



データマイニング

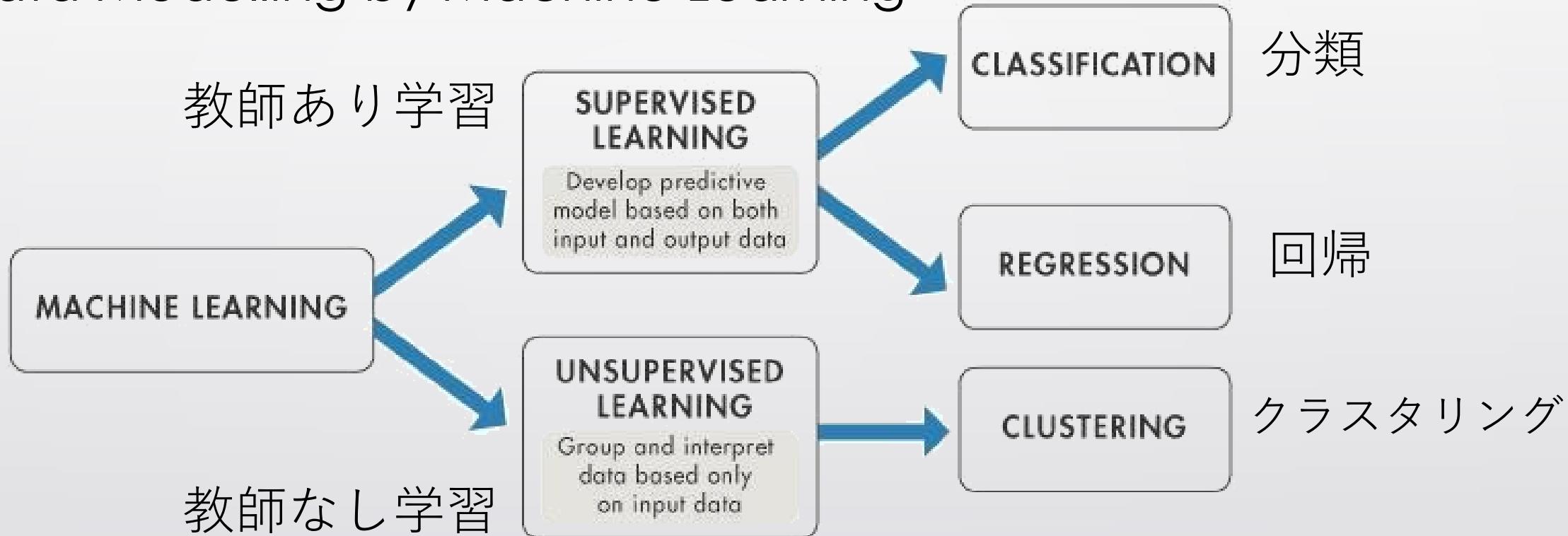
- 1: データマイニングの全体像
- 2: データセキュリティとデータサイエンスの倫理

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

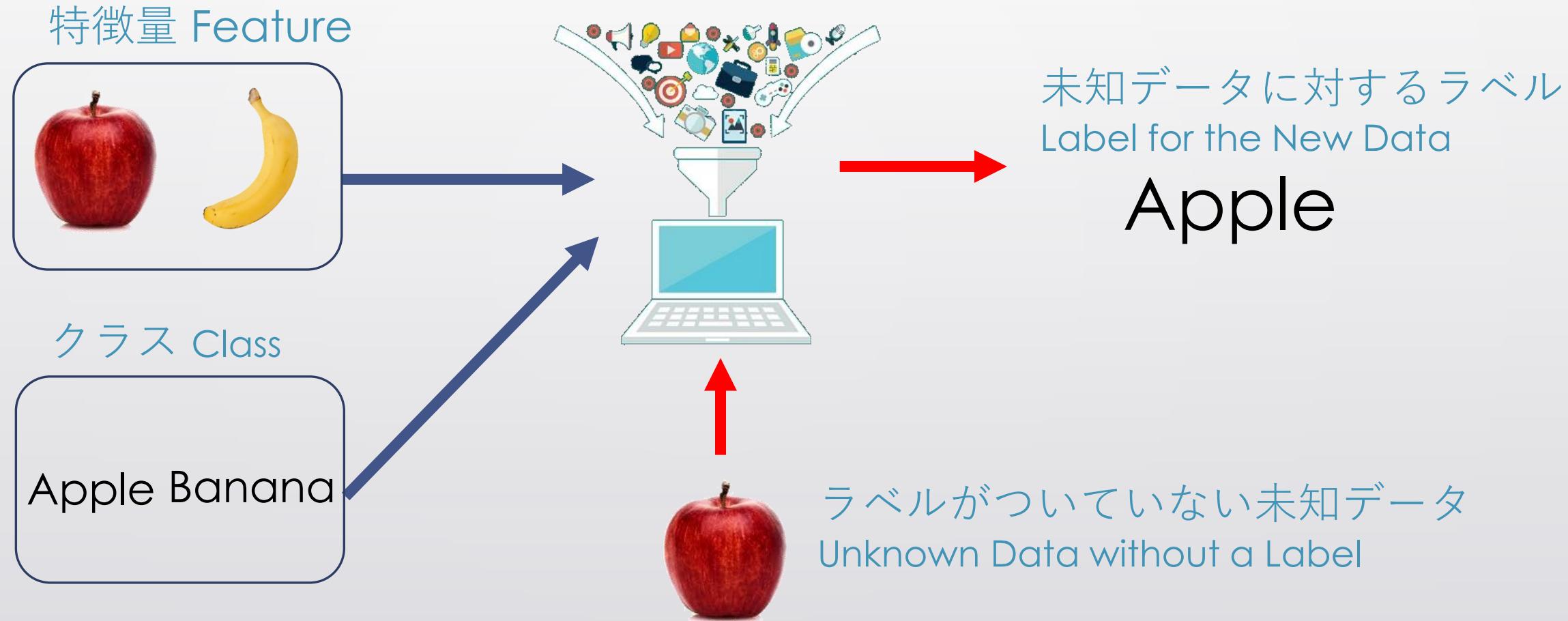
機械学習によるモデル化

Data Modelling by Machine Learning

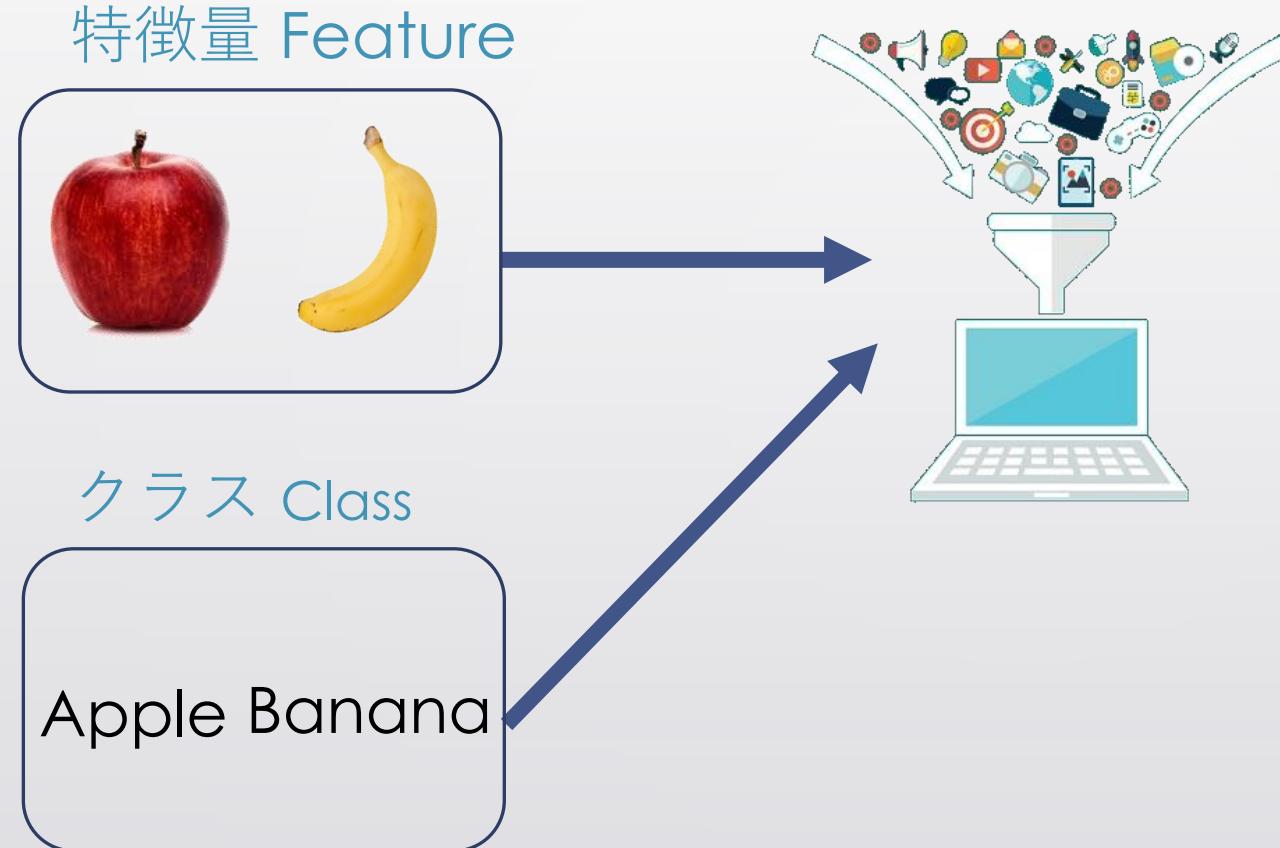


Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

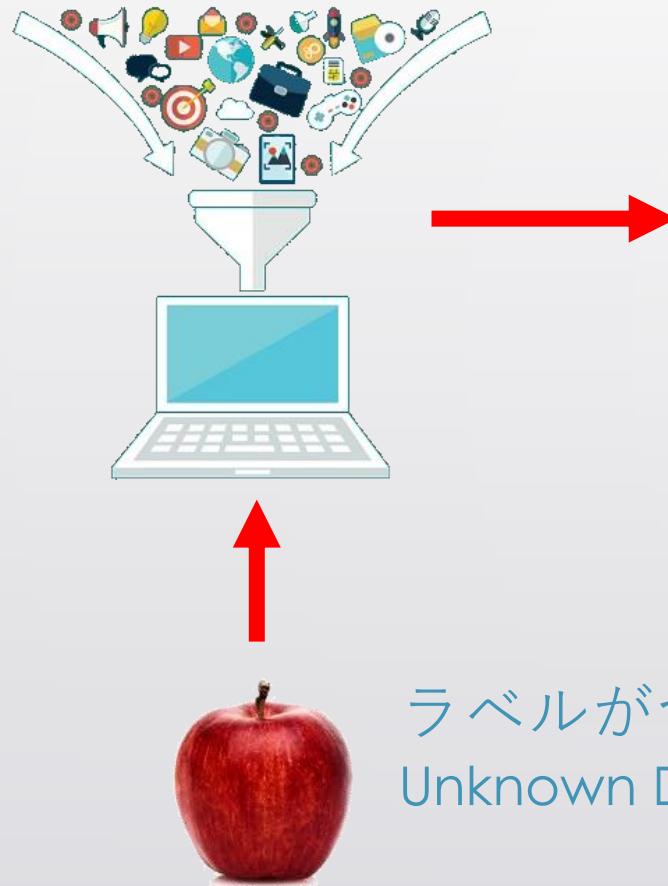
教師あり学習 Supervised Learning



モデルの訓練（トレーニング） Model Training



予測 Prediction



Apple

量を予測する回帰に対して、分類ではカテゴリー/クラスを予測する

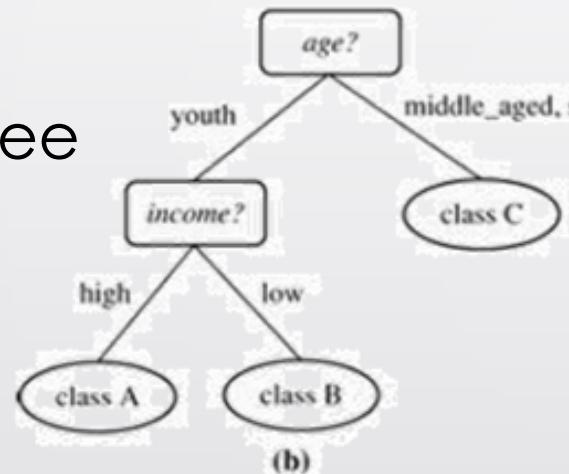
Classification algorithms predict category/class while regression predicts quantity.

※量とクラスの予測の両方に対応したアルゴリズムも多い

There are many algorithms that can make predictions about both quantity and category.

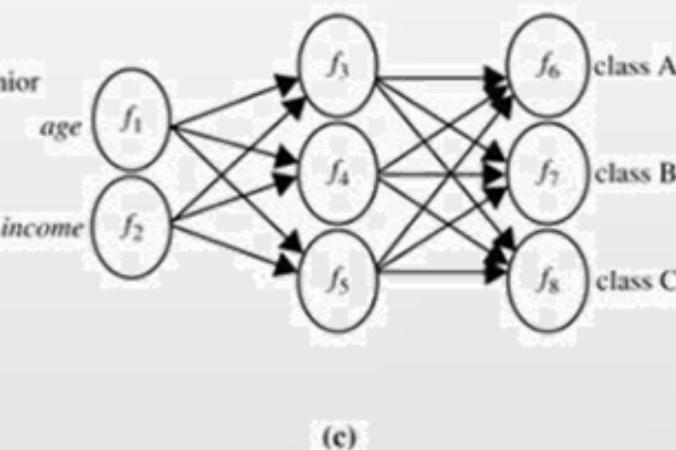


決定木 Decision Tree



$age(X, \text{"youth"}) \text{ AND } income(X, \text{"high"}) \longrightarrow class(X, \text{"A"})$
 $age(X, \text{"youth"}) \text{ AND } income(X, \text{"low"}) \longrightarrow class(X, \text{"B"})$
 $age(X, \text{"middle_aged"}) \longrightarrow class(X, \text{"C"})$
 $age(X, \text{"senior"}) \longrightarrow class(X, \text{"C"})$

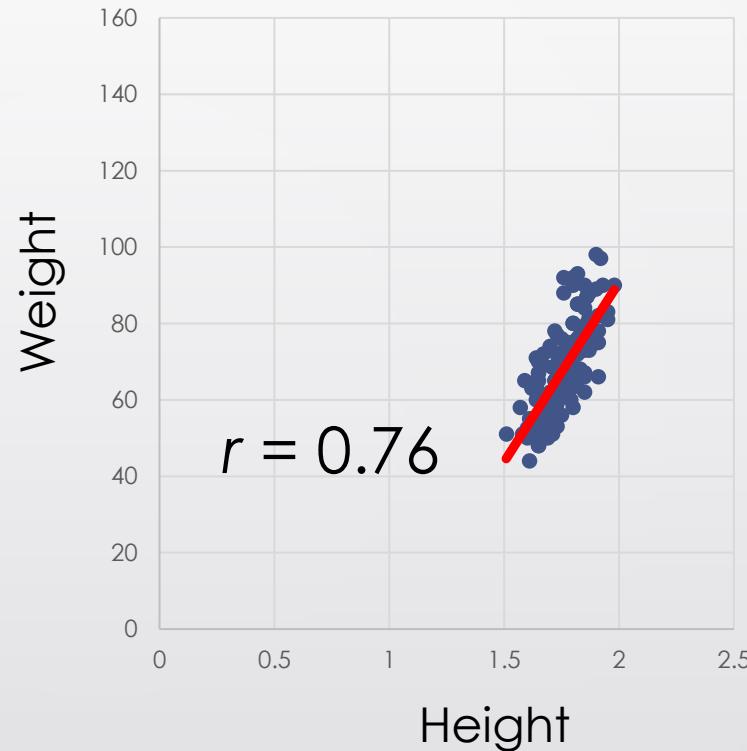
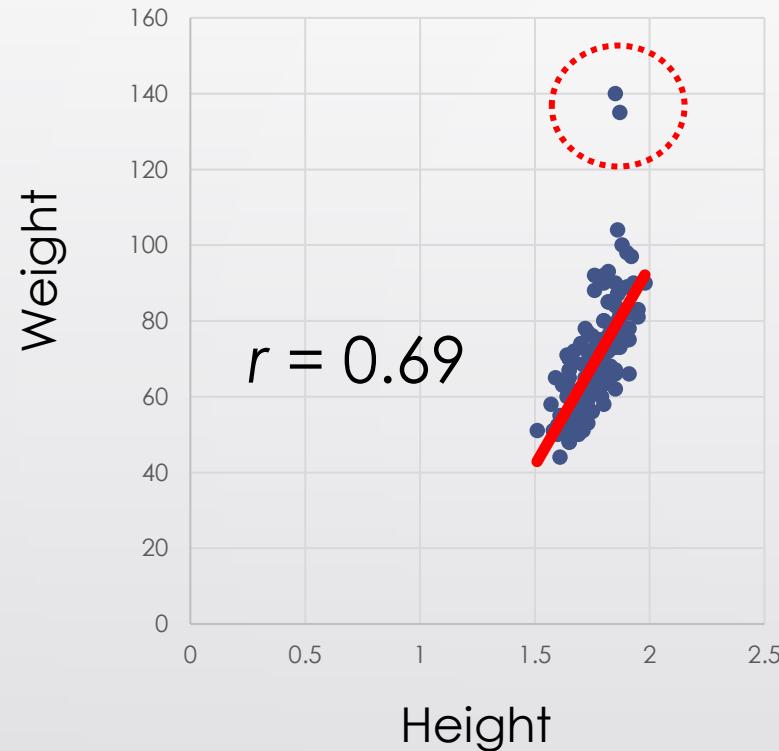
(a)



ニューラル
ネットワーク
Neural Network

FIGURE 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

外れ値 Outlier



結果に影響を与える可能性がある外れ値の存在をチェックする必要がある
Data set should be checked for the existence of outliers that can influence the results

外れ値分析による不正検知
Fraud Detection by Outlier Analysis

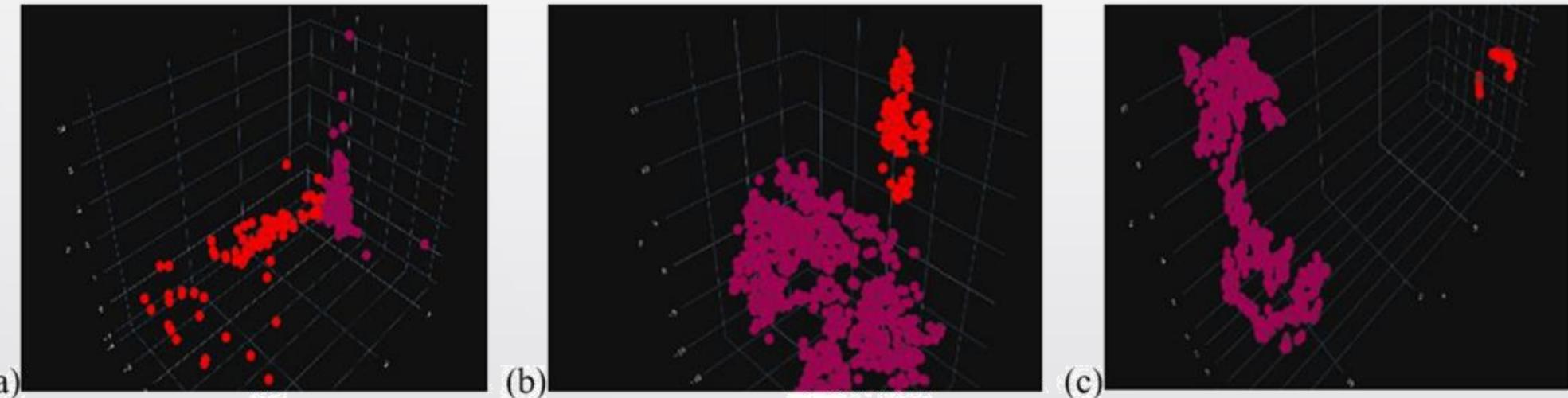


Fig. 8 The performance of different dimensional reduction algorithms on credit card fraud dataset. Feature size is reduced to dimension 3 by **a** PCA, **b** t-SNE and **c** UMAP



データマイニングの流れ Steps in Data Mining

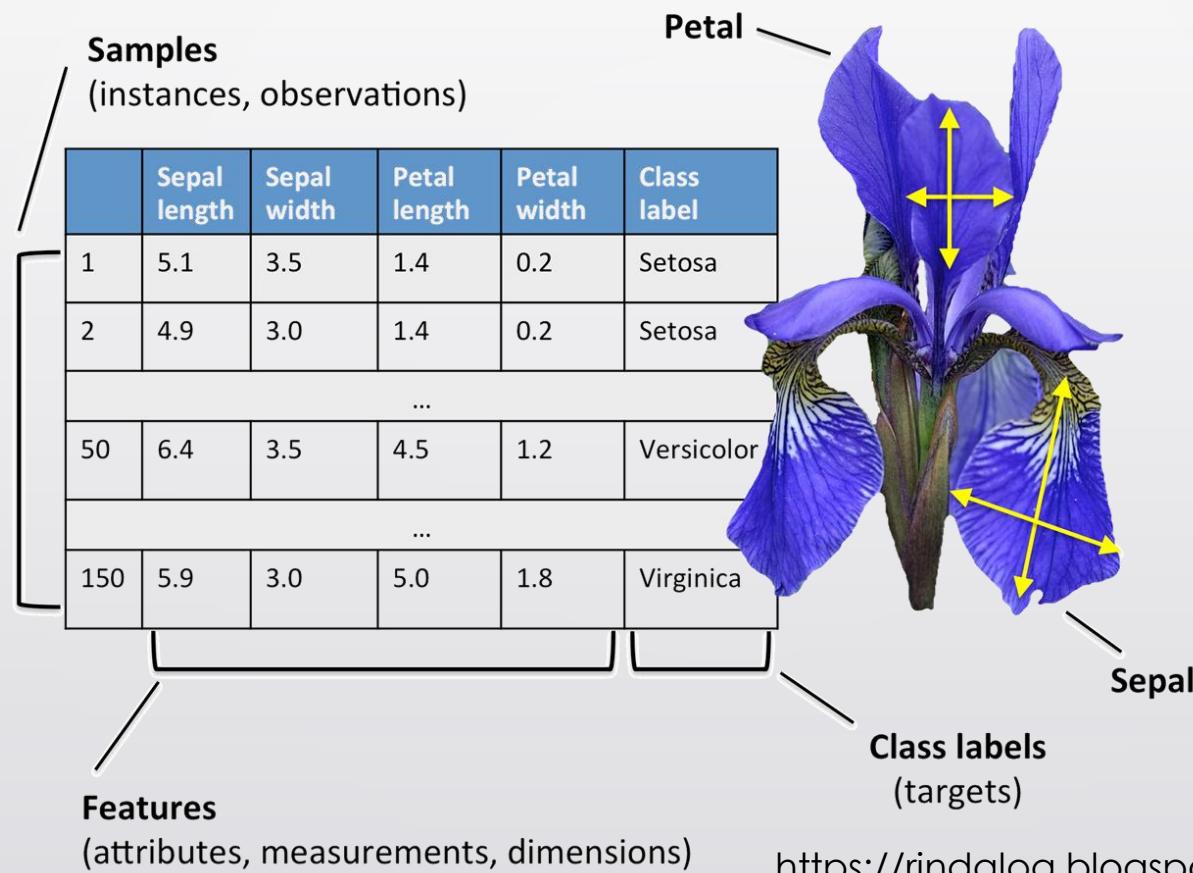
1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis・モデリング Modeling
6. 性能評価 Performance Evaluation
7. (ディプロイメント Deployment)



目標設定 Goal Setting

1. 調査の目的を設定する Verify the purpose of the survey
2. どのようなデータを集めるかを決める Decide what data to collect for our purposes
3. 目的を達成するには、どのようなデータ分析を行うのが適切かを、前もって考えておく Consider what type of data analysis is appropriate to achieve the goal

アヤメからの特徴量抽出 Feature Extraction from Iris



アヤメデータセット Iris Dataset

ヒオウギアヤメ



setosa

ブルーフラッグ



versicolor

アイリス・バージニカ



virginica

Samples
(instances, observations)

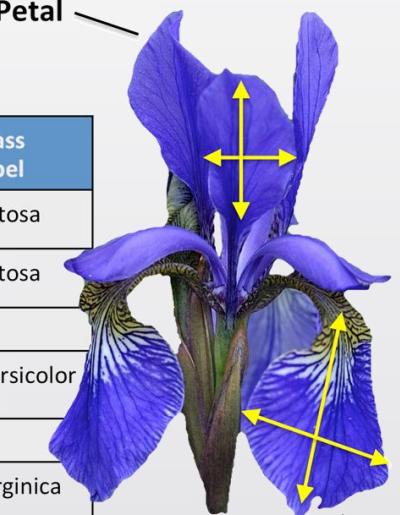
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features

(attributes, measurements, dimensions)

花弁

Petal



がく

Class labels
(targets)

データの読み込み Loading Data

Iris plants dataset

Data Set Characteristics:

:Number of Instances: 150 (50 in each of three classes)

:Number of Attributes: 4 numeric, predictive attributes and the class

:Attribute Information:

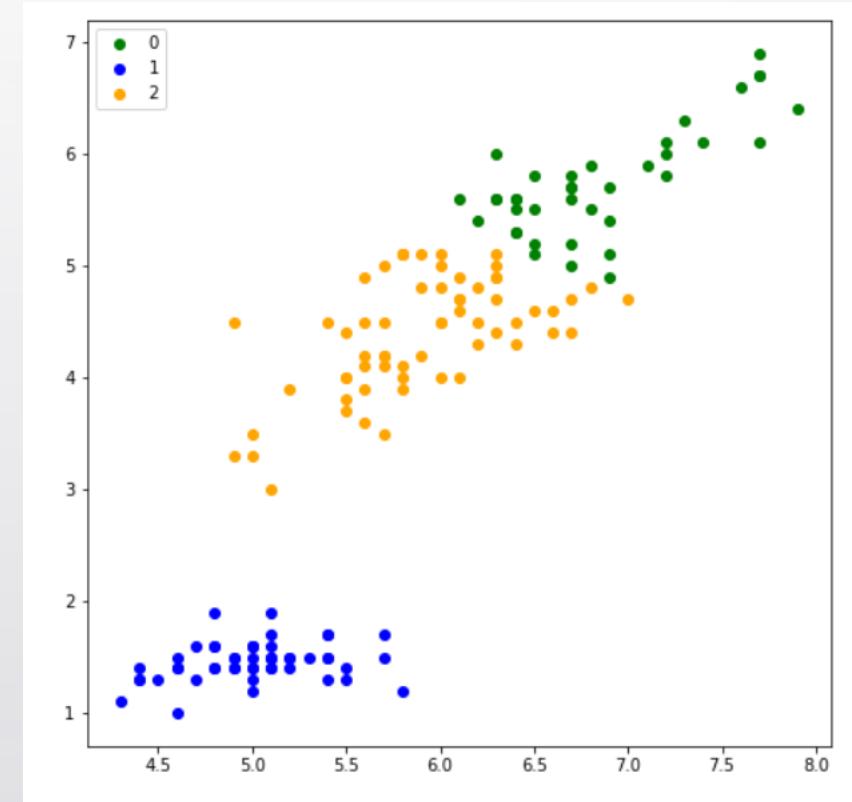
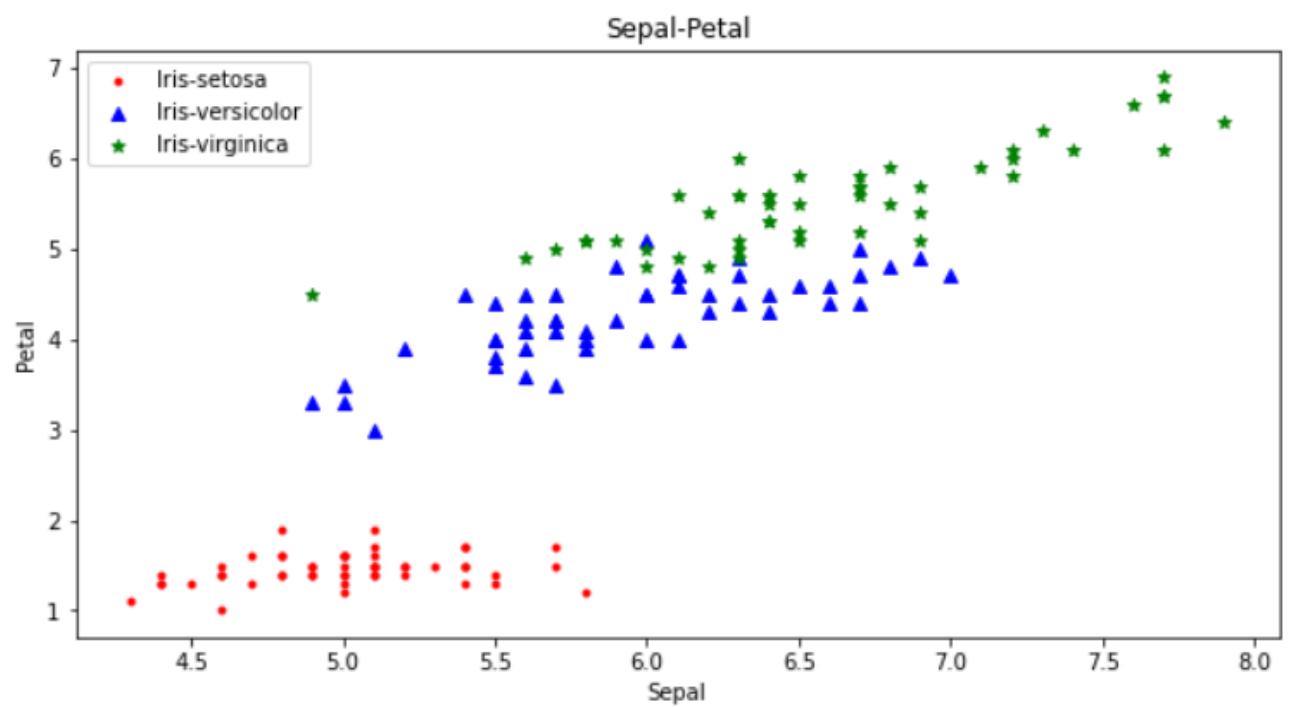
- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class:
 - Iris-Setosa
 - Iris-Versicolour
 - Iris-Virginica

記述統計 Descriptive Statistics

In [15]: df.describe()

Out[15]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000





	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

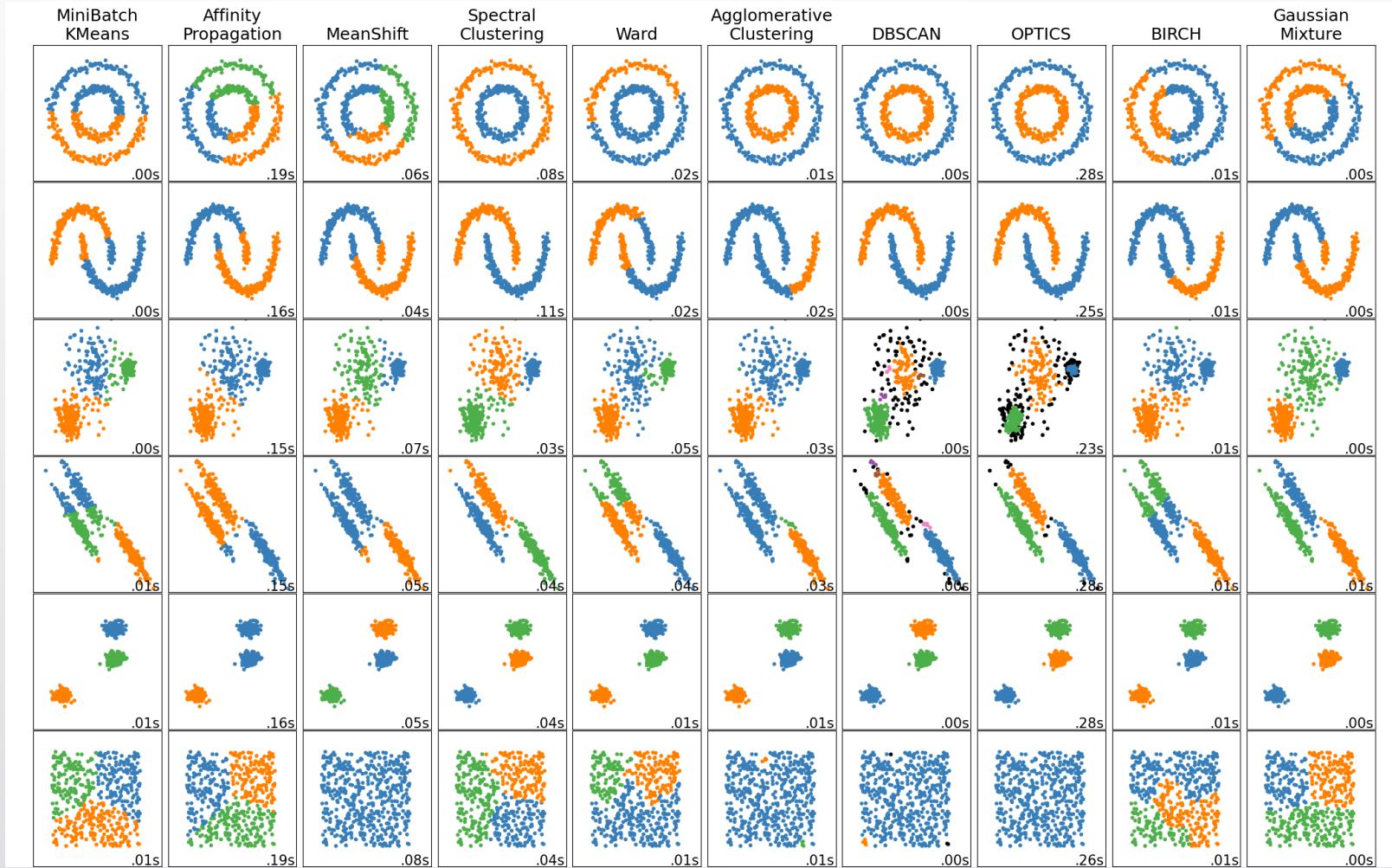
クラスタリング

Final result of clustering
depends on

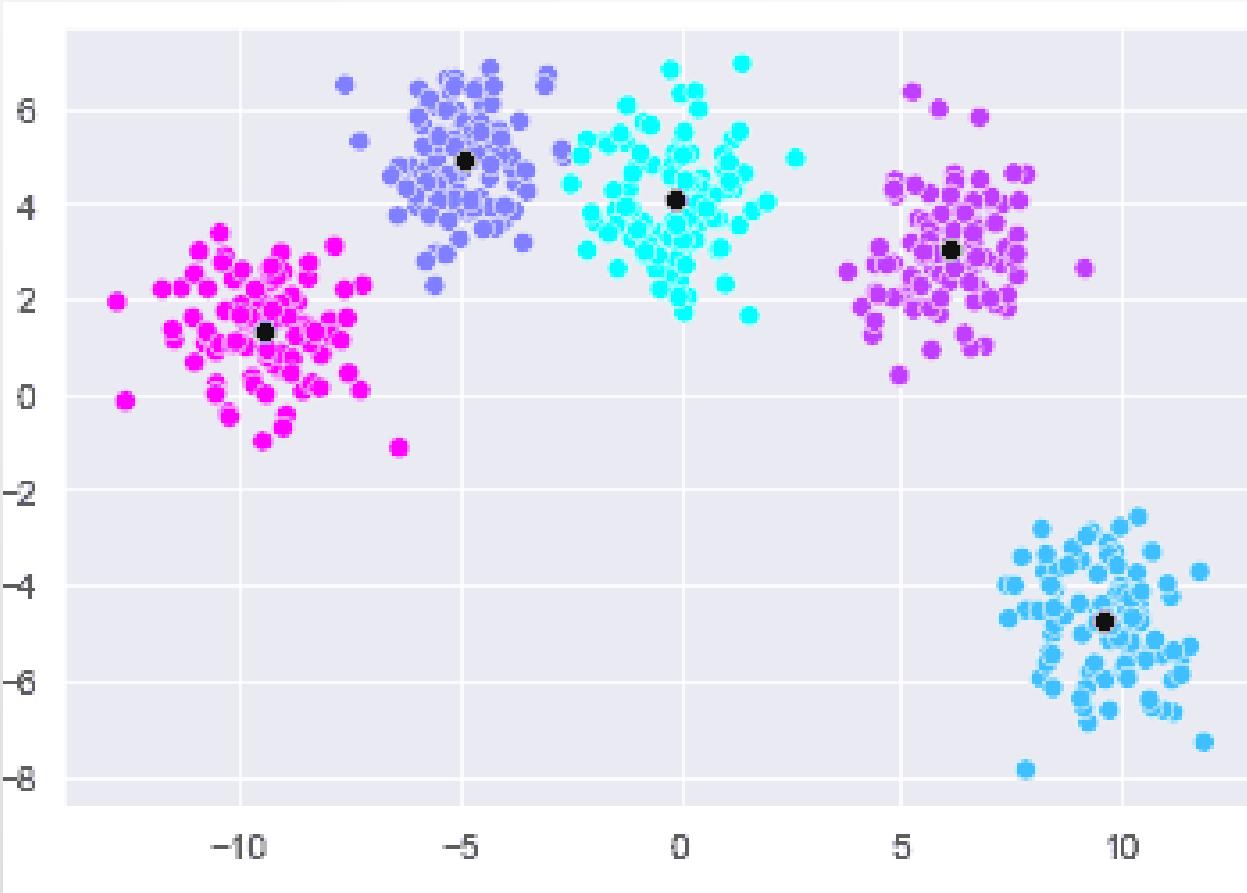
Type of algorithm
アルゴリズムの種類

Parameter Setting
パラメータ設定

<https://scikit-learn.org/stable/modules/clustering.html>



K-Means Clustering



クラスターの数を指定しなくては
いけない

You have to specify the number of
clusters, k .



