

Machine Learning from Scratch with Python

본 강의는 [TEAMLAB](#)과 [Inflearn](#)이 함께 구축한 데이터 사이언스 과정의 두 번째 강의인 [밀바닥](#) 부터 시작하는 머신러닝 입문 입니다. [밀바닥](#)부터 시작하는 머신러닝 입문 은 [Part I](#)과 [Part II](#)로 구성되어 있습니다.

본 강의는 [TEAMLAB](#)과 [Inflearn](#)이 함께 준비한 [WADIZ](#) 펀딩의 지원을 받아제작되었습니다. 아래 목록에 대한 강의를 개발할 예정입니다.

- [데이터 과학을 위한 파이썬 입문](#) - 개발 완료
- Machine Learning from Scratch with Python Part I - 본과정
- Machine Learning from Scratch with Python Part II

또한 기존 K-MOOC 과정은 아래 목록을 참고하시기 바랍니다.

- [K-MOOC: 데이터 과학을 위한 파이썬 입문](#)
- [Operation Research with Python Programming](#)

Course overview

- 강좌명: 밀바닥 부터 시작하는 머신러닝 입문(Machine Learning from Scratch with Python)
- 강의자명: 가천대학교 산업경영공학과 최성철 교수 (sc82.choi@gachon.ac.kr, Director of [TEAMLAB](#))
- Facebook: [Gachon CS50](#)
- Email: teamlab.gachon@gmail.com

Course Info

- 본 과정은 머신러닝에 대한 기초개념과 주요 알고리즘들에 대해 이해하고 구현하는 것을 목적으로 함
- 본 과정을 통해 수강자는 데이터 과학에서 사용되는 다양한 용어에 대한 기본적인 이해를 할 수 있음
- 본 과정의 기본적인 구성은 알고리즘에 대한 설명, [Numpy](#)를 사용한 사용자 구현, [Scikit-Learn](#)을 사용한 패키지 활용 으로 이루어 져 있음
- 수강자는 머신러닝에서 주로 사용되는 알고리즘을 구현하기 위해 고등학교 수준의 통계학과 선형대수학의 이해가 필요함
- 수강자는 본 과정을 통해 [Numpy](#), [Pandas](#), [Matplotlib](#), [Scikit-Learn](#) 등 데이터 분석을 위한 기본적인 파이썬 패키지를 이해하게 됨

Course Contents

Chapter 0 - Environment Setup

- 가상환경과 Package 활용하기 - [강의영상](#), [강의자료](#)
- Python Ecosystem for Machine Learning - [강의영상](#)
- Pycharm 설치(Mac) - [강의영상](#)
- How to use Jupyter Notebook

Chapter 1 - Introduction to Machine Learning

Lecture

- Machine Learning Overview
 - 머신러닝이 무엇인가 - 짤막한 개괄
 - 현재 머신러닝으로 무엇을 할 수 있는가?
- An Understanding of the Data Keyw ords - [강의영상_테스트](#), [강의자료](#)
- How to Learn Machine Learning - [강의영상_테스트](#), [강의자료](#)
- Types of Machine Learning - [강의영상_테스트](#), [강의자료](#)
- A History of Data Analysis: In Perspective of Business
 - 정보시스템의 등장~~~부터 현재까지의 이야기

Chapter 2 - Warm Up Section: An understanding of data

Lecture

- The concepts of a feature - [강의영상_테스트](#), [강의자료](#)
- Data types - [강의영상_테스트](#), [강의자료](#)
- Loading a data with pandas - [강의영상_테스트](#), [강의자료](#)
- Representing a model with numpy - [강의영상_테스트](#), [강의자료](#)

Chapter 3 - Pandas Section

Lecture

- Series - [강의영상_테스트](#), [강의자료](#), [code](#)
- DataFrame - [강의영상_테스트](#), [강의자료](#), [code](#)
- Data Cleaning Problem Overview - [강의영상_테스트](#), [강의자료](#)
- Missing Values - [code](#)
- Categorical Data Handling - [code](#)
- Feature Scaling - [강의영상_테스트](#), [강의자료](#), [code](#)
- Pivot Handling - [강의영상_테스트](#),
- Operation Function

Chapter 4 - Numpy Section

Lecture

- Axis 이해하기 - [강의영상_테스트](#) -

Supplements

- TF-KR 첫 모임: Zen of NumPy - [발표자료](#), [강의영상](#) (하성주, 2016)

Chapter 5 - Linear Regression

Lecture

- Probability overview - [강의자료](#)
- Overfitting - bias vs. variance
- Generalization - L1 and L2
- Implementation of generalization

