밑바닥부터 시작하는 데이터 과학

1회차

- 데이터 과학이란 ?
- 파이썬 실습
- 데이터 시각화
- Pandas

1. 데이터 과학이란?

데이터 과학이란

- 데이터를 수집하고 분석하여 활용하기 위한 모든 기술의 집합을 말하며,
- **컴퓨터 사이언스, 수학, 통계학, 머신 러닝(machine learning)**, 영상 및 신호 처리 등 다양한 학문 분야가 만나는 영역이다.
- 프로그래밍을 포함하는 **컴퓨터 사이언스**는 실제로 데이터를 다루기 위한 필수 기술이며,
- **수학과 통계학**은 데이터 분석 모형의 기반에 깔린 핵심적인 개념을 구체화하는 언어이다.
- **머신 러닝**은 이러한 분석 결과를 활용하여 지금까지 인간이 해오던 각종 분석과 의사 판단을 대신하고자 하는 노력이다.

2. 데이터 과학 학습

- **기초 수학 이론**
  - 선형대수
  - 미분과 적분
  - 최적화
  - 확률론

- **데이터 분석 이론**
  - 확률 모형
  - 검정 및 추정
  - 회귀 분석과 분류, 클러스터링

- **컴퓨터 관리 및 프로그래밍 기술**
  - 리눅스 운영체제 사용법
  - 프로그래밍 언어
  - 데이터베이스 시스템
  - 병렬처리, 가상화, 클라우드 사용법

- **해당 분야에 대한 전문 지식**
  - 해당 분야의 정보를 이해하고 분석 결과가 올바른지 판단할 수 있는 능력
  - 이미지 처리, 음성/음향 처리, 텍스트 처리 등의 자료 전처리 기술

3. 데이터 과학 활용

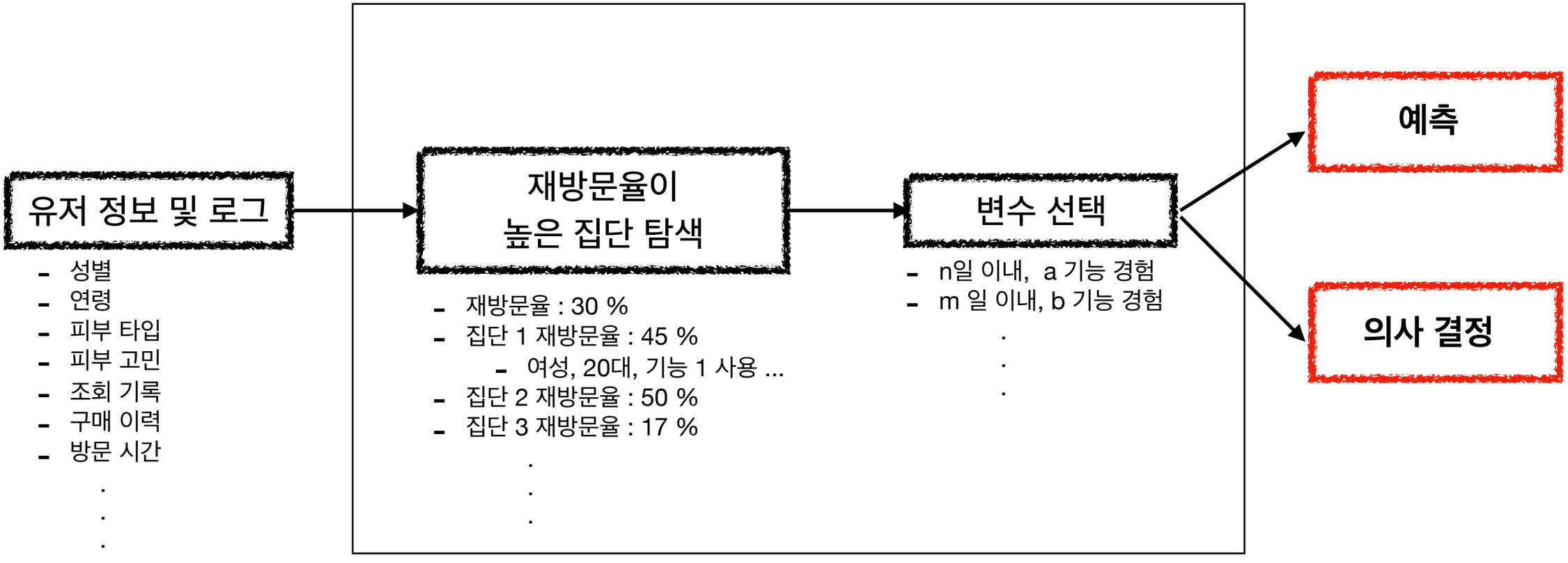
이 많은 방법론들을 공부하면 우리는 무엇을 할 수 있을 까요?

