LeCun, et al. (2015), "Deep learning", Nature, 521(7553), 436–444.
6. Thomas et al. (2012), "Data Scientist: The Sexiest Job of the 21st Century", Harvard Business Review.