

머신러닝의 이해



파이썬 기반 머신러닝 이해하기



빅데이터 정의

서버 한 대로 처리할 수 없는 규모의 데이터
기존의 소프트웨어로는 처리할 수 없는 규모의
데이터



빅데이터 정의

 3V(Value, Velocity, Variety)

주로 컨설팅회사들이 많이 사용하는 정의

Value(규모)

데이터의 크기가 대용량인가

Velocity(속도)

데이터가 얼마나 빠르게 생성되는가

Variety(다양성)

데이터가 구조화/비구조화된
데이터를 다 포함하는지 여부

빅데이터의 예

웹 검색엔진 데이터

- 웹페이지 데이터
- 검색어 로그와 클릭 로그 데이터

디바이스에서 생성되는 데이터

스마트폰, 스마트 tv, 보잉 제트기(매
30분마다 10TB 데이터 생성), 스마트 미터

소셜미디어의 데이터

페이스북, 트위터, 링크드인, 포스퀘어
등

빅데이터 시스템의 구성



빅데이터 처리

흐름

데이터 소스

수집

저장

분석

표현

내부데이터



외부데이터



로그
수집기

데이터
Integration

웹로봇

RSS
Feed

Open
API

배치처리

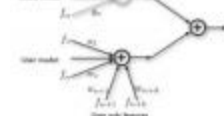
점점

실시간&배치

분산스토리



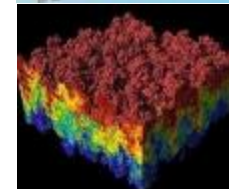
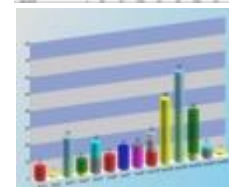
분석알고리
즘



스크립트엔
진

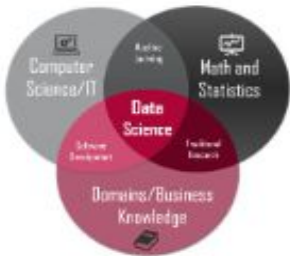

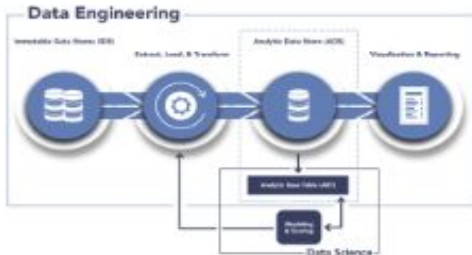



분산병렬처
리



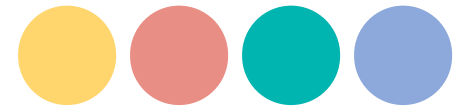
데이터 사이언스 / 데이터 엔지니어링

🎯 데이터 사이언스와 데이터 엔지니어링의 차이

		사용 툴	커리어패스
데이터 사이언스	 <p>‘도메인 빅데이터’에 ‘통계’와 ‘IT기술’을 접목해 문제를 해결</p>		<ul style="list-style-type: none">Chief Data ScientistSenior Data ScientistData ScientistJunior Data Scientist
데이터 엔지니어링	 <p>‘대규모의 데이터’를 효율적으로 ‘관리’하고 ‘처리’하여 분석에 적합한 형태로 가공</p>		<ul style="list-style-type: none">Data ArchitectBI ArchitectSenior Data EngineerData Engineer

※ 출처 <https://www.analyticsvidhya.com>
Copyright by Mullicampus Co., Ltd. All right reserved.

