

Statistical Learning for Data Science 2 (229352)

ผู้สอน: ดลภาค พรนพรัตน์ (บอม)

Statistical Learning for Data Science 2

229352

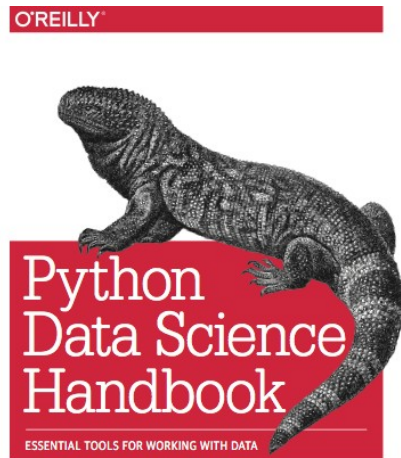
- Instructor: Donlapark Ponnoprat (<https://donlapark.pages.dev/>)
- Lectures: Tu/F 9:00-10:30pm
 - Tuesday at SCB 4405-06
 - Friday at STB205
- Mango (<https://mango-cmu.instructure.com/courses/8025>)
 - Labs

Prerequisites

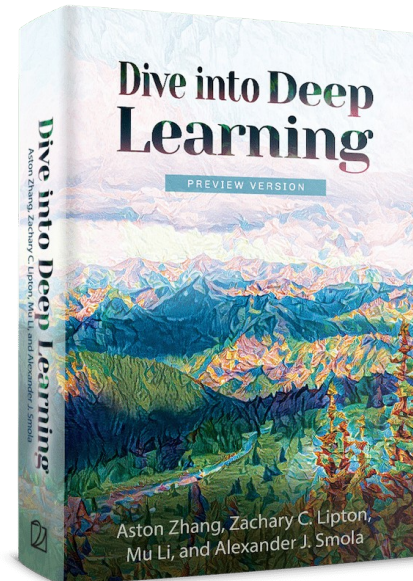
- On paper
Stat. Learning for Data Sci 1 (229351)
- In reality
Familiarity with probability, linear algebra and calculus,
programming (python)

Textbook

- No textbook required for this course.
- Recommended:
 - Python data science handbook (<https://jakevdp.github.io/PythonDataScienceHandbook>)
 - Dive into Deep Learning (<https://d2l.ai>)



Jake VanderPlas



Textbook

- No textbook required for this course.
- Recommended:
 - Stackexchange (<https://stackexchange.com>)

▲ How could I split randomly a data matrix and the corresponding label vector into a X_train, X_test, X_val, y_train, y_test, y_val with Sklearn? As far as I know, `sklearn.cross_validation.train_test_split` is only capable of splitting into two, not in three...

126


▼

machine-learning scikit-learn

51

Share Improve this question Follow

asked Nov 15 '16 at 14:55

 Hendrik
6,997 ● 16 ● 36 ● 52

Add a comment

12 Answers

Active

Oldest

Votes

▲ You could just use `sklearn.model_selection.train_test_split` twice. First to split to train, test and then split train again into validation and train. Something like this:

161



```
X_train, X_test, y_train, y_test  
= train_test_split(X, y, test_size=0.2, random_state=1)
```



```
X_train, X_val, y_train, y_val  
= train_test_split(X_train, y_train, test_size=0.25, random_state=1) # 0.25 x 0.8
```



Grading

- Labs: 30%
 - Lab with the lowest score will be dropped.
- Midterm 35% TBD.
- Final 35% 27 October 2567 15:30 – 18:30
 - Closed-book
 - Two a4 papers with notes on both sides is allowed
 - A calculator is allowed

Overview

What is machine learning, and why should we care?

- Someone gives you data, you write a method that learns from it.
 - Learns the distribution of the data
 - Learns the pattern of the data
 - Learns the relationship between features and labels

What is machine learning, and why should we care?

- Someone gives you data, you write a method that learns from it.
 - Learns the distribution of the data
 - Learns the pattern of the data
 - Learns the relationship between features and labels
- Machine learning has many applications
 - Google search
 - Email autocomplete
 - Netflix movie recommendation

Types of learning

Back then...

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Types of learning

Right now

- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Online learning
- Semi-supervised learning
- Weakly supervised learning
- Self-supervised learning
- Continual learning/Lifelong learning

Supervised learning



Chihuahua or muffin?

