

Data Prediction Model and Machine Learning

Online course #1

Kaggle

Kaggle Notebook

- 노트북은 캐글에서 제공하는 데이터 분석용 프로그래밍 환경
- 웹상의 코드 편집기에 코드를 작성하면 서버에서 해당 프로그램을 실행해 결과를 변환하는 SaaS(Software as a Service) 환경

성능

- 4 Core CPU, 16GB RAM
- 2 Core CPU+GPU, 13GB RAM

특징

- Private or Public 선택가능
- Once public → Apache 2.0 license

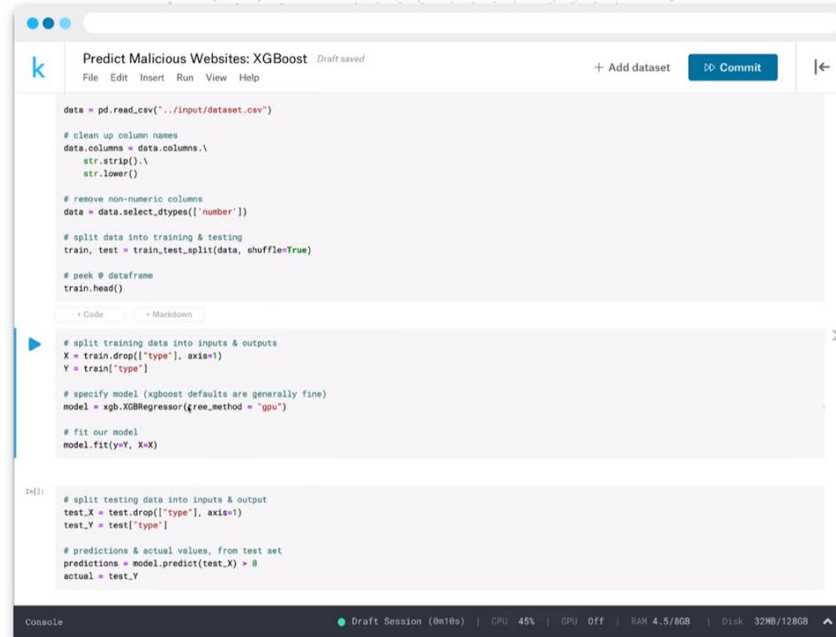


Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email



```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()

# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(ree_method = "gpu")

# fit our model
model.fit(y=Y, X=X)

# split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]

# predictions & actual values, from test set
predictions = model.predict(test_X) * 0
actual = test_Y
```

Console: Draft Session (0m10s) | CPU 45% | GPU Off | RAM 4.5/8GB | Disk 32MB/128GB

Inside Kaggle you'll find all the code & data you need to do your data science work. Use over 19,000 public [datasets](#) and 200,000 public [notebooks](#) to conquer any analysis in no time.

