Machine Learning Engineer Course
Day7

- Basics of Machine Learning -

DIVE INTO CODE

Thursday, 22 April 2021

DIOP Mouhamed

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What did you learn in the previous week?
(Anything is fine.)

Agenda

1. Please Note
2. Today’s Word
3. Today's Objective
4. Quick Review
5. Class Assignment
6. Sample Code
7. ToDo by next class

🚨 Please Note

How to proceed with this course and precautions
You will be the leader in the IT industry in Vietnam.

❶ Advance at top speed

We do not bottom up

❷ Promote autonomous self-propelled

We do not accept unexplained questions

❸ Focus on problem-solving ability

We do not give lectures on building up the foundation

Today’s Word

The future is something we aim for and something we create.

Kazuto Ataka

"Shin Nihon": Rebirth of Japan and Human Resource Development in the Age of AI x Data
（ https://www.mof.go.jp/pri/research/conference/fy2017/inv2017_04_02.pdf ）

Today's Objective

Purpose of learning. Purpose clarifies a person's role and the learning required. Clear learning leads to a sense of growth and confidence.

	Objective	NOT Objective
1	**Learn how to think about the program with your peers**	Memorize lots of functions
2	**Use the basic elements of the program**	Complete assignments quickly
3	**Feel like a fresh business person**	

Today's Objective

“Understanding the basics of machine learning”

What are the fundamentals of machine learning in the first place?
It is not to be concerned with model building.

- **Iterate the implementation and execution of the machine learning sequence**

Today's Objective

“Feel like a fresh business person”

This is an important attitude to keep in mind when using analysis tools. Try to have the following image.

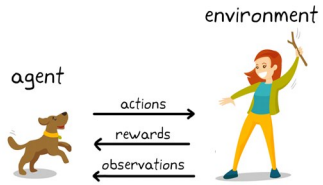
- **I am a fresh business person**
  - ○ **No domain knowledge**
  - ○ **I have data**
- **Report, communicate, and consult with your seniors and superiors at work**
  - ○ **Have a business goal in mind**

Review (Machine Learning)

The process of making a program which allows a computer to learn from data (could be anything, images – audio – texts)