Large Language Models (LLM)



How can I pass a Machine Learning course?



To pass a Machine Learning course, here are some general guidelines and tips:

1. **Attend lectures and review materials:** Actively participate in lectures, take detailed notes, and review the course materials regularly. Make sure you understand the fundamental concepts, algorithms, and techniques taught in the course.
2. **Complete assignments and projects:** Stay on top of your assignments and projects. Allocate sufficient time to work on them, and try to complete them well before the deadlines. If you encounter difficulties, seek help from your instructor, teaching assistants, or classmates.
3. **Practice coding:** Machine Learning involves a significant amount of programming. Develop your coding skills in relevant languages such as Python or R. Implement algorithms from scratch, solve programming exercises, and practice data manipulation and visualization.

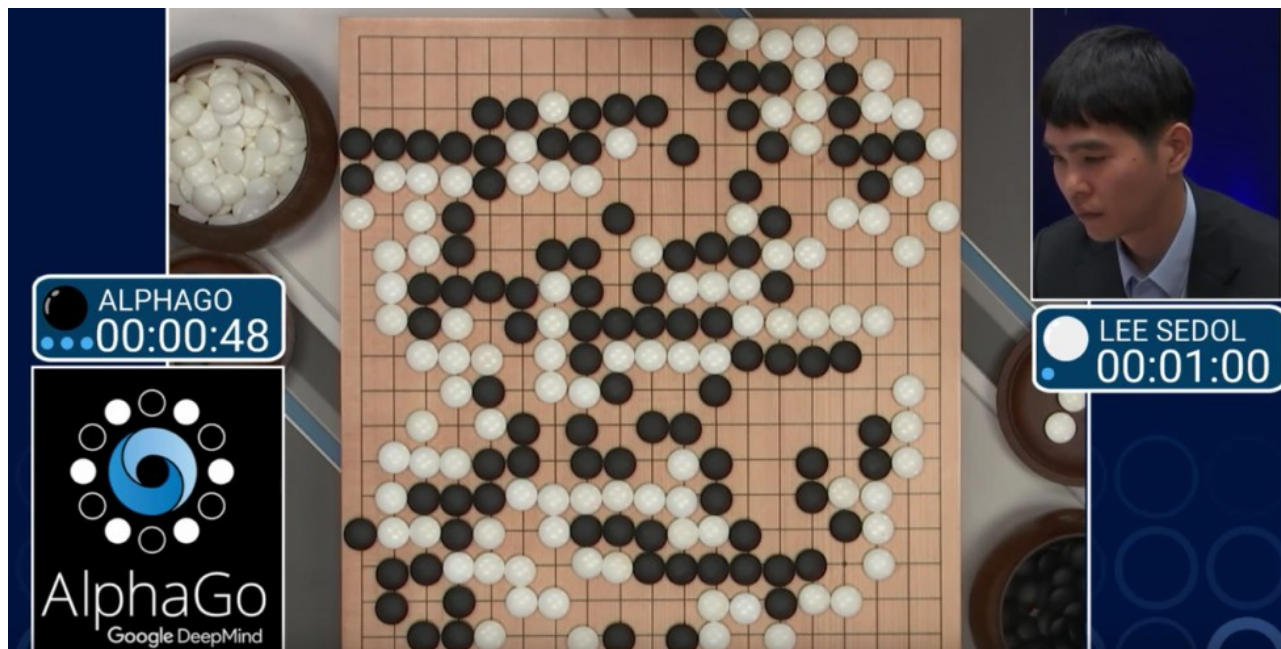
Generative models

Prompt: cat in Picasso flower



Reinforcement learning

- Games, Changing environment



AlphaGo

Course topics

- Classification
 - K-nearest neighbors
 - Naive Bayes
 - Decision tree
 - Support vector machine

Course topics

- Neural networks
 - Multilayer perceptron
 - Convolutional Neural networks
 - Recurrent neural networks
 - Transformer, Bert and GPT
 - Generative adversarial networks (GANs)
 - Denoising diffusion models

Course topics

- Unsupervised learning
 - K-mean clustering
 - Mixture models

Course topics

- Unsupervised learning
 - K-mean clustering
 - Mixture models
- Possible special topic(s)
 - Online learning
 - Recommender system
 - Reinforcement learning
 - Interpretable models
 - Causal inference