

Data Mining (Machine Learning)

- Course Overview

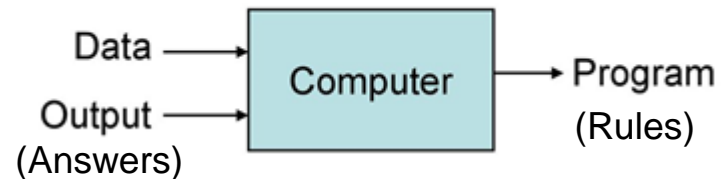
머신러닝(Machine Learning)이란 ?

- Limitations of explicit programming
 - Spam filter: many rules
 - Automatic driving: too many rules
- Machine learning
 - "Field of study that gives computers the ability to learn without being explicitly programmed", Arthur Samuel (1959)

Traditional Programming



Machine Learning

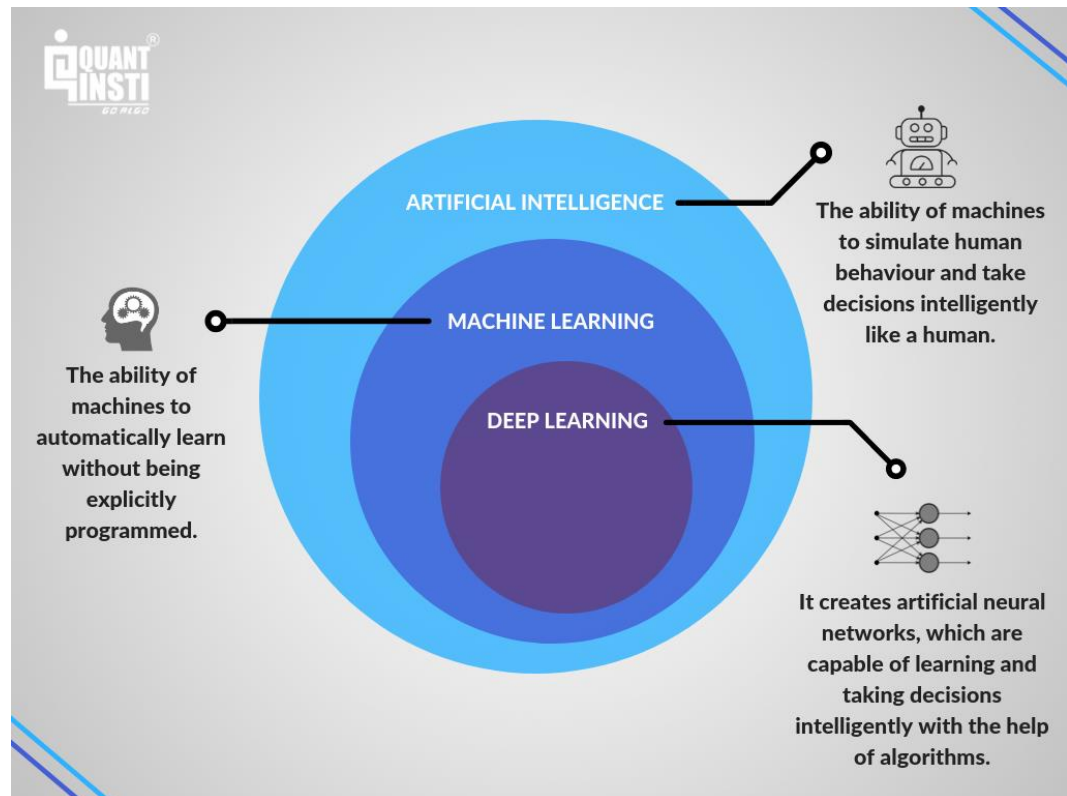


혼동되어 사용되는 용어

- Data Mining

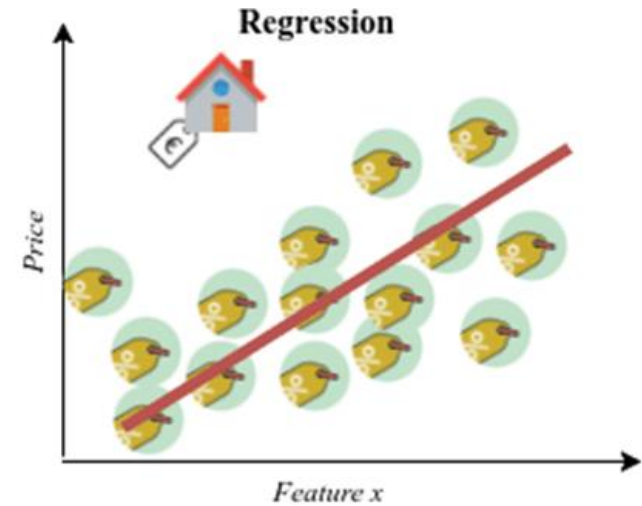
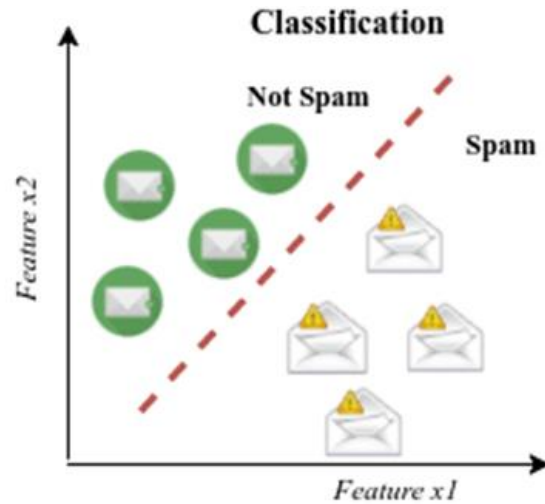
- Data analysis processes that apply ML techniques to solving real world problems