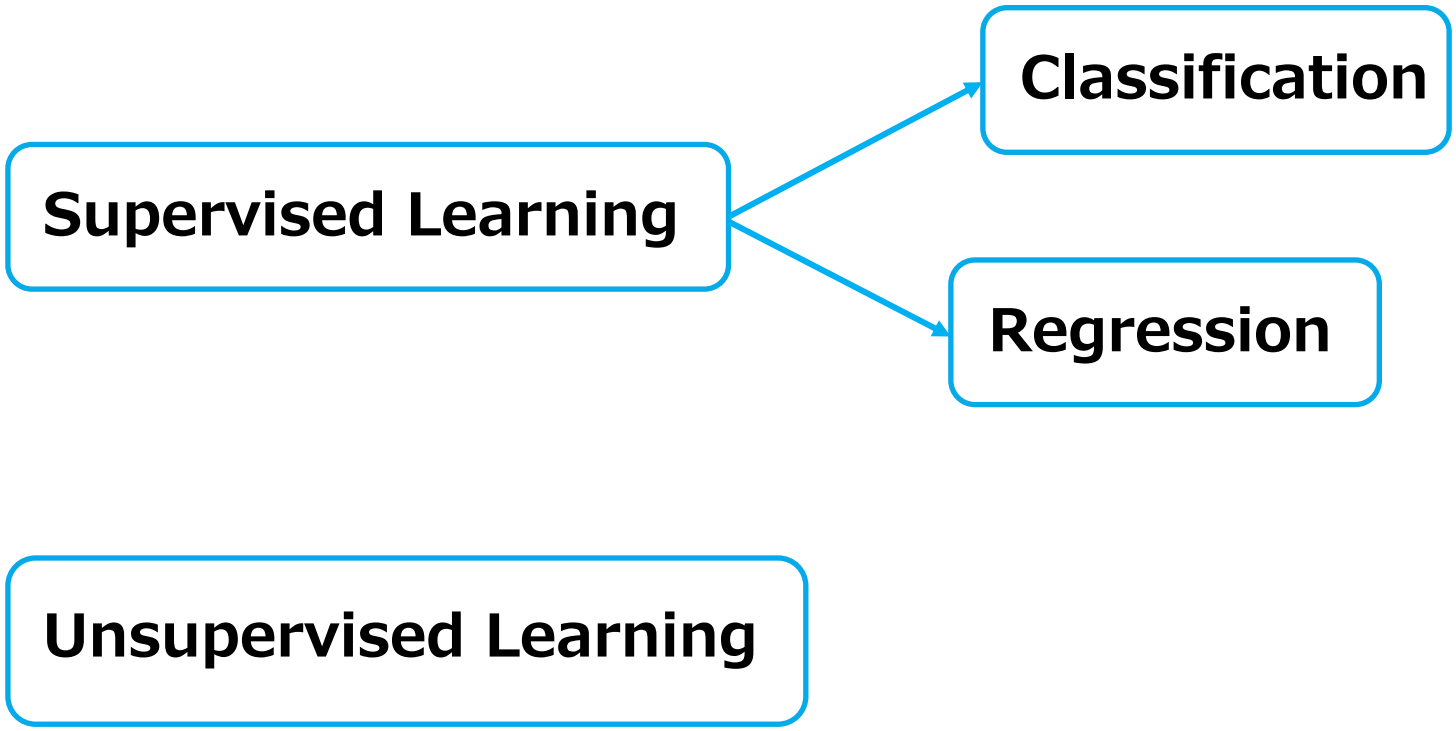
Common Use Cases

- **Computer Vision:** scanning documents using your phone's camera, making self-driving cars, image recognition (damages in products, cancer)
- **Natural Language Processing:** recognizing sounds
- **Reinforcement learning**

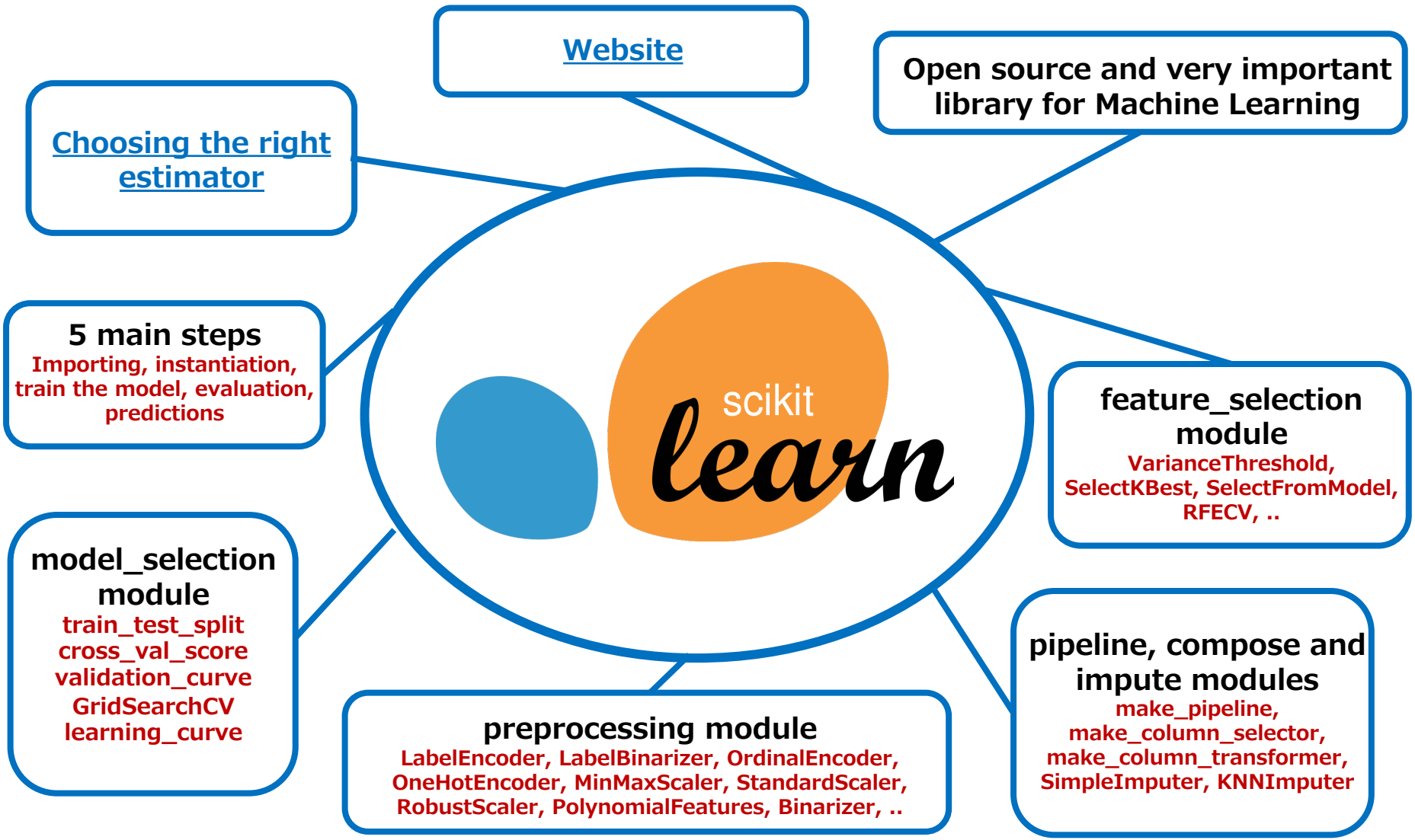


- **Recommendation Systems:** giving users recommendations

Review (Machine Learning)



Review (Scikit-learn)



Week 4 Assignments

Explanations about each one of them have been given except for the Class Assignment.