データセキュリティ

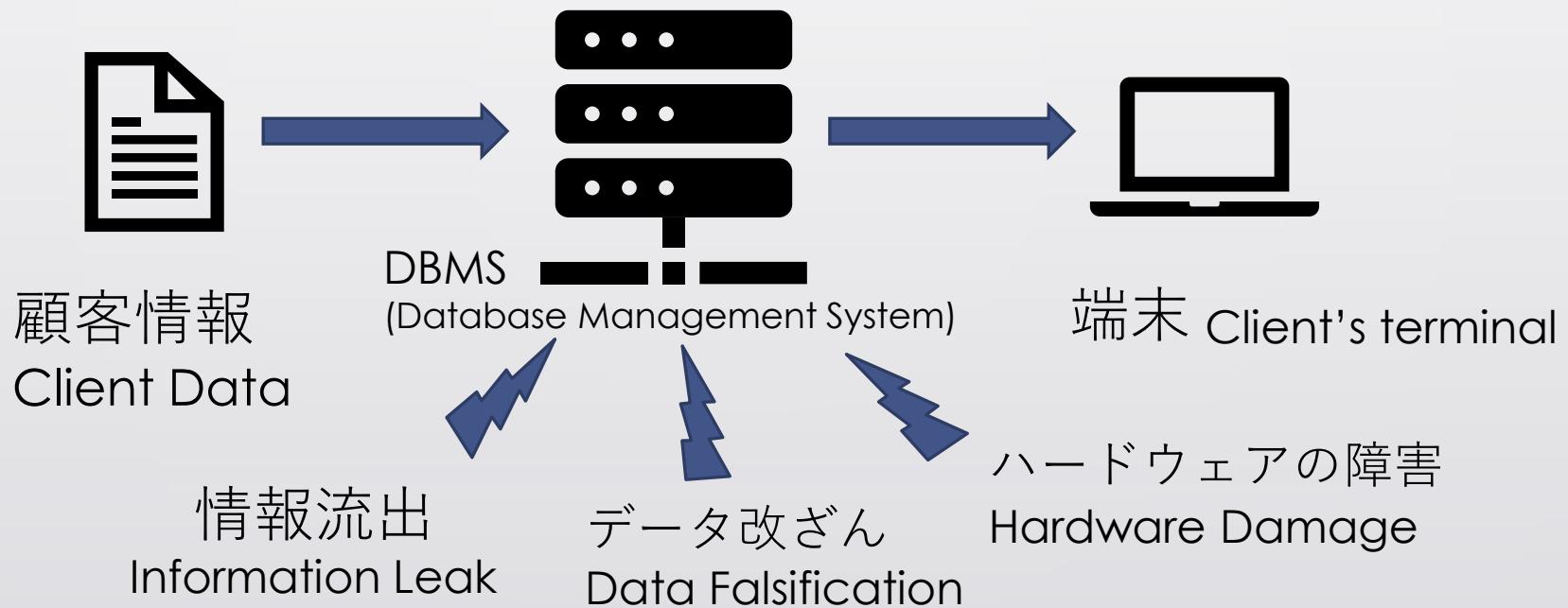
Data Security



情報資産 Information Asset

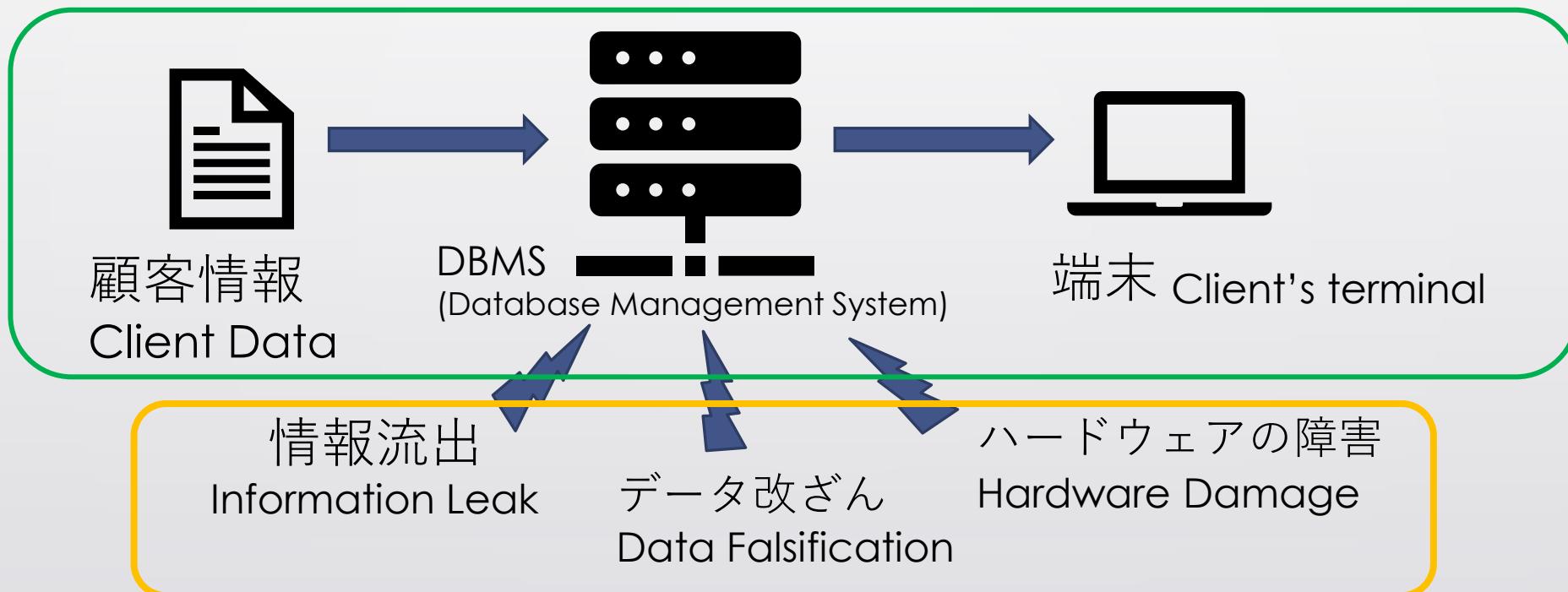
個人や企業にとって、価値のある情報や情報システム

Information or the integrity of information system valuable for an individual or corporation.



リスクと脆弱性 Risk and Vulnerability

脆弱性：情報システムの弱点 Weak point of an information system



リスク：情報システムが損なわれる可能性 Possibility that an information system is damaged

機密性 Confidentiality

許可された人のみが情報にアクセスできる

Only persons granted permission can access information



Confidentiality

アクセス制限 Access Control

サーバールームの施錠 Locked Server Room

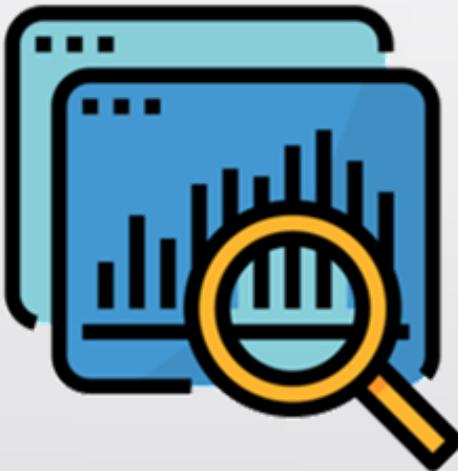
暗号化 Encryption

<https://www.simplilearn.com/what-is-information-security-article>

完全性 Integrity

改ざんなどされることなく、情報が完全に保たれている

Information is being kept intact without falsification etc



Integrity

アクセスログ・操作 Logging Access and Operation

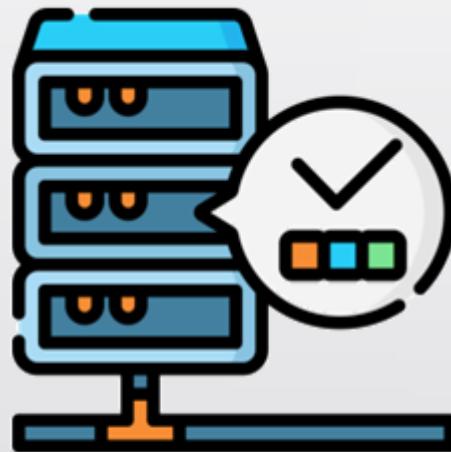
直感的なユーザー・インターフェース

Intuitive User Interface

<https://www.simplilearn.com/what-is-information-security-article>

可用性 Availability

許可された人が、必要な時に、いつでも情報にアクセスできる
Any persons granted permission can access information whenever necessary



Availability

複数箇所におけるデータ保存
Data storage in multiple sites

定期的なバックアップ
Periodical BackUp

<https://www.simplilearn.com/what-is-information-security-article>

UPDATE 1-Sony may face global legal scrutiny over breach

Tom Hals, Leigh Jones

5 分で読む



* U.S. lawyers focused on data breaches eye legal action

* British government watchdog launches investigation

* U.S. state attorneys general also discussing incident (Recasts with comments from U.S. legislators)

WILMINGTON, Delaware/NEW YORK, April 27 (Reuters) - Sony Corp 6758.T could face legal action across the globe after it belatedly revealed one of the biggest online data breaches ever.

In the United States, several members of Congress seized on the breach, in which hackers stole names, addresses and possibly credit card details from users of Sony's PlayStation Network, to push for tougher laws protecting personal information.



AP Photo/Shizuo Kambayashi

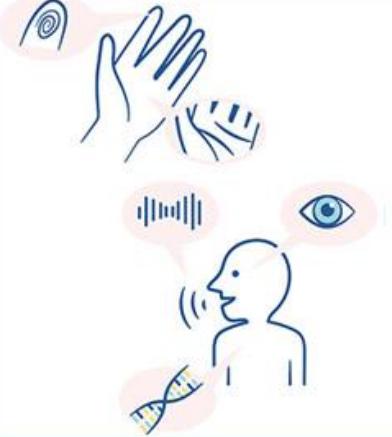
個人情報 Personal Information

個人情報とは



特定の個人を識別できるもの

Information that identifies particular individual



個人の身体のデータ

Data extracted from individual's body



個人に割り振られる公的な番号

Official number assigned to each individual
政府広報オンライン

「個人情報保護法」

Personal Information Protection Law

生存する個人に関する情報で、氏名、生年月日、住所、顔写真などにより**特定の個人を識別できる情報**

Information concerning living individual...
Information by which a particular individual is identifiable

データはあらかじめ決まった目的でしか利用しない
Use information only for pre-defined purposes

流出を防止するため、データを安全に管理する
Should be managed securely to avoid data leakage



第三者とデータを共有する場合は、関係者の同意を得る
Consent must be obtained from relevant individuals
when their data is shared with the third party.

請求があった場合には、関連情報を開示する
Should disclose any relevant information
whenever requested

Samarati & Sweeney, 1998

マサチューセッツ州のGIC (Group Insurance Commission)が、氏名を削除

したうえで、医療保険に関する情報を民間企業に販売

GIC in Massachusetts sold to private companies the data relevant to medical insurance after “anonymization”, i.e. deleting individuals’ names

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath



Samarati & Sweeney, 1998

マサチューセッツ州ケンブリッジの選挙人名簿は20ドルで購入可能だった

Voters' list in Cambridge, Massachusetts could be purchased for 20 dollars

Voter List

Samarati & Sweeney, 1998

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

個人情報の例: 表1

Example of Personal Information: Table 1

Identification Number	User ID	Name	Gender	Age	Income
識別ナンバー	ユーザーID	氏名	性別	年齢	年収
339829Q	sanapon	真田昌幸	男	26	411万円
905473R	oggi1985	荻野吟子	女	33	536万円
099878L	murachan	紫式部	女	39	681万円
013214H	shozan.s	佐久間象山	男	23	309万円

個人情報の例: 表2

Example of Personal Information: Table 2

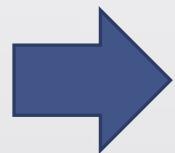
User ID	Product	Price	Date	Store
ユーザーID	購買物品	購買価格	購買日時	購買店舗
murachan	人参	100円	2021/2/3 18:09	Cマート 代田2丁目店
oggi1985	バナナ	150円	2021/2/3 21:13	Cマート 小石川店
oggi1985	ダイエットコーク	160円	2021/2/4 21:15	Cマート 小石川店
murachan	粉ミルク	980円	2021/2/4 21:16	Cマート 代田2丁目店
murachan	紙おむつ	1700円	2021/2/4 21:16	Cマート 代田2丁目店

名前が削除されても、表に含まれた情報がら、個人を特定できる場合がある

Even if name is deleted, sometimes, an individual can be identified from information contained in a table.

仮名化 Pseudonymization 匿名化 Anonymization

識別ナンバー	ユーザーID	氏名	性別	年齢	年収	住所
339829Q	sanapon	真田昌幸	男	26	411万円	・・・
905473R	oggi1985	荻野咲子	女	33	536万円	・・・
099878L	murachan	紫式部	女	39	681万円	・・・
013214H	shozan.s	佐久間象山	男	23	309万円	・・・



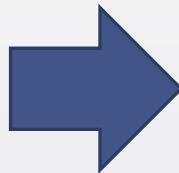
ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

個人識別情報を削除する Deleting person identifiable information



k-匿名化 k-Anonymization

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円



性別	年齢	年収	
男	[20-29]	[300-499]	万円
女	[30-39]	[500-699]	万円
女	[30-39]	[500-699]	万円
男	[20-29]	[300-499]	万円

データの中に、同じ属性を持つ人が、少なくとも k (≥ 2) 人含まれる

At least k (≥ 2) persons with the same attributes are included in a table.



データサイエンスの倫理

Ethics in Data Science

一般データ保護規則 General Data Protection Rule (GDPR)

EU/UKに住む人々の個人情報を対象

個人情報を保護し、プライバシーを守る

What ?

Data Protection regulation that applies to processing personal data of **EU/UK residents**

Which ?

Any information relating to EU/UK citizens whether they can be identified directly or indirectly

Why?

To protect personal data from **misuse** and to ensure **data privacy**



How?

More **obligations** on Data Controller & provide **rights** to data owners to control their data

Global

Applies **globally** to any organization processing information on EU/UK residents



Penalty

Penalties up to **4%** (or **€20m** whichever is higher) for major breaches

個人が、自分の個人情報をコントロールできるようにする

EU/UKに住む人々のデータを扱う、**世界中の機関**に適用される

忘れられる権利 The Right to be Forgotten

13 May 2014 - The court's decision comes by appeal of Mario Costeja Gonzalez, a Spanish man who sought to remove evidence of his home's repossession ...

Some results may have been removed under data protection law in Europe. [Learn more](#)

Goooooooooooooogle >

1 2 3 4 5 6 7 8 9 10

Next

● Unknown - Use precise location - [Learn more](#)

[Help](#)

[Send feedback](#)

[Privacy & Terms](#)

Retrieved on 2023/01/28 from
<https://www.cima.ned.org/publication/right-to-be-forgotten-threat-press-freedom-digital-age/>

透明性とトラスト Transparency and Trust

透明性 Transparency :

データマイニングシステムやAIが、結論を導くプロセスが、人々に説明可能になっていること

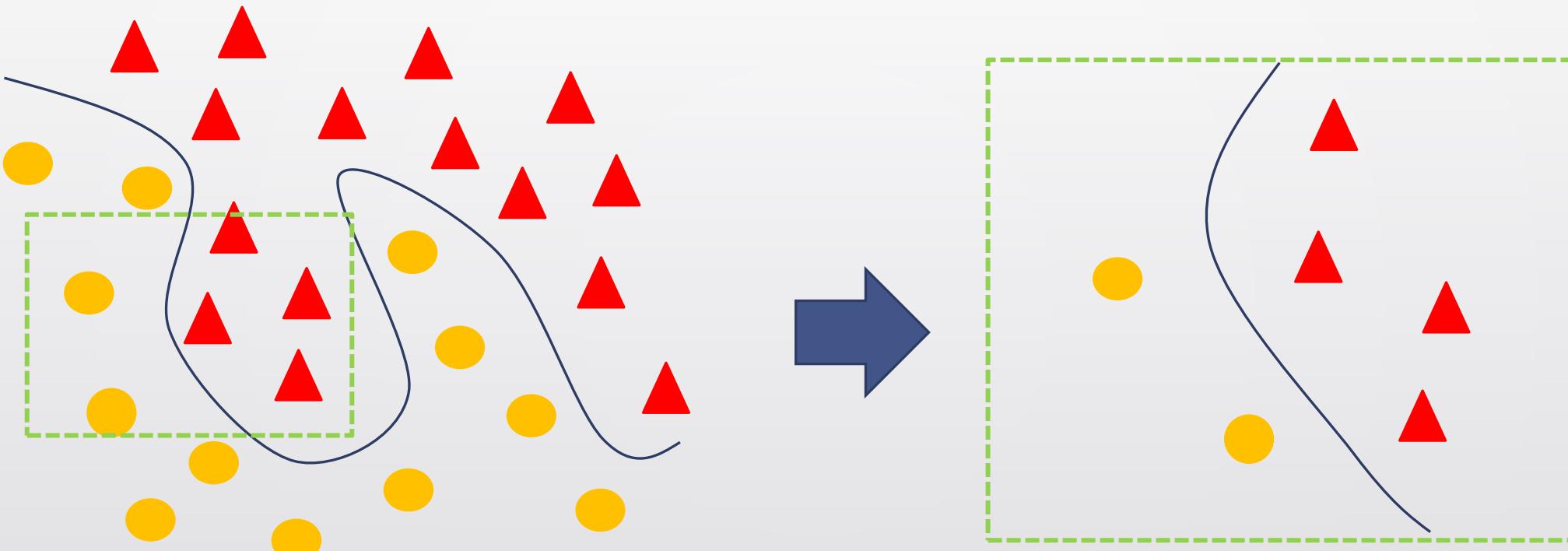
To make the process of reaching conclusions by data-mining system and AI understandable and accessible to ordinary people

トラスト Trust :

人々が、そのデータ処理プロセスを理解したうえで、データマイニングシステムやAIを信頼すること

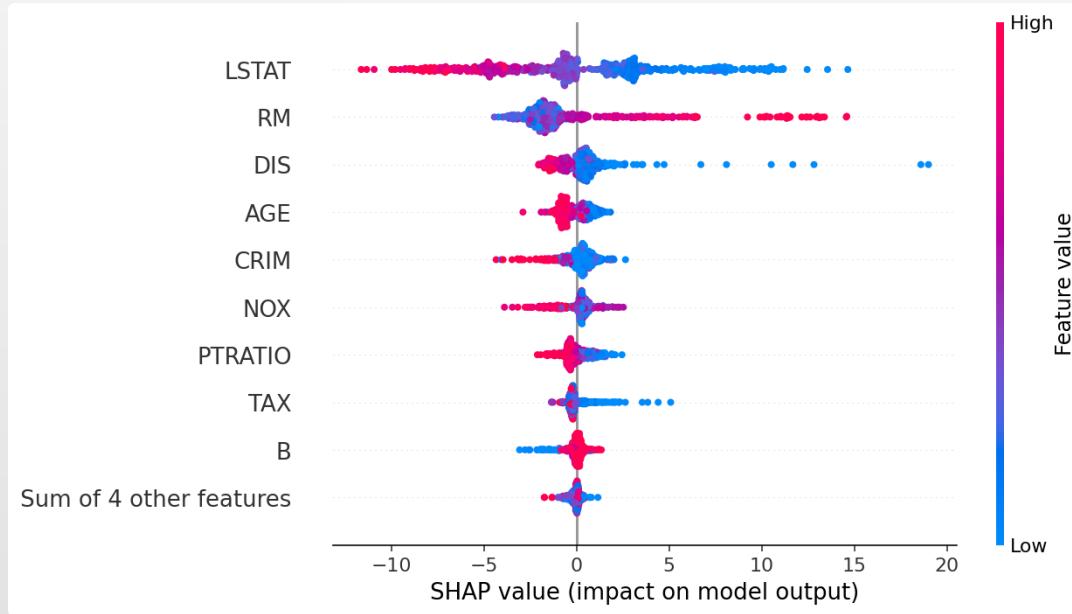
People trust data-mining system and AI after understanding their process of data processing

説明可能なAI Explainable AI



説明可能なAI Explainable AI

SHAP value



<https://github.com/slundberg/shap>

Grad CAM (Gradient-weighted Class Activation Mapping)

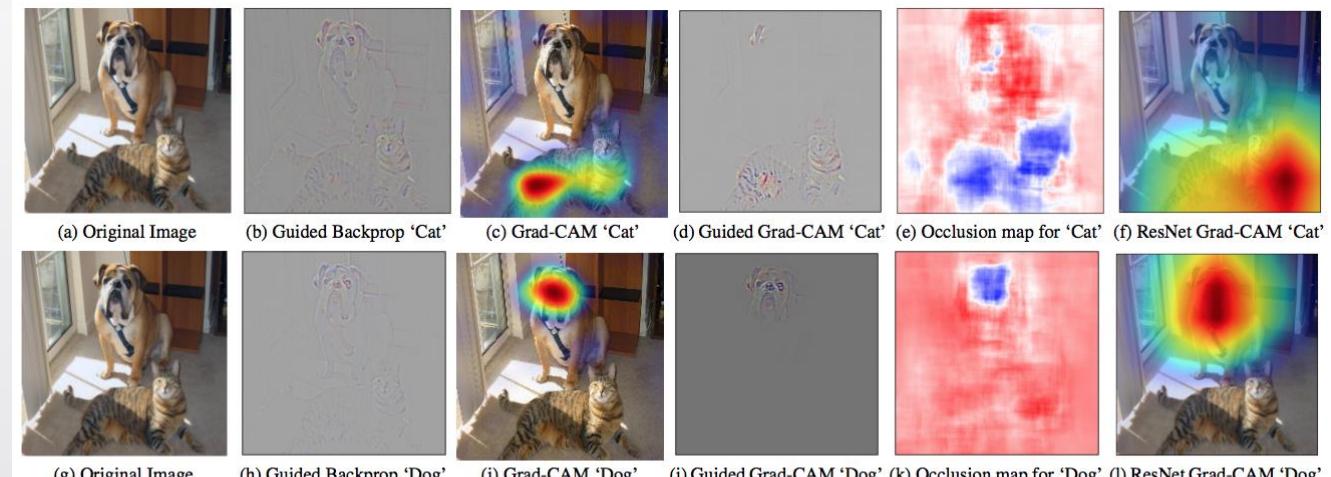


Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG and ResNet. (b) Guided Backpropagation [46]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (d, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.



データマイニング

Data Mining

3: データの要約・前処理 4: 次元削減

3: Data Summarization, Preprocessing 4: Dimension Reduction

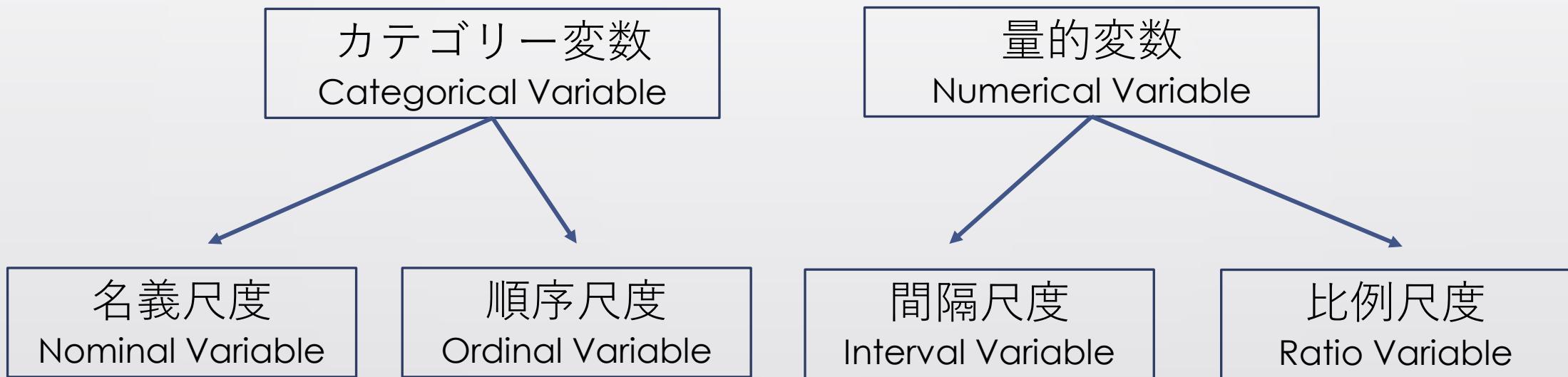
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology



データの要約
Data Summarization
&
Getting to know your data

変数の種類 Types of Variables



変数の種類 Types of Variables

名義尺度

Nominal Variable

あるカテゴリーを、別のカテゴリーと区別するために用いられる、数値自体には意味がない変数

Variables, whose number has no numerical value, often used to discriminate multiple categories

順序尺度

Ordinal Variable

順序を表す変数。変数間の間隔には意味がない。

Variables representing ordering of categories. Intervals between variables do not have any meanings.

変数の種類 Types of Variables

間隔尺度

Interval Variable

変数間の差の値に意味がある変数

Variables whose difference have meanings

比例尺度

Ratio Variable

間隔尺度と似ているが、原点がある点が違う。

Similar to interval variables, but ratio variables have clearly defined point of zero.

