The 23rd Summer Course
Behavior Modeling in Transportation Networks

Statistical Estimation with Machine Learning

Junji Urata

September 12, 2024

University of Tsukuba

Outline

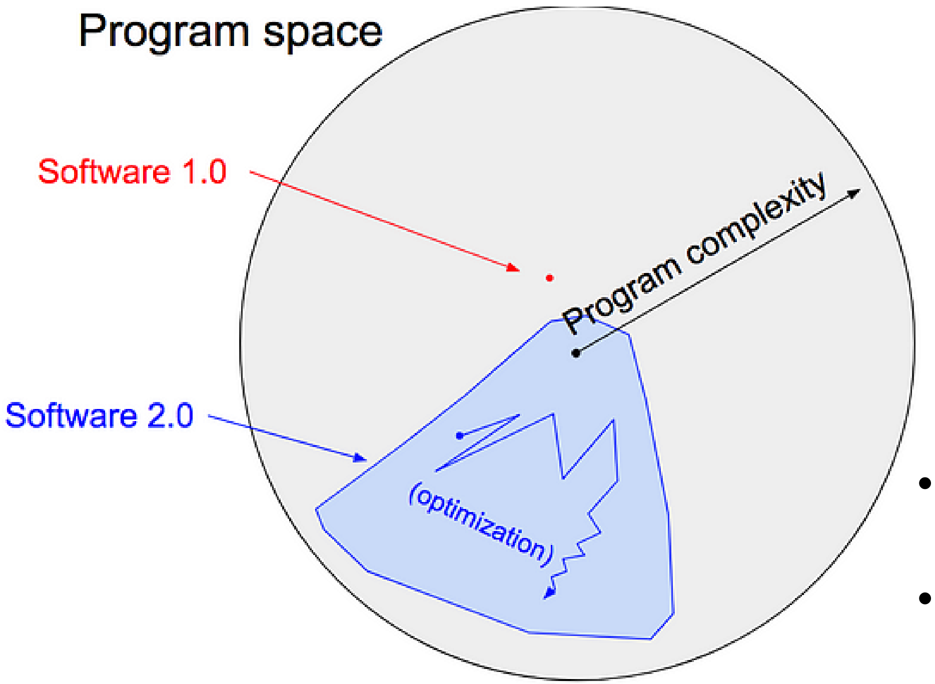
1. Introduction
2. Un-supervised learning
3. Supervised learning
    1. Data preparation
    2. Evaluation
    3. Estimation
    4. Model
        1. Neural Network
        2. Support Vector Machine Classification

What is Machine Learning?

Arthur Samuel (1959)

> "Field of study that gives computers the ability to learn without being explicitly programmed"

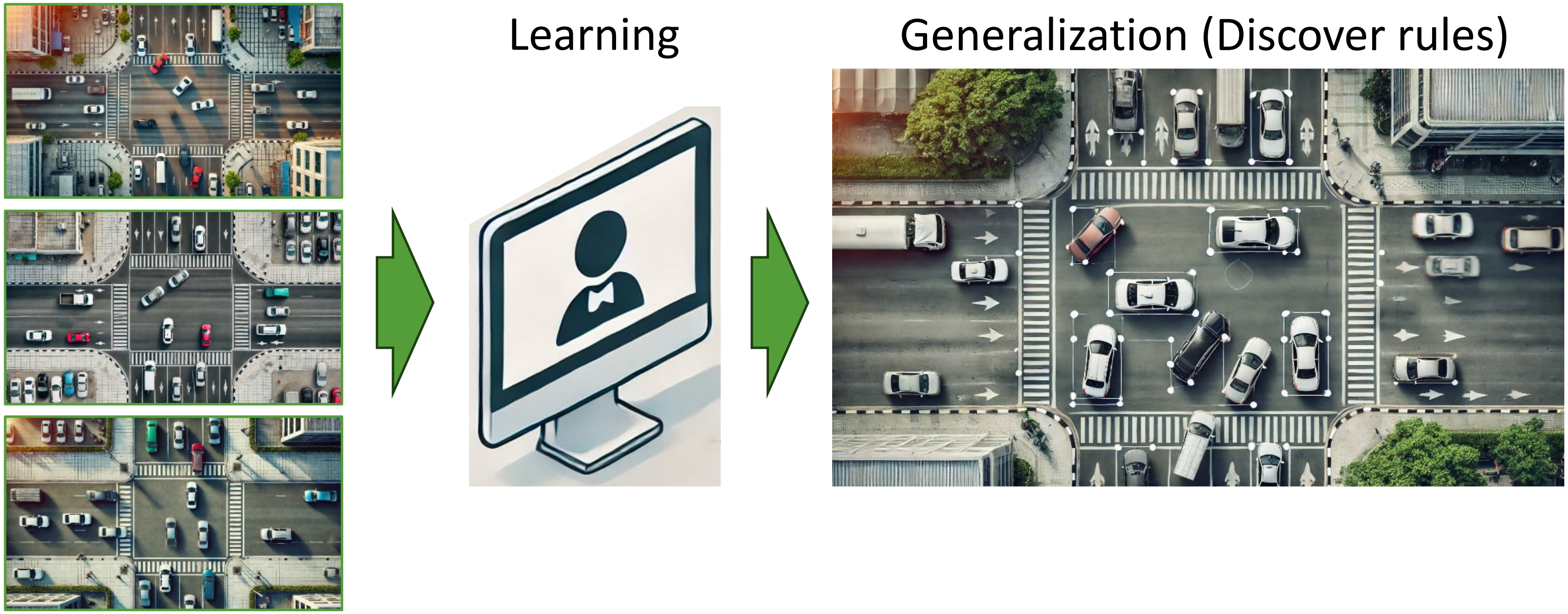
Software 2.0 *Andrej Karpathy (2017)*



- In the realm of program complexity and data, there are countless possible models.
- Software 2.0 is about searching for the optimal model from the complexity.

What is Machine Learning?

- In Machine learning, we program the method of learning itself so that the computer can discover rules from accumulated data
- What we program is not how to recognize something, but how to learn.
- The goal of machine learning is to correctly predict results even for unknown data that is not in the training data



Training Data

Learning

Generalization (Discover rules)

* Images are made in ChatGPT 4o

Applying Machine learning for what?