1. **Intro to ML - Pre-Class Assignment 1 Classification of Irises**
2. **Intro to ML - Pre-Class Assignment 2 Predicting House Prices**
3. **Intro to ML - Pre-Class Assignment 3 Utilization of Object Orientation**
4. **Intro to ML – Class Assignment Learning Credit information**

Please work on your own after class and submit your assignments on DIVER.

Class assignment

Use the skills you've acquired in the DIVER pre-class assignments to tackle more practical problems!

1. Learning about credit information
    a. Verifying Competition Details
    b. Learning and Verification
    c. Estimation on Test Data
    d. Feature Engineering

Reference
- HomeCredit_columns_description.csv
  https://www.kaggle.com/c/home-credit-default-risk/data

Class assignment

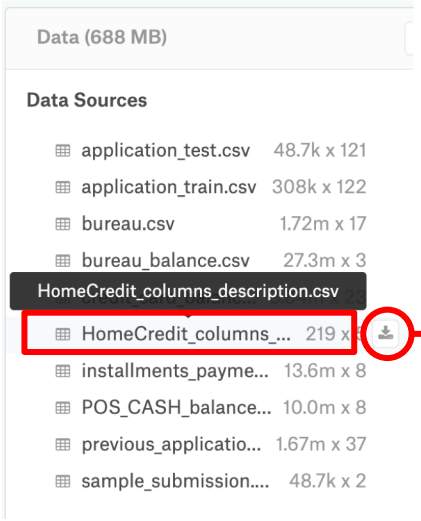
The flow of working on a Kaggle competition

1. Understanding the Problem Statement
2. Understand the evaluation indicators
3. Identify the ratio of Public to Private
4. EDA (Data Analysis)
5. Make First Submission
6. Create a function of indicator values
7. Preprocessing
8. Do Feature Engineering
9. Training
10. Evaluate with index values
11. Submit

Class assignment

https://www.kaggle.com/c/home-credit-default-risk/data

[Note] Take a look at HomeCredit_columns_description.csv to see the description of the columns in the dataset.