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

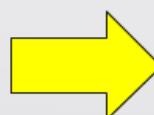


名義尺度の数値的表現

Numerical representation of nominal variable

ダミー変数 Dummy Variable 男 ⇒ 1, 女 ⇒ 2

One-hot Encoding



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow			

<https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook>



記述統計

Descriptive Statistics

代表值 Representative value

平均 Mean, 中央値 Median, 最頻値 Mode

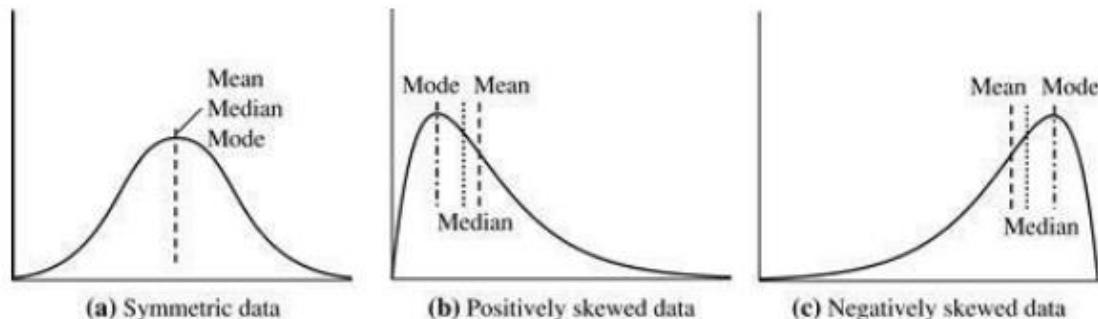


FIGURE 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

算術平均

Arithmetic Mean

$$M = \frac{\sum_{i=1}^n x_i}{n}$$

中央値 順番に並べた時, 中央に位置する数値

Value lying at the midpoint when a number sequence is ordered in ascending/descending order

最頻値

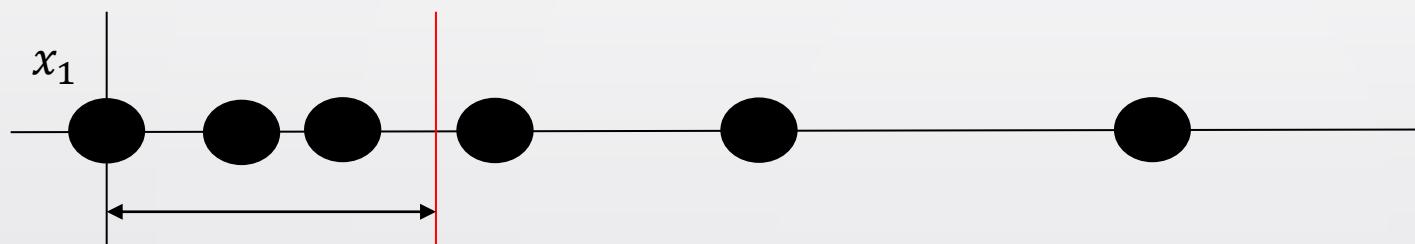
最も頻繁に現れる数値

The value that appears most frequently in the sequence



分散と標準偏差 Variance and Standard Deviation

平均 Mean



$$V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

データ x_1 と平均のズレ

Deviation of x_1 from the arithmetic mean

$$\text{標準偏差} = \sqrt{\text{分散}}$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$



分散と行列 Variance and Matrix

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad \bar{X} = \begin{bmatrix} \bar{x} \\ \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \\ \bar{x} \end{bmatrix} \quad X - \bar{X} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{bmatrix} \quad (X - \bar{X})^T = [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}]$$

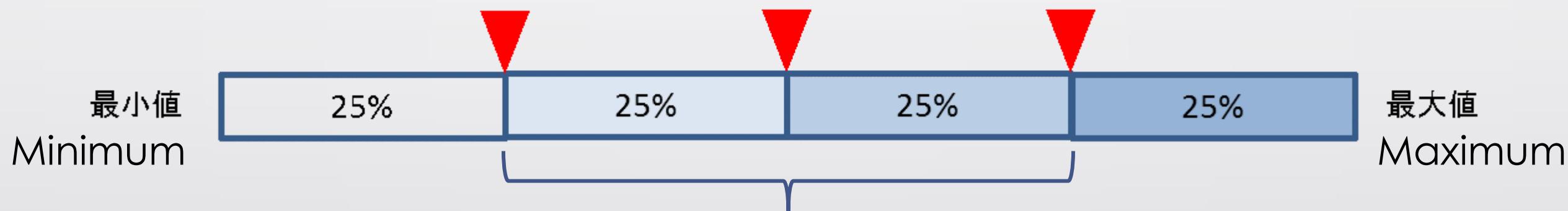
$$V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{1}{n} [x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}] \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_{n-1} - \bar{x} \\ x_n - \bar{x} \end{bmatrix} = \frac{1}{n} (X - \bar{X})^T (X - \bar{X})$$



四分位数 Quartile

Second Quartile (Q2)
First Quartile (Q1) = Median Third Quartile (Q3)

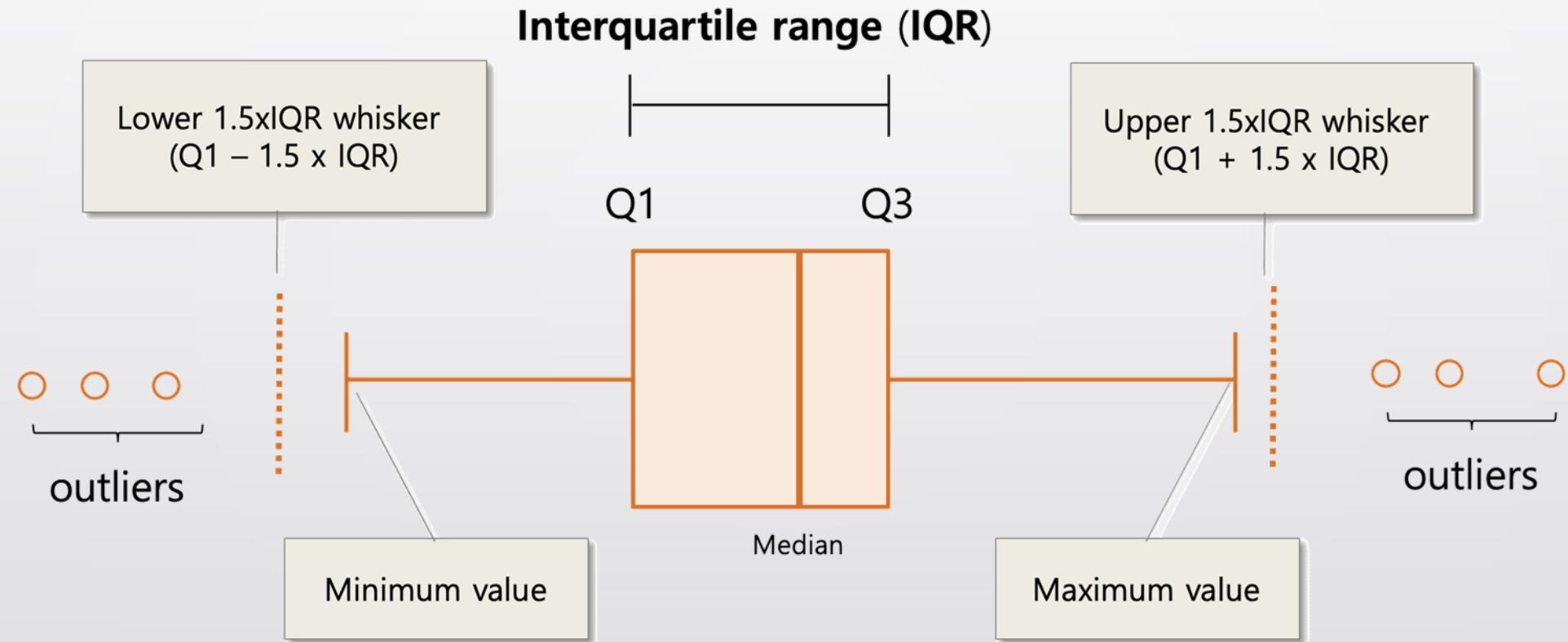
第一四分位数 第二四分位数 第三四分位数



四分位範囲 Interquartile Range (IQR)

$$= Q3 - Q1$$

箱ひげ図 Boxplot



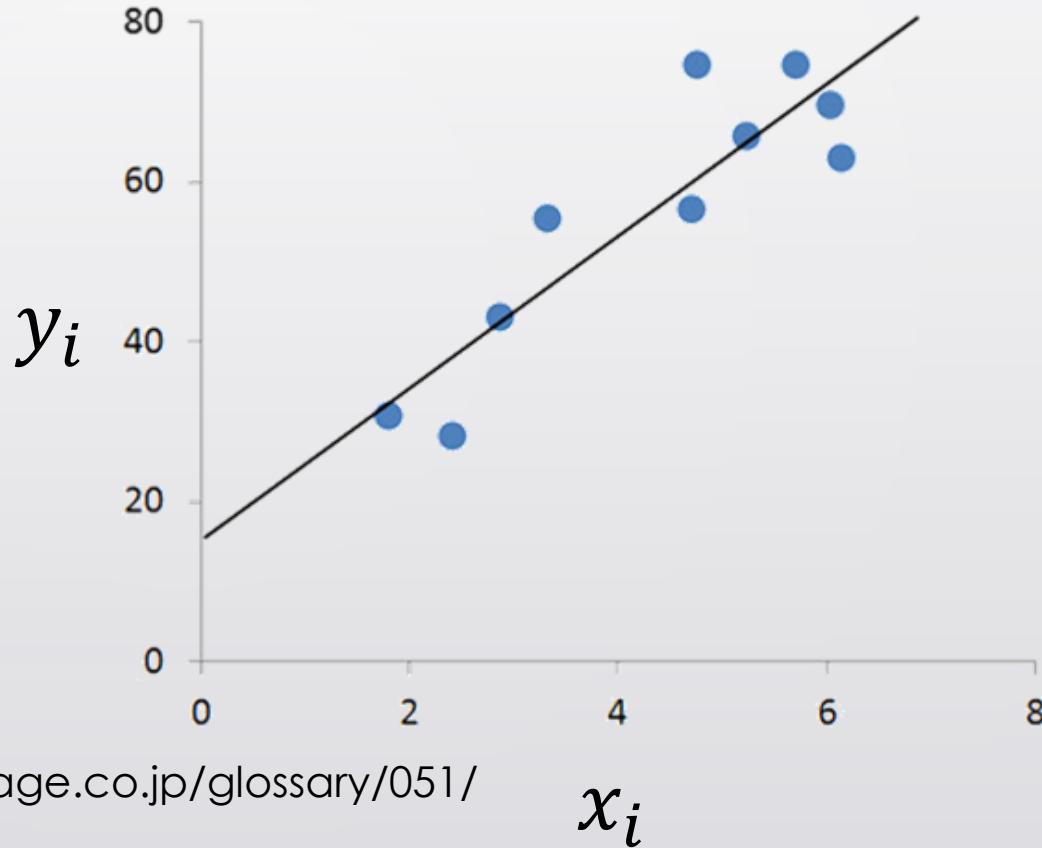
データの可視化: 散布図

Data Visualization: Scatter Plot

(x_i, y_i)

x_i : 家族の人数
Number of Family Member

y_i : 購入数
Number of purchased Items





データの可視化:ヒストグラム

Data Visualization: Histogram

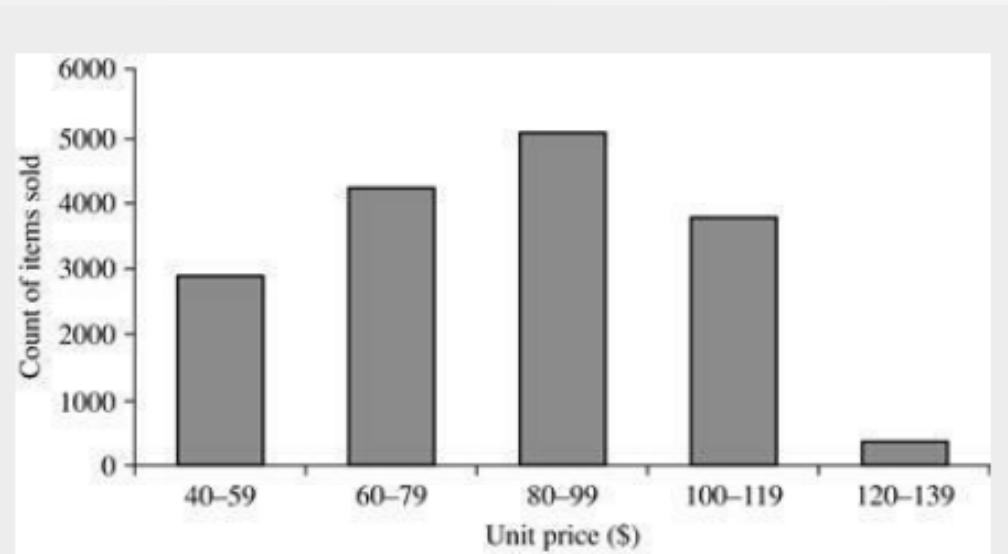


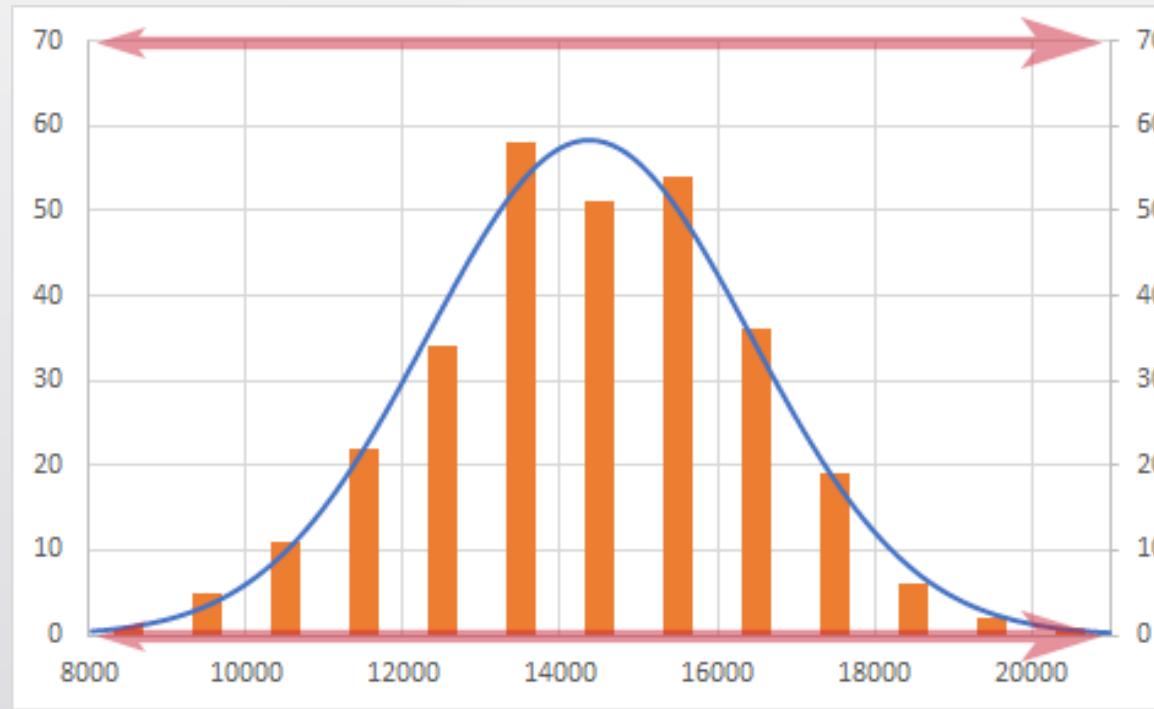
FIGURE 2.6 A histogram for the **Table 2.1** data set.

ヒストグラムを描くことで、
データの分布形状を把握できる
We can grasp data distribution by drawing a histogram

データの分布 Data Distribution

標本数が多ければ、度数分布は曲線で近似できる

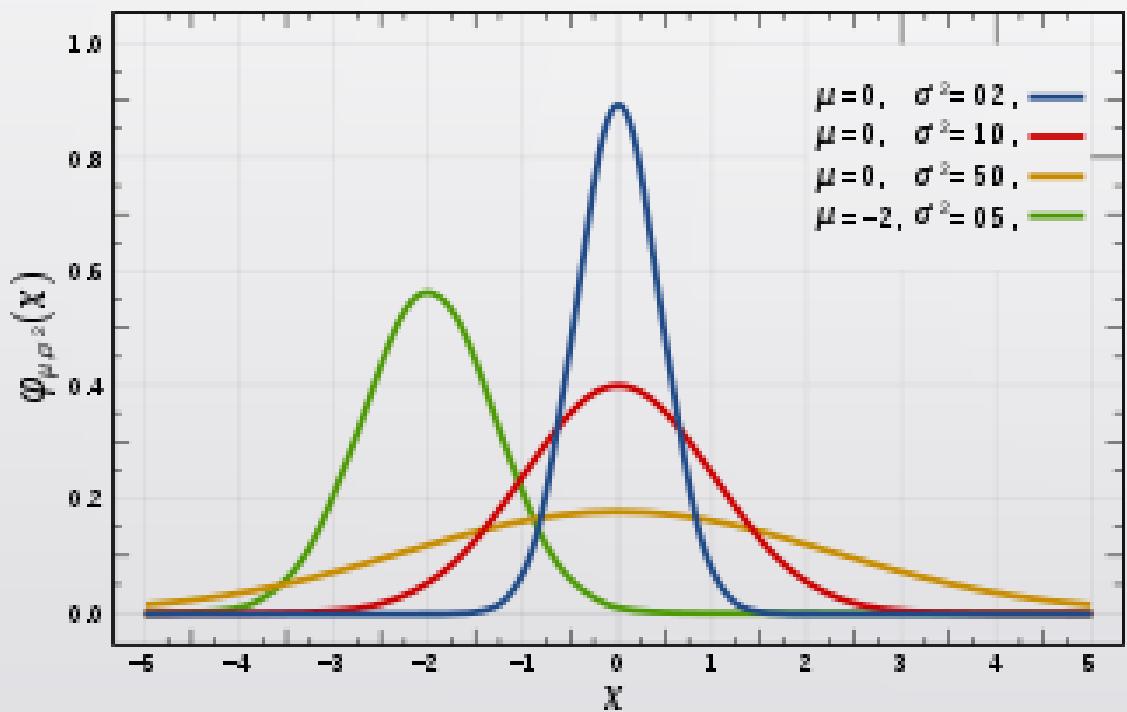
With sufficient number of samples, the frequency distribution can be approximated by a curve



<https://bdastyle.net/tools/histogram/page6.html>



正規分布 Normal Distribution



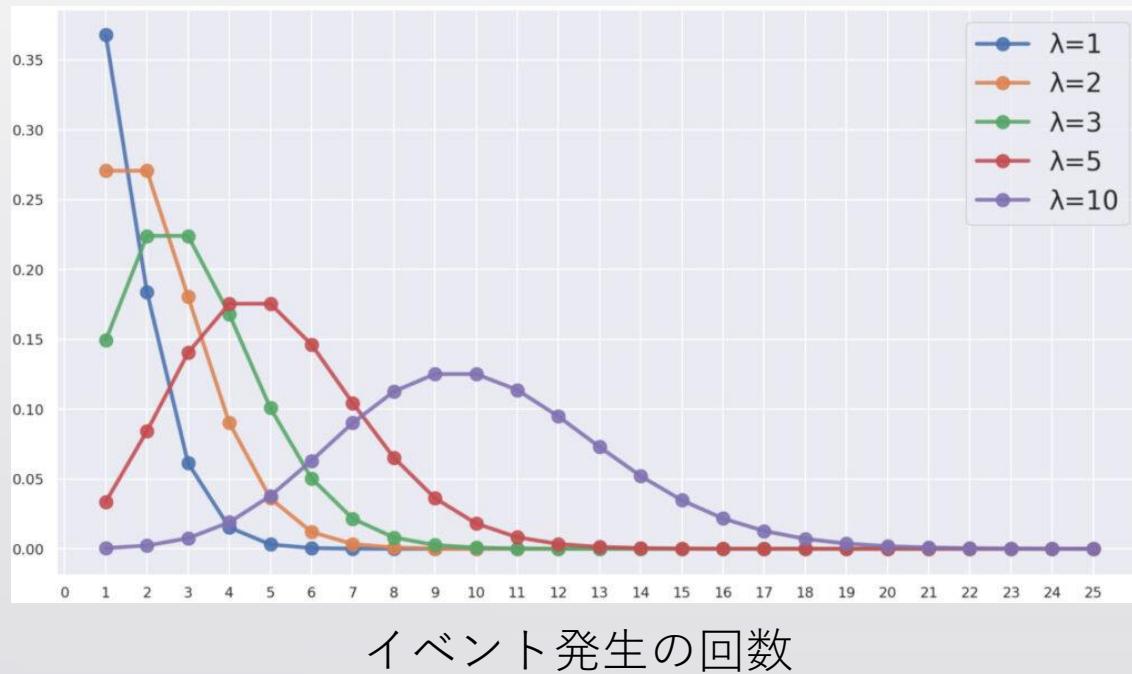
確率密度関数 Probability Density Function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

$\mu = 0, \sigma = 1$ の時は、標準正規分布

Standard normal distribution when $\mu = 0, \sigma = 1$

ポアソン分布 Poisson Distribution

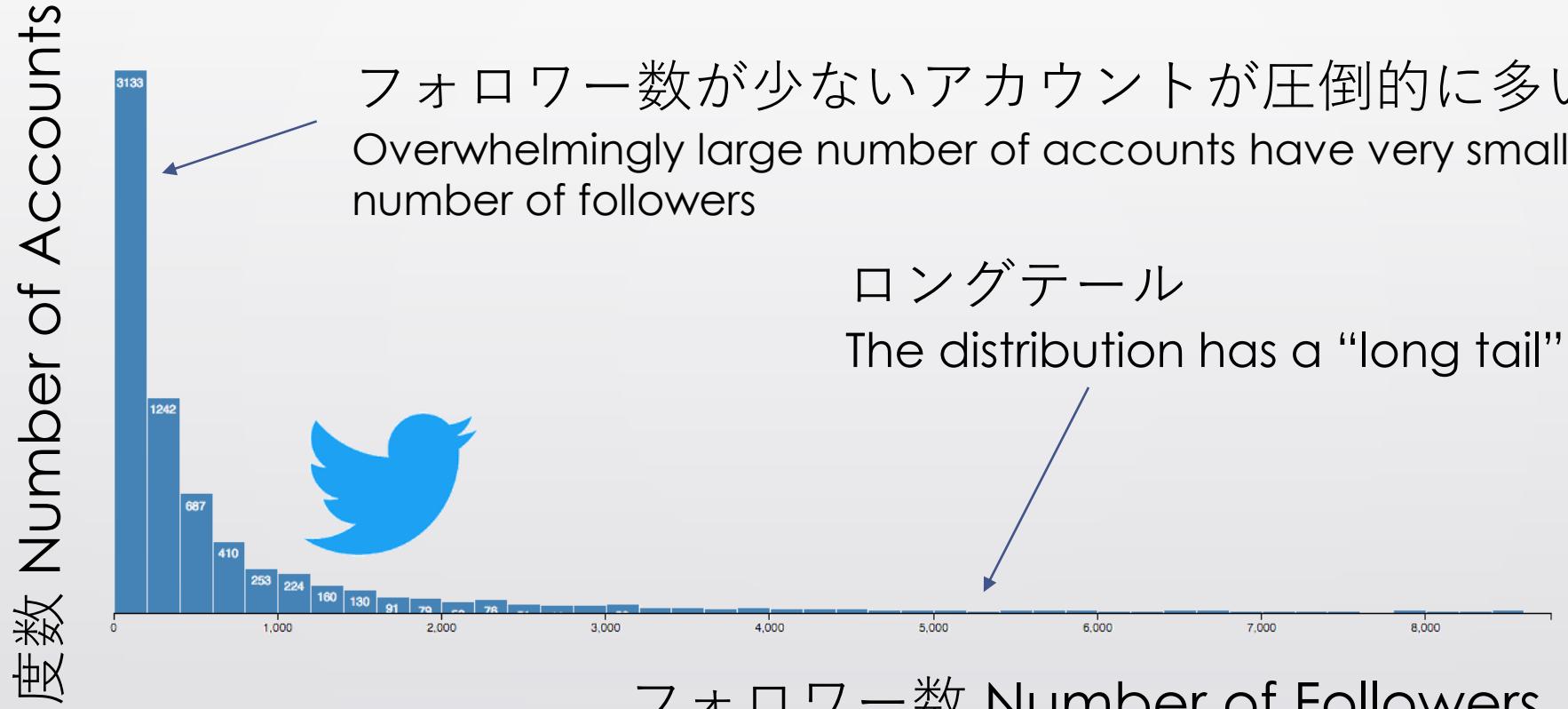


あるイベントが一定時間内に発生する確率を表す

Gives the probability with which a rare event happens within a certain time-window

<https://mathlandscape.com/poisson-distrib/>

指数分布 Exponential Distribution



<https://ultrasaurus.com/2015/05/distribution-of-twitter-followers/>



ロングテール戦略 Long Tail Strategy



<https://blogs.ubc.ca/kathzhang/2014/10/05/the-long-tail-theory-with-examples/>

Q-Q プロット Q-Q Plot

QQプロットにより、2つのデータセットが同じ分布に従うかどうかを検証できる

Q-Q plot tells us whether two dataset conforms to identical distribution

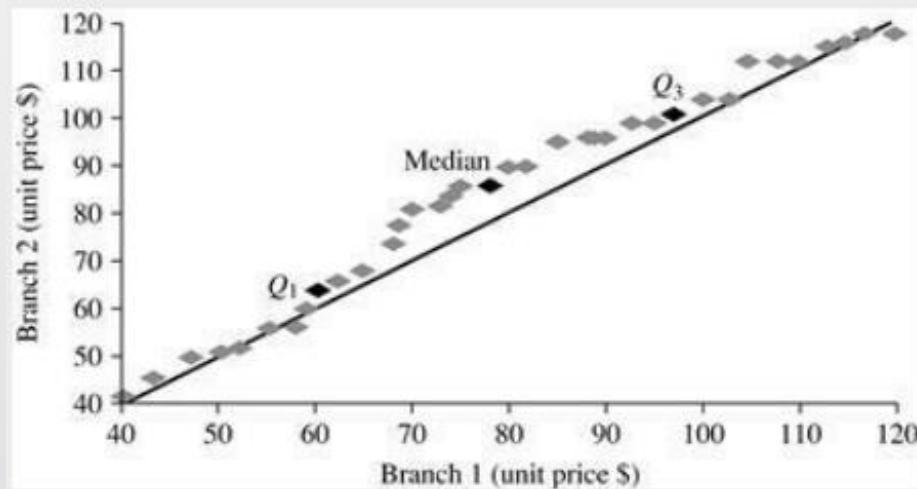
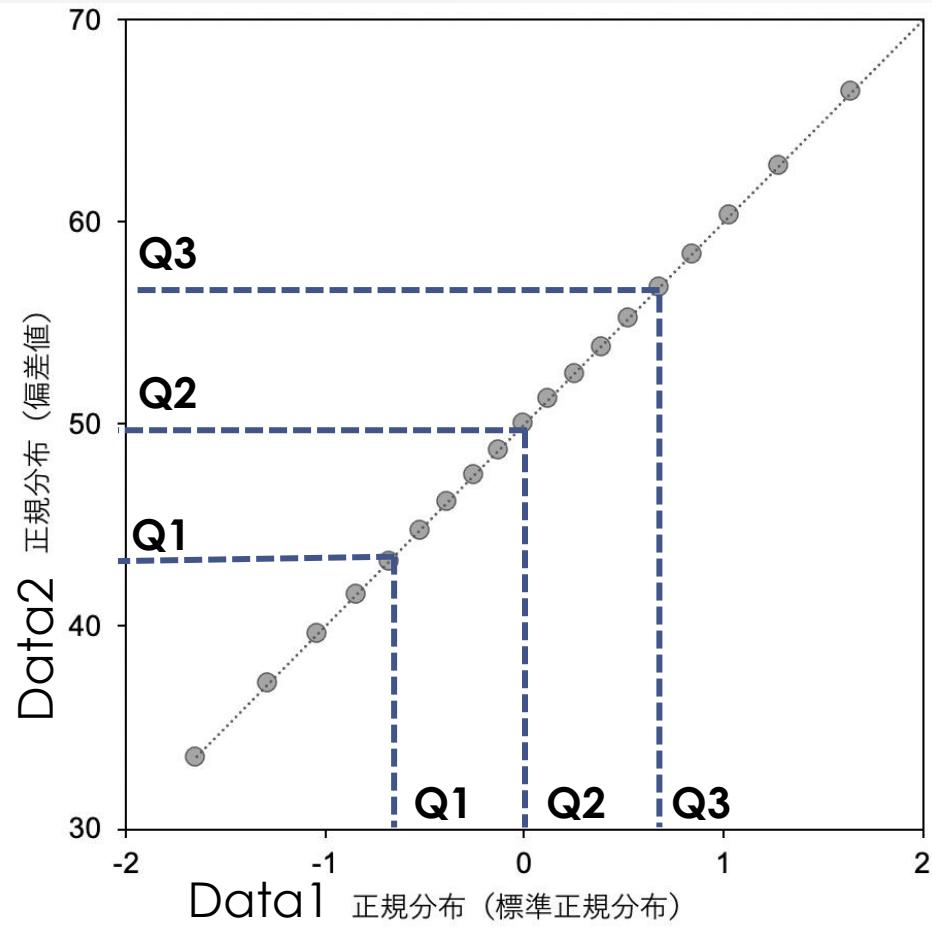


FIGURE 2.5 A q-q plot for unit price data from two AllElectronics branches.

Q-Q プロット Q-Q Plot



Data1 は標準正規分布に従う

Data1 comes from standard normal distribution

Data2 は平均50, 標準偏差10の正規分布に従う

Data2 comes from normal distribution with $\mu = 50, \sigma = 10$

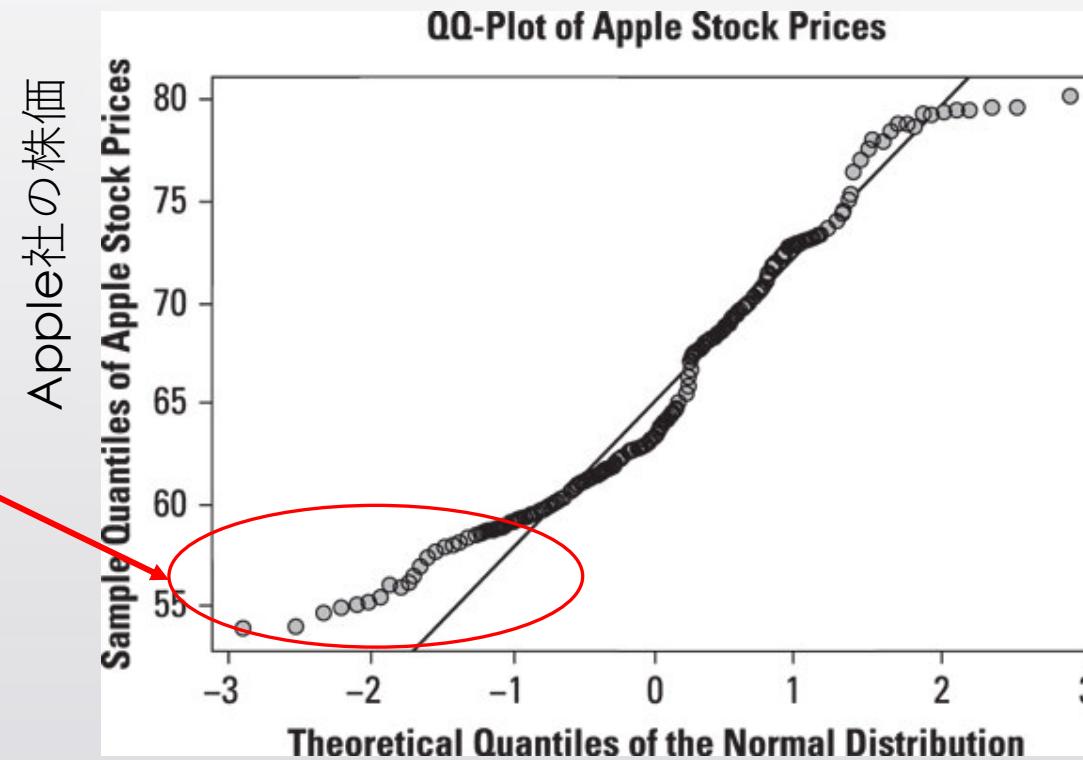
データ分布が同じだと、スケールに関係なく、直線上に点がプロットされる

Points of datasets generated under identical distribution are plotted on a straight line irrespective of data scale

<https://best-biostatistics.com/ezr/q-q-plot-ezr.html>

Q-Q プロット Q-Q Plot

QQプロットは、しばしば、ある変数が正規分布に従うかどうかを検証するために使われる
QQ plot is often utilized to check if certain variable conforms to normal distribution



株価の底値は、正規分布で予想される価格よりも高い

Lowest stock price is lower than the price expected based on normal distribution

<https://www.dummies.com/article/technology/information-technology/data-science/big-data/quantile-quantile-qq-plots-graphical-technique-for-statistical-data-141221/>



前处理

Preprocessing

データマイニングの流れ Steps in Data Mining

1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis・モデリング Modeling
6. 性能評価 Performance Evaluation
7. (ディプロイメント Deployment)

データ前処理 Data Preprocessing

アルゴリズムが扱いやすい形式にデータを変換する

Transform dataset into a format easy for an algorithm to handle

欠損の補完 Interpolation of missing value

外れ値・重複除去 Deletion of outliers and redundancy

ノイズ除去 Noise Cleansing

離散化 Discretization

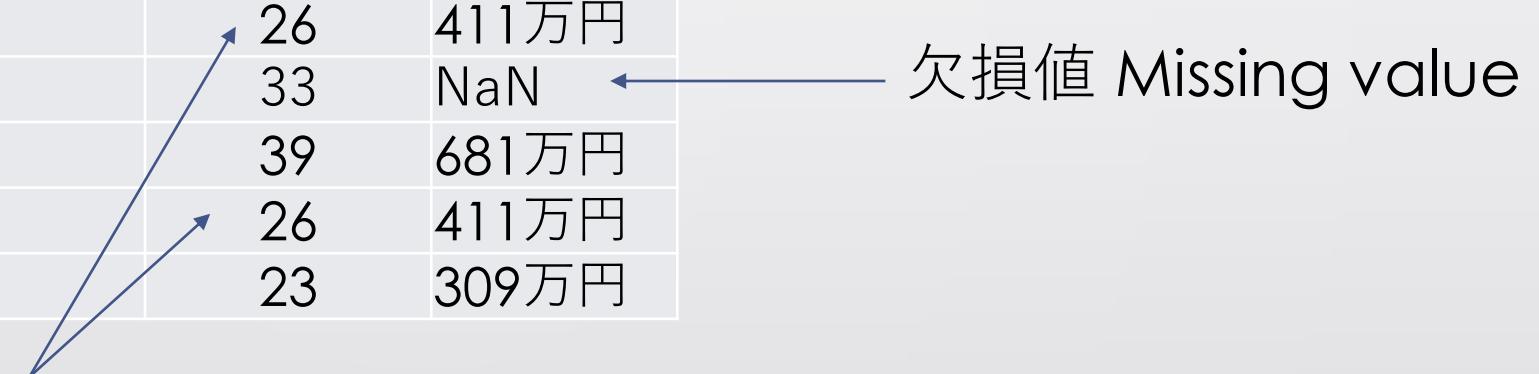
スケーリング Scaling

次元圧縮 Dimension Reduction

欠損値と重複の扱い

Handling of Missing Value and Redundancy

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	NaN
murachan	女	39	681万円
sanapon	男	26	411万円
shozan.s	男	23	309万円



重複 Redundant record

重複がある場合は、特別なケースを除いて、一方を削除する

When there is redundancy, delete either one of the record except for special cases

欠損値の補完 Interpolation of Missing Value

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	NaN
murachan	女	39	681万円
sanapon	男	26	411万円
shozan.s	男	23	309万円

削除することが多い

In many cases this row is deleted entirely

対応するデータを探し出す Somehow retrieve the corresponding value

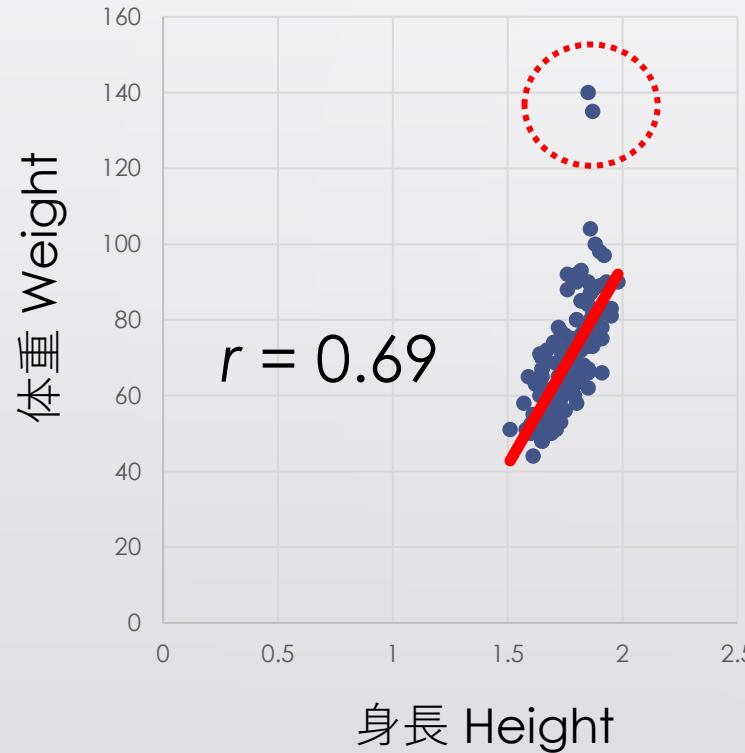
同じクラスの中央値や平均値で代替する Replace NaN with mean or median of the same class

回帰等の方法で補完する Interpolate the missing value by regression etc

外れ値 Outlier

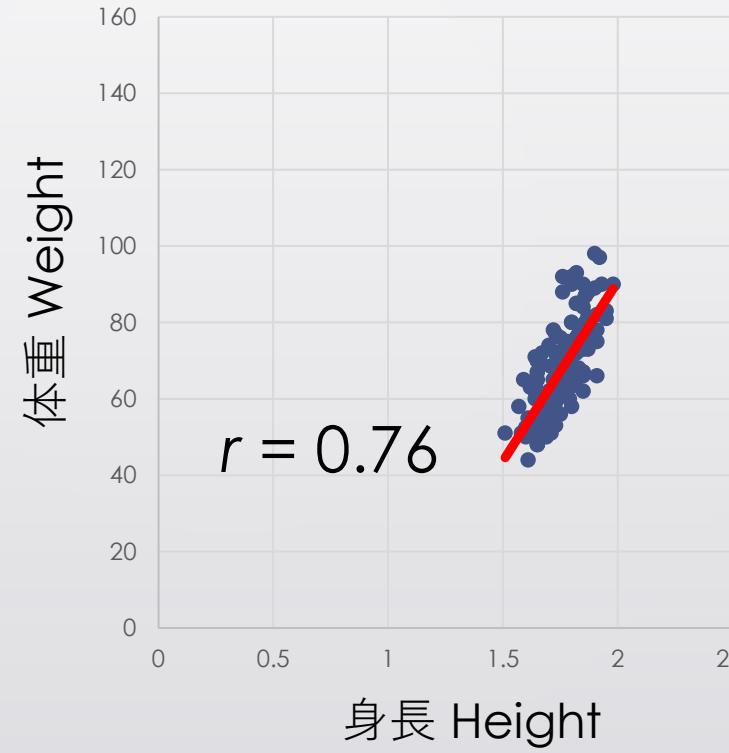
外れ値除去前

Before Outlier Deletion



外れ値除去後

Before Outlier Deletion



結果に影響することがあるので、特にサンプルサイズが小さい場合は、除去がある

Since outliers potentially influence final results, they are sometimes deleted from the dataset especially when the sample size is relatively small

離散化 Discretization

数量を階級やカテゴリーに分割する

Replace numeric values with classes or categories

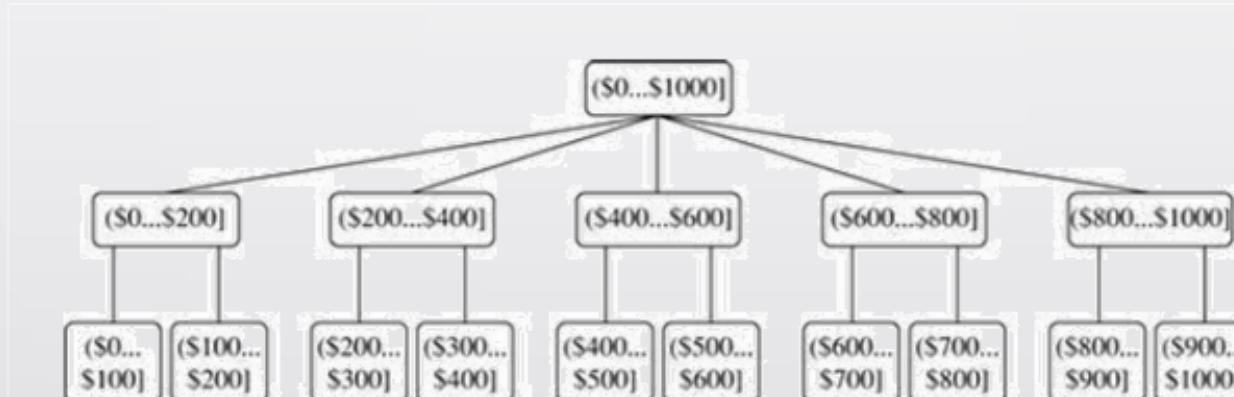


FIGURE 3.12 A concept hierarchy for the attribute *price*, where an interval $(\$X \dots \$Y]$ denotes the range from $\$X$ (exclusive) to $\$Y$ (inclusive).