대표적 빅데이터 성공 모델



넷플릭스
영화추천
서비스

이베이
쿼리로그
마이닝

트위터
대용량
머신러닝
시스템

페이스북
메시지
시스템



머신러닝

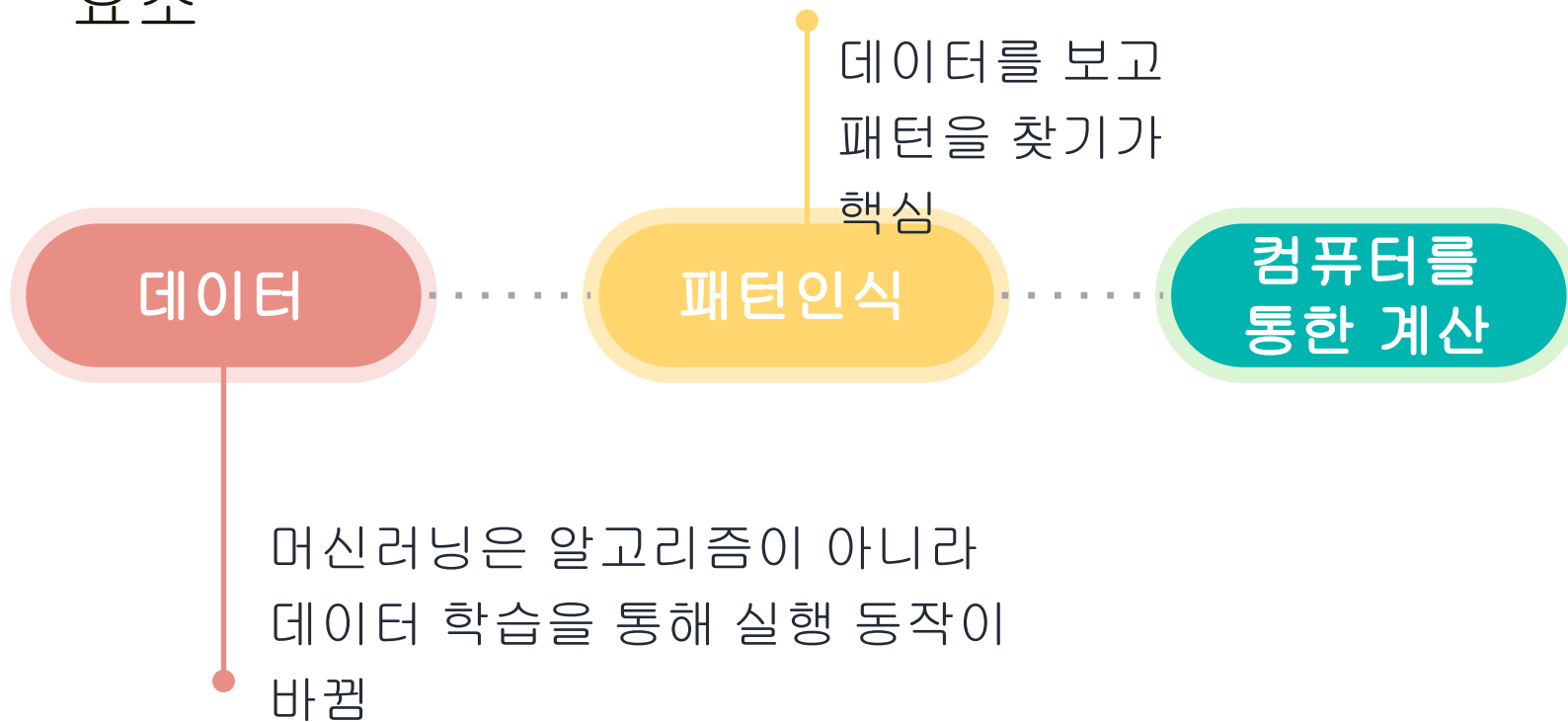
데이터를 이용해서 명시적으로 정의되지 않은
패턴을
컴퓨터로 학습하여 결과를 만들어내는 학문 분야