- AI & Deep Learning

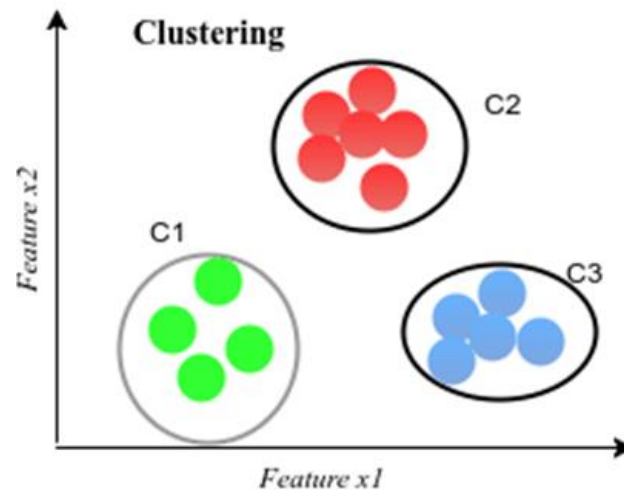


머신러닝의 유형

Supervised Learning (지도학습)



Unsupervised Learning (비지도학습)





Learning Objectives

머신러닝 응용역량 개발

- 머신러닝 핵심 이론의 이해
- 머신러닝의 유용성과 한계, 통계 기반의 데이터분석과의 차이를 이해
- 전체 머신러닝 프로세스의 이해
- 최신 머신러닝 기법과 기술을 파악하여 비즈니스 기회를 포착할 수 있는 역량 개발

오픈 소스 활용능력 확보

- scikit-learn, keras, numpy, pandas, matplotlib, gensim 등 중요 Python 머신러닝 라이브러리 활용능력 확보
- Github 등에 공개된 머신러닝 관련 오픈 소스를 수정하여 실무에 활용할 수 있는 능력 확보

분석 프로젝트 수행경험 체득

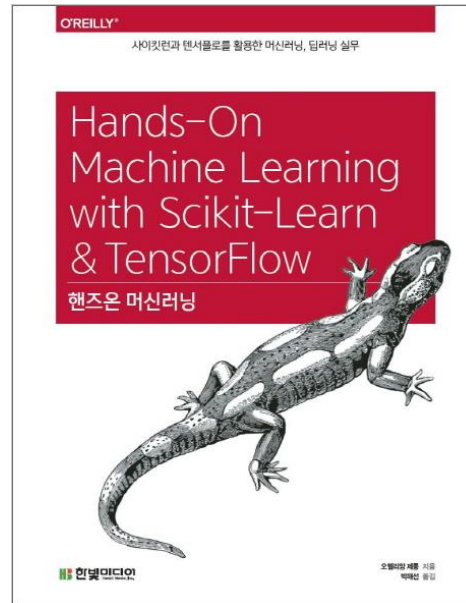
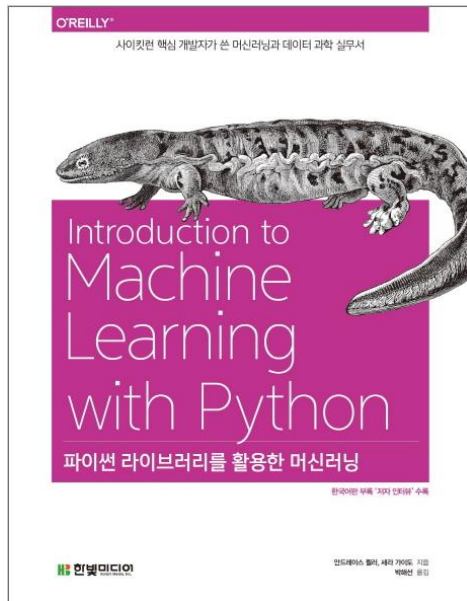
- End-to-End 머신러닝 프로젝트 수행 경험 확보
- 문제를 정의하고 이슈사항을 도출하여 분석 목표 및 프로젝트 계획을 수립하는 능력 확보
- Kaggle Competition에서 사용된 문제 해결 접근방식과 분석기법을 경험