スケーリング Scaling

データの単位やレンジの違いが結果に影響することを防ぐため、

To mitigate potential influences of the difference in unit and range among variables,

データを標準化 /正規化する

Variables are sometimes standardized/normalized

Z-スコア化 Z-score normalization

$$x' = \frac{x - \mu}{\sigma}$$

Min-Max normalization

$$v'_i = \frac{v_i - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A.$$

ロバストZスコア Robust Z-Score

Z-score化はデータが正規分布することを前提としている

Z-score transformation rests on the presumption that data conforms to normal distribution

ロバストZスコア化は正規分布の仮定を必要としない

Robust z-scoring does not presume normal distribution

$$\text{robust z score} = \frac{x - \text{median}}{\text{NIQR}}$$

$$\text{NIQR} = \frac{\text{IQR}}{1.3489}$$

データが正規分布するとき,
 $\text{IQR} \approx 1.3489\sigma$

次元削減 Dimension Reduction

出来るだけ多くの情報を残しながら、多次元データを、低次元のデータに変換すること

Transform/compress multidimensional data into data with lower dimensions while retaining as much information as possible

計算・データ処理の高速化 Acceleration of computation and data processing

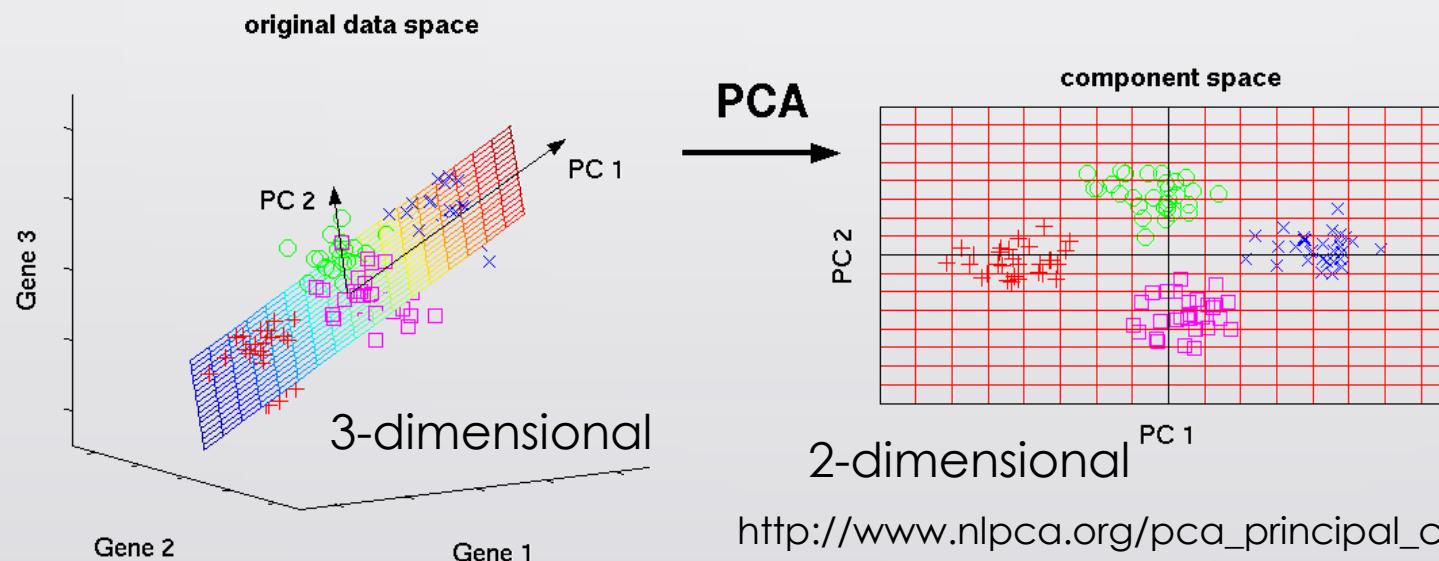
変数の直交化 Orthogonalization of variables

データを解釈しやすくする Enhance interpretability of the data (sometimes...)

主成分分析 Principal Component Analysis (PCA)

変数の線型和により新たな変数(主成分)を合成することで、次元削減を行う手法

Method of dimension reduction with which new composite variables, principal components, are created by linear combination of multiple variables

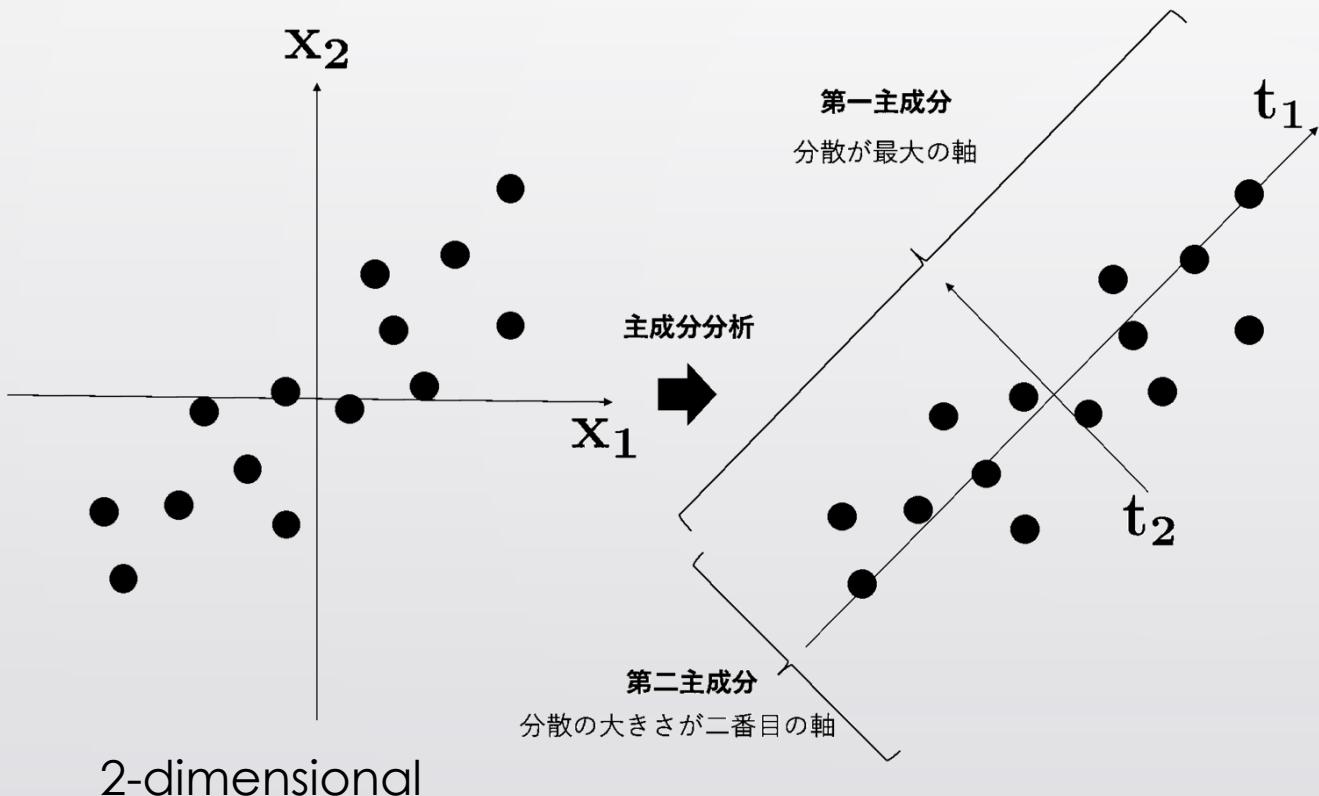


3次元データが、2つの主成分で構成される2次元空間内に表現されている

Three dimensional data are projected onto two-dimensional space defined by two principal components

http://www.nlPCA.org/pca_principal_component_analysis.html

主成分 Principal Components



第1主成分軸は、データの分散が最大化される方向を向いている

The first PC axis is oriented in the direction along which variance of projected data is maximized

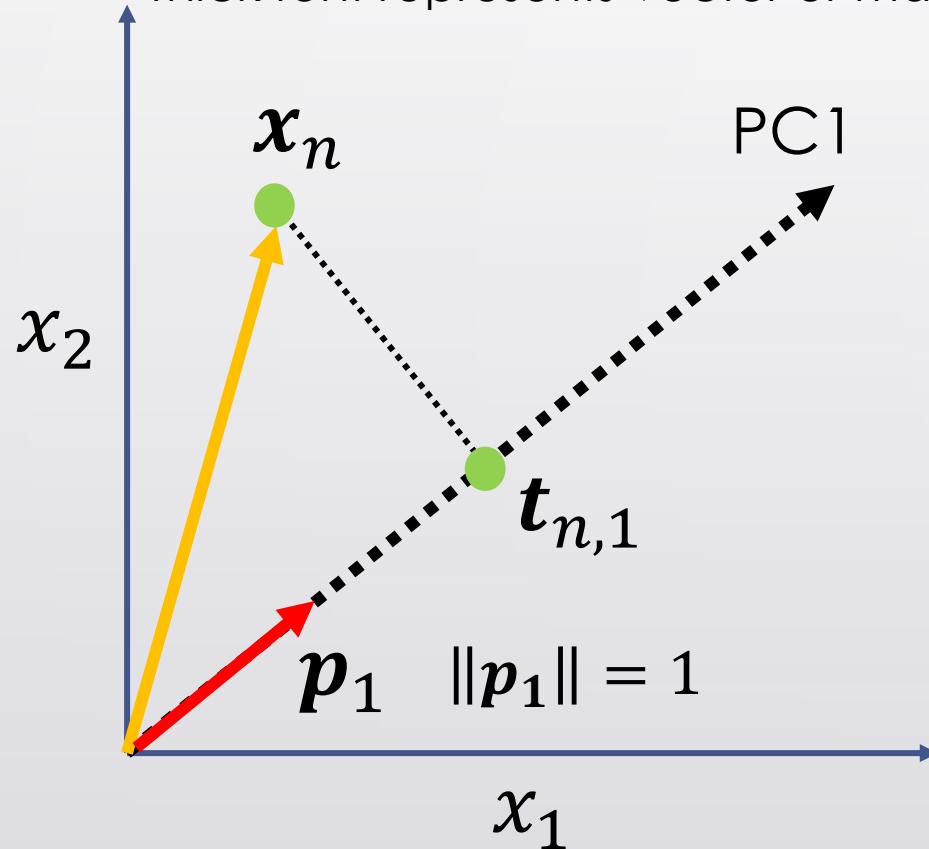
第 j 主成分軸は、データの分散が j 番目の大きさになる方向を向いている

The j -th PC axis is oriented in the direction along which projected data has j -th largest variance

第1主成分の計算 Computation of PC1

太字はベクトルか行列

Thick font represents vector or matrix



変数を中心化しておく Center the variables

観測データ x_n の第1主成分軸方向への射影 $t_{n,1}$ を計算する

Compute the projection $t_{n,1}$ of the observed data x_n onto the first PC axis

$t_{n,1}$ は x_n と p_1 の内積

$t_{n,1}$ is dot(inner) product of x_n and p_1

第1主成分の計算 Computation of PC1

$t_{n,1}$ は x_n と p_1 の内積 $t_{n,1}$ is dot product of x_n and p_1

$$\begin{aligned} \mathbf{x}_n &= [x_{n,1} \ x_{n,2} \ \dots \ x_{n,M}] && \text{データは } M \text{ 次元で } N \text{ 個の観測値 (データ) がある} \\ \mathbf{p}_1 &= [p_{1,1} \ p_{1,2} \ \dots \ p_{1,M}] && \text{Data is } M\text{-dimensional and there are in total of} \\ &&& N \text{ observations (Data points)} \end{aligned}$$

$$\mathbf{t}_1 = \begin{bmatrix} t_{1,1} \\ t_{2,1} \\ \vdots \\ t_{N,1} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} \\ p_{2,1} \\ \vdots \\ p_{M,1} \end{bmatrix} = \mathbf{X}\mathbf{p}_1$$

第1主成分の計算 Computation of PC1

t_1 の分散 $s_{t_1}^2$ を計算する Compute variance $s_{t_1}^2$ of t_1

$$Variance = \frac{1}{n} (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})$$

$$s_{t_1}^2 = \frac{1}{N} (\mathbf{t}_1 - mean(\mathbf{t}_1))^T (\mathbf{t}_1 - mean(\mathbf{t}_1)) = \frac{1}{N} \mathbf{t}_1^T \mathbf{t}_1 = \frac{1}{N} (\mathbf{X}\mathbf{p}_1)^T (\mathbf{X}\mathbf{p}_1) = \frac{1}{N} \mathbf{p}_1^T \mathbf{X}^T \mathbf{X} \mathbf{p}_1$$

変数は中心化されているので $mean(t_1) = 0$

$mean(t_1) = 0$ since variables are centered

$$\mathbf{t}_1 = \mathbf{X}\mathbf{p}_1$$

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

第1主成分の計算 Computation of PC1

t_1 の分散 $s_{t_1}^2$ を計算する Compute variance $s_{t_1}^2$ of t_1

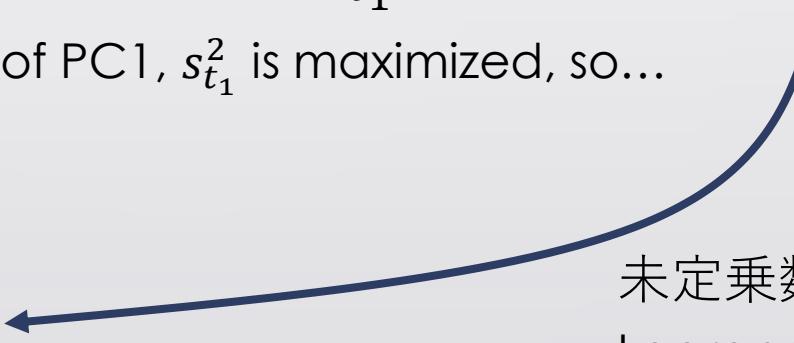
$$s_{t_1}^2 = \frac{1}{N} \mathbf{p}_1^T \mathbf{X}^T \mathbf{X} \mathbf{p}_1 = \mathbf{p}_1^T \mathbf{V} \mathbf{p}_1 \quad \mathbf{V} = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$

\mathbf{p}_1 が第1主成分軸の方向を向いている時 $s_{t_1}^2$ が最大化されるので...

When \mathbf{p}_1 is oriented in the direction of PC1, $s_{t_1}^2$ is maximized, so...

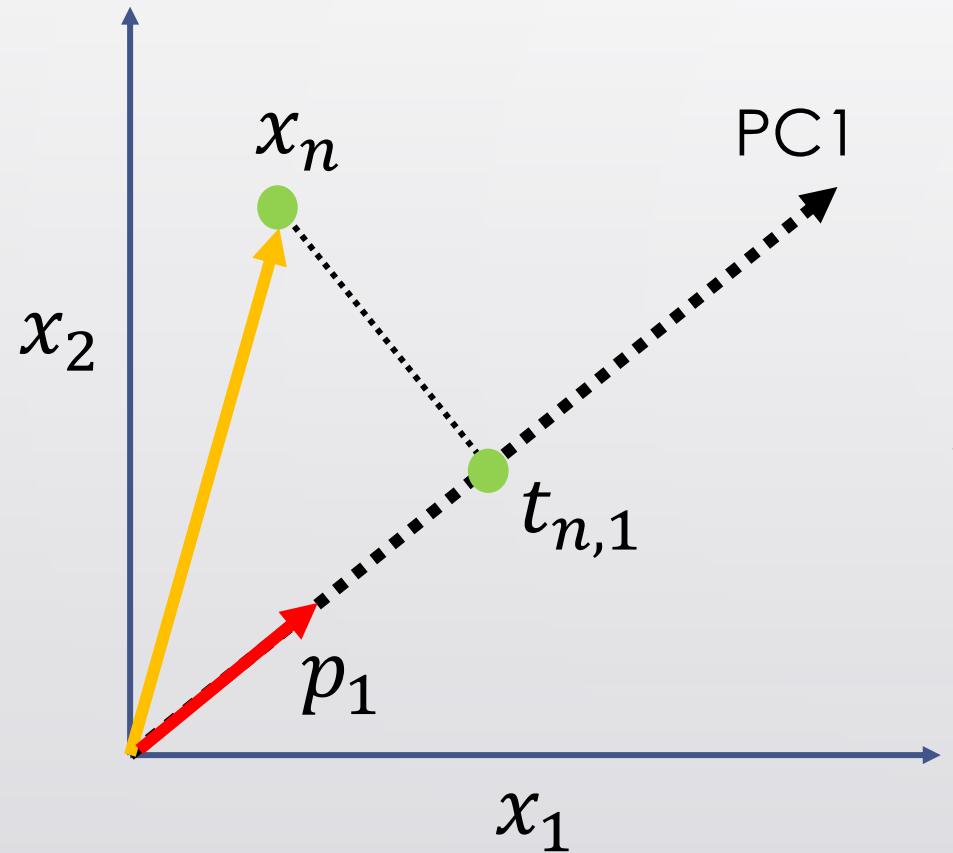
$$\mathbf{V} \mathbf{p}_1 = \lambda \mathbf{p}_1$$

\mathbf{p}_1 は \mathbf{V} の固有ベクトルである
 \mathbf{p}_1 is eigenvector of \mathbf{V}



未定乗数法
Lagrange method of multiplier

第1主成分の計算 Computation of PC1

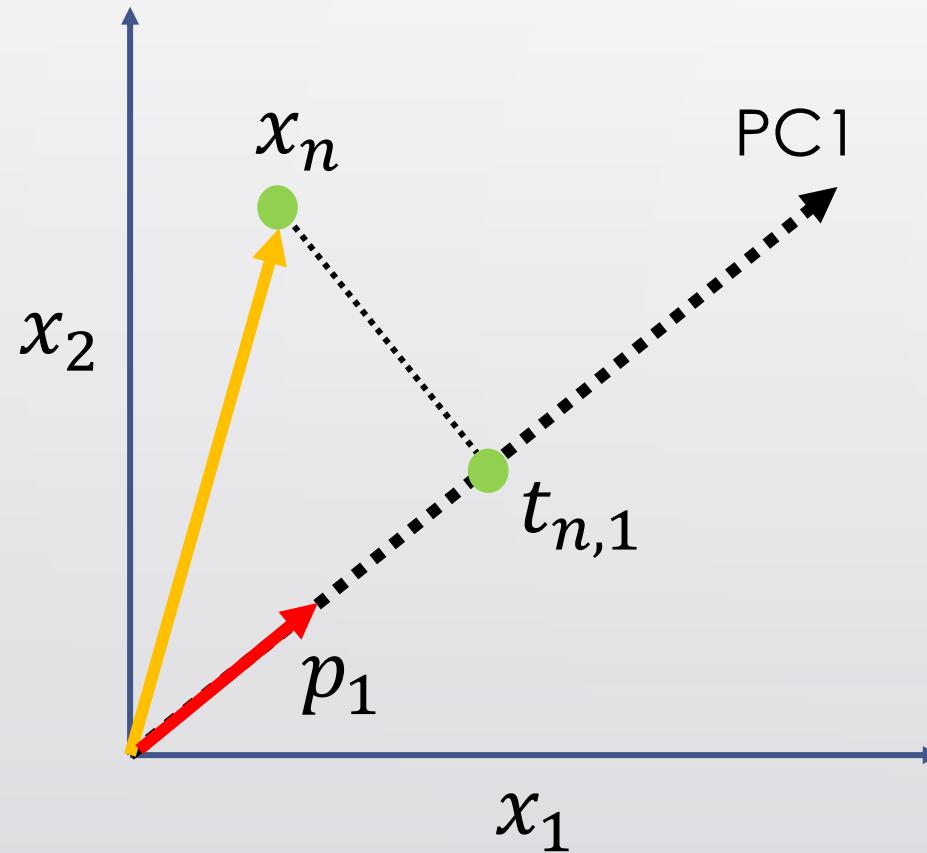


$$Vp_1 = \lambda p_1 \quad V = \frac{1}{N} X^T X$$

p_1 は V の固有ベクトルである p_1 is eigenvector of V

では、固有値 λ は何を表しているのだろうか？
Then, what does eigen value λ represent?

第1主成分の計算 Computation of PC1



では、固有値 λ は何を表しているのだろうか？
Then, what does eigen value λ represent?

$$s_{t_1}^2 = \mathbf{p}_1^T V \mathbf{p}_1 = \lambda \mathbf{p}_1^T \mathbf{p}_1 = \lambda \|\mathbf{p}_1\|^2 = \lambda$$

$$\begin{aligned} V \mathbf{p}_1 &= \lambda \mathbf{p}_1 \\ \|\mathbf{p}_1\| &= 1 \end{aligned}$$

t_1 の分散は λ に一致する Variance of t_1 equals to λ

第1主成分の計算 Computation of PC1

$V\mathbf{p}_1 = \lambda\mathbf{p}_1$ \mathbf{p}_1 は V の固有ベクトルである \mathbf{p}_1 is eigenvector of V

$V = \frac{1}{N} \mathbf{X}^T \mathbf{X}$ \mathbf{t}_1 の分散は λ に一致する Variance of \mathbf{t}_1 equals to λ

V とは何か？ What is V ？

$$V = \frac{1}{N} \mathbf{X}^T \mathbf{X} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

分散共分散行列 Variance-Covariance Matrix

V は X の分散共分散行列である V is variance-covariance matrix of X

$$V = \frac{1}{N} X^T X = \frac{1}{N} \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,M} & x_{2,M} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \dots & \sigma_{1,M}^2 \\ \sigma_{2,1}^2 & \ddots & & \vdots \\ \vdots & & & \sigma_{M-1,M}^2 \\ \sigma_{M,1}^2 & \sigma_{M,2}^2 & \dots & \sigma_{M,M}^2 \end{bmatrix} \quad \sigma_{i,j}^2 = \frac{1}{N} \sum_{k=1}^N x_{k,i} x_{k,j}$$



主成分分析の手順 Procedure of PCA

1. データ \mathbf{X} の分散共分散行列 \mathbf{V} を計算する

Compute variance-covariance matrix \mathbf{V} of data \mathbf{X}

2. 分散共分散行列 \mathbf{V} を固有値分解する

Eigenvalue decomposition of matrix \mathbf{V}

3. 第 k 主成分の分散は, k 番目に大きな固有値 λ_k

Variance of k -th principal component is k -th largest eigenvalue λ_k of matrix \mathbf{V}

4. 第 k 主成分は, 固有値 λ_k に対応する固有ベクトル \mathbf{p}_k

k -th principal component is eigenvector \mathbf{p}_k that corresponds to eigenvalue λ_k

第k主成分 k -th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X}\mathbf{p}_k$$

主成分同士は直交している Principal components are orthogonal

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 1(i = j) \\ 0(i \neq j) \end{cases}$$

第k主成分 k -th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X}\mathbf{p}_k$$

主成分得点

Factor Score

主成分負荷量

Factor Loading

元データと主成分負荷量の内積で、主成分得点が得られる

Factor score is computed as dot product of observation and factor loading

第k主成分 k -th Principal Component

$$\mathbf{T} = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \mathbf{t}_M] = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,M} \\ t_{2,1} & t_{2,2} & \dots & t_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \dots & t_{N,M} \end{bmatrix} =$$

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,M} \\ p_{2,1} & p_{2,2} & \dots & p_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,M} \end{bmatrix} = X[\mathbf{p}_1 \ \mathbf{p}_2 \ \dots \mathbf{p}_M] = XP$$

固有値が大きい順に固有ベクトルを並べた

Eigenvectors are ordered in a descending order of eigenvalue

次元削減 Dimension Reduction

$$T = [t_1 \ t_2 \ \dots \ t_M] = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,M} \\ t_{2,1} & t_{2,2} & \dots & t_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N,1} & t_{N,2} & \dots & t_{N,M} \end{bmatrix} = \begin{array}{l} \text{第}H+1\sim M\text{主成分を削除する} \\ \text{Deleting } H+1\text{-th to } M\text{-th PCs} \end{array}$$

次元削減 Dimension Reduction

$$T' = [\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_H] =$$

(N, H)

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,H} \\ p_{2,1} & p_{2,2} & \dots & p_{2,H} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,H} \end{bmatrix} = X[\mathbf{p}_1 \ \mathbf{p}_2 \ \dots \ \mathbf{p}_H]$$

(N, M) (M, H)

情報量（分散）が小さい主成分を除くことで、もとのデータが持つ情報を、低い次元で表現できる

Information contained in the original data can be represented in lower number of dimensions by deleting PCs with small amount of information (variance)

次元削減 Dimension Reduction

何番目の主成分まで残すべきか？

How many principal components should we retain in dimension reduction?

3. 第k主成分の分散は, k 番目に大きな固有値 λ_k

第1~H主成分までの情報量の合計は

Sum of amount of information of the first H principal components is

$$\sum_1^H \lambda_k$$



累積寄与率 Cumulative Contribution Ratio

第1~H主成分までの情報量の合計は

Sum of amount of information of the first H principal components is

$$\sum_1^H \lambda_k$$

データ全体の情報量の合計は

Sum of amount of information of the data set is

$$\sum_1^M \lambda_k$$

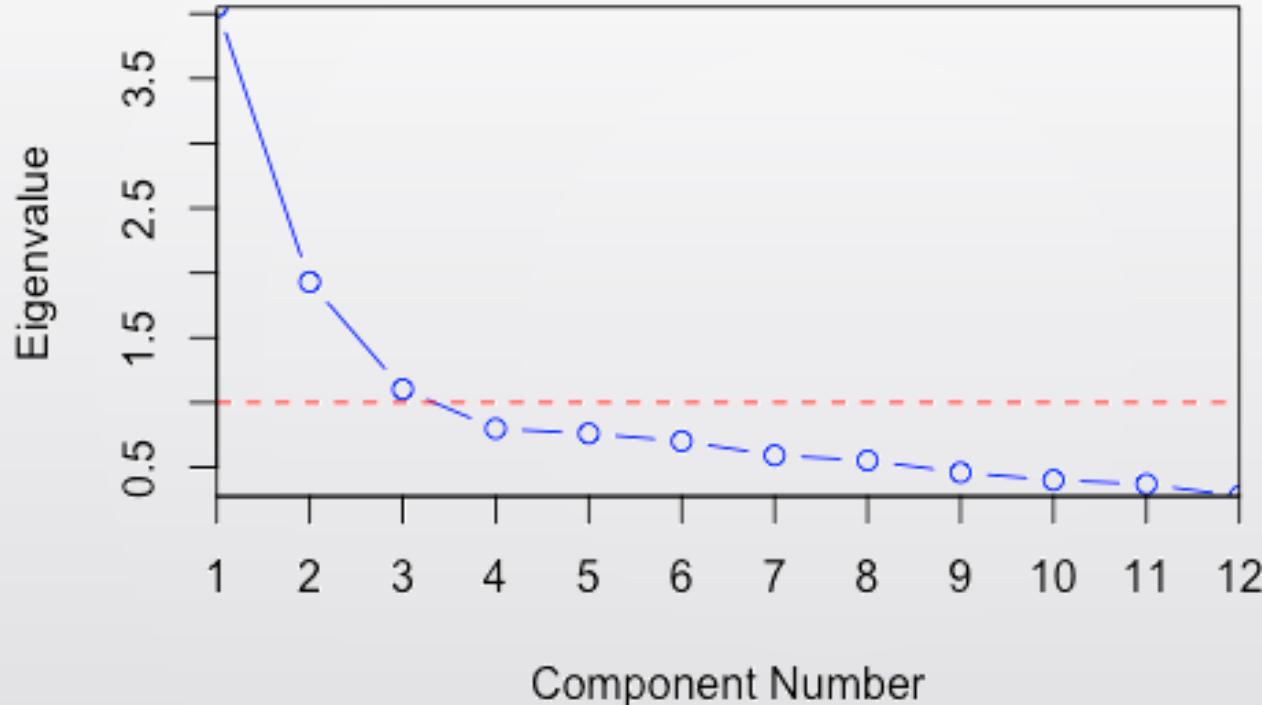
$$\frac{\sum_1^H \lambda_k}{\sum_1^M \lambda_k} \times 100 \geq 80$$

これ以外にも主成分の数を決める基準は色々ある

There are many other customary criteria for determining the number of principal components to be retained

スクリープロット

Scree Plot



https://en.wikipedia.org/wiki/Scree_plot

スクリープロットの肩の位置で、主成分の数を決めることがある

Number of retained PCs is sometimes determined by the location of “shoulder” in scree plot.