Demand forecasting

- The power demand for the day is predicted based on the weather.
- Predicting road traffic demand of each link

Product recommendations

- Displaying recommended products on Amazon based on past purchase data
-Recommending transportation options in Mobility as a Service (MaaS)

Anomaly detection

- Predicting printer failures in advance based on past failure data
- Detecting congestion or traffic accidents

Images with anomaly

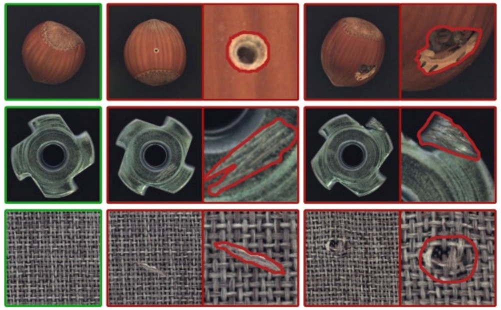
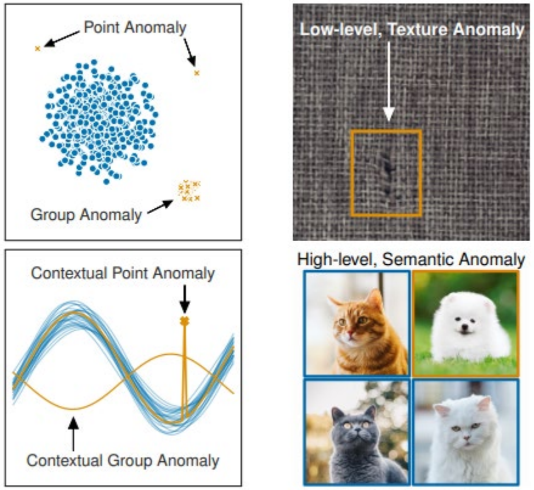


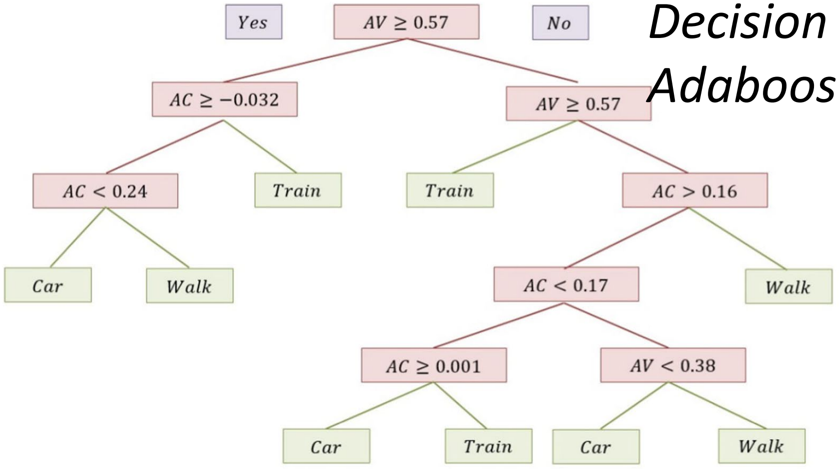
Photo by [Bergmann et al., 2019]



Ruff, Lukas, et al. "A Unifying Review of Deep and Shallow Anomaly Detection." arXiv preprint arXiv:2009.11732 (2020).

Machine learning for what?

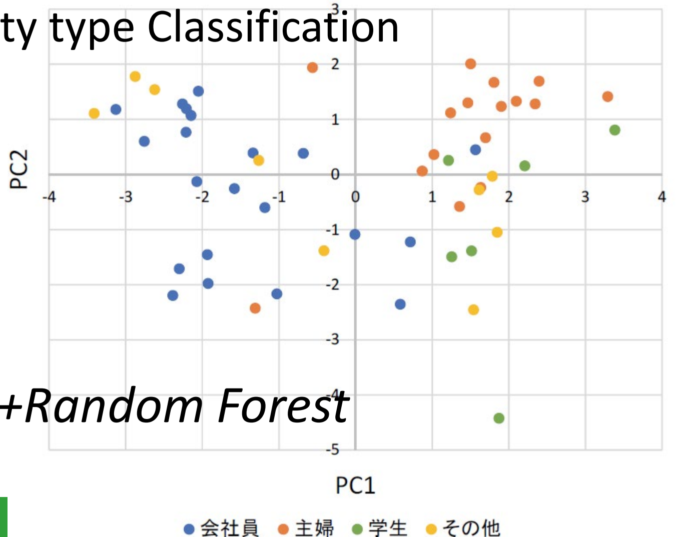
Shafique & Hato (2015): Traffic mode detection



Decision Tree
Adaboost …

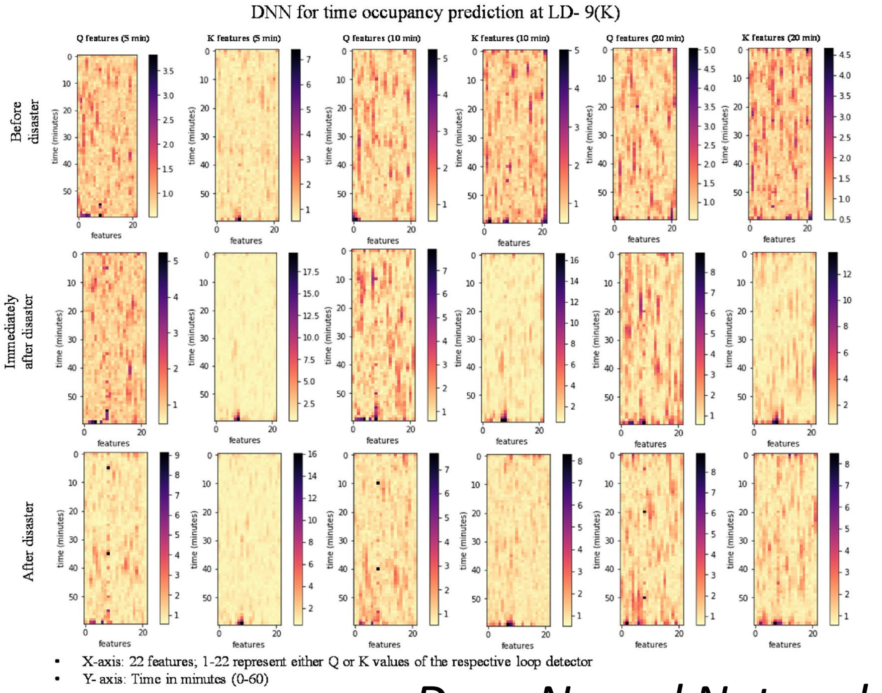
AV = Average acceleration in vertical direction (G)
AC = Average acceleration in cross-wise direction (G)

Sueki, Hara, Sasaki et al. (2018):
Activity type Classification



PCA+Random Forest

Chikaraishi, Varghsee et al. (2020):
Time Occupancy predition



- X-axis: 22 features; 1-22 represent either Q or K values of the respective loop detector
- Y- axis: Time in minutes (0-60)

Deep Neural Network

Machine learning Types

Supervised learning

- To learn the true relationship between input and output from the training data.
- The training data consists of pairs of input data and corresponding output values.