Chapter 6 - Logistics Regression

Lecture

- Logistic regression overview - [강의자료](#), [code](#)
- Sigmoid function - [강의자료](#), [code](#)
- Cost function - [강의자료](#), [code](#)
- Logistic regression implementation - [강의자료](#), [code](#)
- Maximum Likelihood estimation - [강의자료](#)
- Logistic regression with sklearn
- Softmax function for Multi-class classification - [강의자료](#)
- Cross entropy loss function - [강의자료](#)
- Softmax regression - [강의자료](#)

Chapter 7 - Naive Bayesian Classifier

Lecture

- Probability overview - [강의자료](#)
- Bayes theorem - [강의자료](#)
- Single variable bayes classifier - [강의자료](#), [code](#)
- Naive bayesian Classifier - [강의자료](#), [code](#)
- NB classifier with sklearn - [code](#)

- Gaussian Normalization for Naive Bayesian

Chapter 8 - Decision Tree

Lecture

- Decision tree overview - [강의자료](#)
- The concept of entropy - [강의자료](#)
- The algorithm of growing decision tree - [강의자료](#)
- ID3 & Information gain - [강의자료](#)
- CART & Gini Index - [강의자료](#)
- Decision Tree with sklearn - [강의자료](#)
- Handling a continuous attribute - [강의자료](#)
- Decision Tree for Regression - [강의자료](#)
- Tree pruning - [강의자료](#)
- Regression Tree with sklearn - [코드](#)

Chapter 10 - Ensemble Model

Lecture

Chapter 11 - Feature Engineering

Lecture

Chapter 12 - Hyperparameter Search

Lecture

Chapter 13 - Auto ML & Parallel training

Lecture

Chapter 13 - Support Vector Model

Lecture

Chapter 14 - Neural Network

Lecture

지도 학습 (Supervised learning)

- 선형 회귀 (Linear Regression)
 - Lecture: 상관분석 - [강의자료](#)
 - 참고 1 - [상관계수 구하는 법](#) (나부랭이의 수학블로그, 2015)
 - Lecture: 선형 회귀 모델 개요 - [강의자료](#)
 - 참고 1 - [프로그래머를 위한 미분 강의](#) (홍정모, 2016)
 - Lab : 상관분석 - [강의자료](#)
 - Lab : 선형회귀 모델 - [강의자료](#)
 - Lecture: 경사하강법 (Gradient Descent) - [강의자료](#)
 - Lecture: 선형회귀를 위한 경사하강법 - [강의자료](#)
 - Lab: 선형회귀 경사하강법 구현 - [강의자료](#)
 - Lecture: Cost Function Graph
 - Lecture: PyData Package: Tensorflow vs Scikit-learn - [강의자료](#)
 - Lab: Linear Regression Tensorflow - [강의자료](#)
 - 참고 1 : [Linear Regression의 Hypothesis](#) 와 [cost](#) 설명 (김성훈, 2016)
 - 참고 2 : [Tensorflow](#) 로 간단한 [Linear Regression](#)을 구현 (김성훈, 2016)