データマイニング

Data Mining

4: 回帰① Regression 1

土居 裕和 Hirokazu Doi

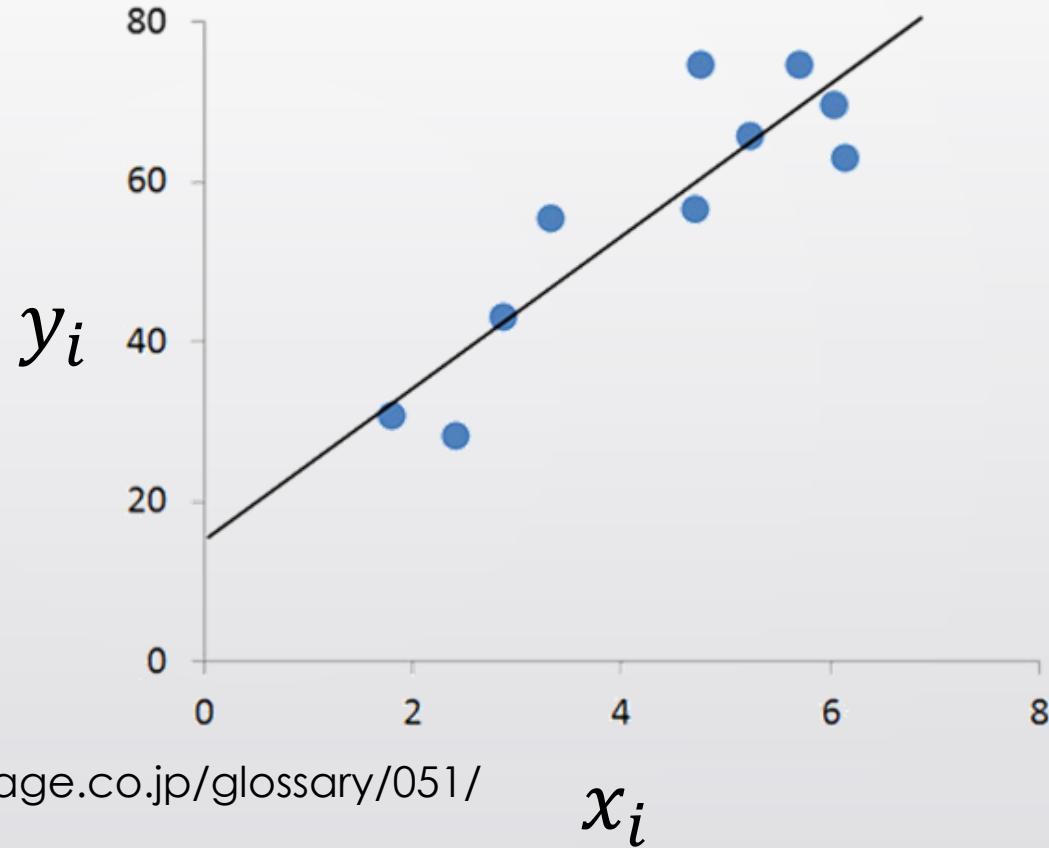
長岡技術科学大学 Nagaoka University of Technology

回帰 Regression

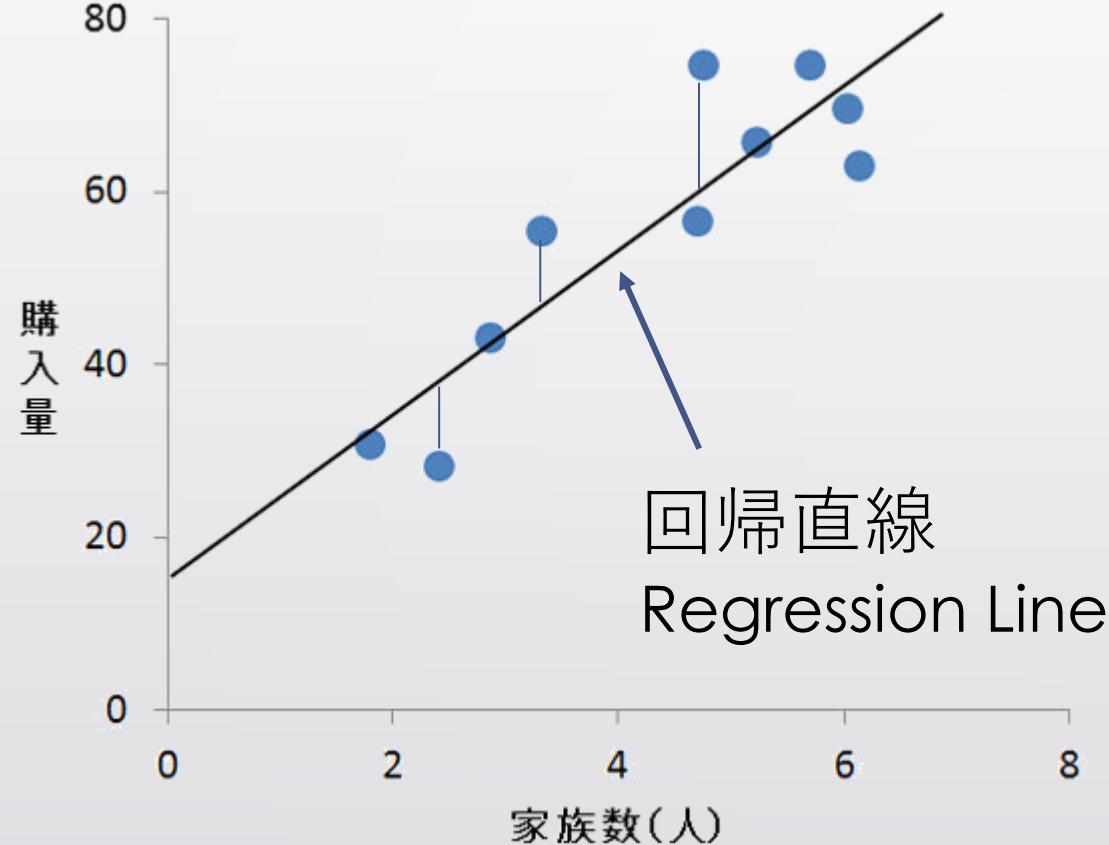
(x_i, y_i)

x_i : 家族の人数
Number of Family Member

y_i : 購入数
Number of purchased Items



線型回帰直線 Linear Regression Line



$$y = kx + y_0$$

傾き
Slope

切片
Intercept

カリフォルニア住宅データセット

California Housing Dataset

California Housing dataset

Data Set Characteristics:

:Number of Instances: 20640

:Number of Attributes: 8 numeric, predictive attributes and the target

:Attribute Information:

- MedInc median income in block
- HouseAge median house age in block
- AveRooms average number of rooms
- AveBedrms average number of bedrooms
- Population block population
- AveOccup average house occupancy
- Latitude house block latitude
- Longitude house block longitude

:Missing Attribute Values: None

This dataset was obtained from the StatLib repository.

<http://lib.stat.cmu.edu/datasets/>

The target variable is the median house value for California districts.

複数の情報に基づいて、住宅価格を予測する

Predict housing price based on multiple information

予測変数 Predictors

ターゲット（目的）変数
Target Variable

重回帰分析 Multiple Linear Regression

複数の変数の線型和によって、ターゲット変数を予測

Predicts target variable by linear combination of multiple variables

Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

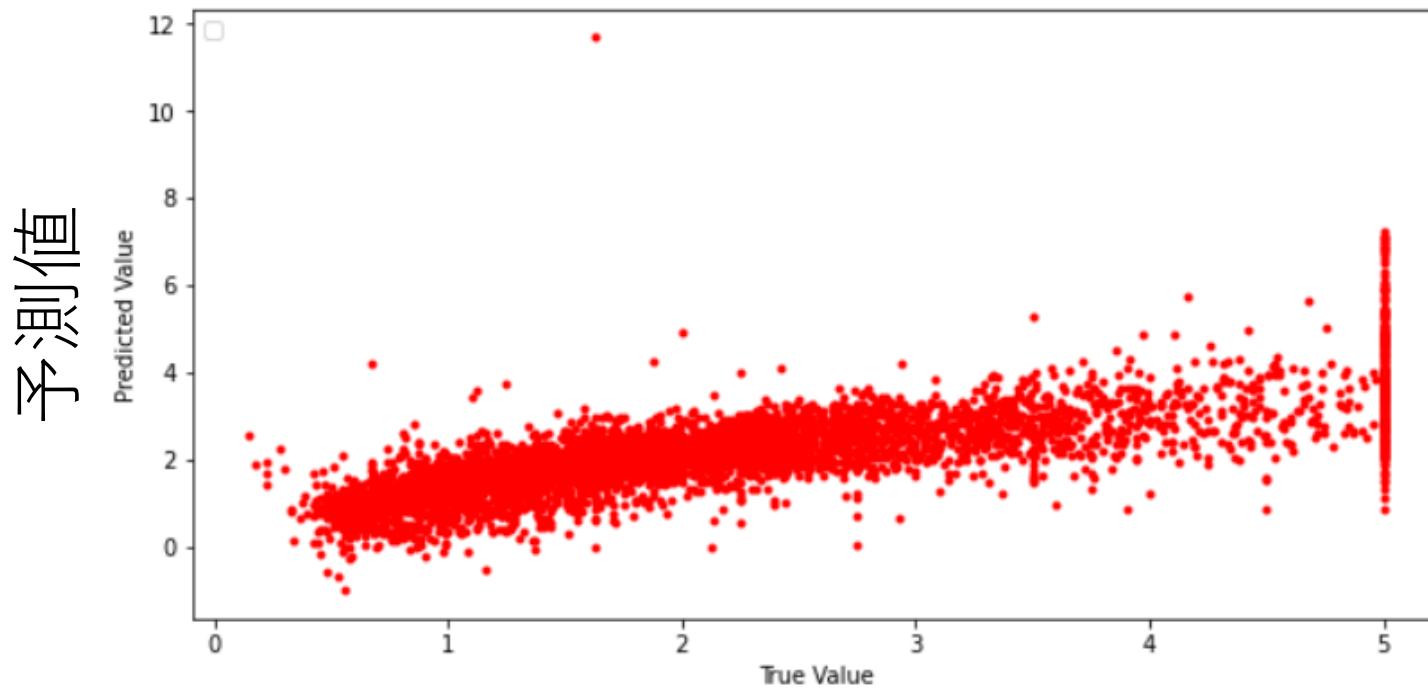
Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

<https://www.i2tutorials.com/difference-between-simple-linear-regression-and-multi-linear-regression-and-polynomial-regression/>

相関係数による性能評価

```
Out[33]: array([[1.          , 0.76907401],  
                 [0.76907401, 1.         ]])
```



正解値



重回帰分析 Multiple Linear Regression

California Housing dataset

Data Set Characteristics:

:Number of Instances: 20640

:Number of Attributes: 8 numeric, predictive attributes and the target

:Attribute Information:

- MedInc	median income in block
- HouseAge	median house age in block
- AveRooms	average number of rooms
- AveBedrms	average number of bedrooms
- Population	block population
- AveOccup	average house occupancy
- Latitude	house block latitude
- Longitude	house block longitude

:Missing Attribute Values: None

This dataset was obtained from the StatLib repository.

<http://lib.stat.cmu.edu/datasets/>

The target variable is the median house value for California districts.

$$House\ Value = \beta_1 MedInc + \beta_2 HouseAge + \dots + \beta_8 Longitude + \beta_0 + \varepsilon$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M + \beta_0 + \varepsilon$$

誤差 Error



重回歸分析 Multiple Linear Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M + \beta_0 + \varepsilon$$

$$= \sum_1^M \beta_i x_i + \beta_0 + \varepsilon = [\beta_1 \ \beta_2 \ \dots \ \beta_M] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} + \beta_0 + \varepsilon$$

$$= \boldsymbol{\beta}^T \boldsymbol{x} + \beta_0 + \varepsilon$$

誤差 Error

最小二乗法 Ordinary Least Squares (OLS) Method

データは M 次元で N 個の観測値（データ）がある

Data is M -dimensional and there are in total of N observations (Data points)

$$\mathbf{x}_n = [x_{n,1} \ x_{n,2} \ \dots \ x_{n,M}] \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

それぞれのデータ \mathbf{x}_n に対応するターゲットは \mathbf{y}_n $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

\mathbf{X} と \mathbf{y} は中心化されている

Note \mathbf{X} and \mathbf{y} are centered



最小二乗法 Ordinary Least Squares (OLS) Method

\mathbf{X} から \mathbf{y} を精度よく予測できる重回帰モデルの $\boldsymbol{\beta}$ を求める

Find $\boldsymbol{\beta}$ so that multiple regression model can predict \mathbf{y} based on \mathbf{X} with good precision

$$\mathbf{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

重回帰モデルによる \mathbf{y} の推定値 Prediction of \mathbf{y} based on multiple regression

観測値 \mathbf{y} と予測値 \mathbf{y}' との差を最小化する $\boldsymbol{\beta}$ を求める

Find $\boldsymbol{\beta}$ that minimizes the difference between observed vector \mathbf{y} and predicted vector \mathbf{y}'



最小二乗法 Ordinary Least Squares (OLS) Method

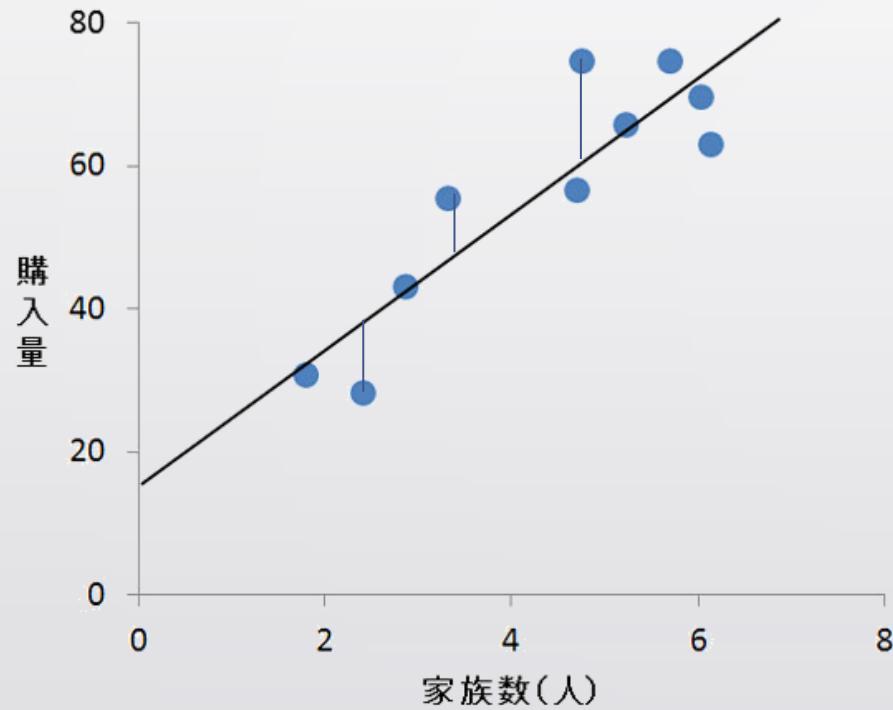
観測値 \mathbf{y} と予測値 \mathbf{y}' との差を最小化する $\boldsymbol{\beta}$ を求める

Find $\boldsymbol{\beta}$ that minimizes the difference between observed vector \mathbf{y} and predicted vector \mathbf{y}'

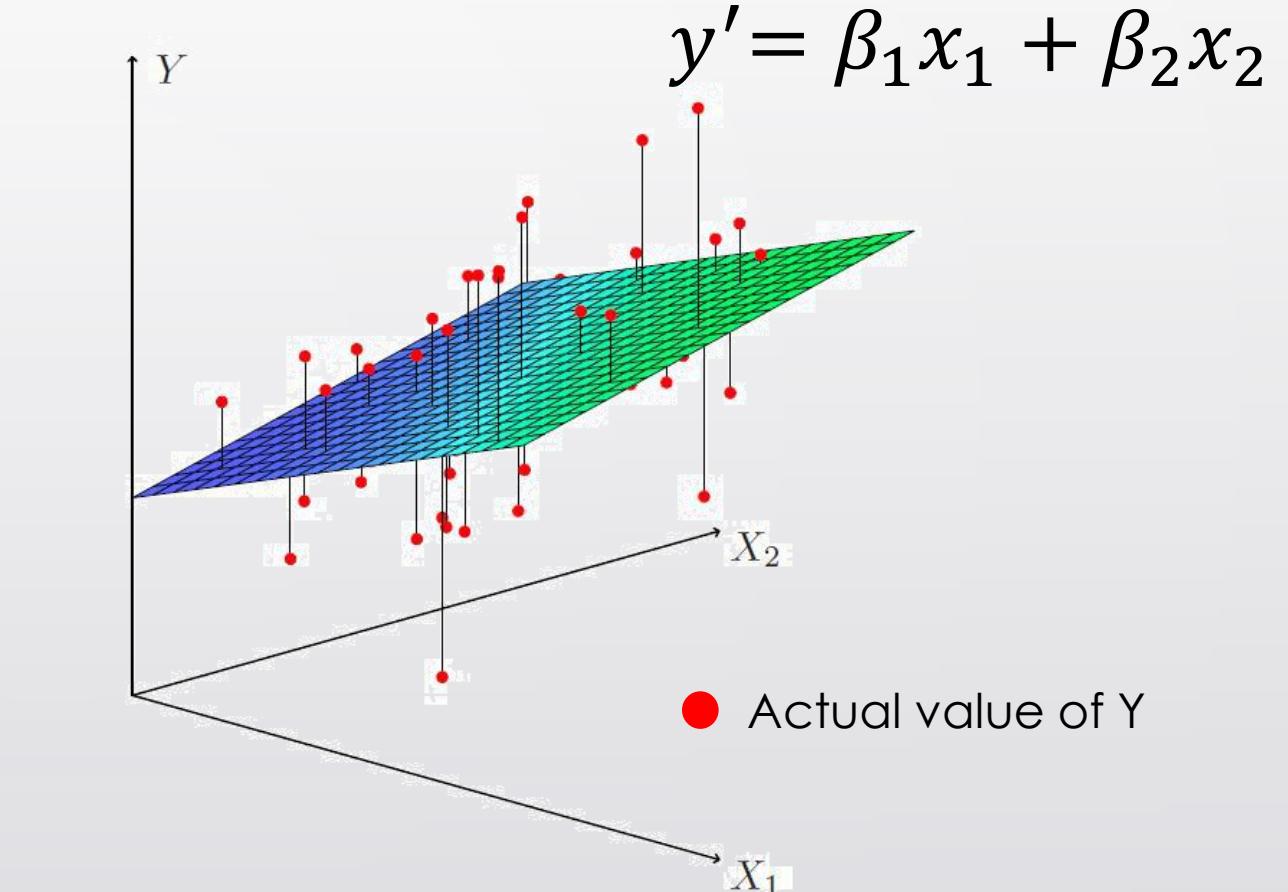
残差二乗和 Residual Sum of Squares (RSS)

$$RSS = \sum_1^N (y_n - y'_n)^2 = [y_1 - y'_1 \ y_2 - y'_2 \ \dots y_N - y'_N] \begin{bmatrix} y_1 - y'_1 \\ y_2 - y'_2 \\ \vdots \\ y_N - y'_N \end{bmatrix} = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}')$$

残差二乗和 Residual Sum of Squares



<https://www.intage.co.jp/glossary/051/>



<https://medium.com/analytics-vidhya/multiple-linear-regression-an-intuitive-approach-f874f7a6a7f9>



最小二乗法 Ordinary Least Squares (OLS) Method

$$RSS = \sum_1^N (y_n - y'_n)^2 = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}')$$

RSSを最小にする β は次の条件を満たす

β that minimizes RSS satisfies the condition below

$$\frac{\partial RSS}{\partial \beta} = 2X^T X \beta - 2X^T y = 0$$

正規方程式 Normal Equation

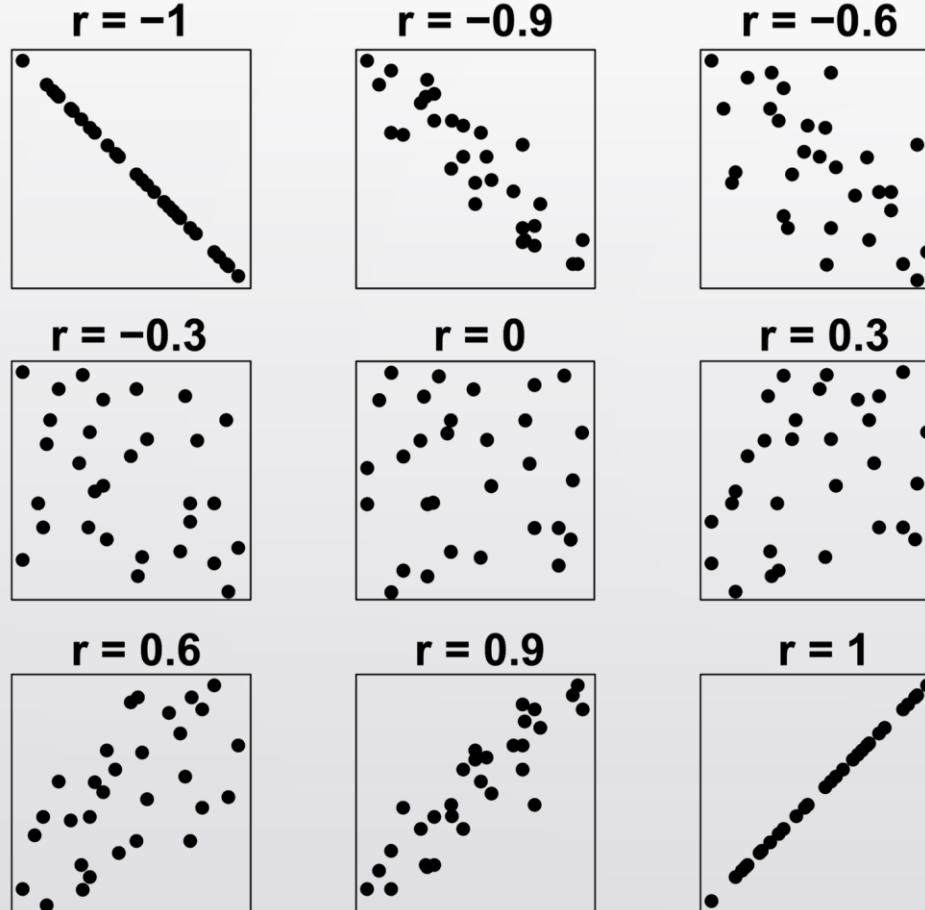
RSSを最小にする β は次の条件を満たす
 β that minimizes RSS satisfies the condition below

$$\frac{\partial \text{RSS}}{\partial \beta} = 2X^T X \beta - 2X^T y = 0$$

$$X^T X \beta = X^T y \longleftarrow \text{正規方程式 Normal Equation}$$

$$\beta = (X^T X)^{-1} X^T y$$

相関係数 Correlational Coefficients



2つの変数の間の関連性の強さを表す

Quantifies the strength of association between two variables

$[-1, 1]$ の間で変動するよう標準化されている
standardized between -1 to 1

共分散 Covariance

$$s_{xy} = \frac{1}{N} \sum_1^N (x_i - \mu_x)(y_i - \mu_y)$$

If

$x_i - \mu_x$ と $y_i - \mu_y$ が共に正

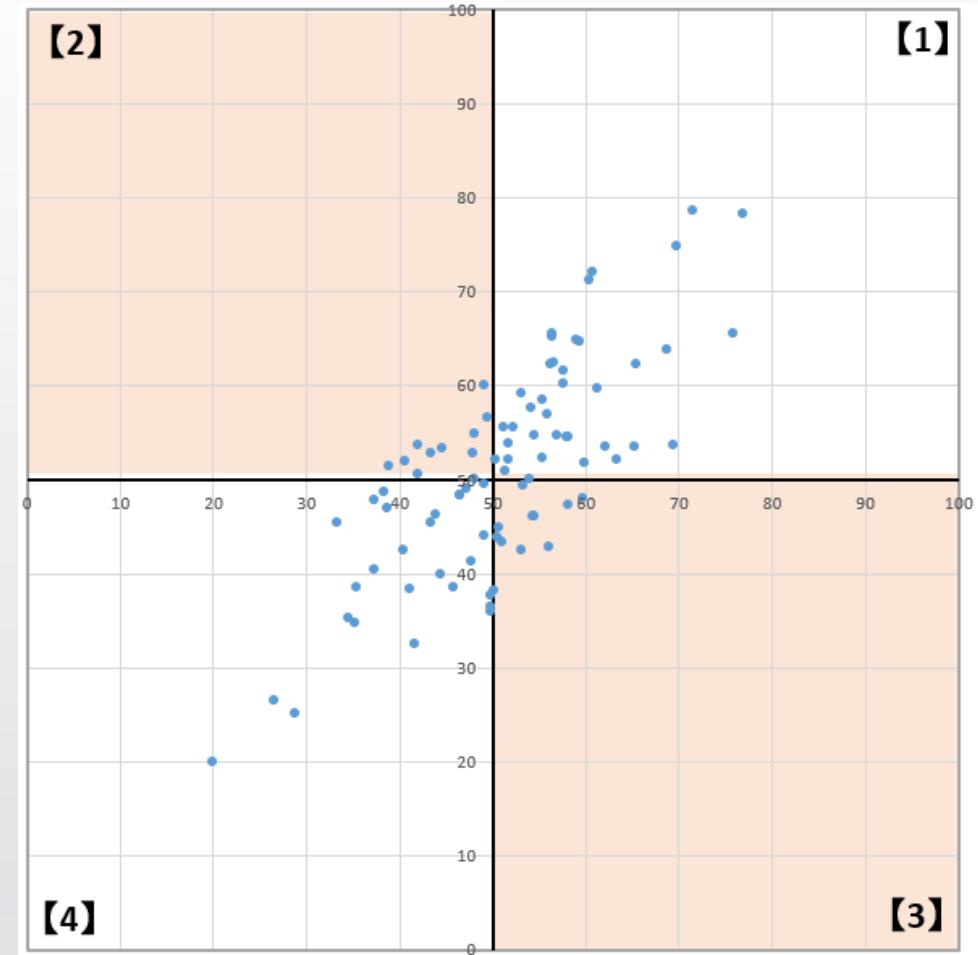
Or

$x_i - \mu_x$ と $y_i - \mu_y$ が共に負

Then

(x_i, y_i) は 【1】 か 【4】 に

(x_i, y_i) belongs to 【1】 or 【4】



<https://datasciencehenomiti.com/post-161/>

相関係数 Correlational Coefficients

$$\begin{aligned} r_{xy} &= \frac{1}{N} \sum_1^N \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\frac{1}{N} \sum_1^N (x_i - \mu_x)^2} \sqrt{\frac{1}{N} \sum_1^N (y_i - \mu_y)^2}} \\ &= \frac{1}{N} \sum_1^N \frac{(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y} = \frac{1}{N} \sum_1^N \frac{x_i - \mu_x}{\sigma_x} \times \frac{y_i - \mu_y}{\sigma_y} \\ &= \frac{1}{N} \sum_1^N \text{zスコア化された } x_i \times \text{zスコア化された } y_i \\ &\quad \text{Z-scored } x_i \qquad \qquad \qquad \text{Z-scored } y_i \end{aligned}$$

相関係数 r_{xy} は zスコア化された x_i と y_i の共分散

Correlational Coefficient r_{xy} is covariance between z-scored x_i and y_i

正規方程式と相関係数

Normal Equation and Correlational Coefficient

$$X = \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

x_i と y_i が共に zスコア化されているとする
Both x_i and y_i are z-scored
 $\|\boldsymbol{x}\|^2 = N, \mu_x = 0, \sigma_x = 1 \quad \|\boldsymbol{y}\|^2 = N, \mu_y = 0, \sigma_y = 1$

$$\boldsymbol{X}^T \boldsymbol{X} = [x_1 \ x_2 \ \dots \ x_N] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \|\boldsymbol{x}\|^2 = N \quad (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \frac{1}{N}$$

正規方程式と相関係数

Normal Equation and Correlational Coefficient

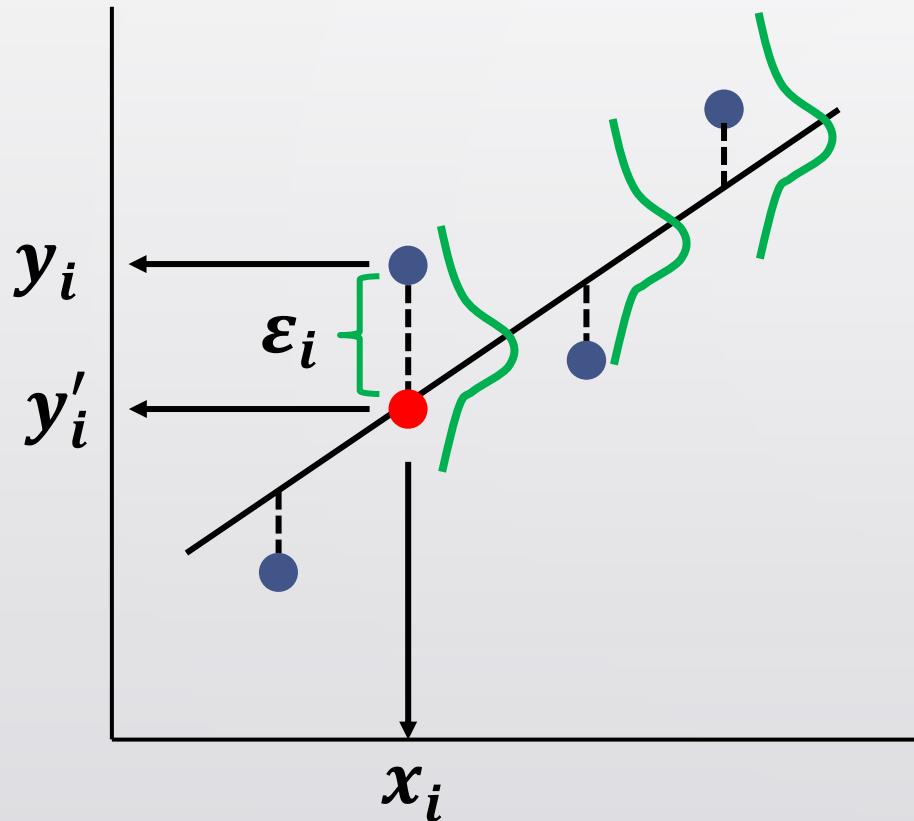
$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \frac{1}{N} [x_1 \ x_2 \ \dots \ x_N] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \frac{1}{N} \left[\frac{x_1 - \mu_x}{\sigma_x} \ \frac{x_2 - \mu_x}{\sigma_x} \ \dots \ \frac{x_N - \mu_x}{\sigma_x} \right] \begin{bmatrix} \frac{y_1 - \mu_y}{\sigma_y} \\ \frac{y_2 - \mu_y}{\sigma_y} \\ \vdots \\ \frac{y_N - \mu_y}{\sigma_y} \end{bmatrix}$$
$$= \frac{1}{N} \sum_1^N \frac{x_i - \mu_x}{\sigma_x} \times \frac{y_i - \mu_y}{\sigma_y} = r_{xy}$$

Zスコア化された x_i と y_i を正規方程式に投入すると、相関係数 r_{xy} が得られる

Correlational coefficient r_{xy} is obtained by entering z-scored x_i and y_i into normal equation

最尤推定による回帰分析

Linear Regression by Maximum Likelihood Estimation



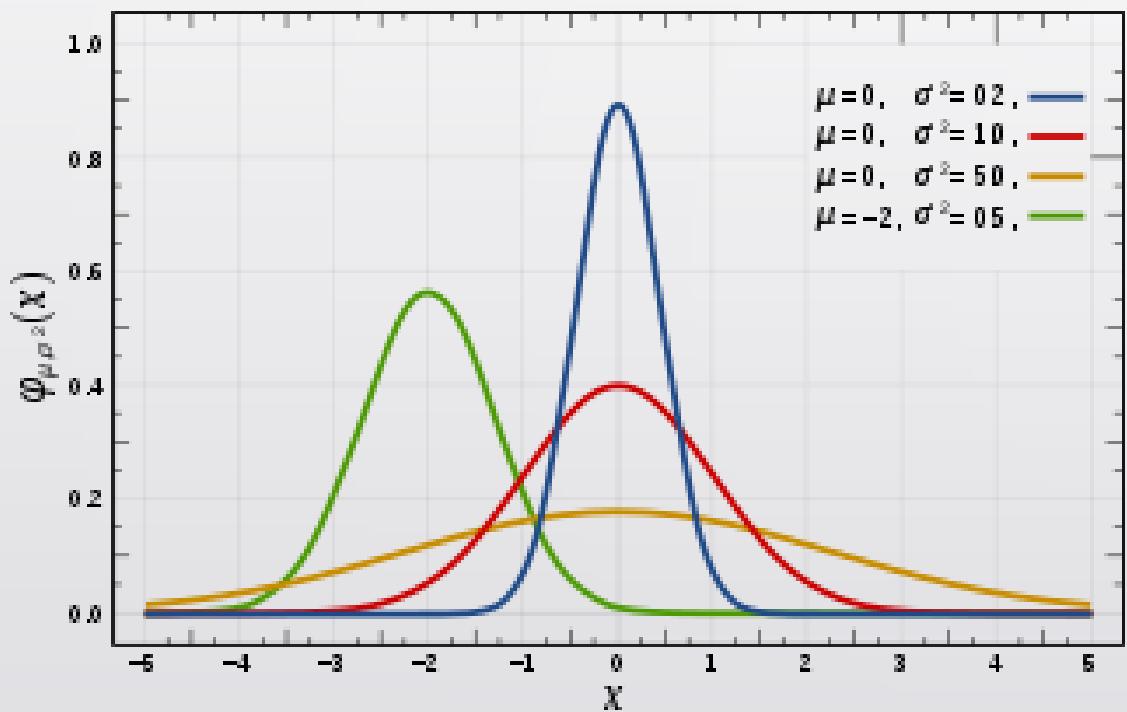
$$y'_i = \beta x_i \quad \varepsilon_i = y_i - y'_i$$

予測誤差が正規分布に従うという前提で、回帰係数 β を推定する

Estimate regression coefficients on the assumption that prediction error conforms to the normal distribution