3. 데이터 과학 활용

수학 및 통계 분석을 공부하면,

수 많은 데이터 속에서 꼭 필요한 데이터만 선택할 수 있게 됩니다.

비즈니스 활용 사례 (재방문율 분석)

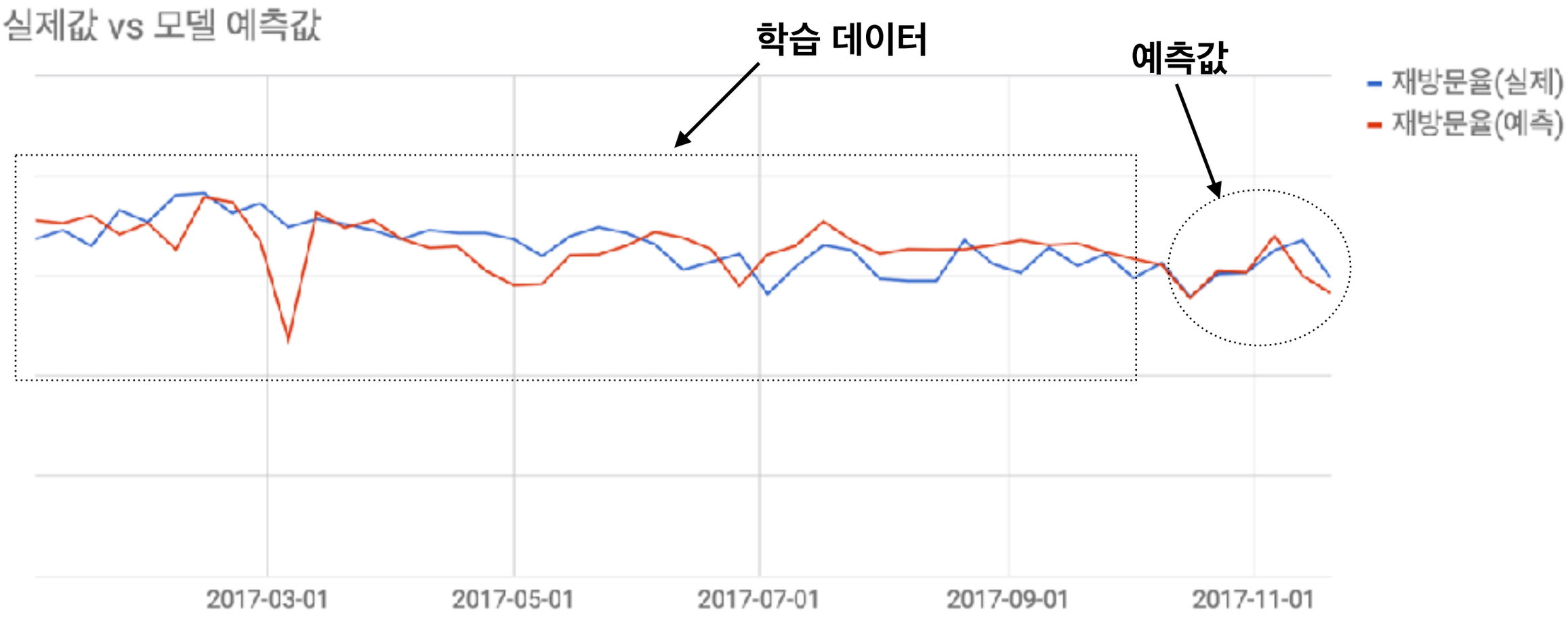


3. 데이터 과학 활용

예측 모델을 공부하면,

수 많은 (과거) 데이터를 바탕으로 비어있는 값 (미래)를 예측할 수 있게 됩니다.