머신러닝



중요
요소



머신러닝 분류



풀고자 하는 목표에 따른
분류

지도학습

주어진 데이터와 레이블 (정답)을 이용해서
미지의 상태나 값을 예측하는 학습 방법



대부분의 머신러닝이 여기에 해당

회귀

숫자값을 예측,
보통 연속된 숫자
(실수)를 예측

분류

입력 데이터들을
주어진 항목들로
나누기

랭킹 / 추천

데이터들의
순위를 예측

머신러닝 분류



풀고자 하는 목표에 따른
분류

비지도 학습

데이터 그 자체에서 유용한 패턴을 찾는 학습방법

군집화 /
토픽 모델링

밀도 추정
데이터 분포
예측 기법

차원 축소
데이터의 차원을
낮추는 기법

머신러닝 분류



풀고자 하는 목표에 따른
분류

강화 학습

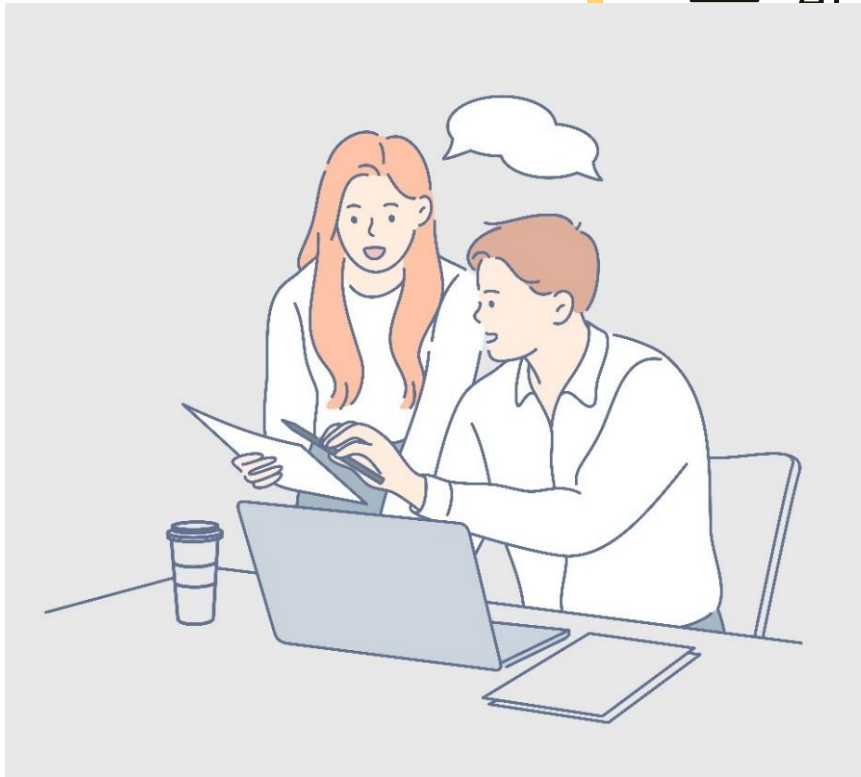
기계(에이전트)가 환경과의 상호 작용(선택과 피드백[보상]
반복)을

통해 장기적으로 얻는 이득을 최대화하도록 하는
예 알파고, 알파제로
학습방법

머신러닝 분류

🎯 기법에 따른
분류

토계



딥러닝

머신러닝 주요 개념

🎯 모델 : 데이터를 바라보는 시점 혹은

가정
▶ 간단한 모델 : 선형 모델(회귀)

▶ 복잡한 모델 : 결정 트리 모델

▶ 구조가 있는 모델

순차 모델

- 연속된 관측값이 서로 연관성이 있을 때 주로 사용
- 문서의 텍스트, 시간과 관계된 데이터 분석에 주로 사용

그래프 모델

- 그래프를 이용해서 순차 모델보다 더 복잡한 구조를 모델링
- 문서 텍스트의 문법 구조(보통 트리 형태)를 직접 모델링하거나 이미지의 픽셀 사이 관계를 그래프로 표현하여 모델링

머신러닝 주요 개념

손실함수(Loss Function)

- ▶ 모델의 수식화된 학습 목표
- ▶ 모델이 실제로 데이터를 바르게 표현했는지 혹은 얼마나 예측이 정확한지 수학적으로 표현하는 방식



머신러닝 주요 개념

손실함수(Loss Function)