正規分布 Normal Distribution



確率密度関数 Probability Density Function

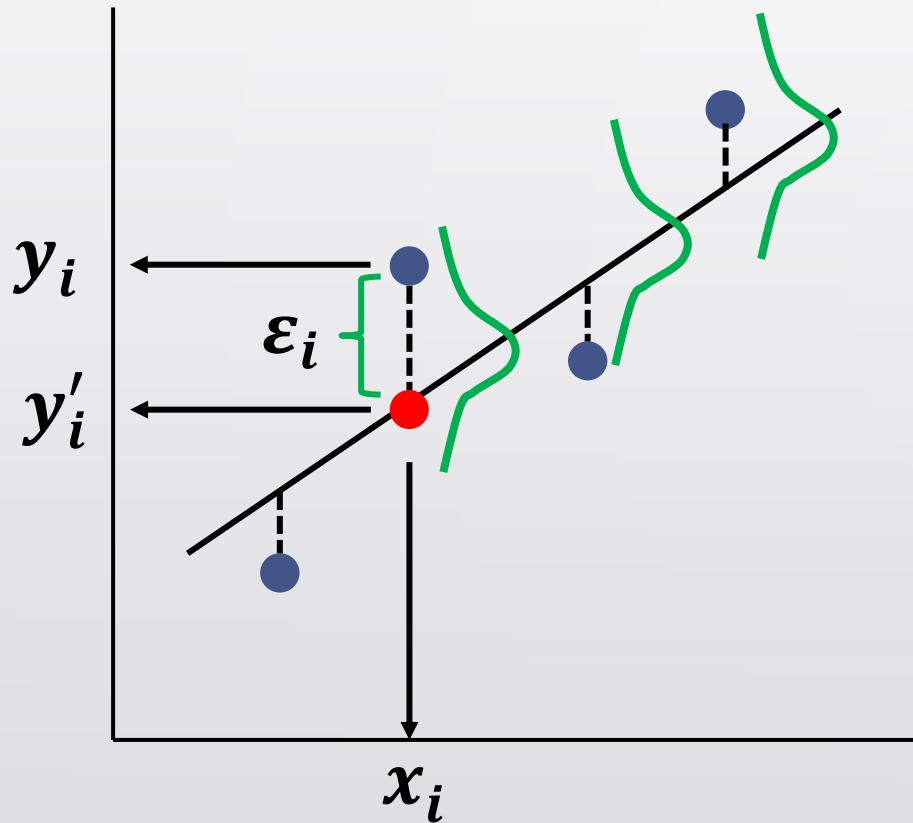
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

$\mu = 0, \sigma = 1$ の時は、標準正規分布

Standard normal distribution when $\mu = 0, \sigma = 1$

最尤推定による回帰分析

Linear Regression by Maximum Likelihood Estimation



$$y'_i = \beta x_i \quad \varepsilon_i = y_i - y'_i$$

$$\varepsilon_i \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - y'_i)^2}{2\sigma^2}\right)$$

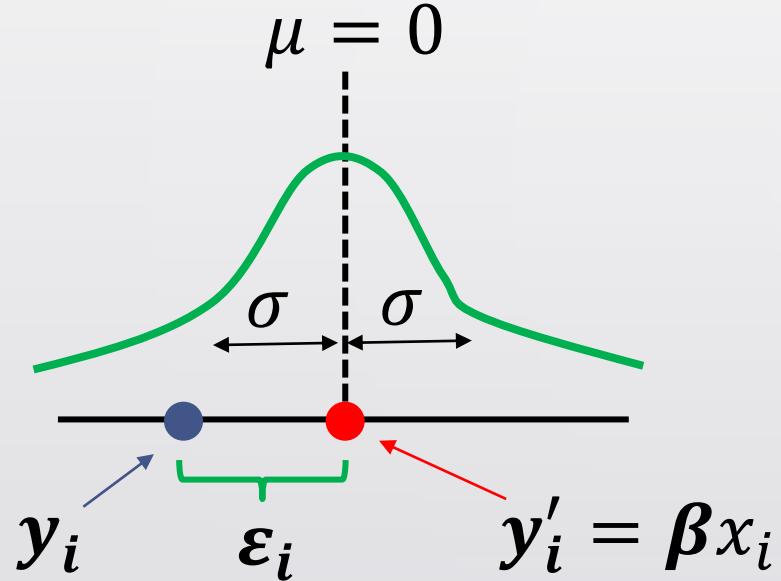
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)$$

誤差 ε_i が0を平均とする正規分布に従うと仮定する

Assume that error ε_i conforms to the normal distribution with mean = 0

最尤推定による回帰分析

Linear Regression by Maximum Likelihood Estimation



$$P(y_i|\boldsymbol{\beta}, \sigma, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{\beta}\mathbf{x}_i)^2}{2\sigma^2}\right)$$

$P(y_i|\boldsymbol{\beta}, \sigma, \mathbf{x}_i)$:

回帰係数が $\boldsymbol{\beta}$ で標準偏差が σ の時、
 x_i に対して、データ y_i が観測される確率

Probability that data y_i is observed for x_i under the condition that regression coefficients are $\boldsymbol{\beta}$ and standard deviation is σ

最尤推定 Maximum Likelihood Estimation

データ列 $\{y_1, y_2 \dots y_{N-1}, y_N\}$ が観測される同時確率は以下のように書ける

The joint probability that data $\{y_1, y_2 \dots y_{N-1}, y_N\}$ is observed can be written as follows

$$\begin{aligned} L &= \prod_1^N P(y_i | \boldsymbol{\beta}, \sigma, x_i) = \prod_1^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{\beta}x_i)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_1 - \boldsymbol{\beta}x_1)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_2 - \boldsymbol{\beta}x_2)^2}{2\sigma^2}\right) \times \dots \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_N - \boldsymbol{\beta}x_N)^2}{2\sigma^2}\right) \end{aligned}$$

最尤推定 Maximum Likelihood Estimation

L は $\{x_1, x_2 \dots x_{N-1}, x_N\}$ に対して $\{y_1, y_2 \dots y_{N-1}, y_N\}$ が観測される同時確率

L is the joint probability that data $\{y_1, y_2 \dots y_{N-1}, y_N\}$ is observed for $\{x_1, x_2 \dots x_{N-1}, x_N\}$

最尤推定では L が最大化されるような β, σ を求める

In maximum likelihood estimation, β, σ are determined so that L is maximized

→ $\log(L)$ を最大化する
Maximize $\log(L)$

最尤推定 Maximum Likelihood Estimation

$$L = \prod_1^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)$$

$$\begin{aligned} \log(L) &= \sum_1^N \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right)\right) \\ &= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_1^N (y_i - \beta x_i)^2 \end{aligned}$$

最尤推定 Maximum Likelihood Estimation

$$\text{Log}(L) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_1^N (y_i - \beta x_i)^2$$

$$\frac{\partial \text{Log}(L)}{\partial \beta} = 0 \quad \text{RSS} = \sum_1^N (y_n - y'_n)^2 \quad y'_i = \beta x_i$$

$\text{Log}(L)$ を最大化する β は, RSS を最小化する

β that maximizes $\text{Log}(L)$ minimizes RSS

2つの重回帰分析 Two Types of Multiple Regressions

最小二乗法

Ordinary Least Squares Method

RSSを最小化 Minimize RSS

$$\beta \downarrow$$

最尤推定

Maximum Likelihood Estimation

誤差 ε が正規分布すると仮定

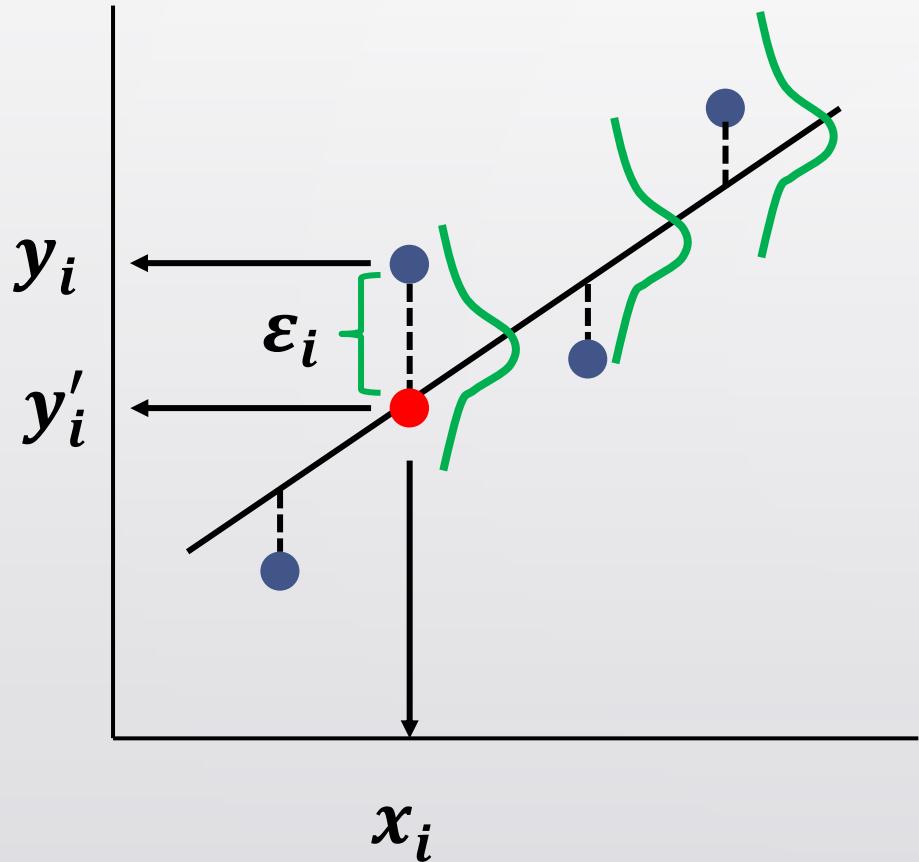
Assume that error ε conforms to
normal distribution

$\text{Log}(L)$ を最大化
Maximize $\text{Log}(L)$

$$\sigma = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2}$$

RSSを最小化 Minimize RSS

一般化線型モデル Generalized Linear Model (GLM)



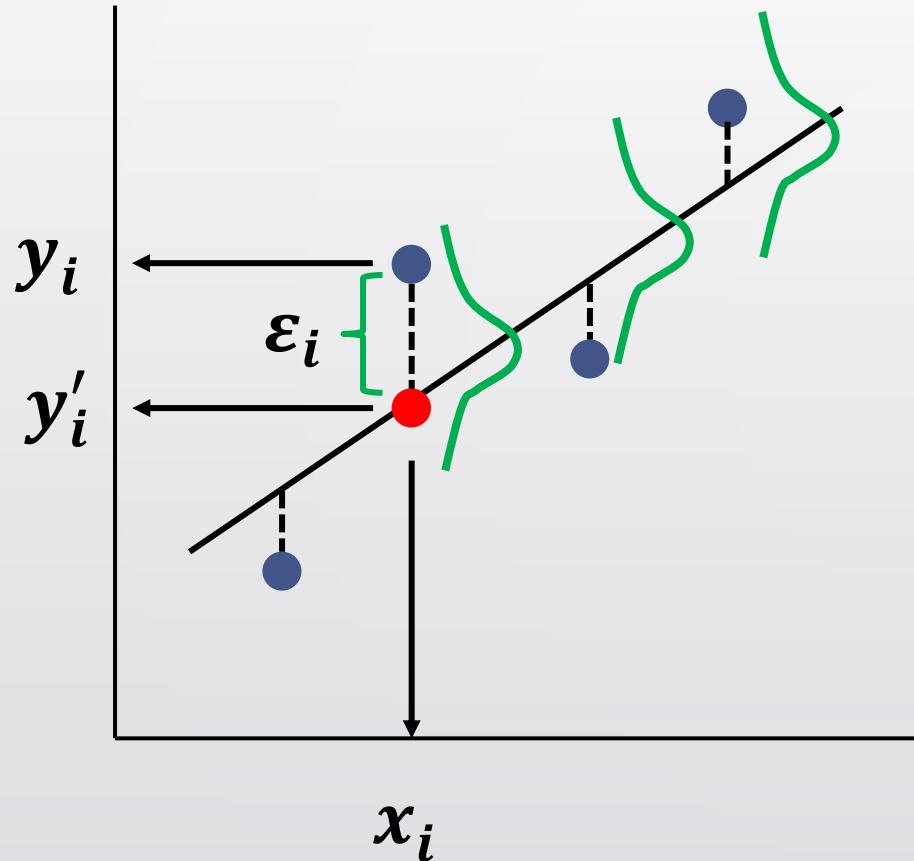
誤差が平均0の正規分布に従う
Error conforms to the normal distribution with $\mu = 0$

y'_i の期待値は βx_i になる
Expected value of y'_i is βx_i

$$g(E[y'_i]) = \beta x_i$$

$$g(\mu) = \mu$$

一般化線型モデル Generalized Linear Model (GLM)



$$g(E[y_i]) = \beta x_i$$

$$g(\mu) = \mu$$

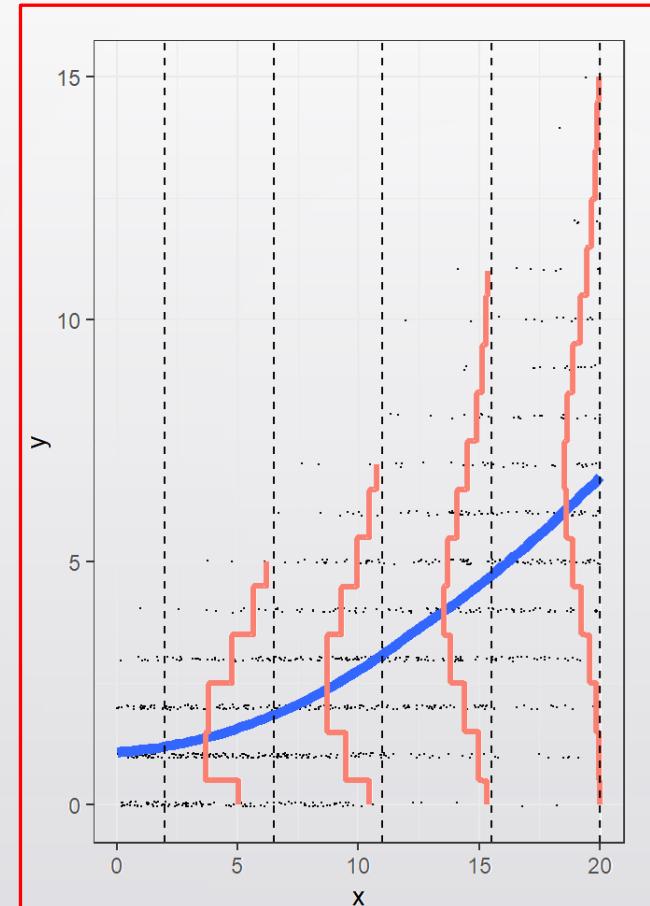
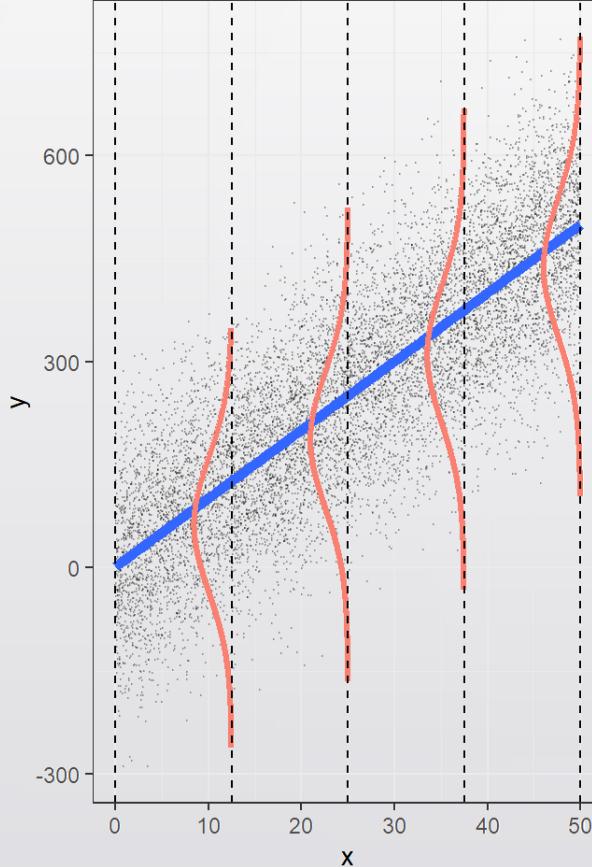
βx_i : 線型予測子 Linear Predictor

g : リンク関数 Link Function

重回帰分析の誤差構造は正規分布である

Error structure of multiple linear regression is normal distribution

ポアソン回帰 Poisson Regression



x_i が大きくなる程, y_i の期待値・分散が大きくなる

As x_i gets larger, so do expected value and variance of y_i

$$g(E[y_i]) = \beta x_i$$

$$g(\mu) = \log(\mu)$$

$$E[y_i] = V[y_i] = e^{\beta x_i}$$



データマイニング

Data Mining

5: 回帰② Regression

土居 裕和 Hirokazu Doi

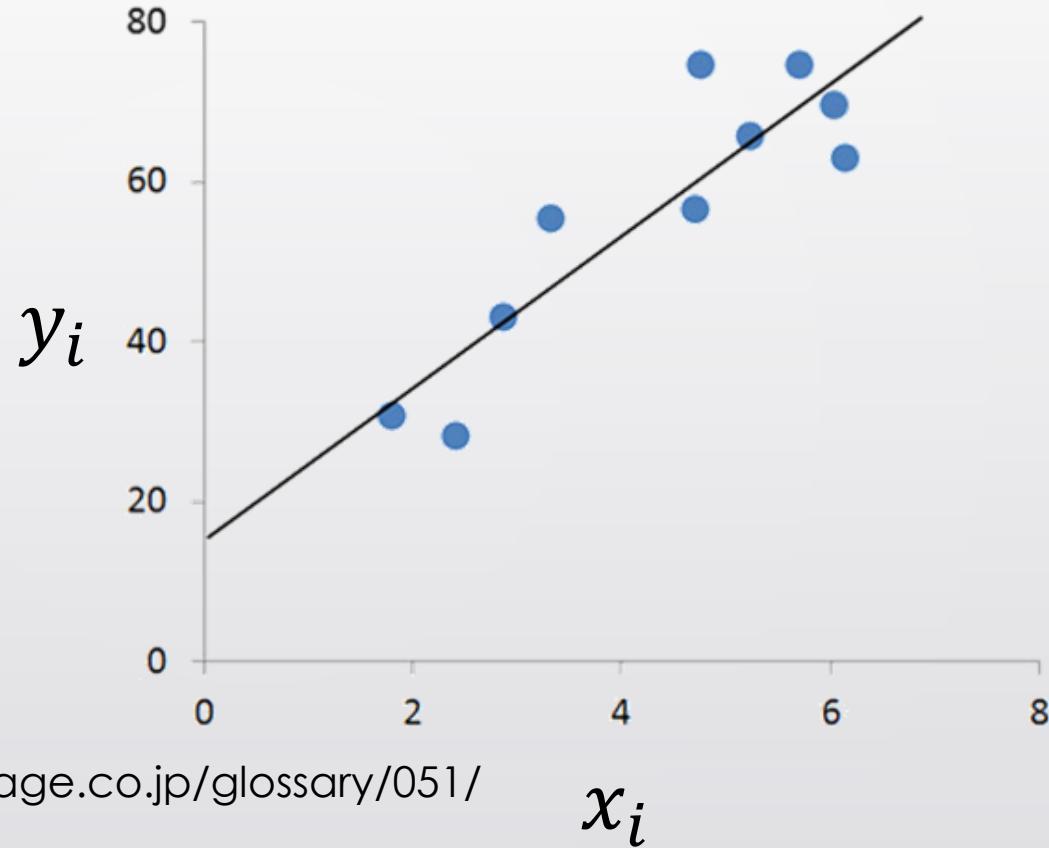
長岡技術科学大学 Nagaoka University of Technology

回帰 Regression

(x_i, y_i)

x_i : 家族の人数
Number of Family Member

y_i : 購入数
Number of purchased Items



重回帰分析 Multiple Linear Regression

複数の変数の線型和によって、ターゲット変数を予測

Predicts target variable by linear combination of multiple variables

Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

<https://www.i2tutorials.com/difference-between-simple-linear-regression-and-multi-linear-regression-and-polynomial-regression/>



最小二乗法 Ordinary Least Squares (OLS) Method

\mathbf{X} から \mathbf{y} を精度よく予測できる重回帰モデルの $\boldsymbol{\beta}$ を求める

Find $\boldsymbol{\beta}$ so that multiple regression model can predict \mathbf{y} based on \mathbf{X} with good precision

$$\mathbf{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

重回帰モデルによる \mathbf{y} の推定値 Prediction of \mathbf{y} based on multiple regression

観測値 \mathbf{y} と予測値 \mathbf{y}' との差を最小化する $\boldsymbol{\beta}$ を求める

Find $\boldsymbol{\beta}$ that minimizes the difference between observed vector \mathbf{y} and predicted vector \mathbf{y}'

2つの重回帰分析 Two Types of Multiple Regressions

最小二乗法

Ordinary Least Squares Method

RSSを最小化 Minimize RSS

$$\beta \downarrow$$

最尤推定

Maximum Likelihood Estimation

誤差 ε が正規分布すると仮定

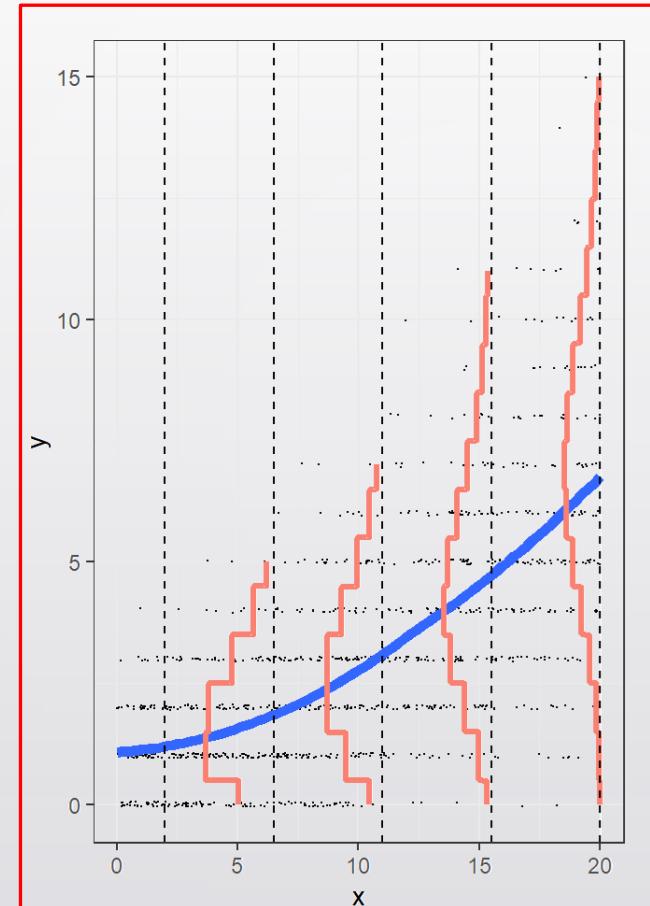
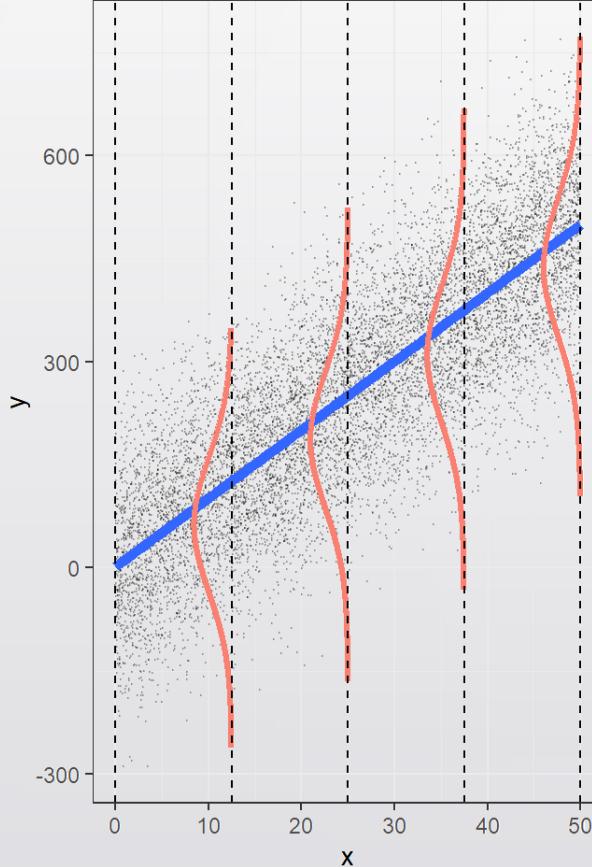
Assume that error ε conforms to
normal distribution

$\log(L)$ を最大化
Maximize $\log(L)$

$$\sigma = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2}$$

RSSを最小化 Minimize RSS

ポアソン回帰 Poisson Regression



x_i が大きくなる程, y_i の期待値・分散が大きくなる

As x_i gets larger, so do expected value and variance of y_i

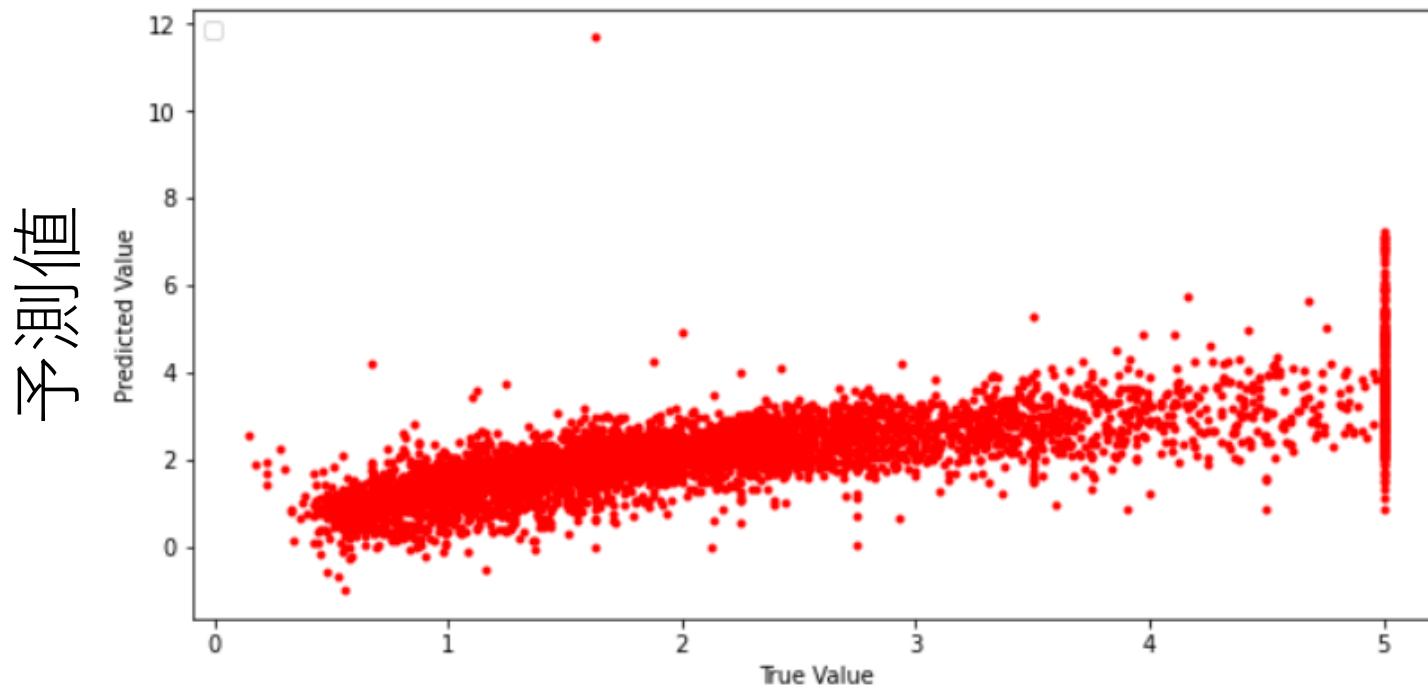
$$g(E[y_i]) = \beta x_i$$

$$g(\mu) = \log(\mu)$$

$$E[y_i] = V[y_i] = e^{\beta x_i}$$

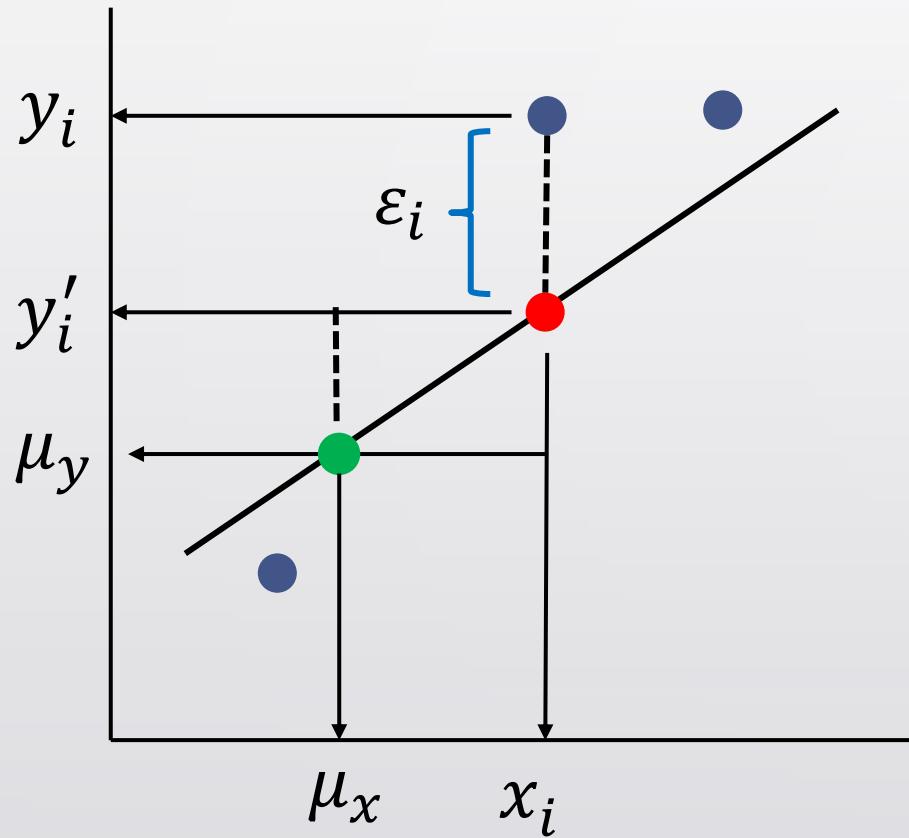
相関係数による性能評価

```
Out[33]: array([[1.          , 0.76907401],  
                 [0.76907401, 1.         ]])
```



正解値

決定係数 R^2 Coefficient of Determination



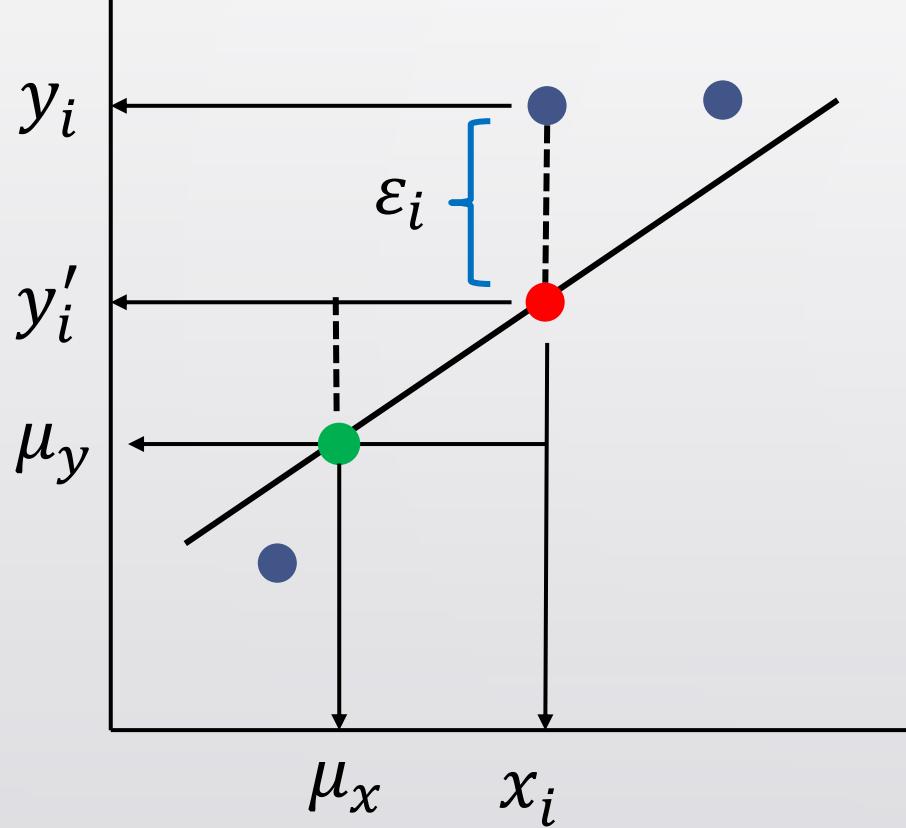
$$S_{total} = \sum_{i=0}^N (y_i - \mu_y)^2$$

$$S_{model} = \sum_{i=0}^N (y'_i - \mu_y)^2$$

$$RSS = \sum_{i=0}^N (\varepsilon_i)^2 = \sum_{i=0}^N (y_i - y'_i)^2$$

残差二乗和 Squared Sum of Residuals

決定係数 R^2 Coefficient of Determination



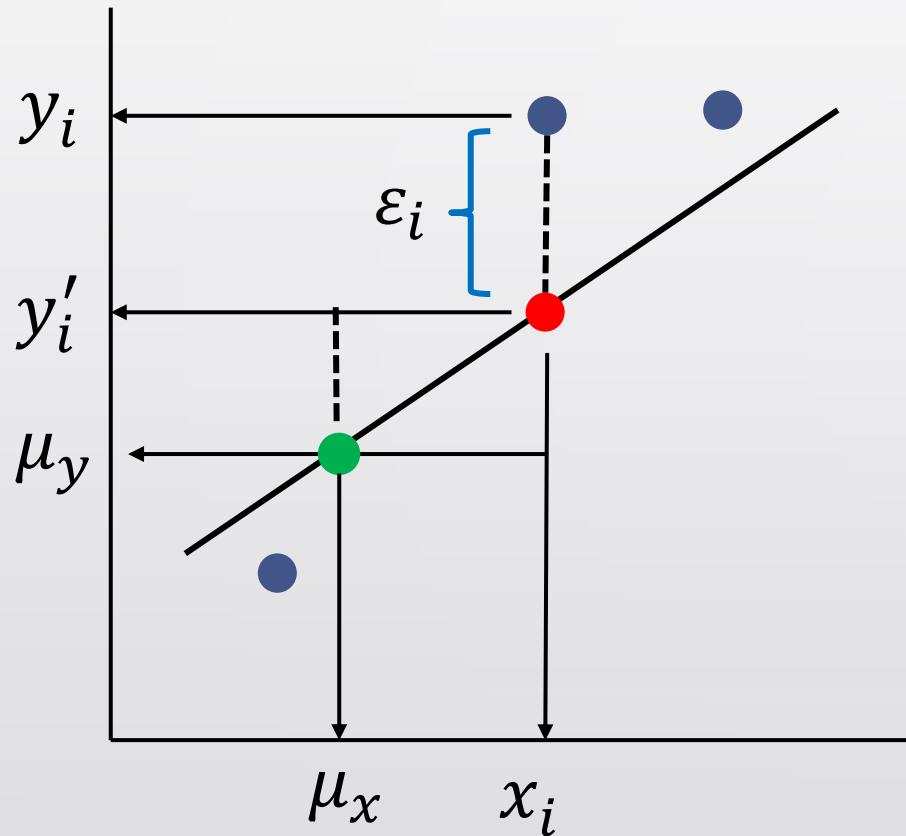
$$y_i - \mu_y = (y'_i - \mu_y) + \varepsilon_i$$