비즈니스 활용 사례 (재방문율 예측)



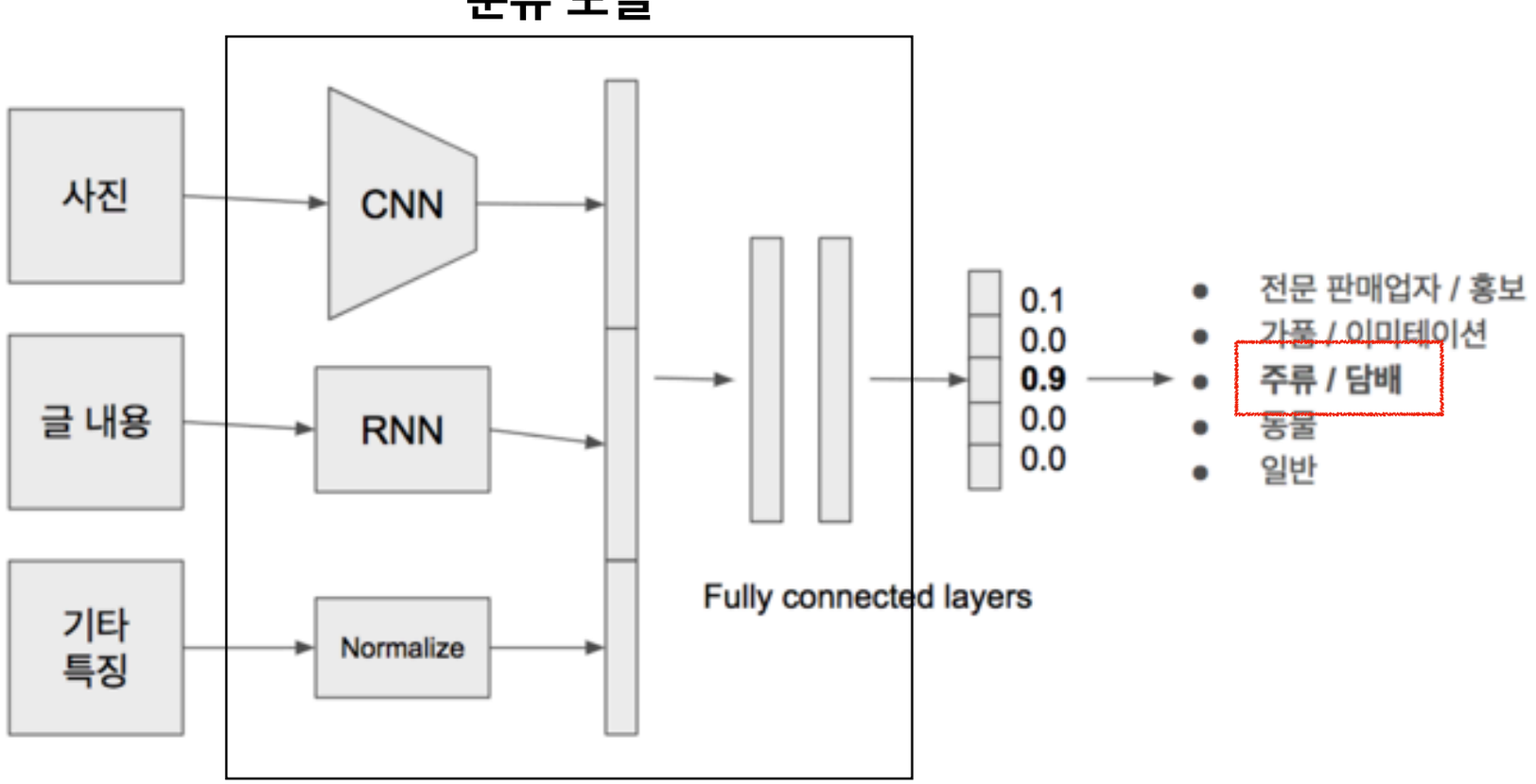
3. 데이터 과학 활용

분류 모델을 공부하면,

수 많은 데이터를 바탕으로 **패턴**을 찾아서, 어떠한 값으로 분류할 수 있게 됩니다.

- **대부분의 머신 러닝 / 딥러닝은 바로 이 분류 문제를 풀기 위한 방법이다.**

비즈니스 활용 사례 (게시글 분류)