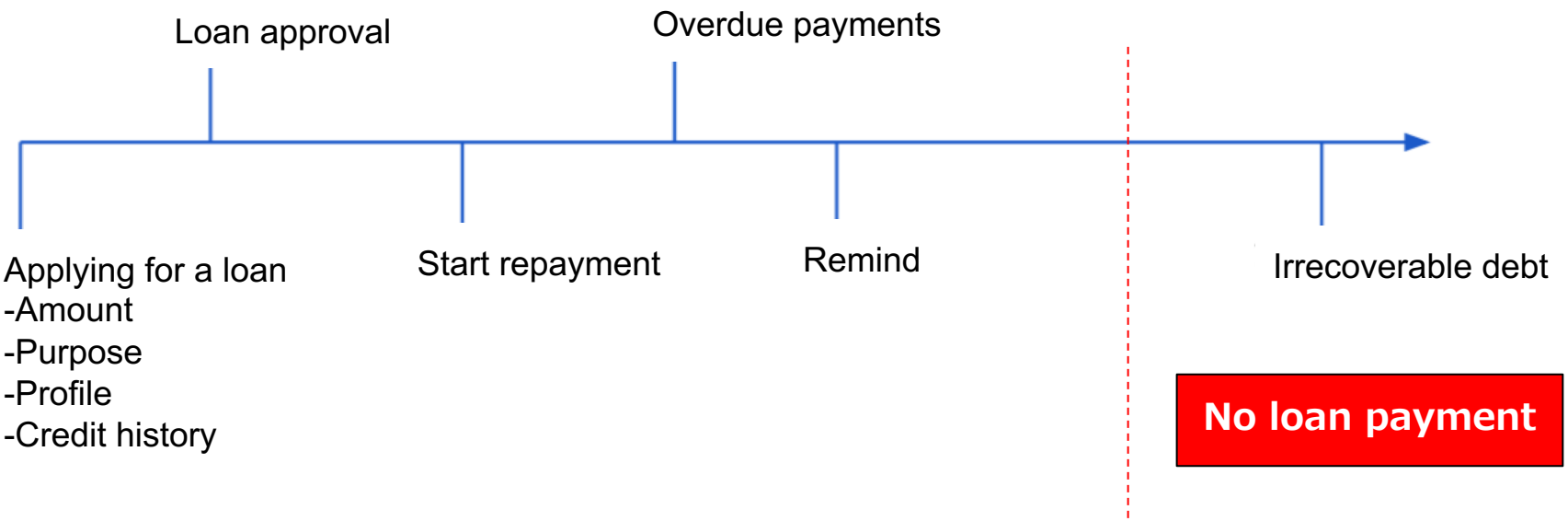


	A	B	C	D
1		Table	Row	Description
2	1	application_{trai	SK_ID_CURR	ID of loan in our sample
3	2	application_{trai	TARGET	Target variable (1 - client with payment difficulties
4	5	application_{trai	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
5	6	application_{trai	CODE_GENDER	Gender of the client
6	7	application_{trai	FLAG_OWN_CAR	Flag if the client owns a car
7	8	application_{trai	FLAG_OWN_REALTY	Flag if client owns a house or flat
8	9	application_{trai	CNT_CHILDREN	Number of children the client has
9	10	application_{trai	AMT_INCOME_TOTAL	Income of the client
10	11	application_{trai	AMT_CREDIT	Credit amount of the loan
11	12	application_{trai	AMT_ANNUITY	Loan annuity
12	13	application_{trai	AMT_GOODS_PRICE	For consumer loans it is the price of the goods fo
13	14	application_{trai	NAME_TYPE_SUITE	Who was accompanying client when he was appl
14	15	application_{trai	NAME_INCOME_TYPE	Clients income type (businessman, working, mat
15	16	application_{trai	NAME_EDUCATION_TYPE	Level of highest education the client achieved
16	17	application_{trai	NAME_FAMILY_STATUS	Family status of the client
17	18	application_{trai	NAME_HOUSING_TYPE	What is the housing situation of the client (renting
18	19	application_{trai	REGION_POPULATION_RELATIVE	Normalized population of region where client live
19	20	application_{trai	DAYS_BIRTH	Client's age in days at the time of application
20	21	application_{trai	DAYS_EMPLOYED	How many days before the application the perso
21	22	application_{trai	DAYS_REGISTRATION	How many days before the application did client
22	23	application_{trai	DAYS_ID_PUBLISH	How many days before the application did client
23	24	application_{trai	OWN_CAR_AGE	Age of client's car
24	25	application_{trai	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)

Class assignment

How to increase the number of features

- **Begin by reading articles in that specialized field**
- **Study the domain knowledge**
- **Interview people in the industry**
- **Try to follow the data over time**

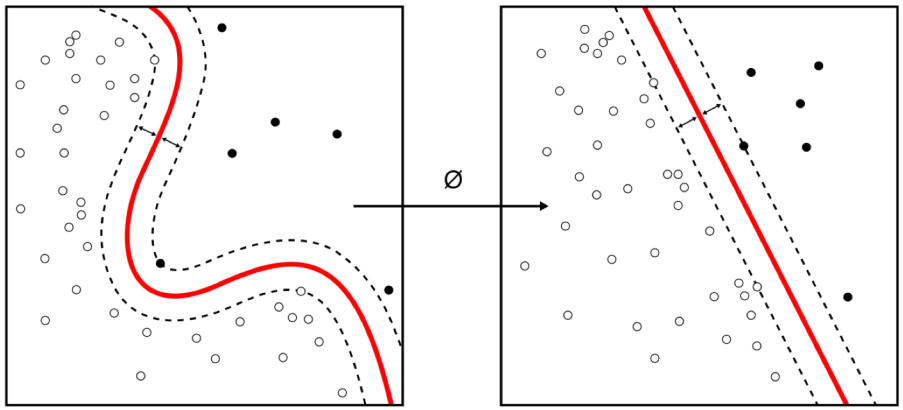


From DataRobot Essentials hands-on training materials

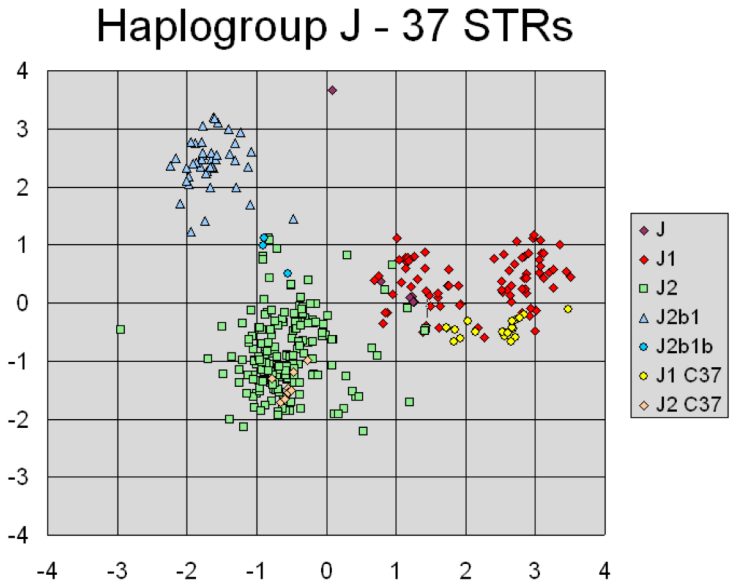
Class assignment

How to reduce the number of features

- **Use unsupervised learning**
- **Perform Principal Component Analysis (PCA)**



From Wikipedia



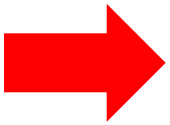
Class assignment

Hypotheses on why we have missing data.