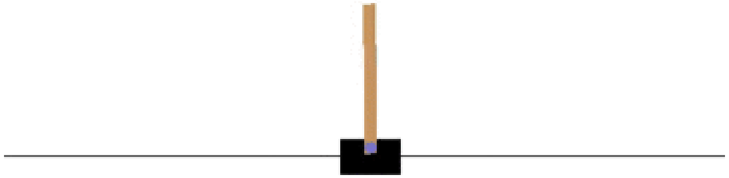
Unsupervised learning

- To discover useful knowledge from the data
- Using only input data in the training process

Reinforcement learning

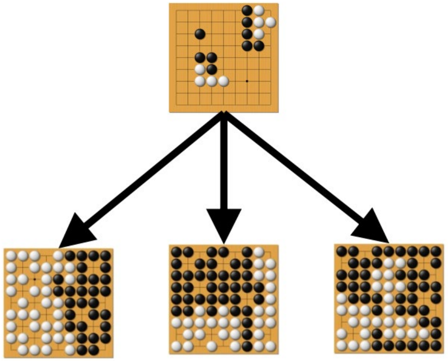
- To learn a set of actions that maximizes the cumulative reward over time.
- Rewards are given for taking specific actions.

CartPole problem



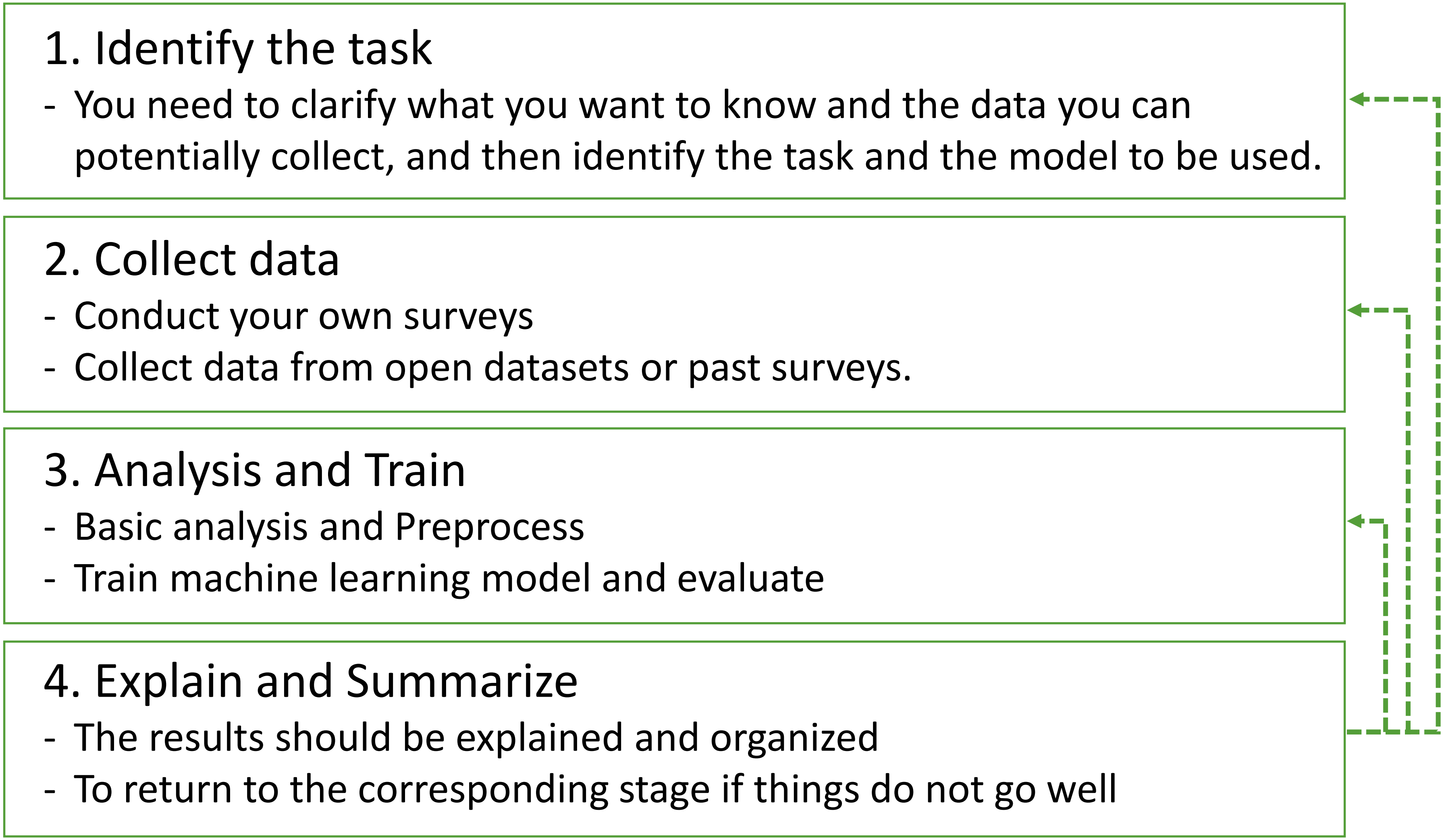
https://blog.brainpad.co.jp/entry/2017/02/24/121500

Igo game



美添（2019）

Flowchart of Machine Learning



1. Identify the task

- You need to clarify what you want to know and the data you can potentially collect, and then identify the task and the model to be used.

2. Collect data

- Conduct your own surveys
- Collect data from open datasets or past surveys.

3. Analysis and Train

- Basic analysis and Preprocess
- Train machine learning model and evaluate

4. Explain and Summarize

- The results should be explained and organized
- To return to the corresponding stage if things do not go well

Outline

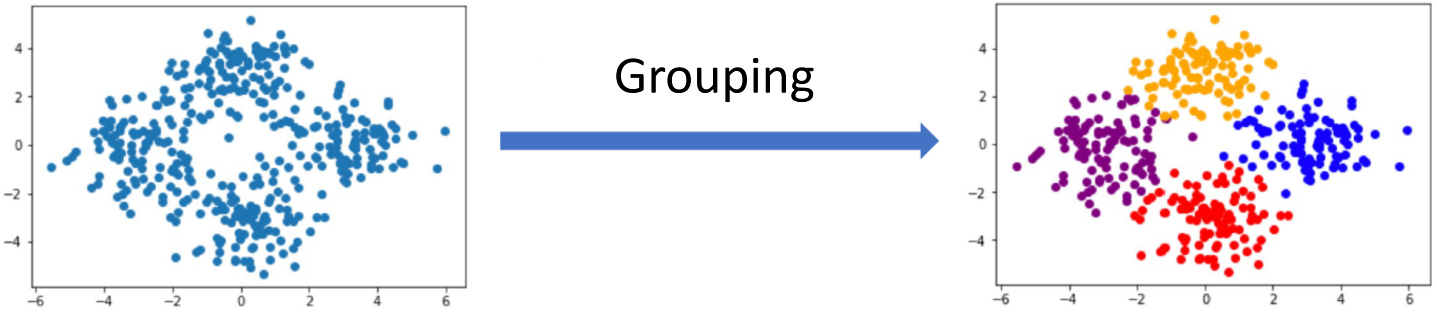
1. Introduction
2. Un-supervised learning
3. Supervised learning
    1. Data preparation
    2. Evaluation
    3. Estimation
    4. Model
        1. Neural Network
        2. Support Vector Machin

Un-supervised Learning

- Unsupervised learning aims to estimate the underlying structure or processes that generate the input data
- The input data consists of a collection of feature vectors

Ex.