게시글 분류 모델 네트워트

출처 : https://medium.com/n42-corp/당근마켓에서-딥러닝-활용하기-3b48844eba62

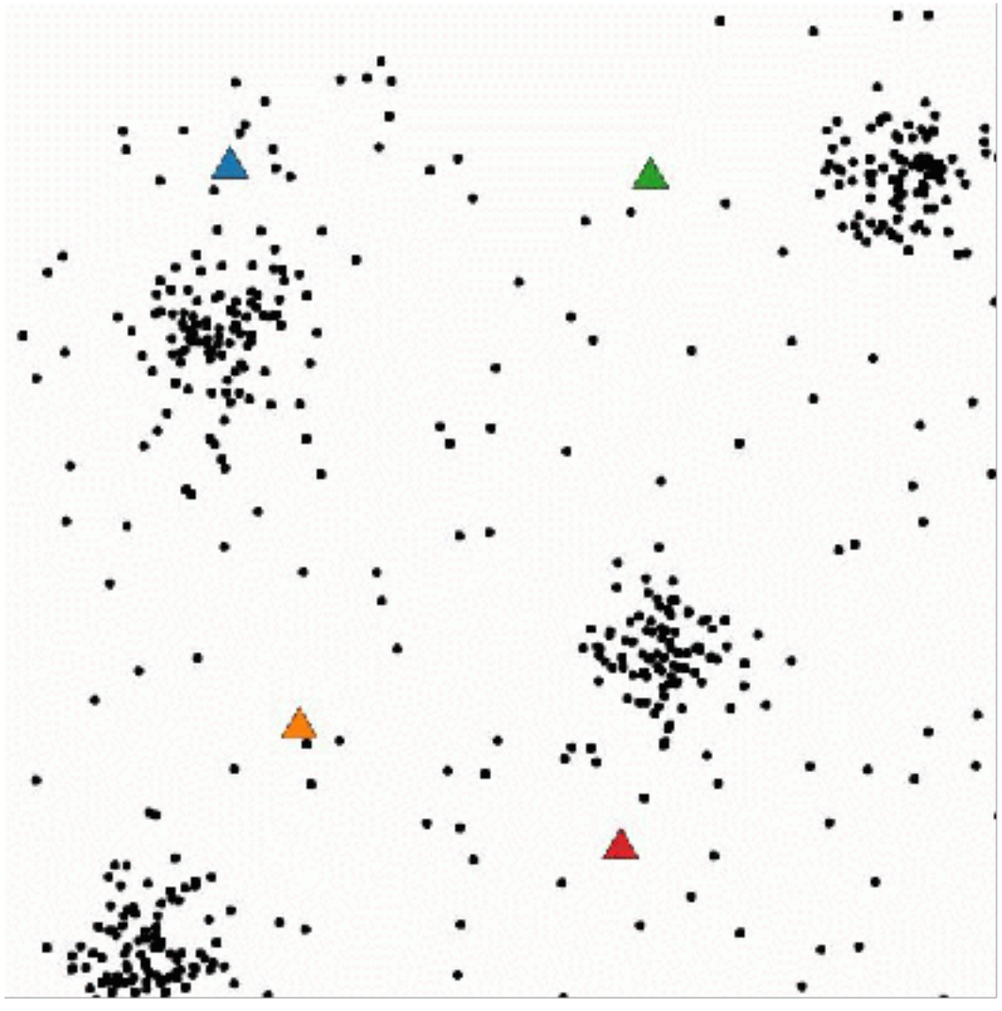
3. 데이터 과학 활용

클러스터링을 공부하면,

수 많은 데이터들을 몇 개의 그룹으로 묶을 수 있습니다.

게임 유저 클러스터링 분석

k-means 클러스터링 알고리즘이 클러스터를 나눠가는 과정



(출처 : 위키피디아 Incheol, CC BY-SA 4.0)