산술 손실 함수

모델로 산술값을 예측하고
데이터에 대한 예측값과
실제 관측값을 비교하는 함수

확률 손실 함수

모델로 항목이나 값에 대한
확률을 예측하는 경우에 사용

랭킹 손실 함수

모델로 순서를 결정할 때 사용

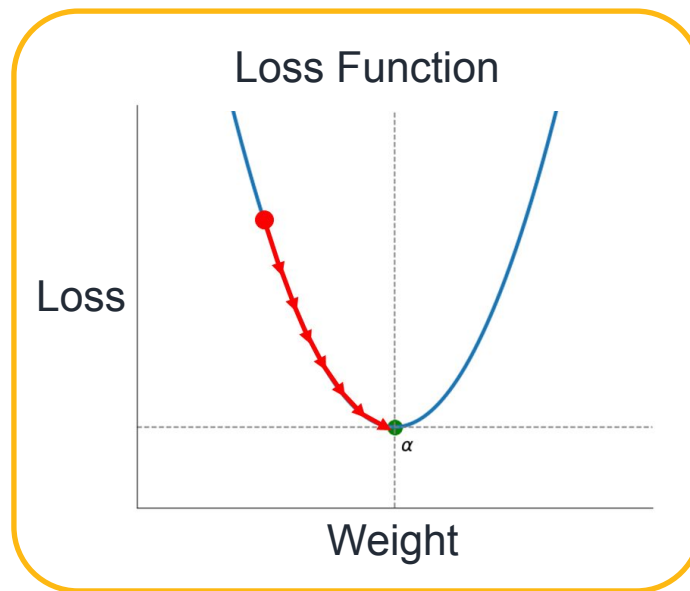
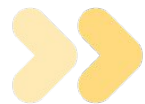
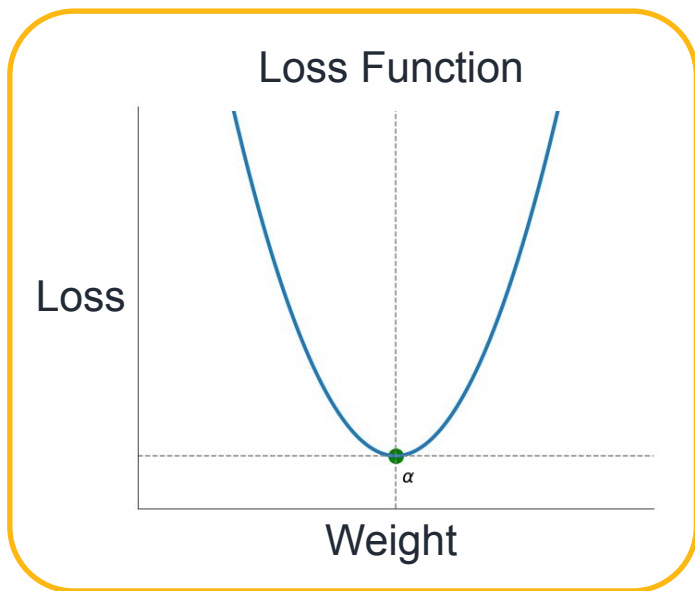
모델 복잡도와 관련된 손실함수