Clustering is a method that groups data based on similarities



Grouping

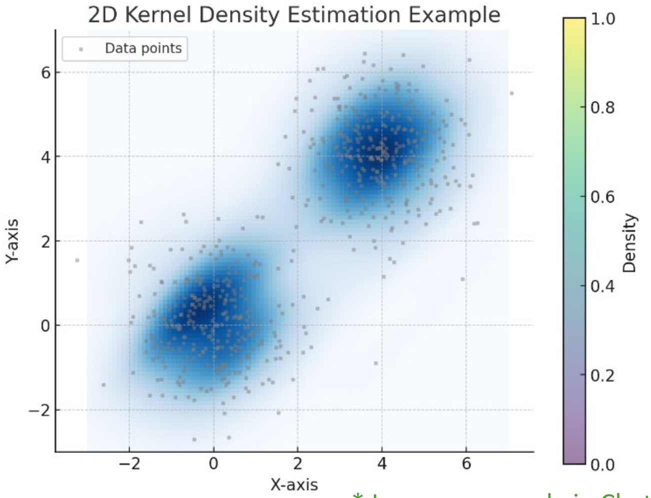
you can cluster

- people based on previous behaviors and personal attributes

- destination zones based on attributes of the facilities and people who stay.

Examples - Un-supervised Learning

Kernel Density Estimation (KDE)



* Images are made in ChatGPT 4o

- non-parametric method that estimates the probability density function based on data, without assuming a specific distribution
- KDE can be applied to anomaly detection or detecting rare phenomena.

Self-Organizing Maps (SOM)

https://swdrsker.hatenablog.com/entry/2016/12/08/171356

- the nonlinear structure of high-dimensional data onto a 2D grid while preserving the relationships in the data
- Using neural networks, SOM projects data patterns and distributions onto a 2D grid

Outline

1. Introduction
2. Un-supervised learning
3. Supervised learning
    1. Data preparation
    2. Evaluation
    3. Estimation
    4. Model
        1. Neural Network
        2. Support Vector Machine Classification

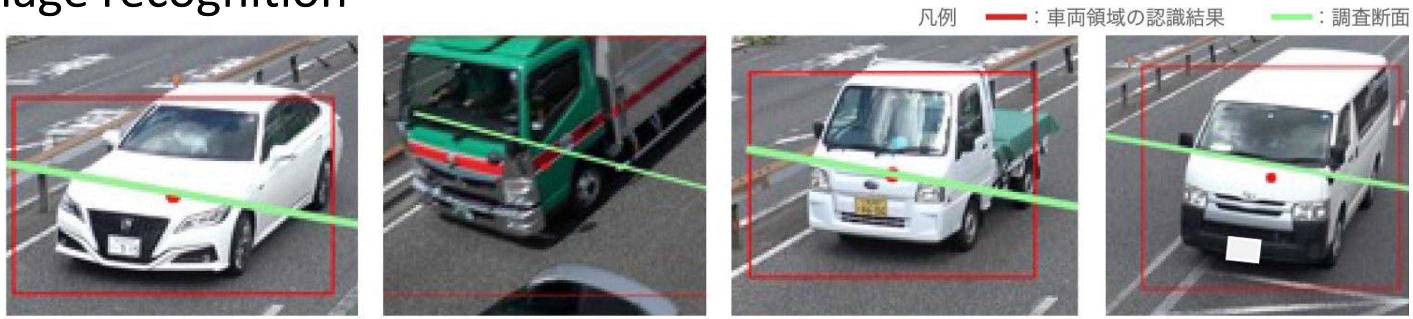
Purpose and Data

The goal is to build a model that can correctly predict y from **x** by learning from this input data.

The data in supervised learning consists of the pair of

- **x**: features of the data
- y: target value.

Ex. Image recognition



x

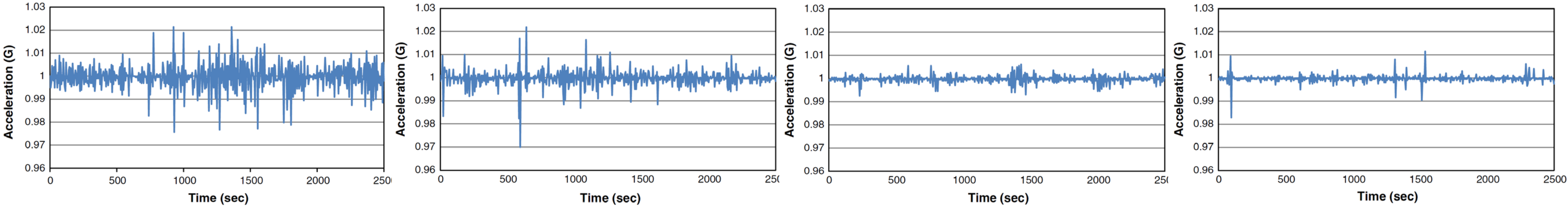
y Small vehicle Large vehicle Large vehicle Small vehicle

https://www.intelligentstyle.co.jp/product/aitraffic/

Ex. Traffic Mode detection

Shafique & Hato (2015)