1. Random stuff

2. It's practically deficient
    a. No data, not open for business

3. It's arbitrarily missing
    a. Not completing the annual income survey

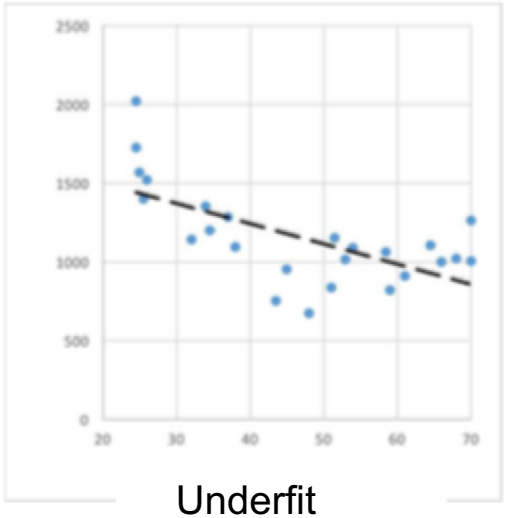


Consider whether to use dummies or not, depending on the reason for the deficiency

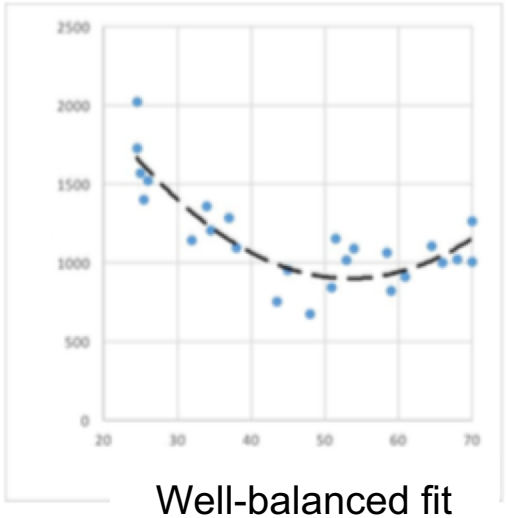
Class assignment

It is preferable to be in a state where it is learned for general use. Prevent overtraining and improve generalization performance.

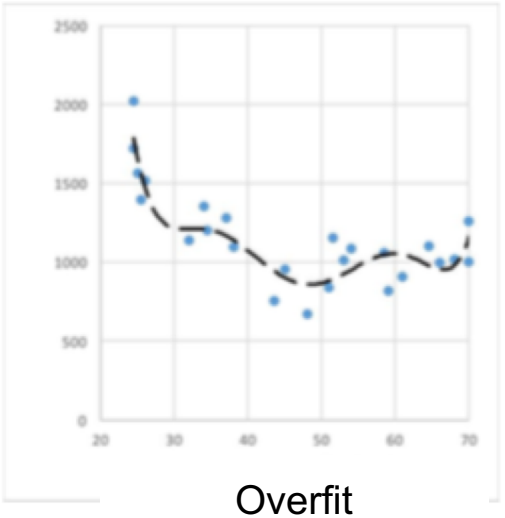
Insufficient learning ← **Moderate** → **Indiscriminate memorizing**



Underfit



Well-balanced fit



Overfit

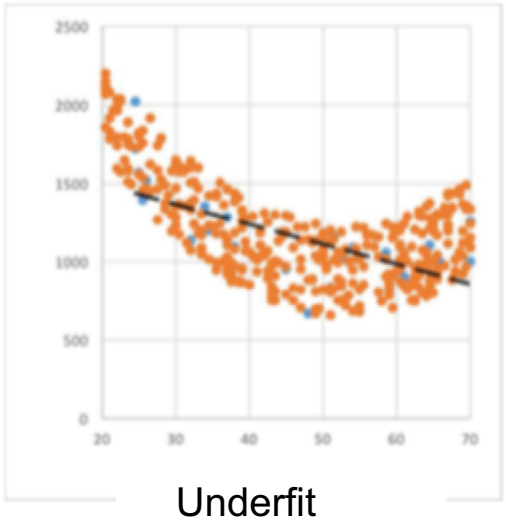
From DataRobot Essentials hands-on training materials

Alias: Over Learning

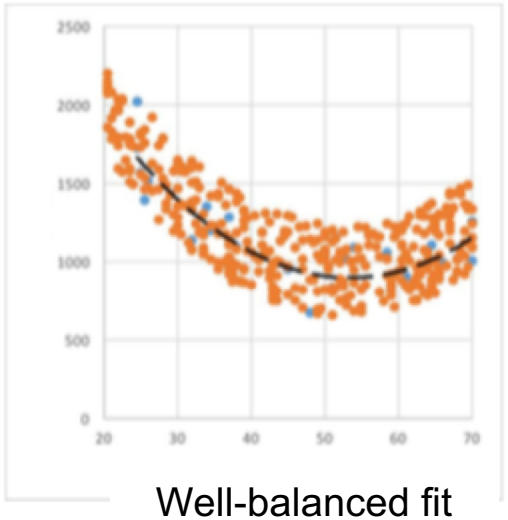
Class assignment

In prediction using unknown data, the accuracy of prediction will decrease, whether it is due to lack of learning or rote learning.