3가지 손실함수들과 결합하여
모델이 필요 이상으로
복잡해지지 않도록 방지하는
손실함수

머신러닝 주요 개념

🎯 최적화

- ▶ 실제로 학습을 하는 방법
- ▶ 손실함수의 결과값을 최소화 하는 모델의 인자를 찾는 과정



머신러닝 주요 개념

최적화

경사하강법

임의의 지점에서 시작해서
경사를 따라 내려갈 수 없을 때까지
반복적으로 내려가며 최적화를
수행

뉴턴/준 뉴턴 방법

확률적 경사하강법

뉴턴/준 뉴턴 방법
(계산시 자원을 많이 사용)을 보완

역전파

딥러닝에서 많이 사용하는 방법

머신러닝 주요 개념



모델

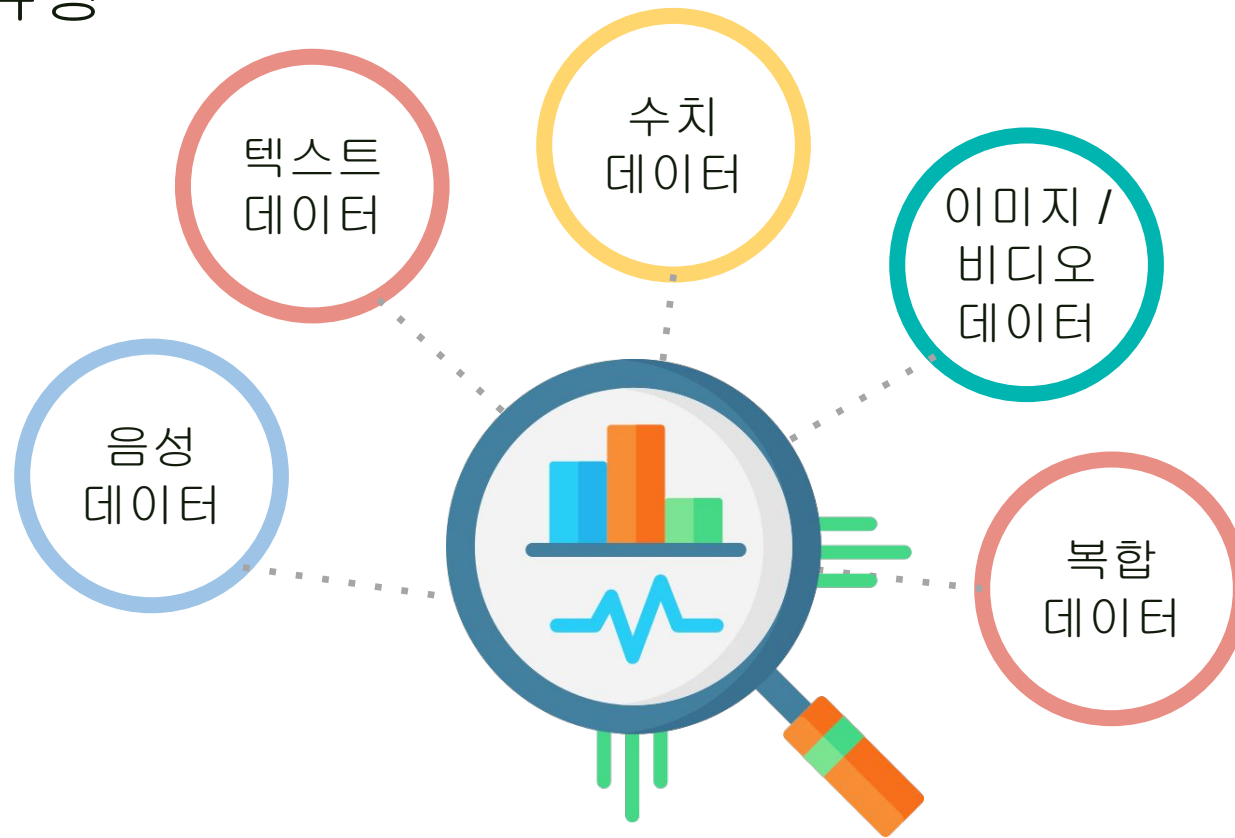
평가

- ▶ 실제 활용에서 성능을 평가
- ▶ 학습 데이터 뿐만 아니라 새로운 데이터가 들어왔을 때도 잘 동작하는지 평가



머신러닝 주요 개념

데이터의 유형



머신러닝 주요 개념

문제 유형과 해결 방법

회귀 문제 (가장 기본)

입력을 받아 가장 적합한 숫자값을 예측하는 문제

➡ 선형 회귀, 가우시안 프로세스 회귀, 칼만 필터

분류 문제 (기본)

보통 손실함수를 직접 최적화해서 푸는 방법을 많이 사용

➡ 로지스틱 회귀, 서포트 벡터 머신, 신경망

머신러닝 주요 개념

문제 유형과 해결 방법

군집화 문제

➡ k-평균 군집화, 토픽 모델링, 평균 이동 군집화

표현형 학습(임베딩 학습)

풀고자 하는 문제에 적합한 표현형 (representation)을
데이터로부터 추출하는 방법

➡ word2vec 모델과 그 파생모델, 행렬 분해

수업 도구 소개



Jupyter notebook

<https://jupyter.org/>

Google Colab

https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index



파이썬 라이브러리 소개



출처 : <https://numpy.org/>

- ➡ Python에서 벡터, 행렬 등 수치 연산을 수행하는
선형대수 (Linear algebra) 라이브러리



출처 : <https://pandas.pydata.org/>

- ➡ 데이터 조작 및 분석을 위해 Python 프로그래밍
언어로 작성된
소프트웨어 라이브러리

Data Science Process

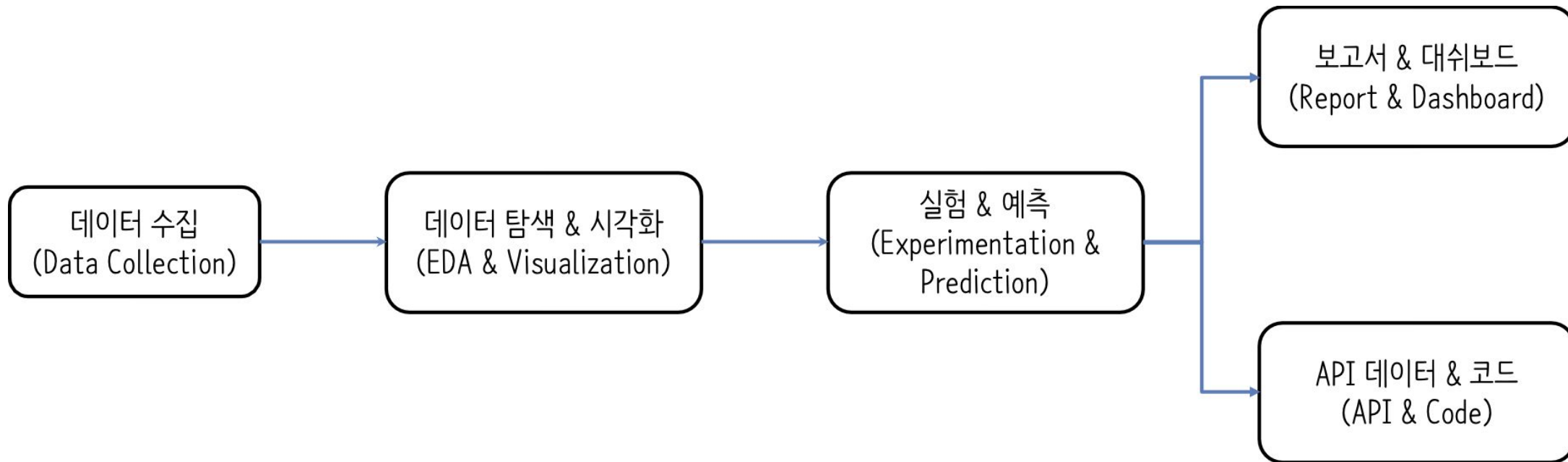
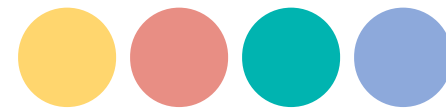
The background features a white field with a yellow triangle in the top-left corner and a green triangle in the bottom-right corner. A faint, grey silhouette of a city skyline is visible in the lower half of the image.