Kaggle



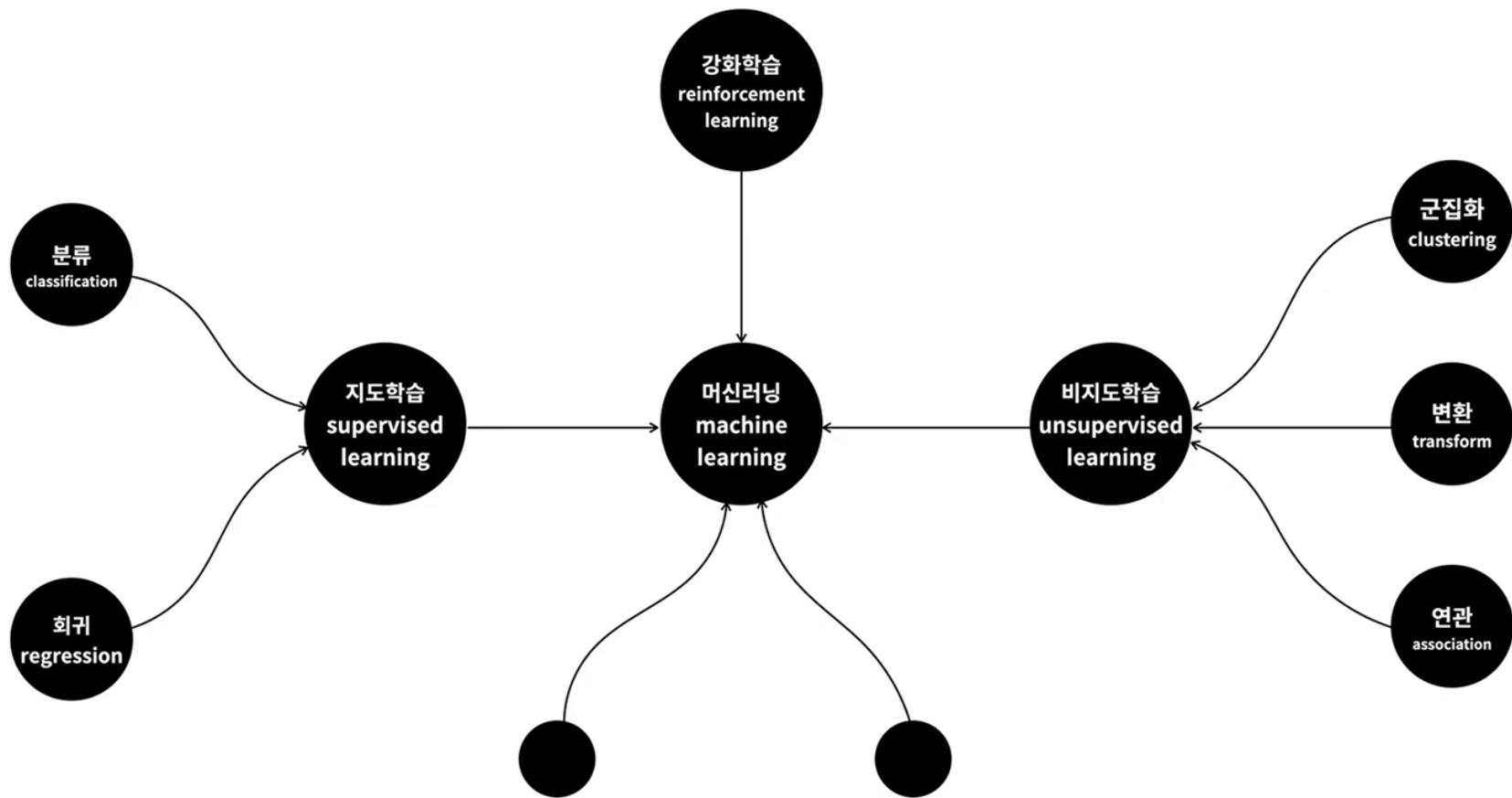
캐글 가이드 전 세계 데이터 과학자

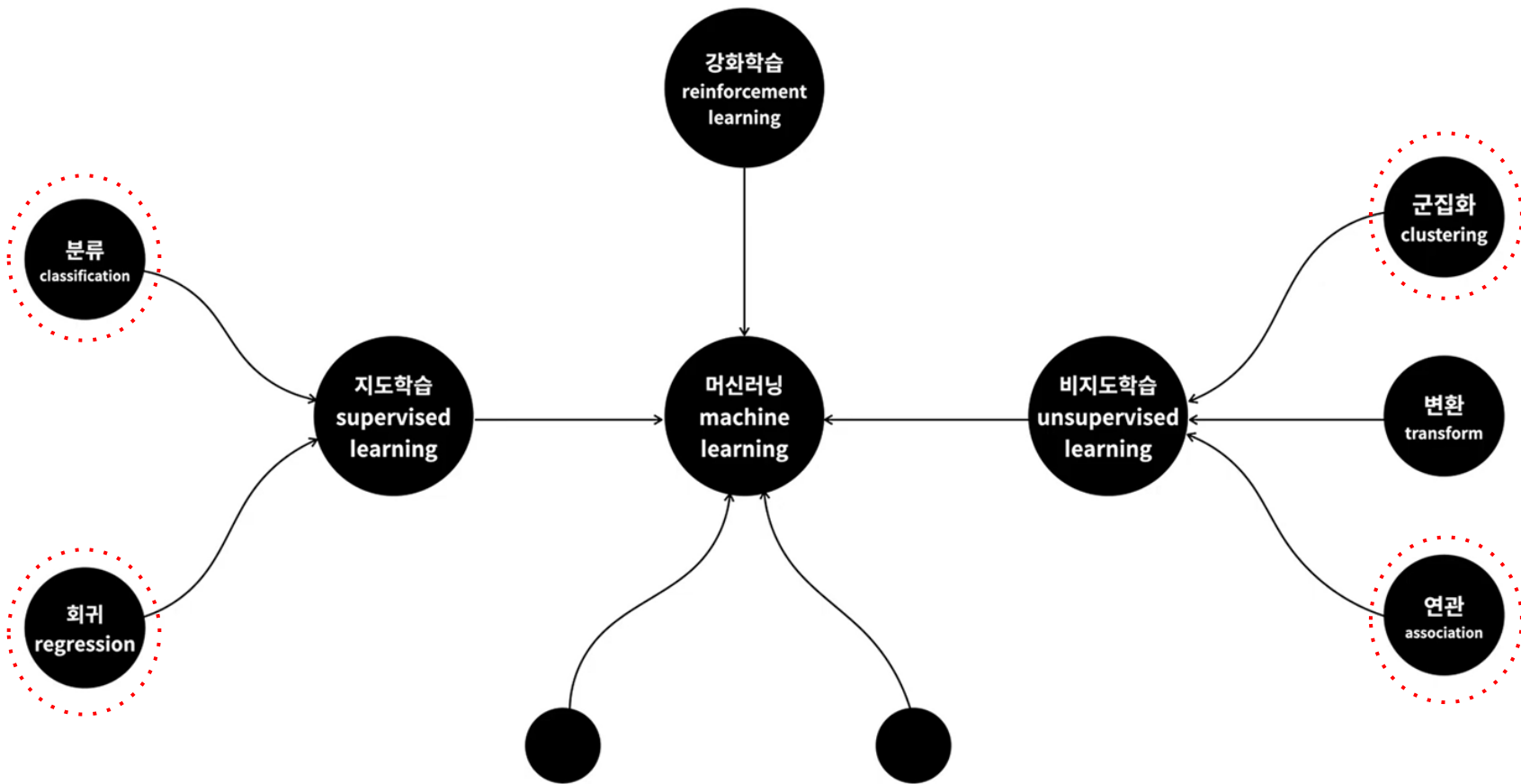
와 소통하고, 경쟁하고, 성장하기
시카모토 도시유키 저/박광수 역 | 동양북스(동양 books)