誤差 ε_i が平均0の正規分布に従うとき

When error ε_i conforms to the normal distribution $N(0, \sigma^2)$

$$S_{total} = S_{model} + RSS$$

決定係数 R^2 Coefficient of Determination



$$S_{total} = S_{model} + RSS$$

$$R^2 = 1 - \frac{RSS}{S_{total}} = \frac{S_{model}}{S_{total}}$$

単回帰の場合、決定係数は相關係数の二乗に一致する

In the case of simple regression, coefficient of determination equals to squared coefficient of regression

決定係数 R^2 Coefficient of Determination

調整可能なモデルパラメータが多い程, 決定係数は大きくなる

The larger the number of adjustable model parameters, the larger the coefficient of determination becomes

重回帰分析の場合は、自由度調整済み決定係数を用いる

Adjusted coefficient of determination as described below is used in the case of multiple regression

$$Adjusted R^2 = 1 - \frac{RSS}{S_{total}} \frac{n - 1}{n - p - 1}$$

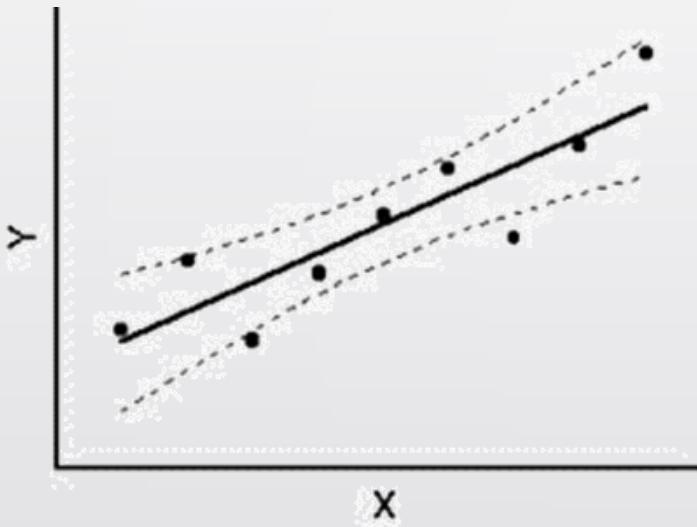
n : サンプル数
Sample Size

p : 予測変数の数
The number of predictors

信頼区間 Confidence Interval

信頼区間は、回帰モデルの不確実性を表す

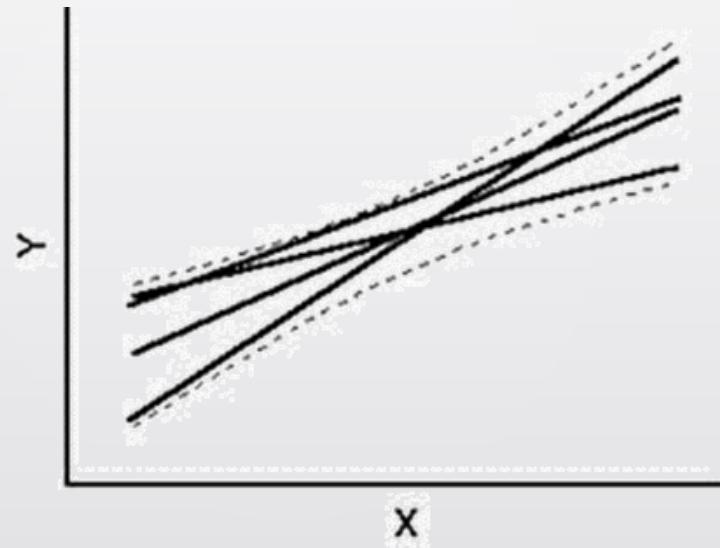
Confidence interval represents uncertainty about regression model



点線は95%信頼区間
Dotted lines represent 95%
confidence interval



回帰直線は95%の確率で点線内のどこかに引かれる
Regression line falls somewhere inside the region defined by dotted
lines with 95% probability

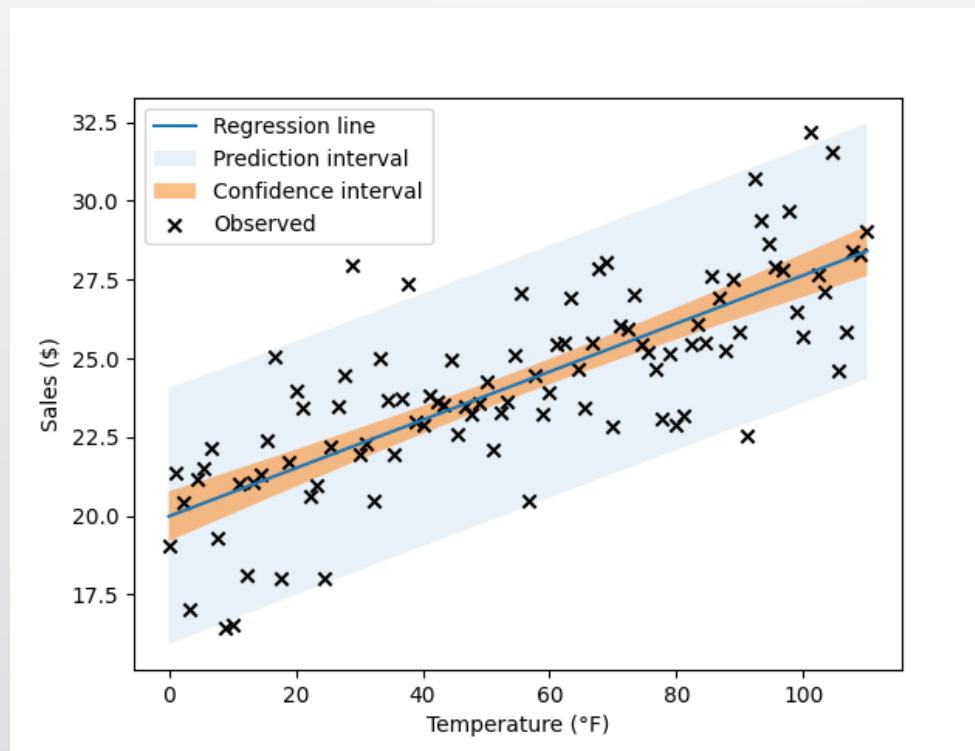


<https://real-statistics.com/regression/confidence-and-prediction-intervals/>

予測区間 Prediction Interval

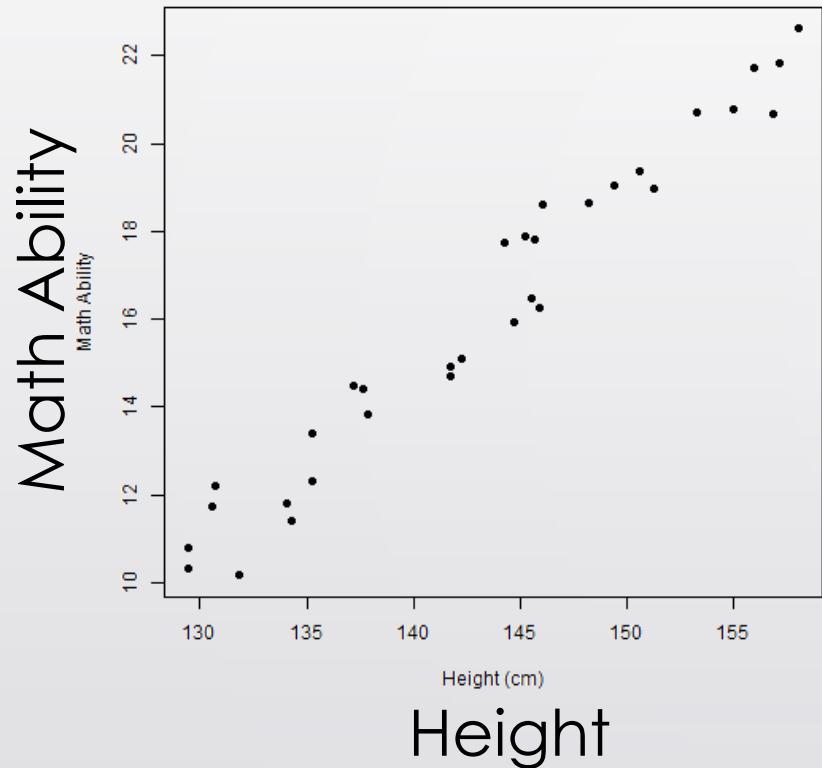
予測区間は、 y の予測値 y' の不確実性を表す

Prediction interval represents uncertainty about estimation of y based on given x

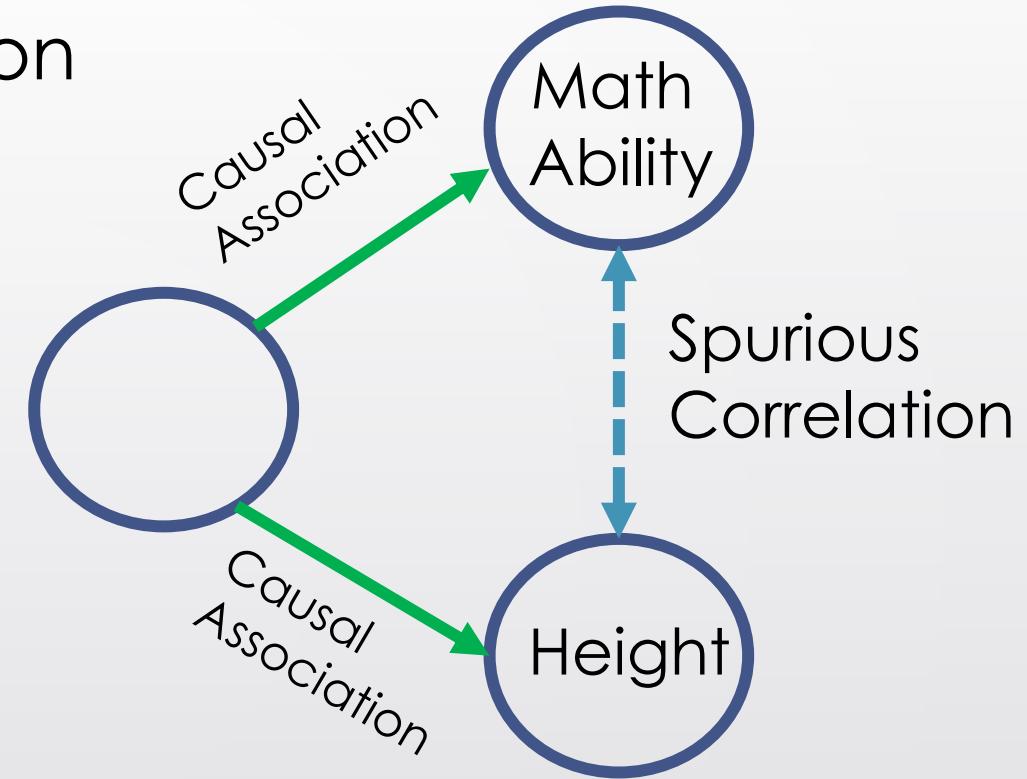


https://lmc2179.github.io/posts/confidence_prediction.html

疑似相関 Spurious Correlation



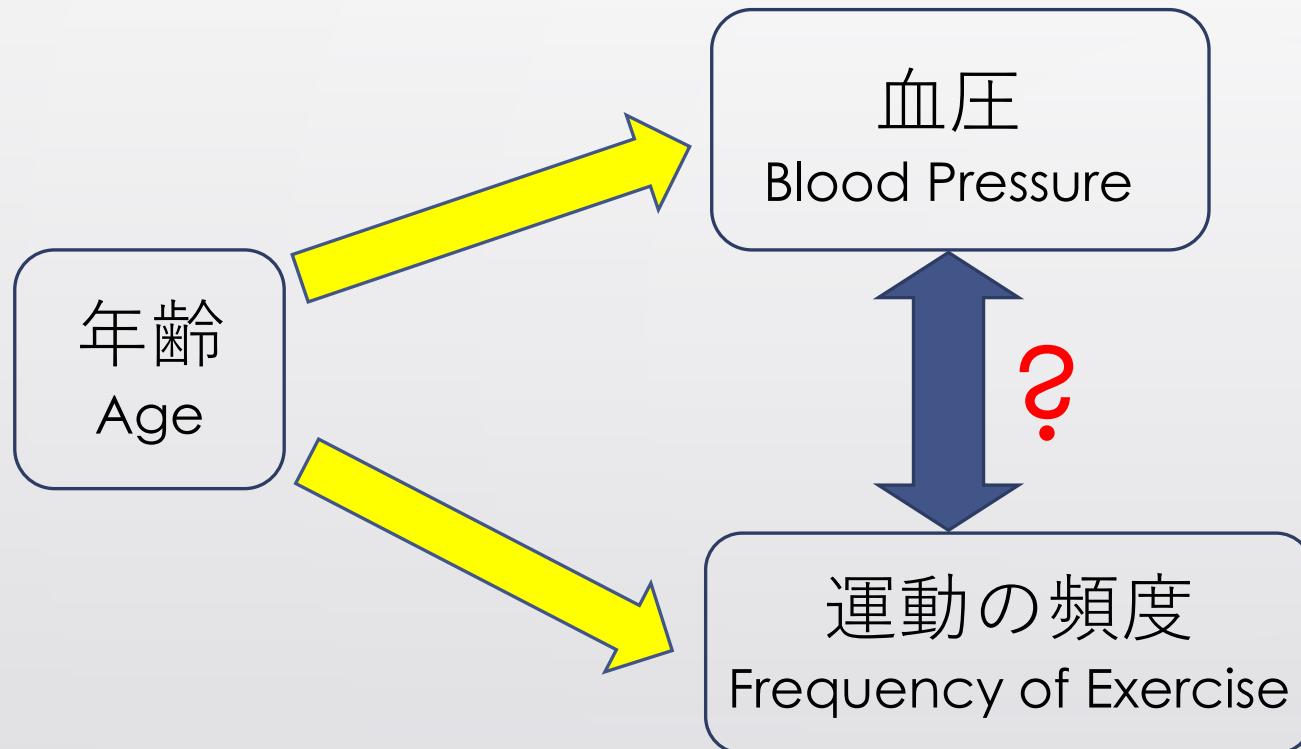
<https://hoxo-m.hatenablog.com/entry/20130711/p1>



データ分析では、見かけの関係性に注意

Be aware of spurious association in any kinds of data analysis

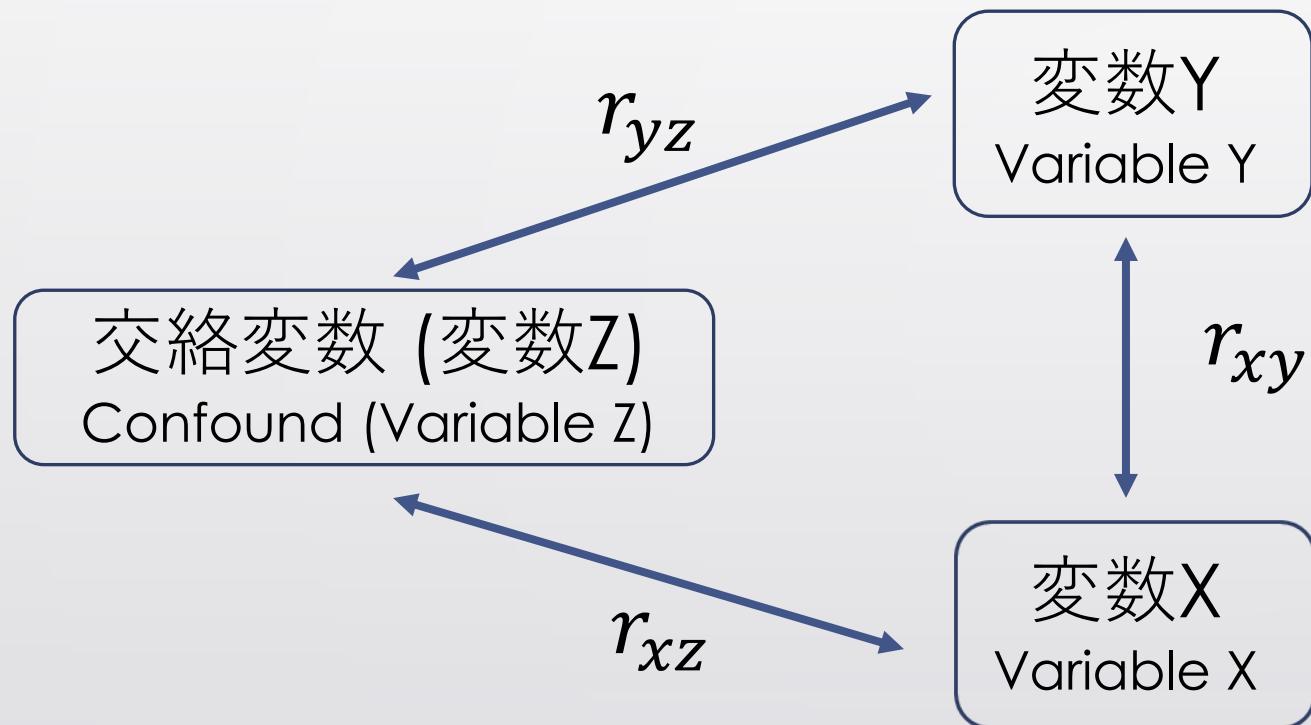
交絡変数 Confound



血圧と運動の頻度の関係には、年齢が交絡している

Association between blood pressure and exercise frequency is confounded by age

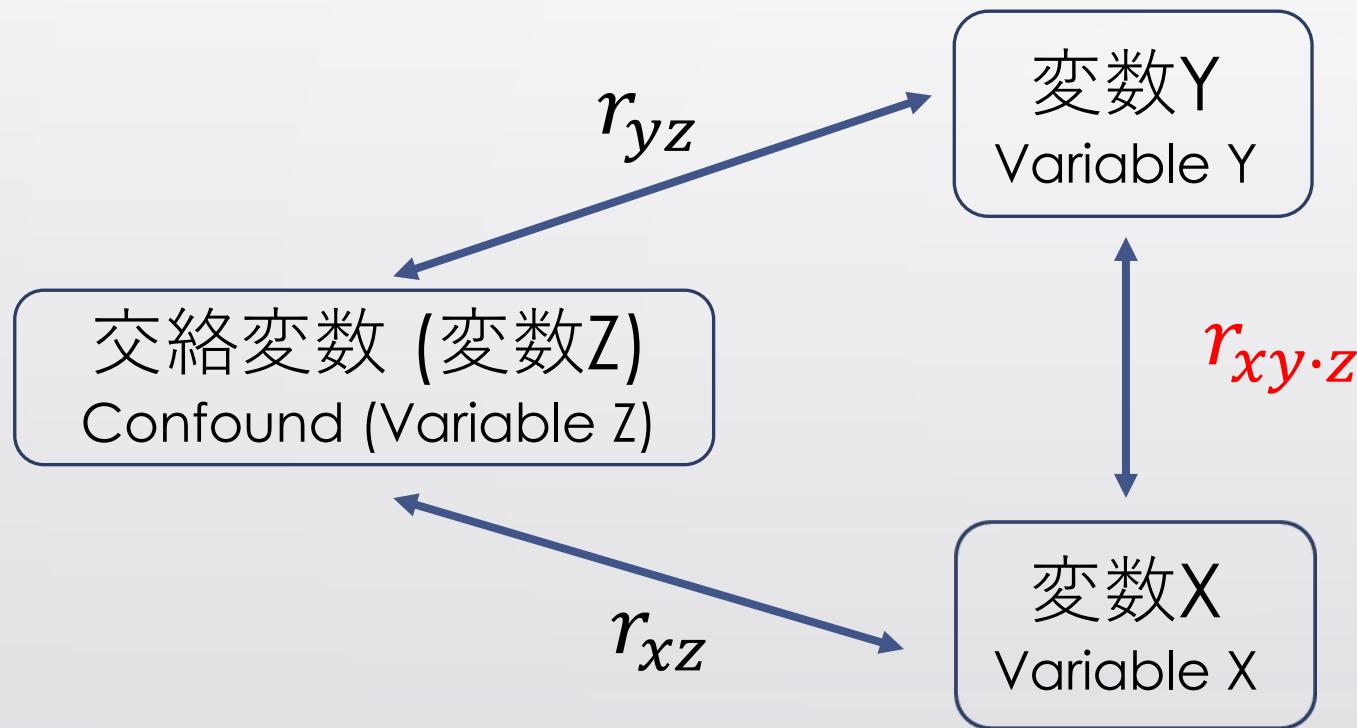
交絡変数 Confound



変数xとyの相関を分析するには、
変数zの影響を除く必要がある

To analyze the association between
variable x and y, the influence (confound)
of variable z should be eliminated

交絡変数 Confound



$$y'_z = Z\beta_{yz}$$

y'_z は変数 **Z** で説明できる **Y** の情報

y'_z is information of **Y** explainable by **Z**

$$x'_z = Z\beta_{xz}$$

x'_z は変数 **Z** で説明できる **X** の情報

x'_z is information of **X** explainable by **Z**



偏回帰係数 Partial Correlation Coefficient

$y - y'_z$: 変数 Z で説明できない Y の情報 Information of Y unexplainable by Z

$x - x'_z$: 変数 Z で説明できない X の情報 Information of X unexplainable by Z

X と Y の偏相関係数 $r_{xy \cdot z}$ は、 $x - x'_z$ と $y - y'_z$ の相関係数

Partial correlational coefficient $r_{xy \cdot z}$ between X and Y is correlational coefficient between $x - x'_z$ and $y - y'_z$

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

正規方程式 Normal Equation

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$X^T X \beta = X^T y \quad \text{← 正規方程式 Normal Equation}$$

$$\beta = (X^T X)^{-1} X^T y$$

多重共線性 Multicollinearity

予測変数の間に強い相関があると、回帰モデルが不安定化する

Regression model becomes unstable when there is strong correlation among predictors

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$



強い相関がある
Strongly correlates with each other

多重共線性 Multicollinearity

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

極端なケース Extreme Case

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 6 \\ 1 & 2 & 1 \end{bmatrix} \quad \det \mathbf{X}^T \mathbf{X} = 0$$

行列式が0なので $\mathbf{X}^T \mathbf{X}$ の逆行列がない
Cannot find inverse matrix of $\mathbf{X}^T \mathbf{X}$ because its determinant is zero



多重共線性 Multicollinearity

現実的な例 More realistic example

$$X_1 = \begin{bmatrix} -1.12 & -0.51 & 0.69 \\ -0.43 & -1.12 & 1.02 \\ 0.37 & 1.10 & -0.98 \\ 1.19 & 0.53 & -0.73 \end{bmatrix} \quad X_2 = \begin{bmatrix} -1.12 & -0.51 & 0.70 \\ -0.43 & -1.12 & 1.02 \\ 0.36 & 1.10 & -0.98 \\ 1.20 & 0.53 & -0.73 \end{bmatrix}$$



とてもよく似た行列 Quite similar matrices

$$0.25 \times (1\text{列目}) - 0.8 \times (2\text{列目}) = (3\text{列目})$$

多重共線性 Multicollinearity

$$y = \begin{bmatrix} 0.40 \\ 1.17 \\ -1.14 \\ -0.42 \end{bmatrix} \quad \boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \boldsymbol{\beta}_1 = \begin{bmatrix} 0.54 \\ 0.24 \\ 1.64 \end{bmatrix} \quad \boldsymbol{\beta}_2 = \begin{bmatrix} -0.42 \\ -2.86 \\ -2.20 \end{bmatrix}$$

データ \mathbf{X}_1 は \mathbf{X}_2 よく似ているのに、回帰係数 $\boldsymbol{\beta}_1$ と $\boldsymbol{\beta}_2$ は全く異なる

Correlational coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are completely different from each other whereas data matrices \mathbf{X}_1 and \mathbf{X}_2 are quite similar

多重共線性があると、データのわずかな違いで、回帰分析の結果が大きく変化する

When there is multicollinearity, slight difference in data results in great change in regression result



分散拡大係数 Variance Inflation Factor (VIF)

VIFをチェックすることで、多重共線性が生じていないかを確認する

See if there is multicollinearity by checking VIF

$$x'_j = \begin{bmatrix} x'_{1,j} \\ x'_{2,j} \\ \vdots \\ x'_{N,j} \end{bmatrix} = \begin{bmatrix} x_{1,1} \dots x_{1,j-1} & \color{red}{x_{1,j+1}} \dots x_{1,M} \\ x_{2,1} \dots x_{2,j-1} & \color{red}{x_{1,j+1}} \dots x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} \dots x_{N,j-1} & \color{red}{x_{N,j+1}} \dots x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \color{red}{\beta_{j-1}} \\ \color{red}{\beta_{j+1}} \\ \vdots \\ \beta_M \end{bmatrix} \quad x_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{N,j} \end{bmatrix}$$

$R_j^2 = x_j$ と x'_j の決定係数



分散拡大係数 Variance Inflation Factor (VIF)

$R_j^2 = x_j$ と x'_j の決定係数

$$VIF_j = \frac{1}{1 - R_j^2}$$

$VIF_j > 10$ を多重共線性の基準とすることが多い

$VIF_j > 10$ is usually taken as threshold of multicollinearity

第k主成分 k -th Principal Component

$$\mathbf{t}_k = \begin{bmatrix} t_{1,k} \\ t_{2,k} \\ \vdots \\ t_{N,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} p_{1,k} \\ p_{2,k} \\ \vdots \\ p_{M,k} \end{bmatrix} = \mathbf{X}\mathbf{p}_k$$

主成分同士は直交している Principal components are orthogonal

$$\mathbf{p}_i^T \mathbf{p}_j = \begin{cases} 1(i = j) \\ 0(i \neq j) \end{cases}$$

主成分回帰 Principal Component Regression

主成分回帰では、予測変数を主成分分析にかけたのち、主成分得点を回帰分析にかける

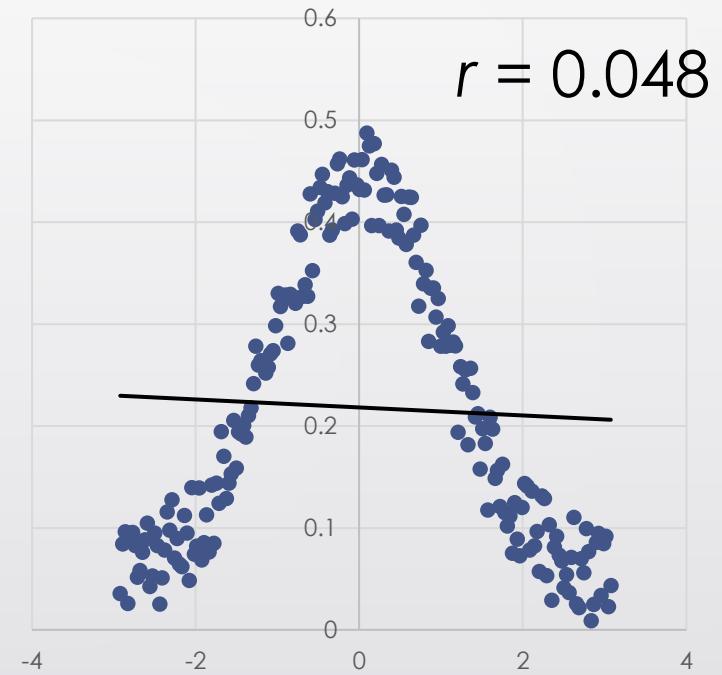
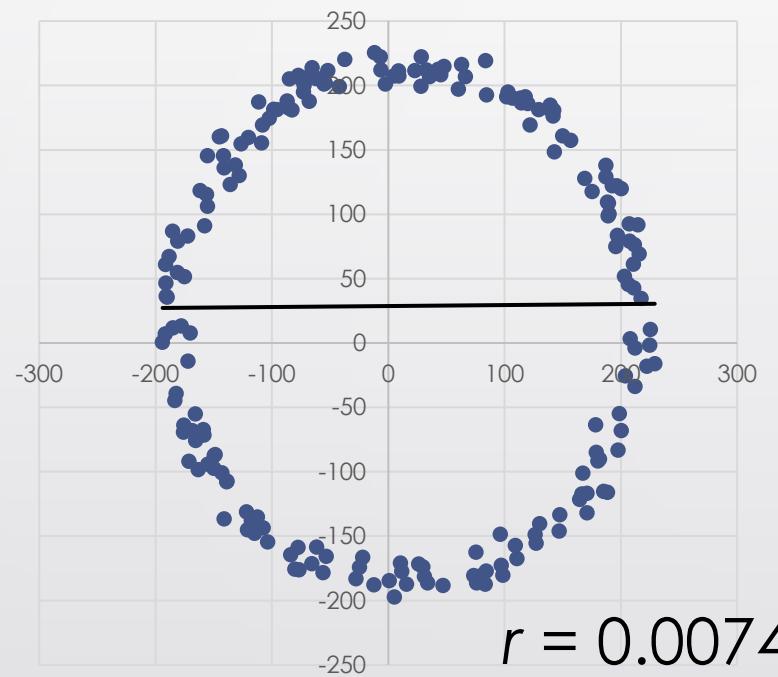
In principal component regression, factor scores are entered in to multiple regression after predictors are submitted to PCA

$$T' = \underbrace{[\mathbf{t}_1 \ \mathbf{t}_2 \ \dots \ \mathbf{t}_H]}_{\begin{array}{l} \text{(N, H)} \\ \text{主成分得点} \\ \text{Factor Scores} \end{array}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{(N, M)} \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,H} \\ p_{2,1} & p_{2,2} & \dots & p_{2,H} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,H} \end{bmatrix}_{(M, H)} = X[\mathbf{p}_1 \ \mathbf{p}_2 \ \dots \ \mathbf{p}_H]$$

↓

重回帰分析

非線形的な関係性 Nonlinear Association



変数間の関係は、適切なモデルで評価する必要がある
Association between variables should be estimated by appropriate model

多項式回帰 Polynomial Regression

Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

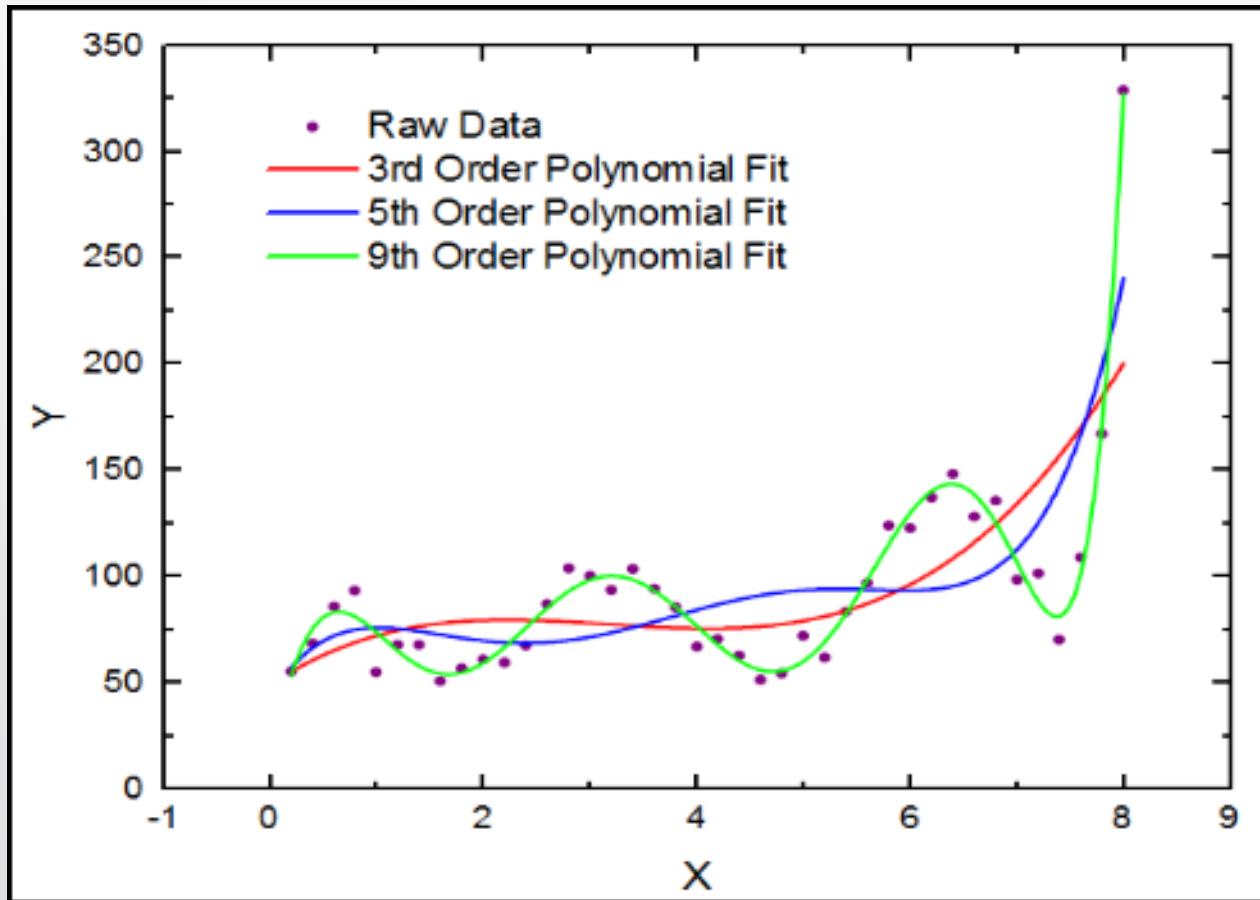
Multiple
Linear
Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Polynomial
Linear
Regression