데이터 과학

데이터를 통해 실제 현상을 이해하고 분석하는데 통계학,
데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념



데이터 과학 작업 흐름도





문제 정의하기

해결하고자 하는 문제를 정의합니다.

이 단계에서는 해결하고자 하는 게 무엇인지, 언제까지 어떤 결과물을 얻을 것인지, 어떤 방식으로 데이터를 활용할 것인지 등을 설정합니다. 아무 목적 없이 데이터를 살펴 보면, 의미 있는 발견을 하지 못합니다.

- 목표 설정
- 기간 설정
- 평가 방법 설정
- 필요한 데이터 설정



데이터 모으기

필요한 데이터를 모을 수 있는 방법을 찾습니다.

누군가 이미 모아 놓은 데이터를 그대로 사용할 수도 있고, 공공 기관 등에서 배포한 자료를 찾아 볼 수도 있고, 혹은 웹사이트에서 직접 데이터를 수집할 수도 있습니다.

- 웹 크롤링
- 자료 모으기
- 파일 읽고 쓰기



데이터 다듬기

데이터의 퀄리티를 높여서 의미 있는 분석이 가능하게끔 합니다.

일반적으로 우리가 수집한 데이터에는 수많은 문제점들이 있습니다. 이런 문제점들로 인해 분석 자체가 불가능할 수도 있고, 혹은 분석을 하더라도 잘못된 결론으로 이어질 수도 있습니다. “쓰레기를 넣으면 쓰레기가 나온다(**garbage in, garbage out**)”라는 표현이 있을 정도입니다.

- 데이터 관찰하기
- 데이터 오류 제거
- 데이터 정리하기



데이터 분석하기

준비된 데이터로부터 의미를 찾습니다.

이 과정은 통계를 이용해서 수치적으로도 할 수도 있고, 수십 가지의 그래프를 그려보면서 탐색할 수도 있습니다. 처음 설계했던 방식대로 데이터를 활용해서 원하는 결과를 도출해야 합니다.

- 데이터 파악하기
- 데이터 변형하기
- 통계 분석
- 인사이트 발견
- 의미 도출



커뮤니케이션

분석 결과를 다른 사람들에게 전달합니다.

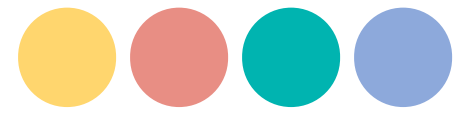
어떤 문제를 해결하려 했는지, 어떻게 데이터를 모았는지, 어떤 방식으로 어떤 인사이트를 얻었는지 등을 다른 사람들에게 전달해야 합니다. 적절한 시각화를 통해 소통을 원활히 할 수 있습니다.