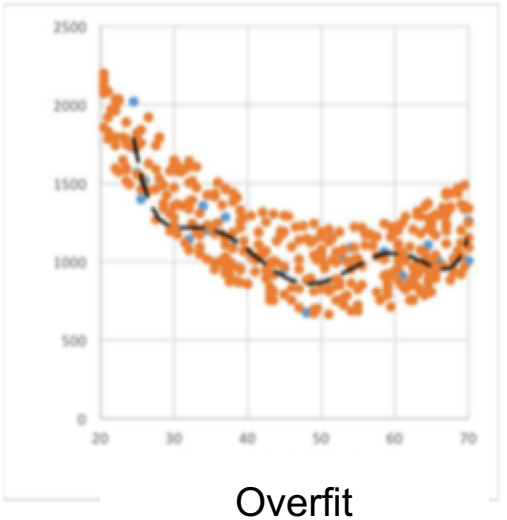
Insufficient learning ← **Moderate** → **Indiscriminate memorizing**



Underfit



Well-balanced fit



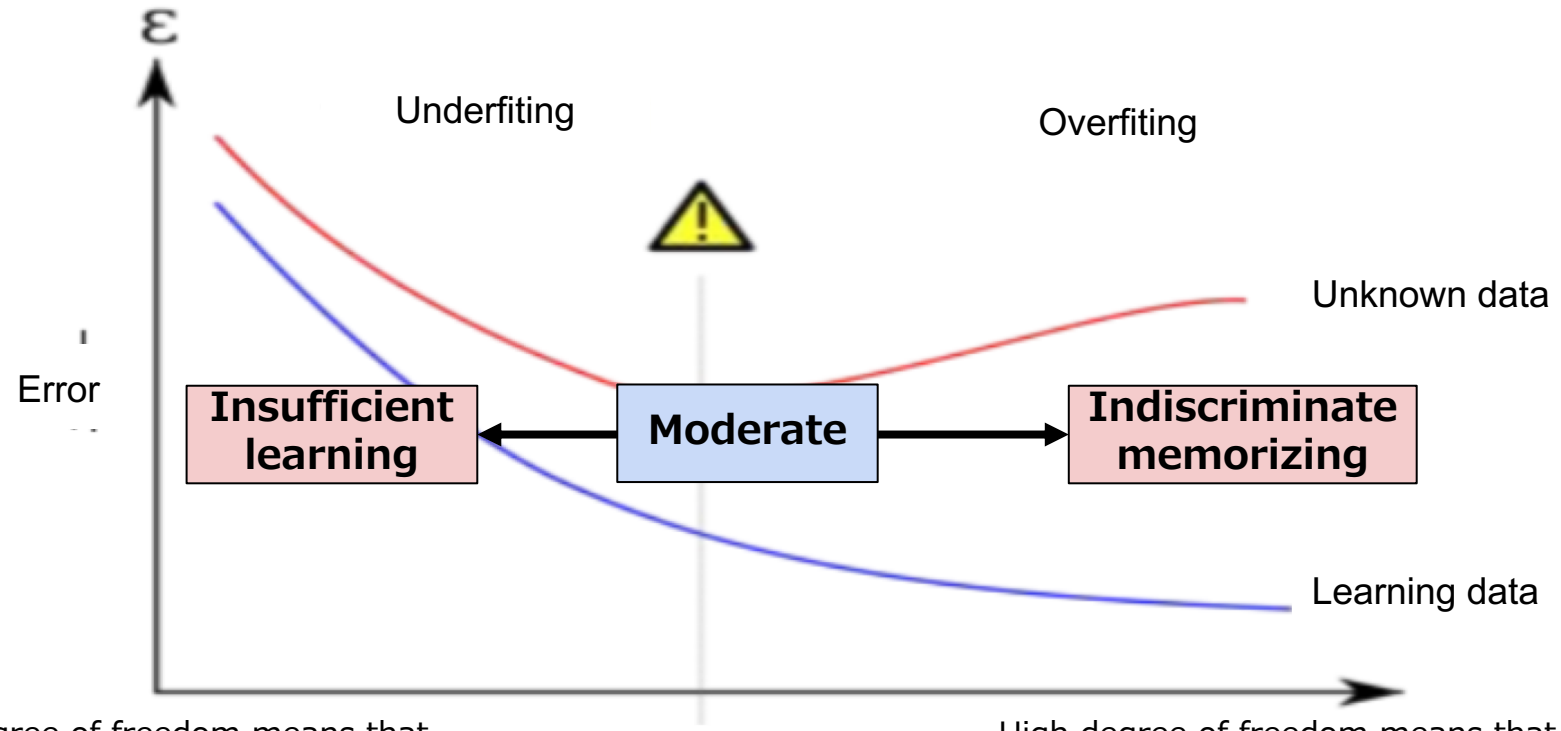
Overfit

"Too much of a good thing is too little of a good thing."
It is important to have generalization performance