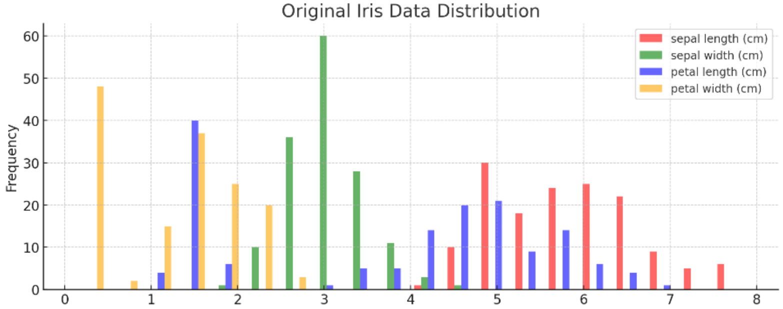
x



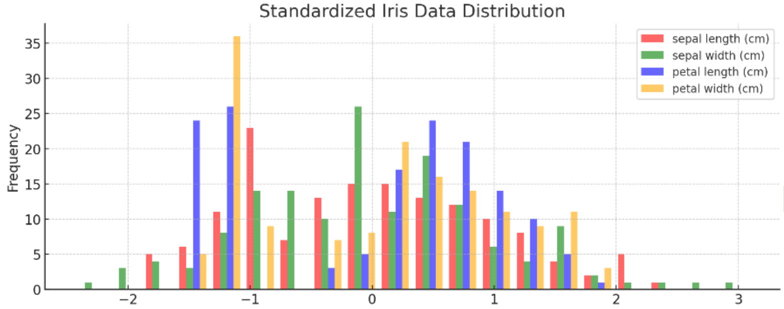
y Walk Bicycle Car Train

Data Preprocess

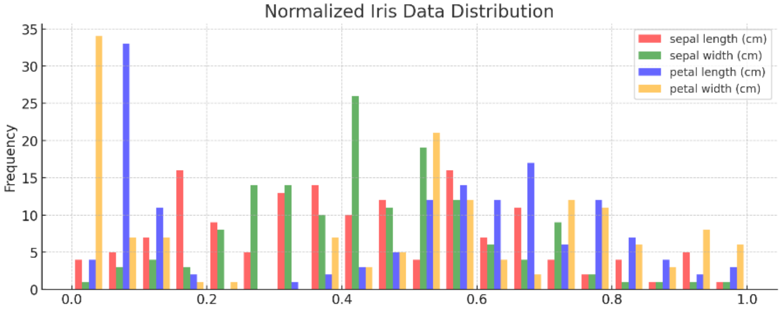
- The input data may contain both continuous data and categorical data
    - Categorical data need to be transformed by creating dummy variable
- Major Preprocessing
    - Standardization: mean of 0 & variance of 1
    - Normalization: range [0, 1] or [-1, 1]
- The purpose is to make the absolute values and distributions of the various variables more comparable. By doing this, the model estimation process proceeds more smoothly.


Original Iris Data Distribution

Standardization


Standardized Iris Data Distribution

Normalization


Normalized Iris Data Distribution

* Images are made in ChatGPT 4o with iris data

Model: Regression and Classification

- If the dependent variable y in the training data is continuous, a regression model is built.
- If y is discrete, a classification model is built.

Ex. Regression

- Sales forecast: y = Net sales of stores/day, x = Store square footage, type of business, weather, day of the week
- Bus demand: y = Number of passengers/h, x = Routes, temperatures, rain, school vacations, day of the week

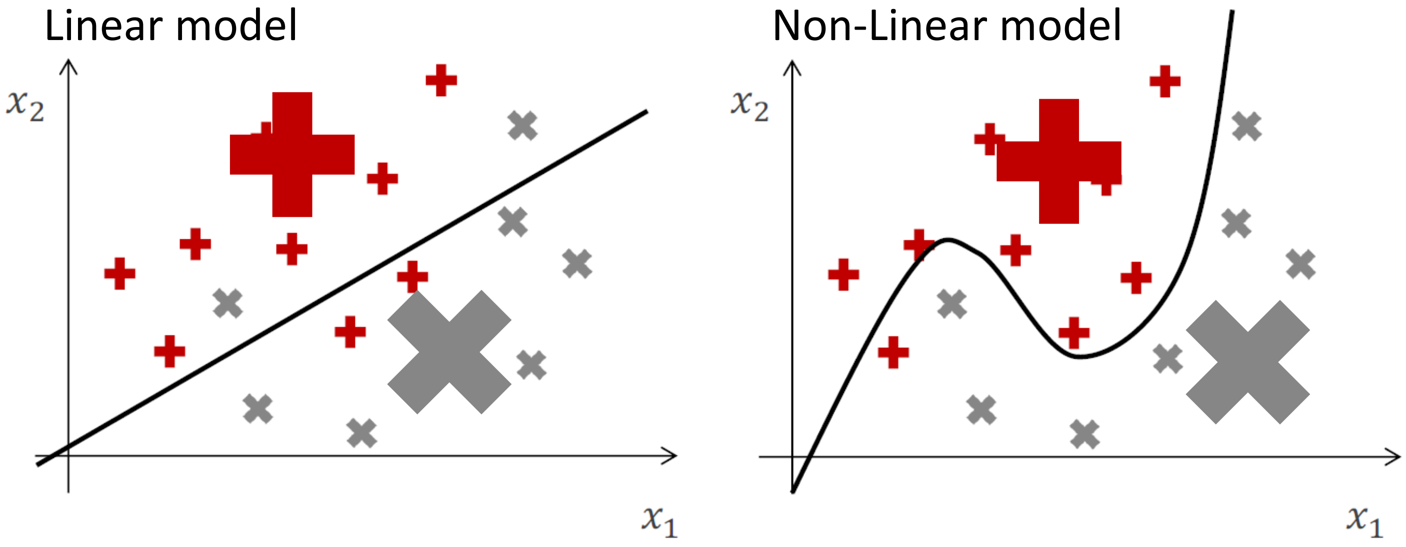
Ex. Classification

- Diseased: y = diseased or not, x = Blood test, saliva test, temperature, facial expression, radiographs
- Traffic mode: y = mode, x = Travel time, fares, access & egress time, temperature, driver license, accompany person

Classification

- Predictions are made using a model that outputs continuous values $f_w(\mathbf{x})$ based on the model's parameters **w**
- The parameter **w** are estimated via training the model

Ex. Classification in Training data


Linear model

Non-Linear model

x_1, x_2: explanatory variable Labels(y): ✚ = 1, ✖ = -1

Objective function (Loss function)

The parameter estimation for the model evaluates how well the model fits the data

Loss function L(**w**)
= Distance between the model output f_w(**x**) and the label y from the training data

Finding parameter **w** by $\min_{\boldsymbol{w}} \sum_i L(\boldsymbol{w}, (x, y)_i)$

Ex. Hinge loss

$$\text{Hinge Loss L} = \max(0, 1\ - y_i\hat{y}_i)$$

y_i: (observed)Label, $\hat{y}_i$: model output
$y_i, \hat{y}_i = (-1, 1)$

Ex. Cross Entropy loss

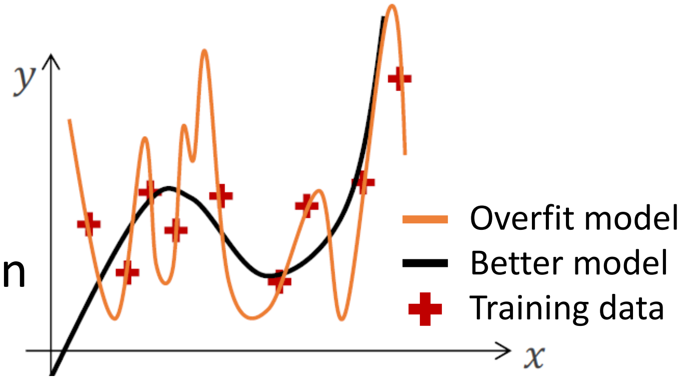
$$\text{Cross Entropy Loss L} = -\left[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right]$$

$\hat{p}_i$: model output (probability)
$\hat{p}_i \in (0, 1)$

Overfit and regularization

What is overfit?

- Overfitting occurs when the model becomes overly complex and fits too well to the training data
- Overfitting makes the model ineffective for prediction



To reduce the complexity

Regularization?