Data Prediction Model and Machine Learning

Online course #2

Modelling and Linear Regression





온도	판매량
20	40
21	42
22	44
23	46

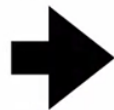
양적



회귀
regression

공부시간	시험결과
20	불합격
21	불합격
22	합격
23	합격

범주형



분류
classification

날짜	요일	온도	판매량
2020.1.3	금	20	40
2020.1.4	토	21	42
2020.1.5	일	22	44

날짜	2020.1.3	2020.1.4	2020.1.5
요일	금	토	일
온도	20	21	22
판매량	40	42	44

열
column

특성 (feature)
속성 (attribute)
변수(variable)
field

날짜	요일	온도	판매량
2020.1.3	금	20	40
2020.1.4	토	21	42
2020.1.5	일	22	44

행
row

개체 (instance)
관측치(observed value)
기록 (record)
사례 (example)
경우 (case)

독립변수

원인

종속변수

결과



인과관계

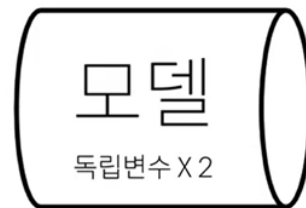
상관관계

- 독립변수는 원인이다.
- 종속변수는 결과다.
- 독립변수와 종속변수의 관계를 인과관계라고 한다.
- 인과관계는 상관관계에 포함된다.

첫번째 과제: 사업계획에 대한 기획서 (환경, 불만족, 꿈)
두번째 과제: 독립변수, 종속변수를 기획서에 반영해서 제출.

독립변수 종속변수

온도	판매량
20	40
21	42
22	44
23	46





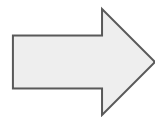
$$F=ma$$

Force = mass x acceleration

$$E = mc^2$$



$$F = G \frac{m^1 m^2}{r^2}$$



머신러닝

공식의 대중화



분류의 좋은 사례

독립변수	종속변수	학습시킬 데이터를 만드는 방법
공부시간	합격 여부 (합격, 불합격)	사람들의 공부시간을 입력받고, 최종 합격여부를 확인한다.
x-ray 사진과 영상 속 종양의 크기, 두께	악성 종양 여부 (양성, 음성)	의학적으로 양성과 음성이 확인된 사진과 영상 데이터를 모은다.
품종, 산도, 당도, 지역, 연도	와인의 등급	소믈리에를 통해서 등급이 확인된 와인을 가지고 품종, 산도 등의 독립변수를 측정하고 기록한다.
키, 몸무게, 시력, 지병	현역, 공익, 면제	키, 몸무게, 시력, 지병들을 토대로 현역, 공익, 면제인 지 확인한다.
메일 발신인, 제목, 본문 내용 (사용된 단어, 이모티콘 등)	스팸 메일 여부	이제까지 받은 메일을 모으고, 이들을 스팸 메일과 일반 메일로 구분한다.
고기의 지방함량, 지방색, 성숙도, 육색	소고기 등급	소고기의 정보를 토대로 등급을 측정한다.

회귀의 좋은 사례

독립변수	종속변수	학습시킬 데이터를 만드는 방법
공부시간	시험점수 (10점, 20점)	사람들의 공부시간을 입력받고 점수를 확인한다.
온도	레모네이드 판매량	온도와 그날의 판매량을 기록한다.
역세권, 조망 등	집 값	집과 역까지의 거리, 수치화된 조망의 평점 등을 집 값과 함께 기록한다.
온실 기체량	기온 변화량	과거에 배출된 온실기체량과 기온의 변화량을 기록한다.
자동차 속도	충돌 시 사망 확률	충돌 시 속도와 사상자를 기록한다.
나이	키	학생들의 나이에 따른 키를 기록한다.

TensorFlow

- TensorFlow™ is an open source software library for numerical computation using data flow graphs.
- Python!



<https://www.tensorflow.org/>