From DataRobot Essentials hands-on training materials

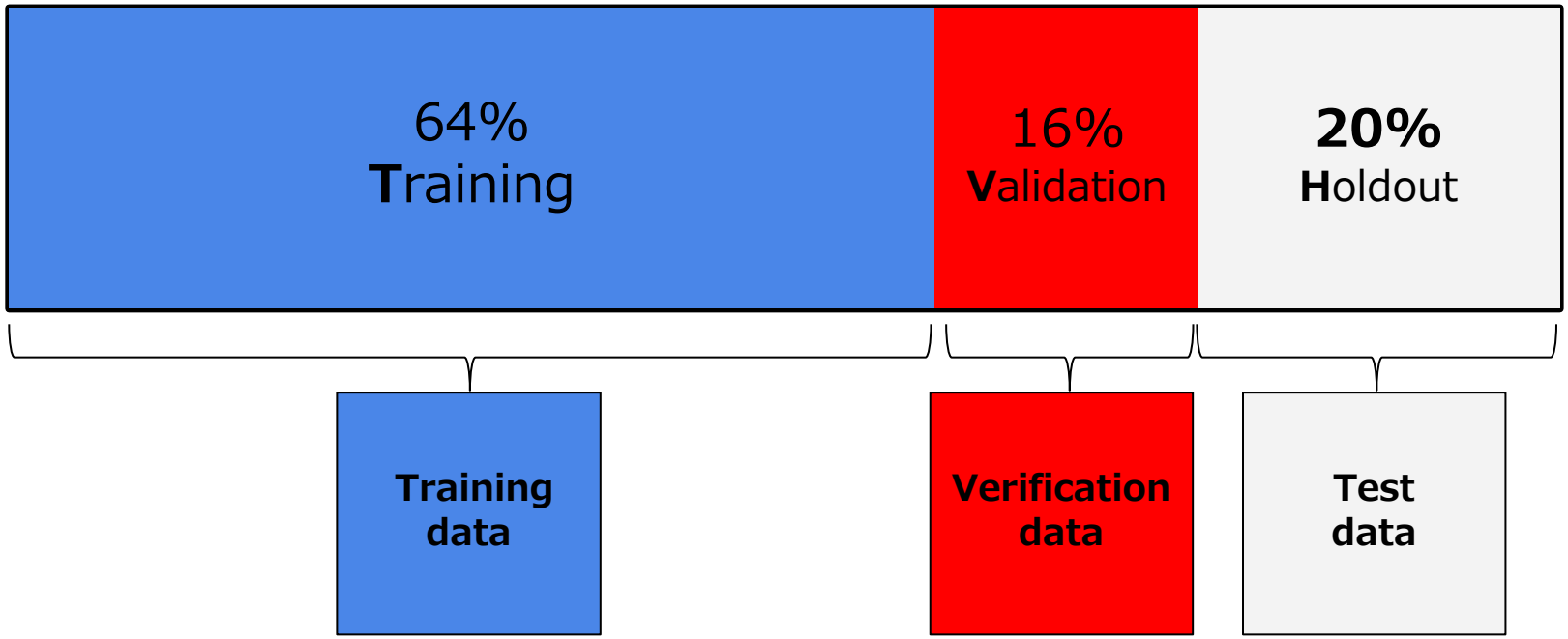
Class assignment

Whether the accuracy is moderate or not can be recognized by the difference in accuracy (error rate) between the data used for training and the unknown data that was tried to be predicted.