- We add the regularization term on the objective function
- The term prevents an excessive number of parameters from taking on non-zero values, allowing the model to remain suitable for making accurate predictions.

$$\min_{\boldsymbol{w}} \sum_i L(\boldsymbol{w}, (x, y)_i) + \lambda R(\boldsymbol{w})$$

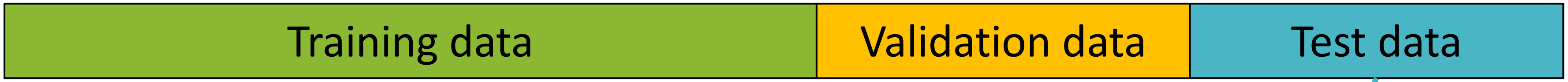
L2 (Ridge regression): $R(w) = \sum_j w_j^2$

L1 (Lasso regression): $R(w) = \sum_j |w_j|$

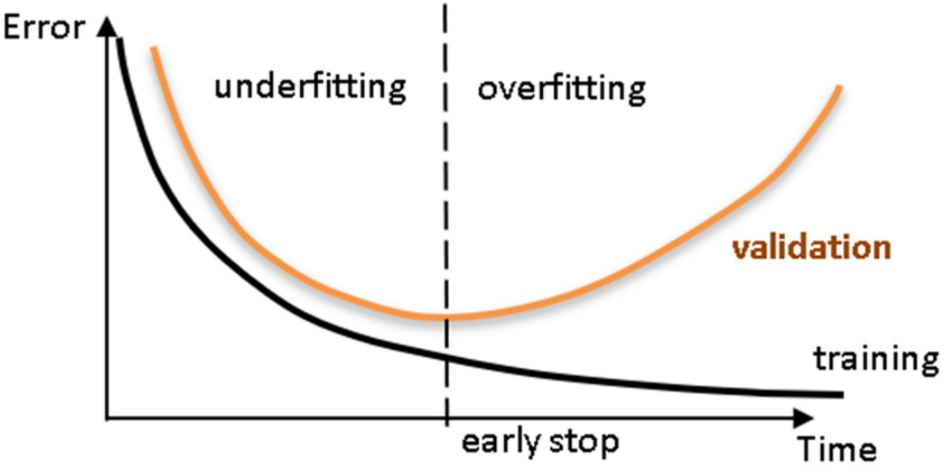
Cross-Validation

Early Stopping

1. The data is split into training and validation sets during parameter estimation
2. The parameters are estimated using the training data
3. To calculate the loss function with the estimated parameters using the validation data
4. Iterate 2 & 3, and then the final model is selected based on the parameters that perform well on the validation data



Training data | Validation data | Test data



The model is ultimately evaluated based on its accuracy on the test data, which is completely independent of the data used for parameter estimation.

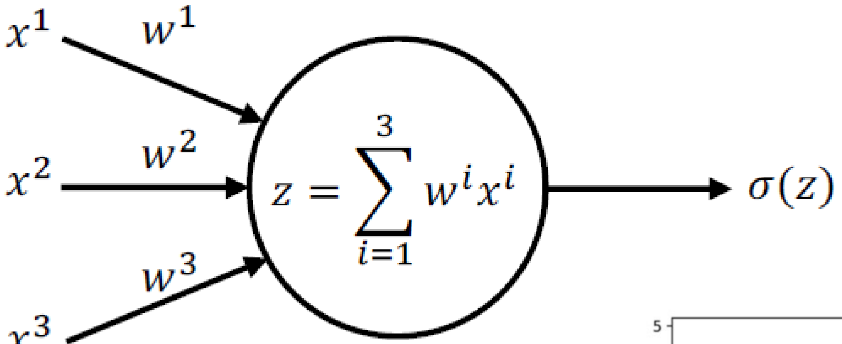
Graph by Abambres, & Lantsoght, Eva. (2019). ANN-Based Fatigue Strength of Concrete under Compression. Materials. 12. 3787. 10.3390/ma12223787.

Outline

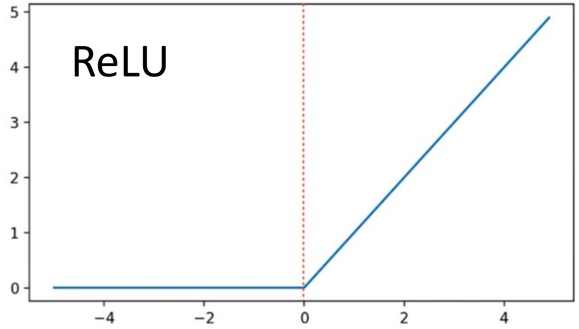
1. Introduction
2. Un-supervised learning
3. Supervised learning
    1. Data preparation
    2. Evaluation
    3. Estimation
    4. Model
        1. Neural Network
        2. Support Vector Machine Classification

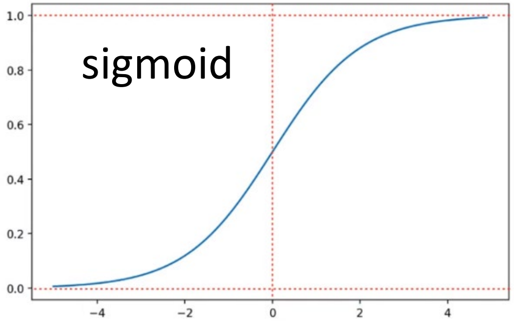
What is Neural Network?

- Neural networks are machine learning models inspired by the brain.
- Neurons (nodes) are connected through weighted edges, forming a network.
- Neurons receive signals z/x from other connected neurons according to the weights w on these edges, and then apply a transformation called an activation function σ to output a signal.
- Activation function – ReLU function, sigmoid function etc



$$\sigma(z) = \begin{cases} \max\{0, z\} & \text{(ReLU)}, \\ \dfrac{1}{1 + e^{-z}} & \text{(sigmoid)}. \end{cases}$$



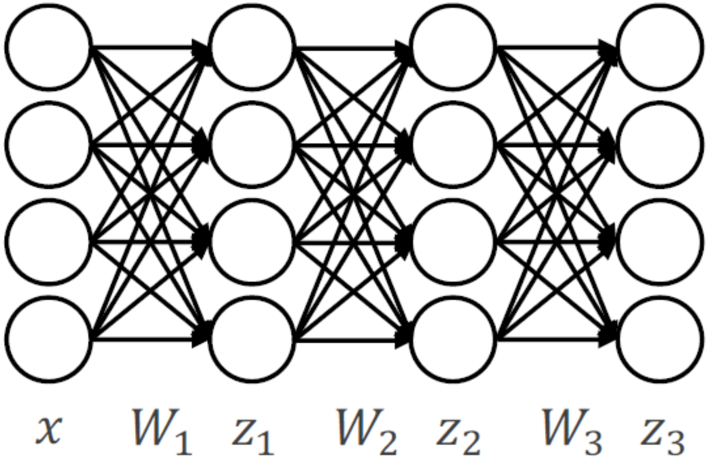


https://onl.tw/r3E6iRE

(Deep) Neural Network

- Neurons are stacked in parallel to form layers.
- By stacking multiple layers of neurons, a deep neural network (DNN) is created.
- By incorporating nonlinear activation functions such as ReLU and sigmoid, complex nonlinear transformations can be achieved.