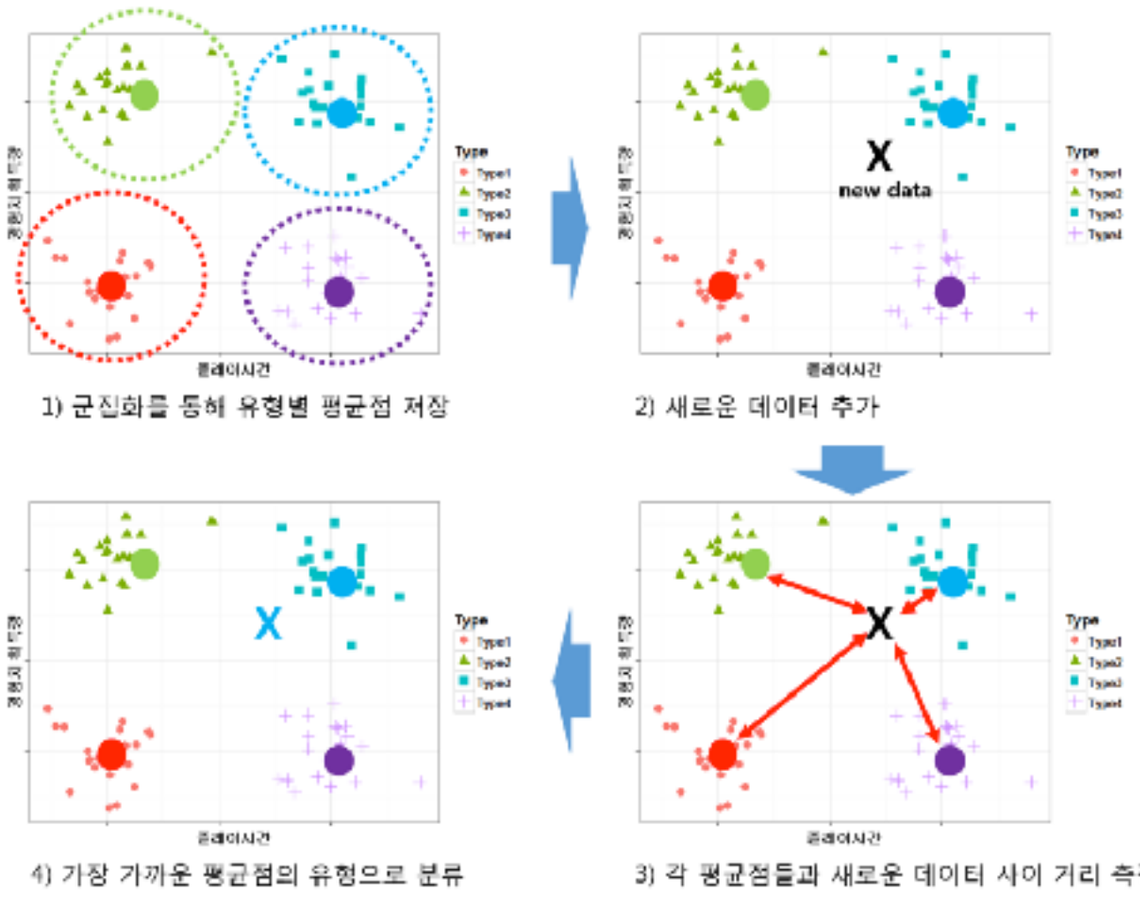
3. 데이터 과학 활용

클러스터링을 공부하면,

수 많은 데이터들을 몇 개의 그룹으로 묶을 수 있습니다.

게임 유저 클러스터링 분석

클러스터별 차이를 분석해,
의사결정에 도움을 준다.



k 평균 군집화로 유형을 분류하는 과정

(출처 : http://blog.ncsoft.com/?p=25333

4. 스터디 소개 - 목표 및 방향

우리는

기초 수학, 통계, 머신러닝 알고리즘까지 직접 코드를 작성하며 이해할 것입니다.

그래서

한 분도 빠짐없이 이론에 기반한 코드를 작성하는 실습 위주로 진행할 것입니다.

스터디가 종료된 후,

데이터 과학을 어떻게 활용할 수 있을지, 쉽게 상상할 수 있습니다 !

그리고, 앞으로 채워나가야할 부분도 좀 더 명확해질 것입니다.

5. 스터디 소개 - 계획

1 회
- 데이터 과학 소개
- Python 속성 강좌
- 데이터 시각화
- Pandas

2 회
- 선형대수
- 최적화
- 차원축소
- 회귀분석 기초

3 회
- 통계와 확률
- 가설과 추론
- 나이브베이즈
- 회귀분석 심화

4 회
- 기계학습 기초
- k-NN
- 로지스틱 회귀분석

5 회
- 군집화
- 의사결정나무
- 모형결합

6 회
- 신경망
- 자연어처리
- 추천 시스템

마지막 시간에는,

풀고 싶은 문제를 정의하고,

그 문제를 풀기 위한 데이터 과학 방법론들을 충분히 상상하고,

데이터 과학을 활용해 꼭 풀수있게 되길 바랍니다.

6. Github

https://github.com/surprisoh/datascience_scratch_2/

7. 실습 환경

Google Colaboratory

1) 구글 계정 생성

2) 크롬 다운로드

3) Colaboratory 접속

- https://colab.research.google.com/

4) 1회차 자료 접속

- https://colab.research.google.com/github/{파일이름}

7. 실습 환경

Jupyter Notebook

1) Anaconda 다운로드

- https://www.anaconda.com/download/

2) Jupyter Notebook 실행

- 공유 자료 참고

https://github.com/surprisoh/datascience_scratch_2/