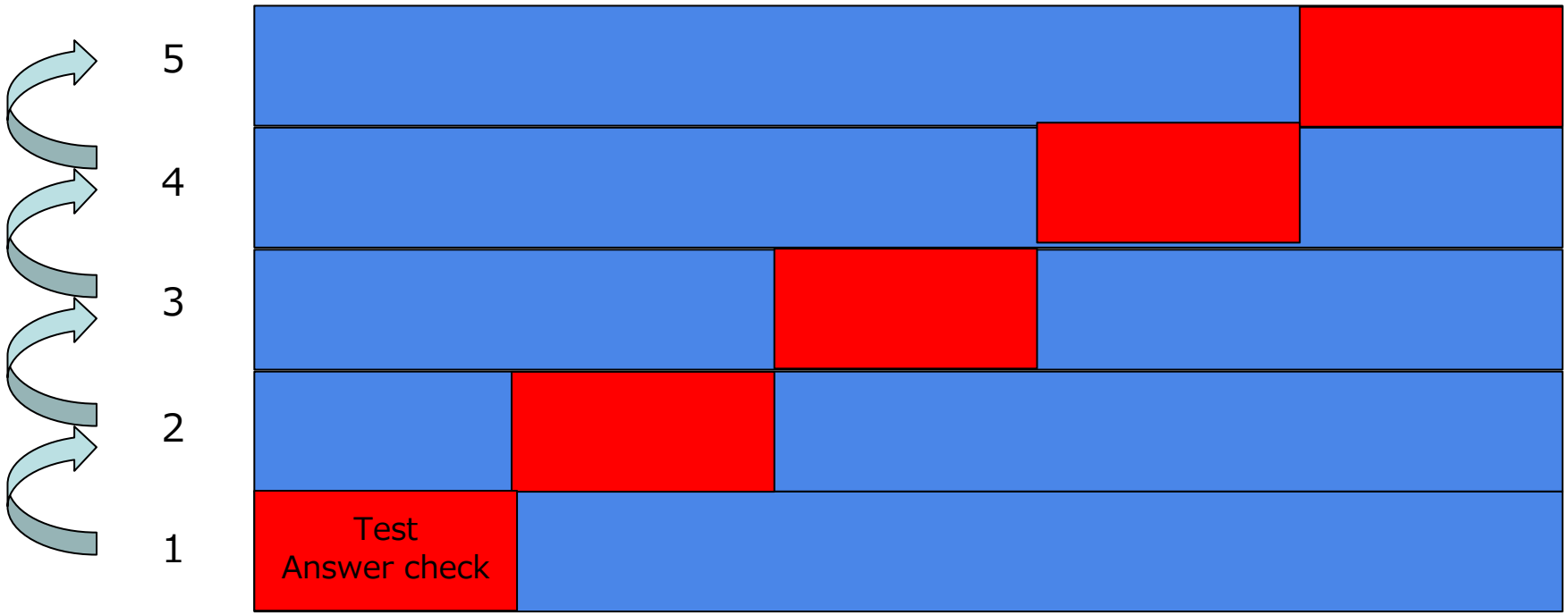
Class assignment

To avoid under- and overfitting, we split the data. The following is an example of splitting: in Kaggle, the test data is separate from the beginning, so a part of the training data is used as the validation data.



Class assignment

Ideally, cross-validation "cross-validation" should be performed to try multiple splits of data for training and validation. (Will be handled in Sprint 1)



Sample code

How to solve the problems

“Learning Credit Information”

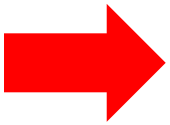
[Problem 1] Confirmation of competition contents
[Problem 2] Learning and verification
[Problem 3] Estimation on test data
[Problem 4] Feature engineering

Start Here: A Gentle Introduction
https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

Week 4 Assignments

Explanations about each one of them will be given but please try them on your own first.

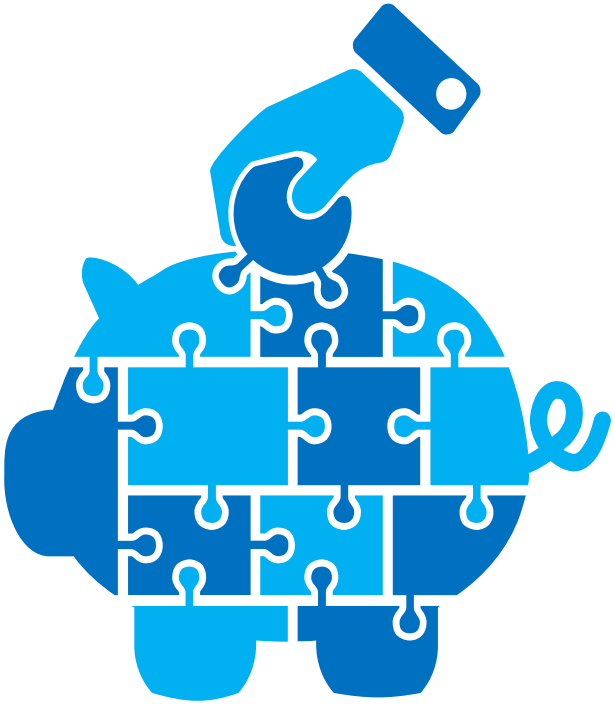
1. **Intro to ML - Pre-Class Assignment 1 Classification of Irises**
2. **Intro to ML - Pre-Class Assignment 2 Predicting House Prices**
3. **Intro to ML - Pre-Class Assignment 3 Utilization of Object Orientation**
4. **Intro to ML – Class Assignment Learning Credit information**



Please work on your own after class and submit your assignments on DIVER.

Done with the Pre-Learning

Congratulations
Pre-Learning



Heading to the Sprints
Machine Learning
Deep Learning

ToDo by next class

Next class will be Zoom : Thursday, 29 April 2021

🚨ToDo: Sprint 1 - Machine Learning Flow
https://diver.diveintocode.jp/curriculums/1642

Check-out

3 minutes Please post the following point to Zoom chat.

Q. Current feelings and reflections
(joy, anger, sorrow, anticipation, nervousness, etc.)

Thank You For Your Attention