- 참고 3 :[Linear Regression](#)의 [cost](#) 최소화 알고리즘의 원리 설명 (김성훈, 2016)
 - Assignment: Tensorflow 로 Linear Regression 구현하기
 - Lab: Linear Regression Scikit-learn - [강의자료](#)
 - Lecture: 다중 선형회귀 개요 - [강의자료](#)
 - Lecture: 다중 선형회귀 구현(w/Gradient Descent) - [강의자료](#)
 - Lab: 다중 선형 회귀 구현(w/Gradient Descent) - [강의자료](#)
 - Lecture: 데이터 정규화
 - Lab: 다중 선형회귀 모델 Tensorflow & Scikit-learn 구현
 - 참고 1 :[Multi-variable linear regression](#) (김성훈, 2016)
 - 참고 2 :[Multi-variable linear regression](#)을 TensorFlow 에서 구현하기 (김성훈, 2016)
- 로지스틱 회귀 (Logistic regression)
 - Lecture: 분류 문제 개요 (Classification Problem Overview) - [강의자료](#)
 - Lecture: 로지스틱 회귀 개요 (Logistic Regression Overview) - [강의자료](#)
 - Lab: 경사하강법으로 로지스틱 회귀 구현 (Pure Python)
 - Lab: Scikit-learn과 Tensorflow 로 로지스틱 회귀 구현
 - 참고 1 :[Logistic Classification](#)의 가설 함수 정의 (김성훈, 2016)
 - 참고 2 :[Logistic Regression](#)의 [cost](#) 함수 설명 (김성훈, 2016)
 - 참고 3 :[TensorFlow](#) 로 [Logistic Classification](#)의 구현하기 (김성훈, 2016)
 - Lecture: 범주형 자료와 다항 로지스틱 회귀 (Categorical data and Multinomial Logistic Regression) - [code](#)
 - Lab: 범주형 자료와 다항 로지스틱 회귀 구현 (Pure Python)
 - Lab: 범주형 자료와 다항 로지스틱 회귀 구현 II (Tensorflow, Scikit-learn) - [강의자료](#), [Code](#)
 - Lecture: 분류 서비스 구현하기 - [강의자료](#), [Modelling code](#), [Service code](#)
- 분석 성능 측정과 개선 (Performance Evaluation)
 - Lecture: 분류/회귀 문제의 성능 측정 - [강의자료](#)
 - RM 1 : Scratch Ch 11(p143~p147)
 - RM 2 : DDS Ch 3(p92), Ch 5(p140~p153)
 - Lab: 분류 문제의 성능 측정 - [Code](#)
 - Lab: 회귀 문제의 성능 측정 - [Code](#)
 - Lecture: 어떻게 성능을 개선할 것인가?
 - 참고 1 :[Overfitting](#) (전상혁, 2014)
 - Lecture: 성능 개선 1 - 벌점 회귀 (Penalized Regression)
 - Lab: 벌점 회귀 구현 I (Numpy)
 - Lab: 벌점 회귀 구현 II (Tensorflow & Scikit-Learn)
 - Lecture: 성능 개선 2 - Feature Engineering
 - Lab: Feature Selection with Pandas
 - Lecture: 성능 개선 3 - 경사하강법 알고리즘의 선택
 - Lab: SGD 알고리즘 구현
- 나이브 베이즈 분류기 (Naive Bayes Classifier)
 - Lecture: 나이브 베이즈 분류기 개요 (Naive Bayesian Classifier Overview) - [강의자료](#)
 - RM 1 : DDS Ch 4(p117)
 - RM 2 : scratch Ch 13
 - Lab: 나이브 베이즈 분류기 구현 (Numpy) - [강의자료](#), [Code](#)
 - Lab: 스팸필터 분류기 (Scikit-Learn) - [Code](#)
 - Lab: Text-mining 뉴스 분류기 (Scikit-Learn & NLTK) - [Code](#)
- 서포트 벡터 머신(Support Vector Machine)
 - SVM
- 뉴럴 네트워크(Neural network)
 - Neural network 개념의 이해
 - 미분 - Chain rule
 - Backpropagation

참고자료

- [Machine Learning](#) (Courera) by Andrew Ng
- [모두를 위한 딥러닝](#) by Sung Kim
- [C++로 배우는 딥러닝](#) by Sung Kim
- Machine Learning From Scratch[<https://github.com/eriklindernoren/ML-From-Scratch>]

Textbooks

- Reading materials
- [밑바닥부터 시작하는 데이터 과학](#)(조엘 그루스, 2016)
- [파이썬 머신러닝](#)(세바스티안 라슈카, 2016)
- [Hands-On Machine Learning with Scikit-Learn and TensorFlow](#) (Aurélien Géron, 2017, [PDF](#))
- [Data Mining: Concepts and Techniques](#)(Jiaw ei Han, Micheline Kamber and Jian Pei , 2011, [PDF](#))
- Supplementary textbooks
- [파이썬 라이브러리를 활용한 데이터 분석](#)(웨스 맥키니, 2013)
- [머신러닝 인 액션](#)(피터 해링턴, 2013)
- [데이터 과학 입문](#)(레이철 슈트 | 캐시 오닐, 2014)
- [머신러닝 인 파이썬](#)(마이클 보울즈, 2015)
- [머신러닝 이론 입문](#)(나카이 에츠지, 2016)

Prerequisites - 수강전 이수 또는 수강중 들었으면 하는 교과들

- 입문 수준의 통계학
 - [세상에서 가장 쉬운 통계학](#)(고지마 히로유키, 2009)
 - [세상에서 가장 쉬운 베이지통계학입문](#)(고지마 히로유키, 2017)
 - [확률과통계](#)(한양대학교 이상화 교수, 2014)
 - Reading Materials: Data Science from the Scratch - Ch.5, Ch.6, Ch.7
- 고교 이과 수준의 선형대수학 (Matrix와 Vector의 기본개념은 Review 필요)
 - [Essence of linear algebra](#)(3Blue1Brown, 2017)
 - [Linear Algebra](#)(Khan Academy)
 - [선형대수학](#)(한양대 이상화 교수, 2013) - Advance Course
 - Reading Materials - Data Science from the Scratch - Ch.4
- 고교 이과 수준의 미적분학 (개념에 대한 이해 필요)
 - [Essence of calculus](#)(3Blue1Brown, 2017)
- 파이썬 기초
 - [데이터 과학을 위한 파이썬 입문](#) (TEAMLAB, 2017)
- Git
 - [Pro Git](#) (스캇 샤콘 | 벤 스트라웁, 2016)
 - [Git & Github](#) (TEAMLAB, 2016)
 - [Git 강의](#) (생활코딩, 2014)