- 다양한 시각화
- 커뮤니케이션
- 리포트

머신러닝 프레임워크 익히기

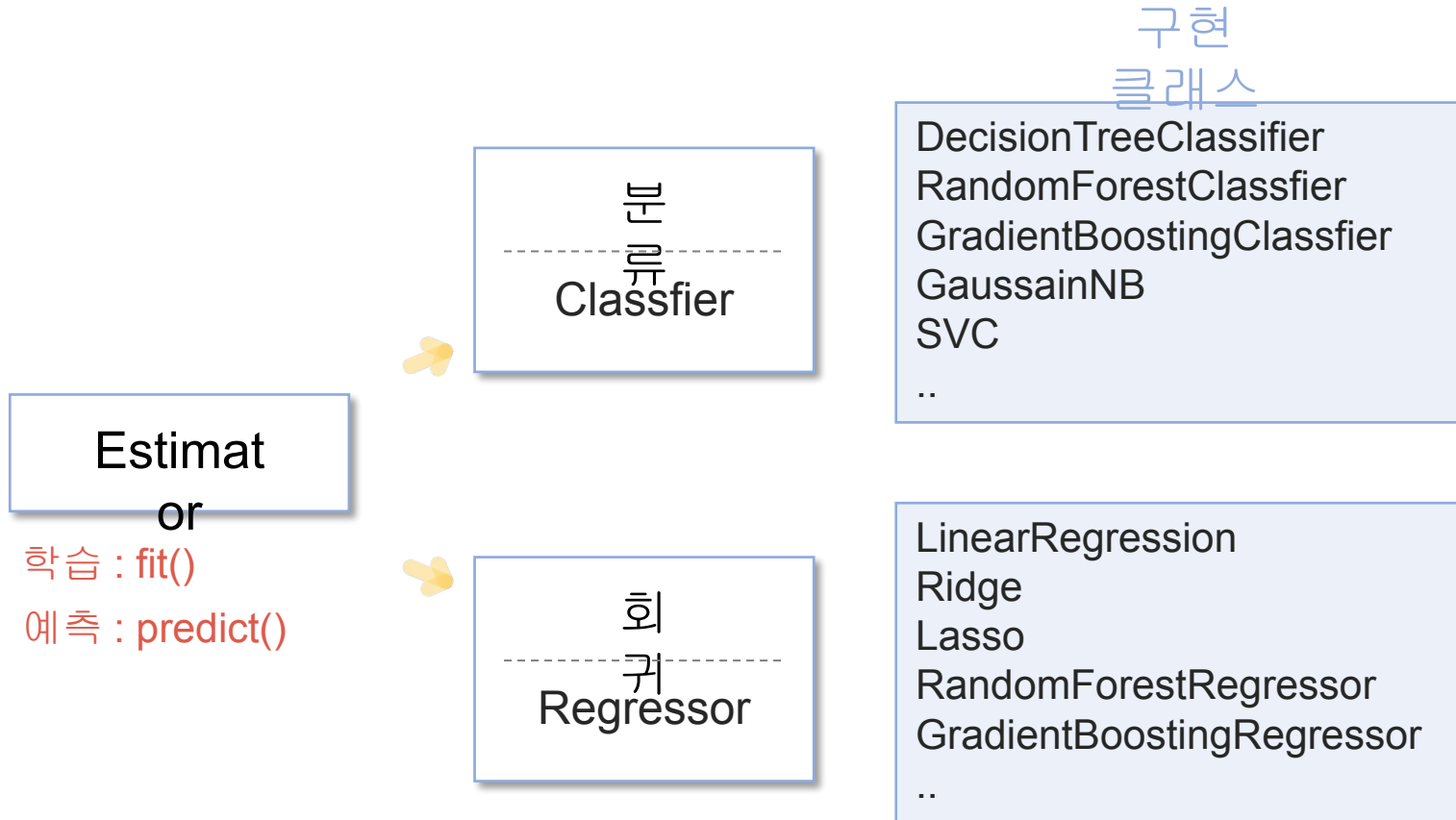
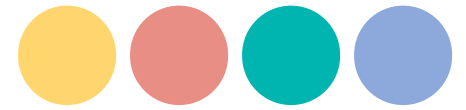