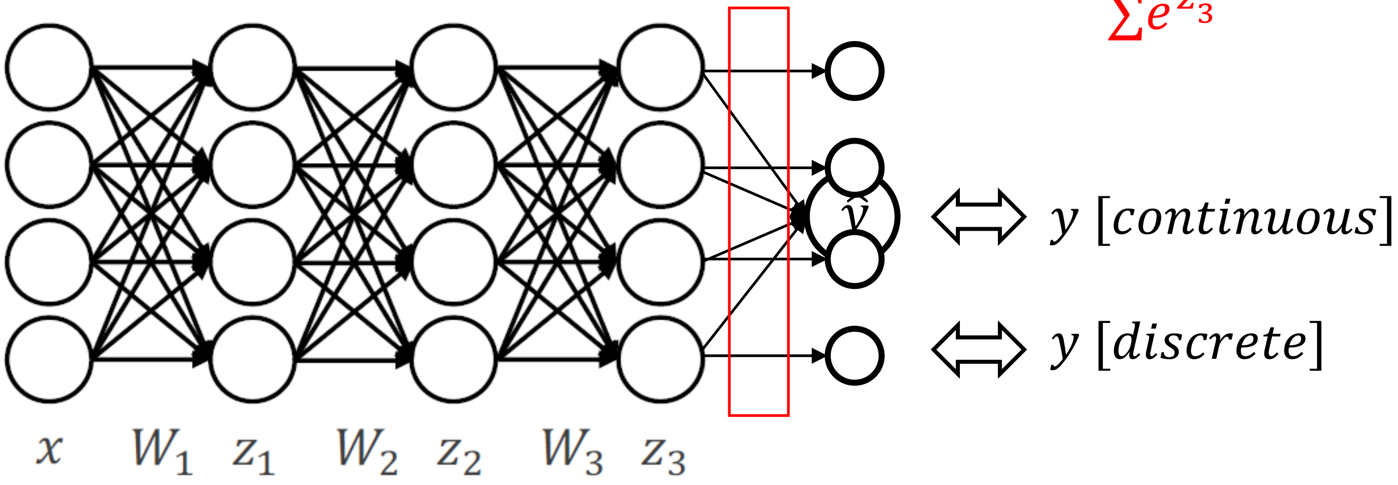
$$x \rightarrow W_1 x \rightarrow \sigma_1(W_1 x) \rightarrow W_2 \sigma_1(W_1 x) \rightarrow \sigma_2(W_2 \sigma_1(W_1 x))$$



Last layer

- The final layer of a neural network is configured according to the task.
- For regression tasks, it is sufficient to apply a transformation that consolidates the output into a single node.
- In classification tasks, the softmax function is commonly used to calculate the probability of belonging to each class.

$$softmax\ (z) = \frac{e^{z_3^j}}{\sum e^{z_3^j}}$$


$\hat{y}$
$\Leftrightarrow$ y [$continuous$]
$\Leftrightarrow$ y [$discrete$]
x
W_1
z_1
W_2
z_2
W_3
z_3

Why you can be DEEP?

→ Applying Backward error propagation method

b: bias (parameter)
y_i: Observed Label
$\hat{y}_i$: model output
$y_i, \hat{y}_i = (0, 1)$

Objective function

Loss function $F(\boldsymbol{w}, b) = \prod_i P(k_i | \boldsymbol{x}_i) = \prod_i \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i}$

Log (loss function) $E(\boldsymbol{w}, b) = -\log F(\boldsymbol{w}, b) = -\sum_n \{y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)\}$

To minimize the loss function, parameters are updated in the direction of the gradient

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \frac{\partial E(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} \qquad b^{t+1} = b^t - \eta \frac{\partial E(\boldsymbol{w}, b)}{\partial b}$$

η: learning rate (hyper parameter)

The derivative calculation is quite simple, allowing for quick parameter updates

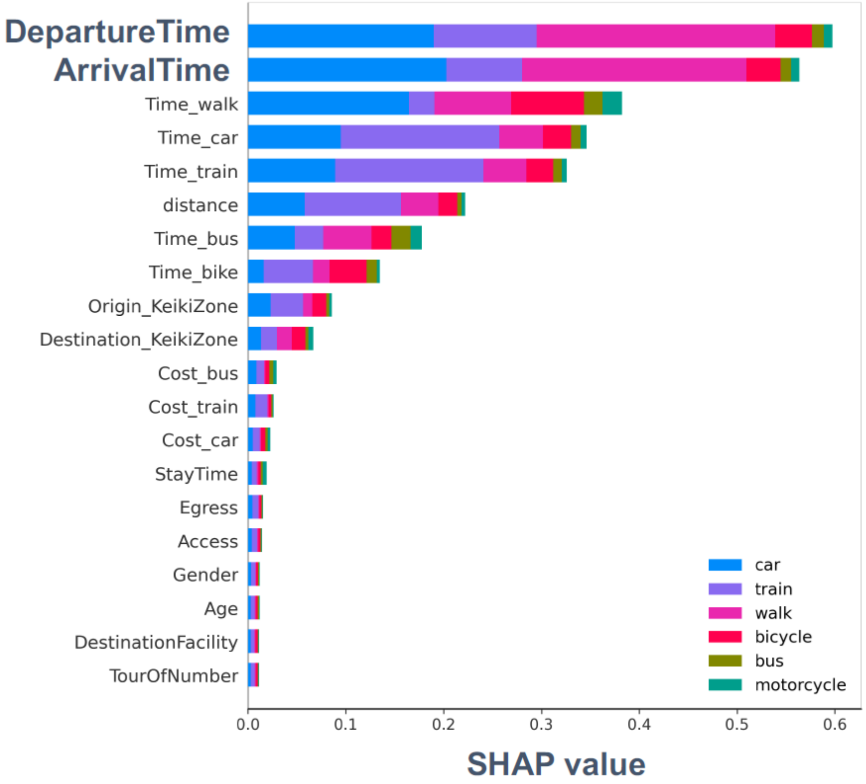
$$\frac{\partial E(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \cdots = -\sum_n (y_i - \hat{y}_i) \boldsymbol{x}_i$$

(In sigmoid function)

In addition, we can use GPUs for parallel computing

[Extra] SHAP (SHapley Additive exPlanations)

- SHAP is used to clarify how each feature contributes to the prediction results.
- It quantifies the impact of each feature on the prediction.
- SHAP makes it possible to interpret machine learning models, which often are regarded as black boxes.
- The values provided by SHAP are similar to elasticity in discrete choice models.