Course Outline

- **Getting started with Machine Learning**
 - Concepts, Methods, & Tools
 - Machine Learning Process
- **Model Evaluation**
 - Measuring Model Performance
 - Cross Validation
- **Performance Improvement**
 - Model Tuning (Hyperparameter Optimization)
 - Ensemble
- **Feature Engineering**
- **Workflow Optimization**
- **Algorithms**
 - Decision Tree, Random Forest, & Gradient Boosting Machines
 - Linear Regression Models & Regularized Linear Models
 - Deep Neural Networks
- **Cluster Analysis**


Textbooks & References



Development Environment

주피터 노트북

Anaconda
(Python 3.6 버전)
설치

jupyter Example8_1_Intro_to_RNN (autosaved)  Logout

File Edit View Insert Cell Kernel Help Not Trusted Python 3

Example 8-1: Introduction to RNN

```
In [52]: from keras.models import Sequential, load_model
from keras import layers
from keras import backend as K
from keras import optimizers

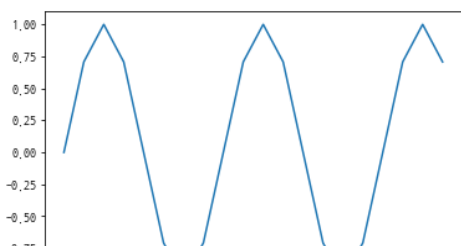
from keras.utils import plot_model
from keras.callbacks import EarlyStopping, TensorBoard, ModelCheckpoint
from keras.preprocessing import image

import numpy as np
import matplotlib.pyplot as plt
from IPython.display import Image
```

Data generation

```
In [53]: from scipy.linalg import toeplitz
s = np.sin(2 * np.pi * 0.125 * np.arange(20))
plt.plot(s)
```

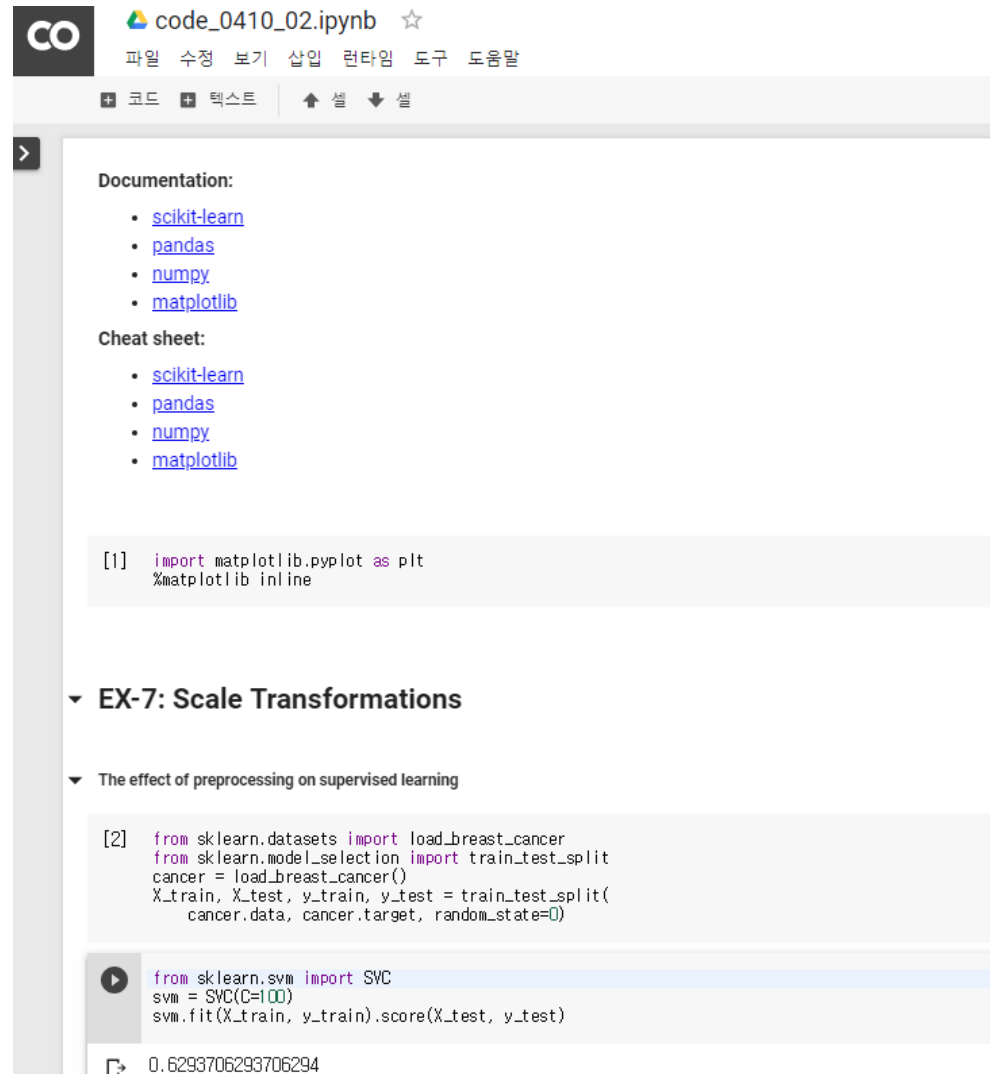
Out [53]: [<matplotlib.lines.Line2D at 0x7f88cc2c5c18>]