Scikit-Learn 프레임워크



출처 : <https://scikit-learn.org/stable/index.html>

Scikit-Learn 프레임워크

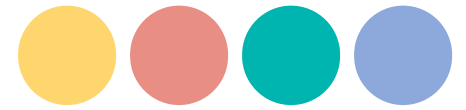


Scikit-Learn 주요 모듈



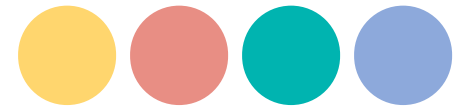
분류	모듈명	설명
예제 데이터	<code>sklearn.datasets</code>	사이키런에 내장 되어있는 예제 데이터 세트
피처 처리	<code>sklearn.preprocessing</code>	데이터 전처리에 필요한 다양한 가공 기능 제공 ex) 문자열을 숫자형 코드 값으로 인코딩 정규화, 스케일링 등
	<code>sklearn.feature_selection</code>	알고리즘에 큰 영향을 미치는 피처를 우선순위대로 선택 작업 수행하는 다양한 기능
	<code>sklearn.feature_extraction</code>	텍스트 데이터나 이미지 데이터의 벡터화된 피처를 추출하는 용도로 사용 ex) 텍스트 데이터에서 Count Vectorizer 또는 Tf-Idf Vectorizer 등을 생성 ex) 텍스트데이터 추출 <code>sklearn.feature_extraction.text</code> ex) 이미지 데이터 피처 추출 <code>sklearn.feature_extraction.image</code>

Scikit-Learn 주요 모듈



피처 처리 & 차원 축소	sklearn.decomposition	차원 축소와 관련한 알고리즘 지원. ex) PCA, NMF, Truncated SVD
데이터 분리, 검증 & 파라미터 튜닝	sklearn.model_selection	교차 검증을 위한 학습용/테스트용 세트 분리. ex) 그리드 서치(Grid Search)로 최적 파라미터 추출 등의 API 제공
평가	sklearn.metrics	분류, 회귀, 클러스터링, 페어와이즈 (Pairwise)에 대한 다양한 성능 측정 방법 제공 ex) Accuracy, Precision, Recall, ROC-AUC, RMSE 등

Scikit-Learn 주요 모듈



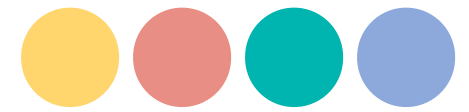
ML 알고리즘	<code>sklearn.ensemble</code>	앙상블 알고리즘 제공 ex) 랜덤 포레스트, 에이다 부스트, 그래디언트 부스팅 등
	<code>sklearn.linear_model</code>	주로 선형회귀, 릿지(Ridge), 라소(Lasso) 및 로지스틱 회귀 등 회귀 관련 알고리즘을 지원. 또한 SGD(Stochastic Gradient Descent) 관련 알고리즘도 제공
	<code>sklearn.naive_bayes</code>	나이브 베이즈 알고리즘 제공. 가우시안 NB. 다항 분포 NB 등
	<code>sklearn.neighbors</code>	최근접 이웃 알고리즘 제공. K-NN 등
	<code>sklearn.svm</code>	서포트 벡터 머신 알고리즘 제공
	<code>sklearn.tree</code>	의사 결정 트리 알고리즘 제공
	<code>sklearn.cluster</code>	비지도 클러스터링 알고리즘 제공 ex) K-평균, 계층형, DBSCAN 등
유틸리티	<code>sklearn.pipeline</code>	피처 처리 등의 변환과 ML 알고리즘 학습, 예측 등을 함께 묶어서 실행할 수 있는 유틸리티 제공

Scikit-Learn 예제 데이터 세트



API 명	설명
<code>datasets.load_boston()</code>	회귀용도. 미국 보스턴 집 피쳐들과 가격에 대한 데이터 세트
<code>datasets.load_breast_cancer()</code>	분류용도. 위스콘신 유방암 피쳐들과 악성/양성 레이블 데이터 셋
<code>datasets.load_diabetes()</code>	회귀용도. 당뇨 데이터 셋
<code>datasets.load_digits()</code>	분류용도. 0에서 9까지 숫자의 이미지 픽셀 데이터 세트
<code>datasets.load_iris()</code>	분류용도. 붓꽃에 대한 피쳐를 가진 데이터 세트

Scikit-Learn fetch 계열 데이터 세트



API 명	설명
fetch_covtype()	회귀 분석용 토지 조사 자료
fetch_20newsgroups()	뉴스 그룹 데이터 자료
fetch_olivetti_faces()	얼굴 이미지 자료
fetch_lfw_people()	얼굴 이미지 자료
fetch_lfw_pairs()	얼굴 이미지 자료
fetch_rcv1()	로이터 뉴스 말뭉치

분류와 클러스터링을 위한 표본 데이터 생성기

API 명	설명
<code>datasets.make_classifications()</code>	분류를 위한 데이터셋 생성기. 높은 상관도, 불필요한 속성 등의 노이즈 효과를 위한 데이터를 무작위로 생성해 줌
<code>datasets.make_blobs()</code>	클러스터링을 위한 데이터 생성기 (무작위) 군집 지정 개수에 따라 여러 가지 클러스터링을 위한 데이터 세트를 쉽게 만들어 줌