Mode choice model with NN (Yaginuma 2023)

Let’s

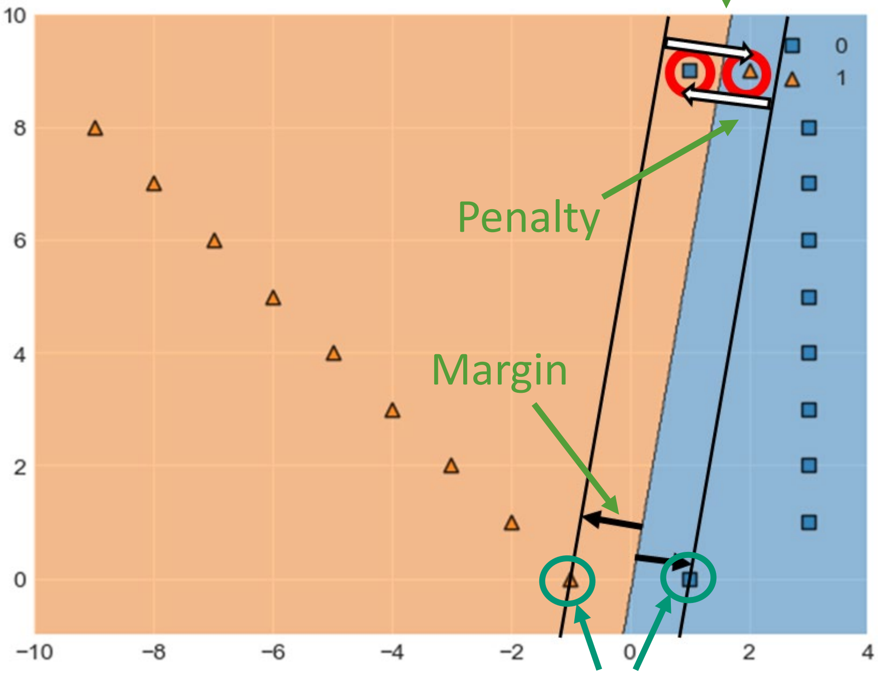
```
pip install shap
```

Support Vector Machine Classification (SVC)

- The decision boundary that classifies the data is learned under the rule of margin maximization.
- The goal is to find the decision boundary that maximizes the margin.
- However, in some cases, the data cannot be perfectly separated by the boundary.
- In such cases, an objective function is set to minimize a penalty term.

Soft-margin SVC

decision boundary



Penalty

Margin

Support Vector

Inverse of margin

<Objective function> $\min\limits_{\boldsymbol{w},\boldsymbol{b},\zeta} \frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{w}\|^{\mathbf{2}} + \boldsymbol{C}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}}\zeta_{\boldsymbol{i}}$

Penalty

<constraint> $\boldsymbol{y_i}\left(\boldsymbol{w^T} \cdot \boldsymbol{x_i} + \boldsymbol{b}\right) \geq \mathbf{1} - \zeta_{\boldsymbol{i}} \quad \forall \boldsymbol{i}$

$\boldsymbol{w}$: Normal vector to the decision boundary

$\boldsymbol{b}$: bias term of the decision boundary

$\boldsymbol{N}$: Number of data

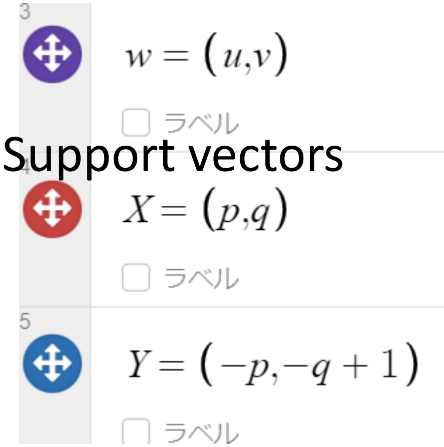
$\boldsymbol{C}$: Weight of parameter

$\zeta_{\boldsymbol{i}}$: Penalty term (usually, apply Hinge function)

$$\zeta_{\boldsymbol{i}} = \boldsymbol{f}(\mathbf{1} - \boldsymbol{y_i}(\boldsymbol{w^t}\boldsymbol{x_i} + \boldsymbol{b}))$$

Demo: Support Vector Machine

Normal vector



3

$w = (u,v)$

□ ラベル

Support vectors

$X = (p,q)$

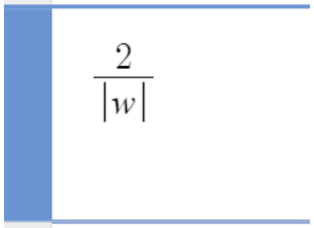
□ ラベル

5

$Y = (-p, -q+1)$

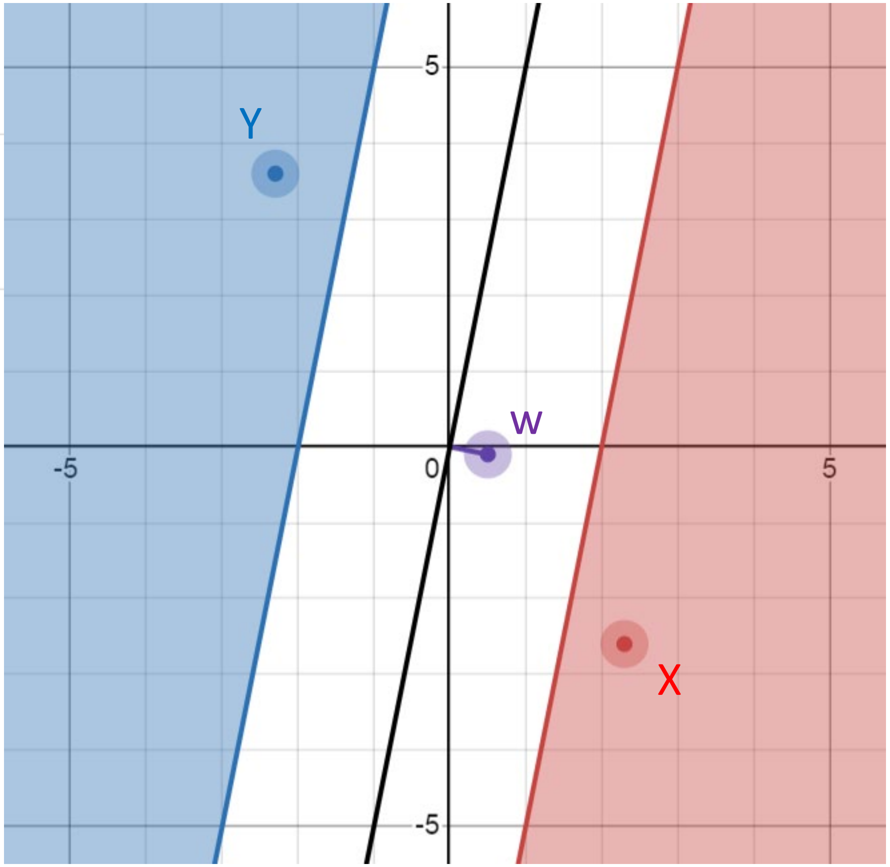
□ ラベル

Margin distance



$\frac{2}{|w|}$

Hard-margin SVC (without penalty function)



decision boundary

There are many advanced model in ML/DL field.

ex. CNN, RNN, Transformer, LLM , GAN

Let’s start to estimate the models, and predict!

Major References [JPN]:
東京大学数理・情報教育研究センター二反田篤史2021
3-3 機械学習の基礎と展望 [Link]
3-4 深層学習の基礎と展望 [Link]