Development Environment (Cont'd)

Google Colab

구글 드라이브에서
무료로
GPU(Nvidia Tesla K80)
사용이 가능한 개발환경



The screenshot displays a Google Colab notebook titled 'code_0410_02.ipynb'. The interface includes a top menu bar with options like '파일', '수정', '보기', '삽입', '런타임', '도구', and '도움말'. Below the menu, there are tabs for '코드', '텍스트', and '셀'. The main content area is divided into sections: 'Documentation' with links to 'scikit-learn', 'pandas', 'numpy', and 'matplotlib'; 'Cheat sheet' with the same links; and a code cell [1] containing the code to import matplotlib.pyplot as plt and set the inline backend. Below this, there is a section titled 'EX-7: Scale Transformations' with a sub-section 'The effect of preprocessing on supervised learning'. This section contains two code cells: cell [2] for loading and splitting the breast cancer dataset, and a cell for training and testing an SVM model. The bottom of the notebook shows a file icon and a unique identifier '0.6293706293706294'.

code_0410_02.ipynb ☆

파일 수정 보기 삽입 런타임 도구 도움말

코드 텍스트 셀 셀

Documentation:

- [scikit-learn](#)
- [pandas](#)
- [numpy](#)
- [matplotlib](#)

Cheat sheet:

- [scikit-learn](#)
- [pandas](#)
- [numpy](#)
- [matplotlib](#)

```
[1] import matplotlib.pyplot as plt
    %matplotlib inline
```

EX-7: Scale Transformations

The effect of preprocessing on supervised learning

```
[2] from sklearn.datasets import load_breast_cancer
    from sklearn.model_selection import train_test_split
    cancer = load_breast_cancer()
    X_train, X_test, y_train, y_test = train_test_split(
        cancer.data, cancer.target, random_state=0)
```

```
from sklearn.svm import SVC
svm = SVC(C=100)
svm.fit(X_train, y_train).score(X_test, y_test)
```

0.6293706293706294

Development Environment (Cont'd)

NAVER
CLOUD PLATFORM

Server

운영체제 설치를 위한 기본 디스크가 제공되며, 추가 디스크는 사용자가 원하는 용량 만큼 직접 생성할 수 있습니다. 월 또는 시간 단위로 서버를 이용할 수 있습니다. 월요금의 경우 1개월 미만으로 이용하면 일할 계산되어 청구됩니다. (Virtual Dedicated Server는 월 단위 요금제만 제공) vCPU 1개, 메모리 2GB 서버는 Linux 계열 운영체제만 사용할 수 있습니다. Micro 서버는 계정 당 1대만 제공됩니다.

(VAT 별도)

타입	제공 사양			이용 요금		비고
	vCPU	메모리	디스크	시간	월	
Micro	1개	1GB	50GB	19원	13,000원	서비스 체험 및 개발 테스트 용도의 서버로써, 신규 가입 후 결제 정보를 등록한 월부터 1년간 무료로 제공되며 무료기간이 지난 후에도 반납하지 않을 경우 자동으로 과금됩니다.



Grading Policy

- 출석 및 참여도 (15%)
 - 지각 2회는 결석 1회로 처리, 4회 결석 시 F학점
 - 팀 발표에 대한 적극적이고 생산적인 질의 및 의견제시 요망
- 개인과제 (15%)
 - 과제 당 5%, 총 3회 과제 부여
 - 제출 시한 내에 제출할 경우만 인정
- 시험 (30%)
 - 중간 또는 기말에 실시
 - 실기 시험
- 팀 프로젝트 (40%)
 - Kaggle (like) Competition 참여
 - 최종 팀 등수에 따라 평가
 - Competition 주제는 9월말까지 결정

* 강의진행 상황에 따라 추후 변경될 수 있음

Q & A