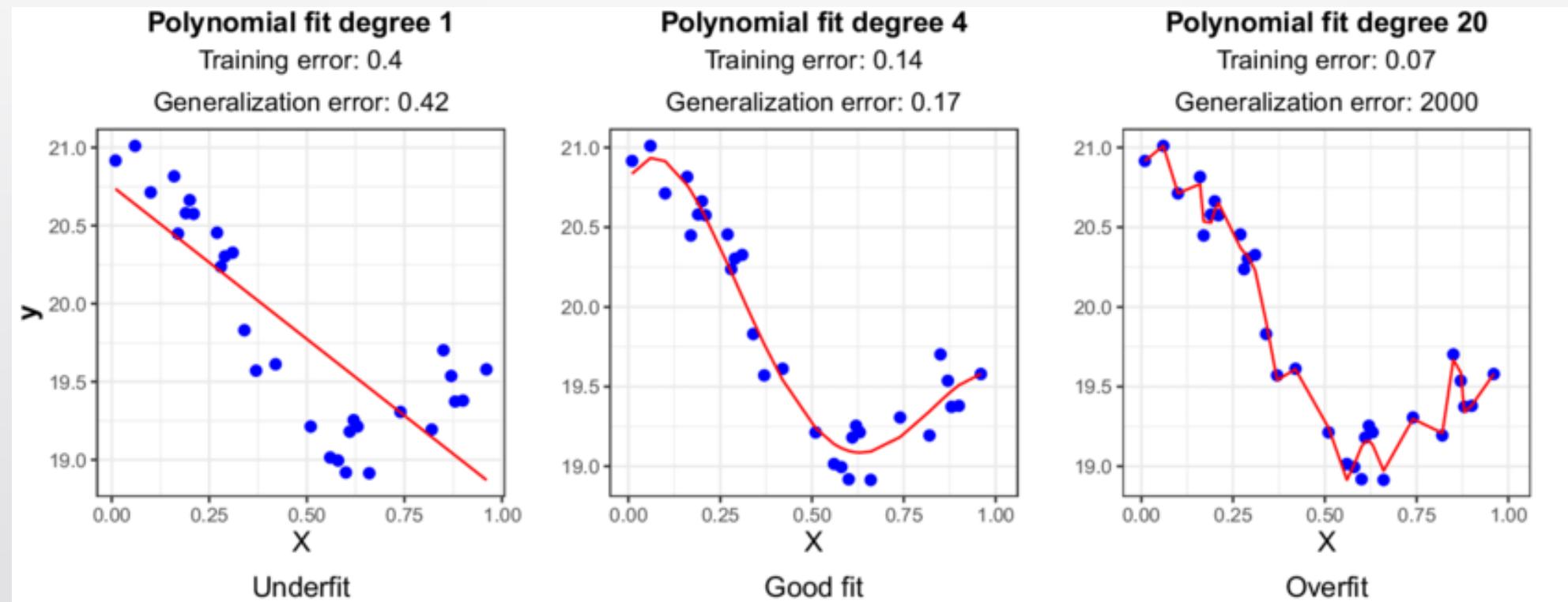
$$y = b_0 + b_1 x_1 + \underline{b_2 x_1^2} + \dots + b_n x_1^n$$

多項式回帰 Polynomial Regression



<https://towardsdatascience.com/polynomial-regression-an-alternative-for-neural-networks-c4bd30fa6cf6>

過学習 Overfitting

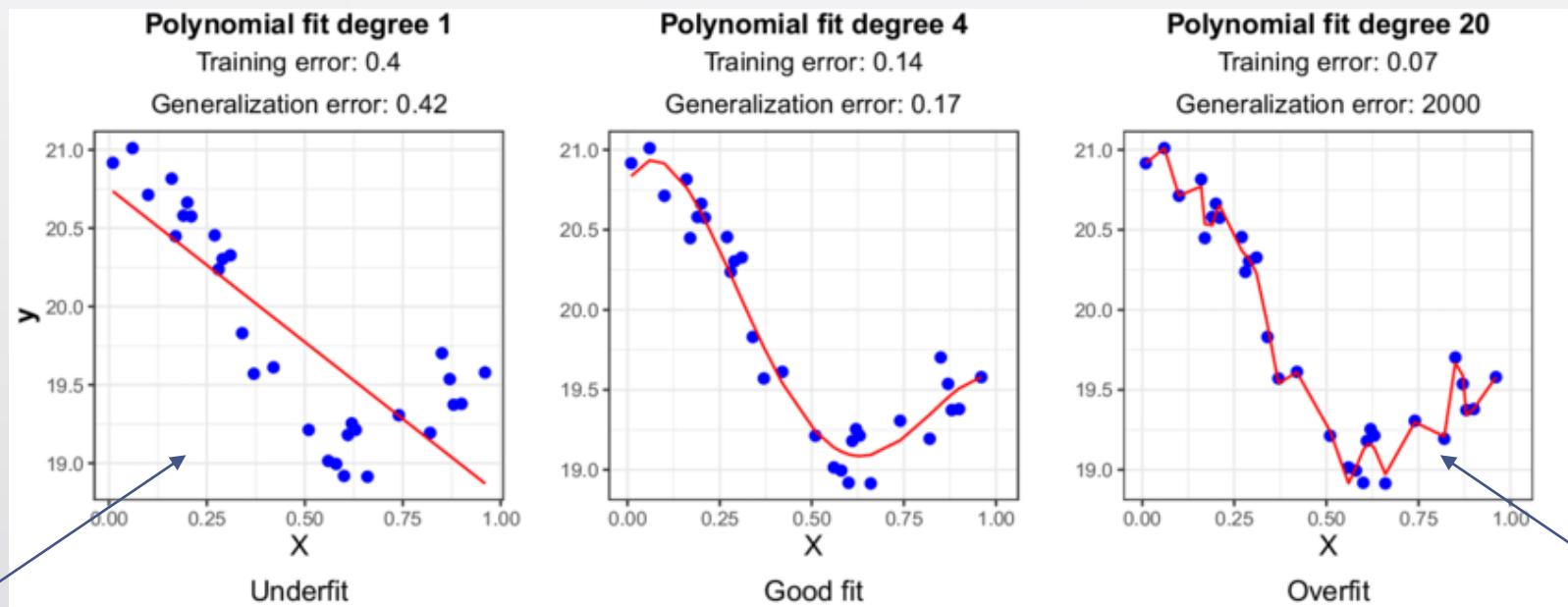


Badillo et al, 2020

良い回帰モデルとは？

What is good regression model?

観測されたデータはノイズを含む Observed data contains random noise



単純すぎてはダメ
Should not be too simplistic

複雑すぎてはダメ
Should not be too complex



赤池情報量基準 Akaike Information Criterion (AIC)

$$AIC = -2\ln(L) + 2k$$

L : 最大尤度 Maximum Likelihood

モデルのもとで、実際に観測されたデータが得られるもっともらしさ
(≈モデルのデータへの当てはまりの良さ)

Likelihood that actually-observed dataset is obtained under the model
(≈ Goodness of fit of the model to the data)

k : モデルのパラメータ数

Number of parameters in the model



赤池情報量規準 Akaike Information Criterion (AIC)

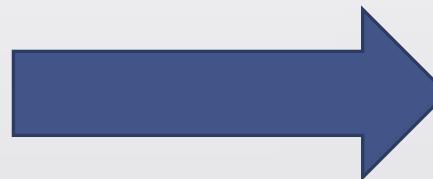
$$AIC = -2\ln(L) + 2k$$

パラメータの数が少ない

The smaller the number of parameters

データへの当てはまりがいい

The better the model fits to the data



AICが小さくなる

The smaller AIC gets



変数選択 Feature Selection

$$AIC = -2\ln(L) + 2k$$

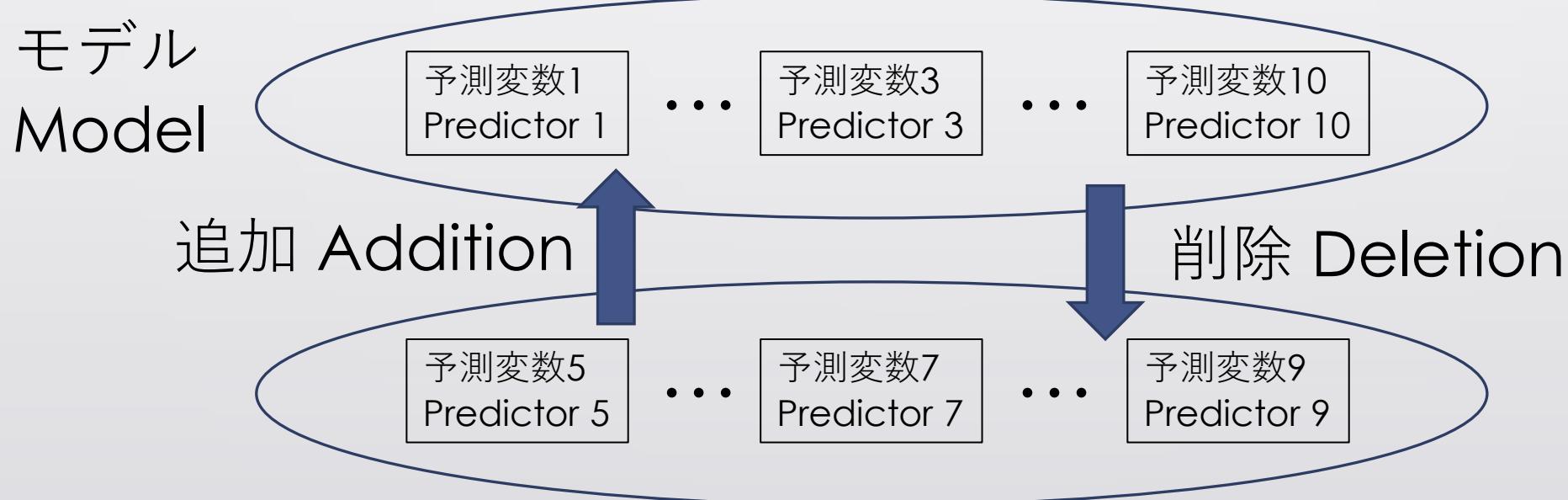
良い回帰モデルは、 AICが小さい
AIC of good regression model is small

AICを基準としてモデルに含める予測変数を選択する
Select predictor variables to be included in the regression model by
using AIC as criteria

ステップワイズ法 Stepwise feature Selection

予測変数を追加したり、削除したりしながら、AICを最小にする予測変数の組み合わせを探す

Find the best combination of predictors that minimizes AIC by adding and deleting variables from the model





ベイズ情報量規準

Bayesian Information Criterion (BIC)

$$AIC = -2\ln(L) + 2k$$

$$BIC = -2\ln(L) + 2k\ln(N)$$

N : サンプルサイズ Sample Size

正則化回帰モデル Regularized Regression Model

$$\mathbf{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_M \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}$$

残差二乗和 Residual Sum of Squares (RSS)
 $RSS = (\mathbf{y} - \mathbf{y}')^T(\mathbf{y} - \mathbf{y}')$

正則化回帰モデルでは、 RSSに正則化項を加えることで、過学習を抑制する
Regularized regression model suppresses overfitting by adding regularization term to RSS

$$L = RSS + \text{Regularization term} = (\mathbf{y} - \mathbf{y}')^T(\mathbf{y} - \mathbf{y}') + \text{Regularization term}$$

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0$$

リッジ回帰 Ridge Regression

RSSを最小にする β
 β minimizing RSS

$\|\beta\|_2 \leq t$ という制約のもとで RSS を最小化する β を見つける

Find β that minimizes RSS under the constraint $\|\beta\|_2 \leq t$

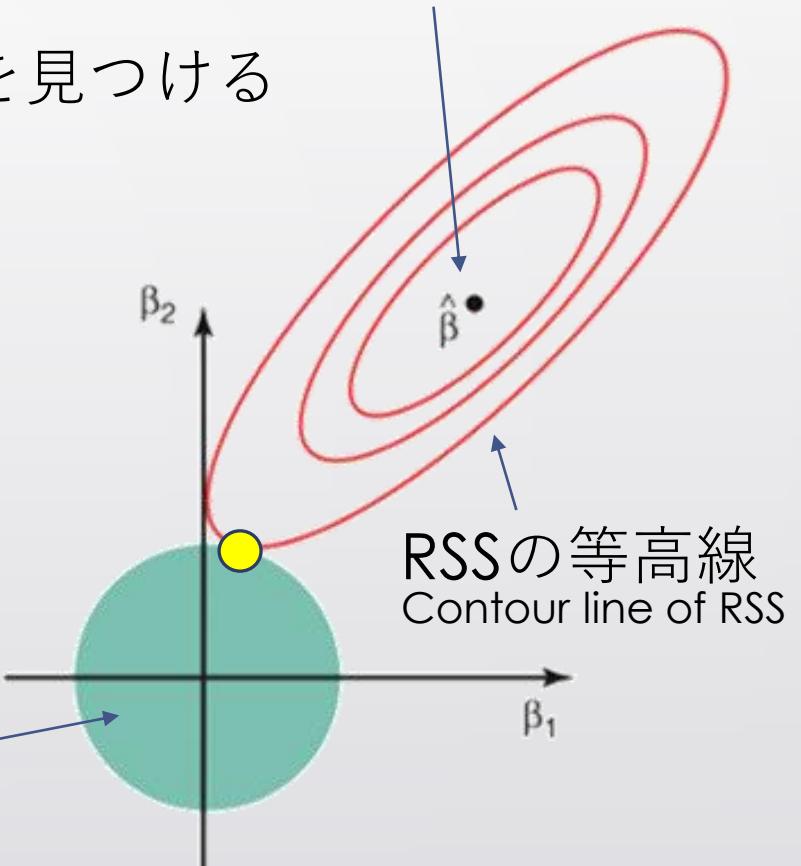
$$L = (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}') + \alpha \|\beta\|_2$$

L2ノルム
L2 Norm

回帰係数の絶対値が小さくなる

Absolute value of regression coefficients tend to get smaller

この領域内の β は制約を満たす
 β inside this region satisfies the constraint



Lasso回帰 Lasso Regression

$\|\beta\|_1 \leq t$ という制約のもとで RSS を最小化する β を見つける

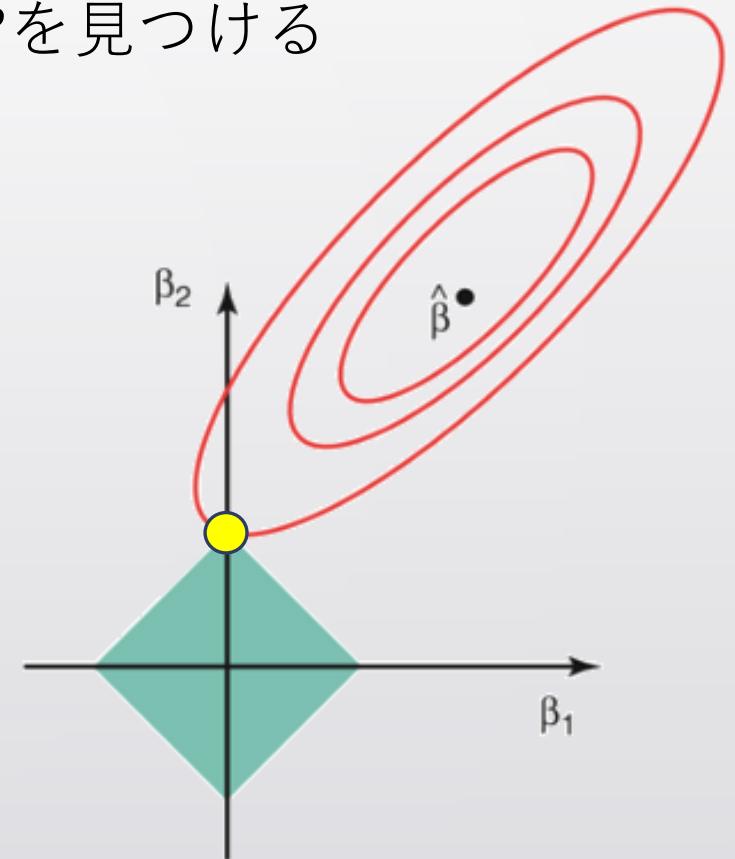
Find β that minimizes RSS under the constraint $\|\beta\|_1 \leq t$

$$L = (\mathbf{y} - \mathbf{y}')^T(\mathbf{y} - \mathbf{y}') + \alpha \|\beta\|_1$$
$$\|\beta\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_n| \quad \begin{matrix} \text{L1ノルム} \\ \text{L1 Norm} \end{matrix}$$

影響が小さな変数の回帰係数は0になる

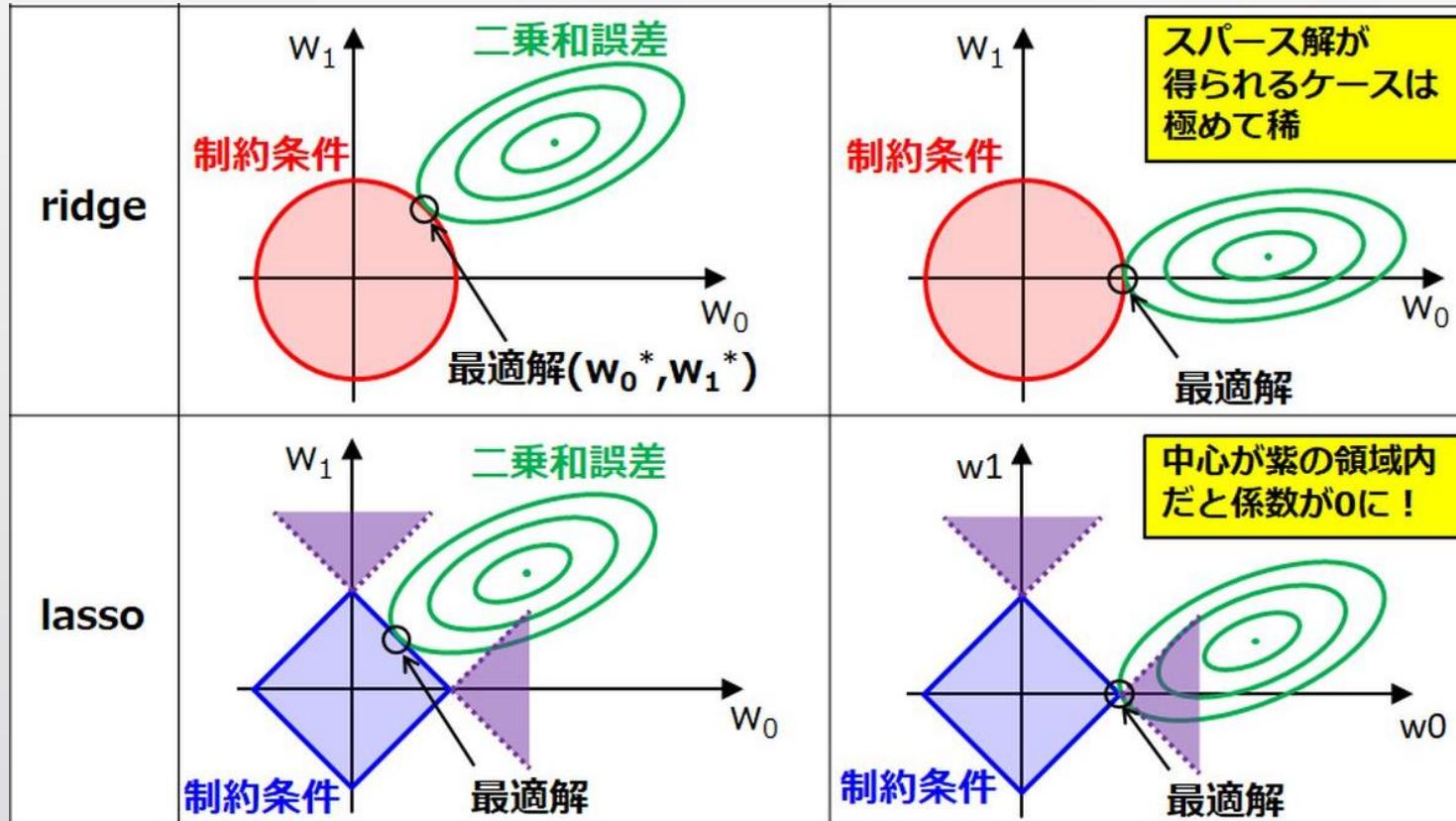
Regression coefficient of variable with little influence becomes zero

→ スparseな（疎な）回帰モデル
Sparse regression model



リッジ回帰とLasso回帰

Ridge Regression and Lasso Regression



Lasso回帰では、回帰係数 β の真の値が紫色の領域にあれば、推定した回帰係数が0になる

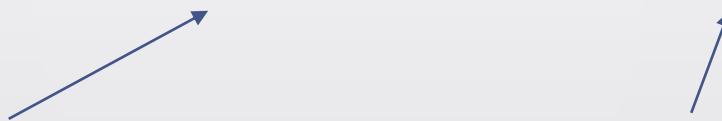
In Lasso regression, estimated regression coefficient becomes zero when true values of β falls within purple-colored regions

<https://yuyumoyuyu.com/2021/01/03/regularizedleastquares/>

Elastic Net

リッジ回帰の特徴とLasso回帰の特徴を組み合わせたもの
Elastic net has characteristics of both Ridge regression and Lasso regression

$$L = (\mathbf{y} - \mathbf{y}')^T(\mathbf{y} - \mathbf{y}') + \alpha \|\boldsymbol{\beta}\|_1 + (1 - \alpha) \|\boldsymbol{\beta}\|_2$$

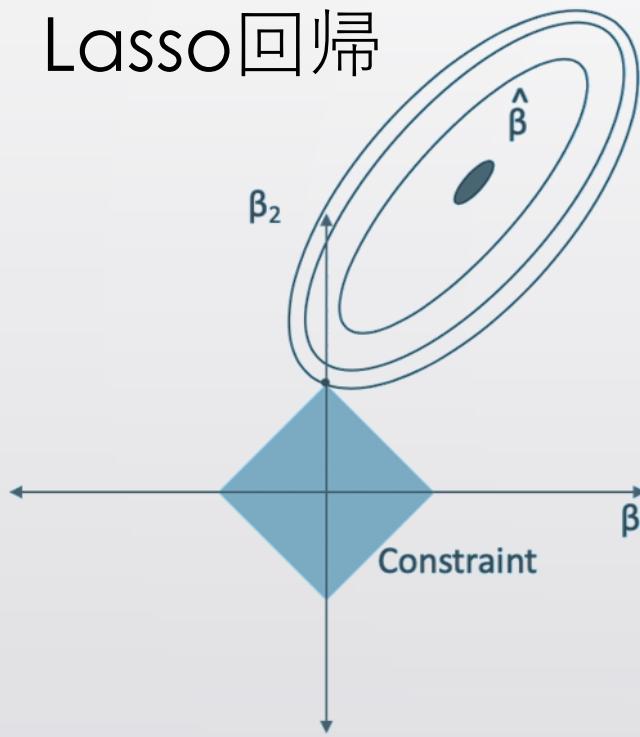


Lasso回帰の正則化項
Regularization term of
Lasso regression

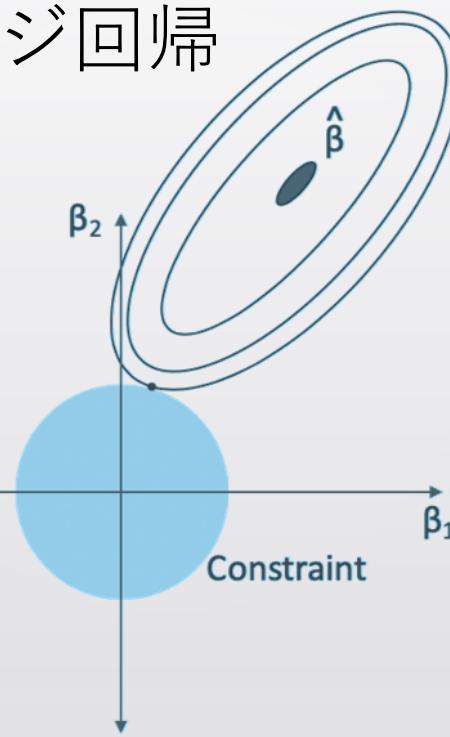
リッジ回帰の正則化項
Regularization term of
Ridge regression

Elastic Net

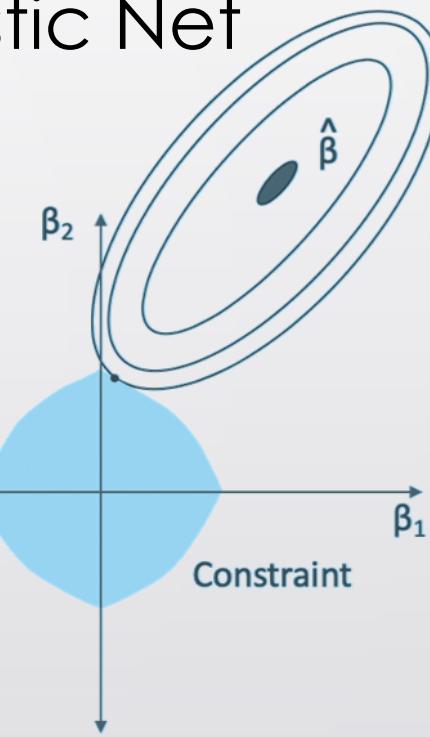
Lasso回帰



リッジ回帰

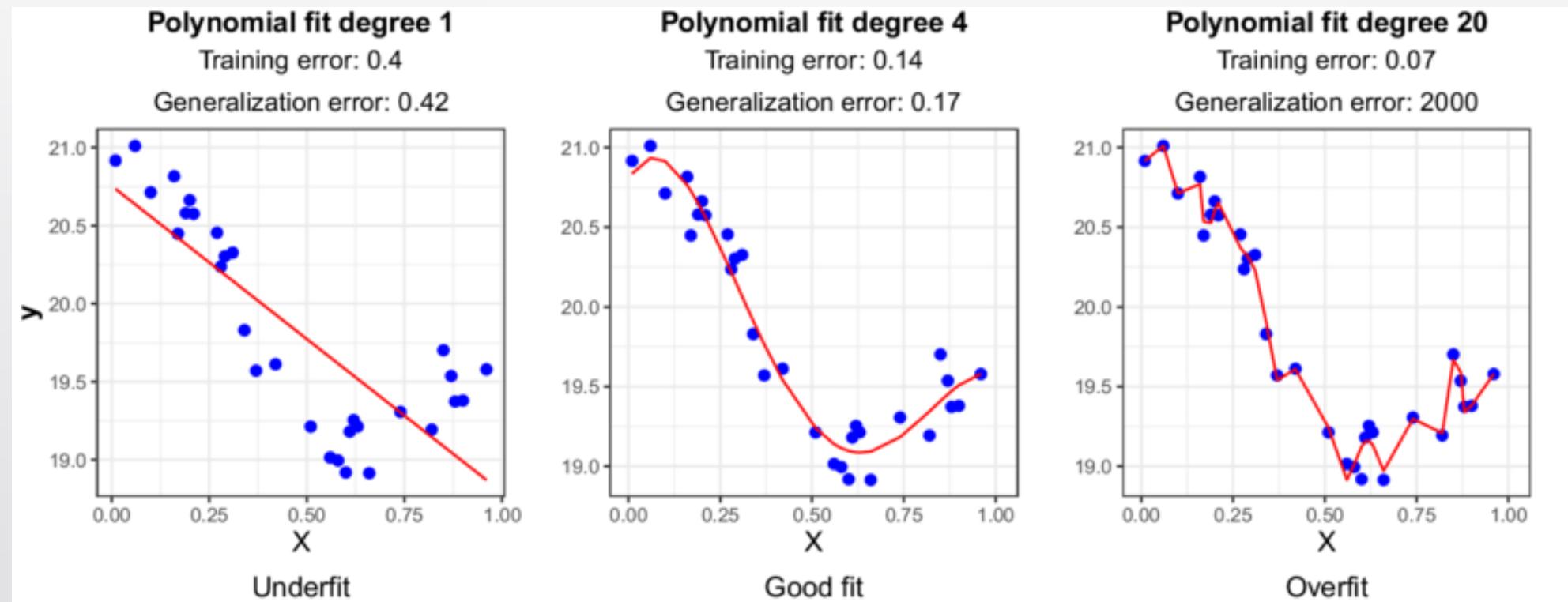


Elastic Net



<https://www.datasklr.com/extensions-of-ols-regression/regularization-and-shrinkage-ridge-lasso-and-elastic-net-regression>

過学習 Overfitting



Badillo et al, 2020



交差検証 Cross validation

1. データを学習(訓練)データとテストデータに分割する

Splitting data into training and test data

2. 学習(訓練)データを使って回帰モデルを作る

Create regression model based on training data

3. 回帰モデルの予測性能をテストデータで検証する

Evaluate prediction performance of regression model using test data

予測性能の指標 Indicator of Prediction Performance

正解値と予測値の相関係数

Correlational coefficient between predicted and actual values

根平均二乗誤差 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2} = \sqrt{\frac{1}{N} (\mathbf{y} - \mathbf{y}')^T (\mathbf{y} - \mathbf{y}')}}$$

2つの重回帰分析 Two Types of Multiple Regressions

最小二乗法

Ordinary Least Squares Method

RSSを最小化 Minimize RSS

$$\beta \downarrow$$

最尤推定

Maximum Likelihood Estimation

誤差 ε が正規分布すると仮定

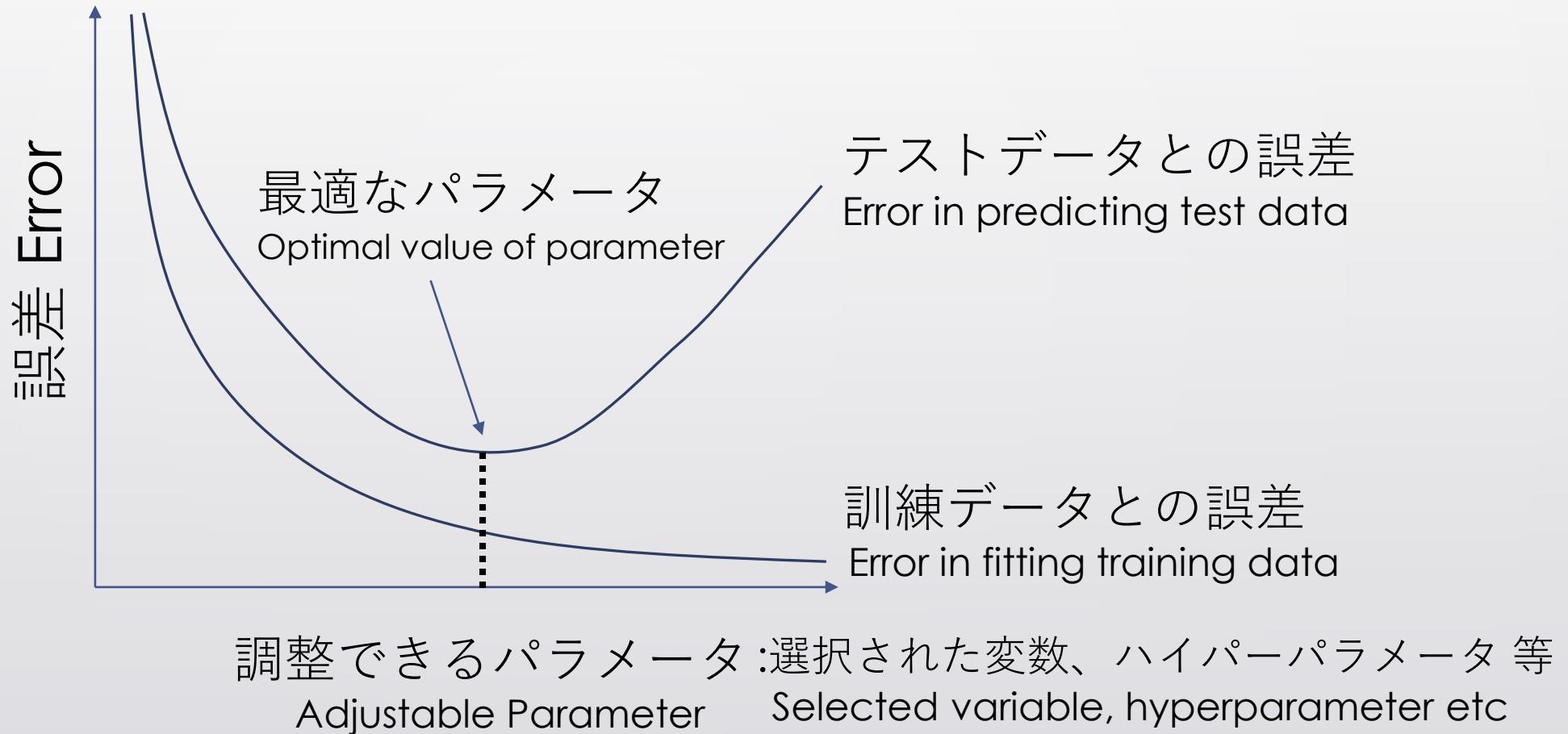
Assume that error ε conforms to
normal distribution

$\text{Log}(L)$ を最大化
Maximize $\text{Log}(L)$

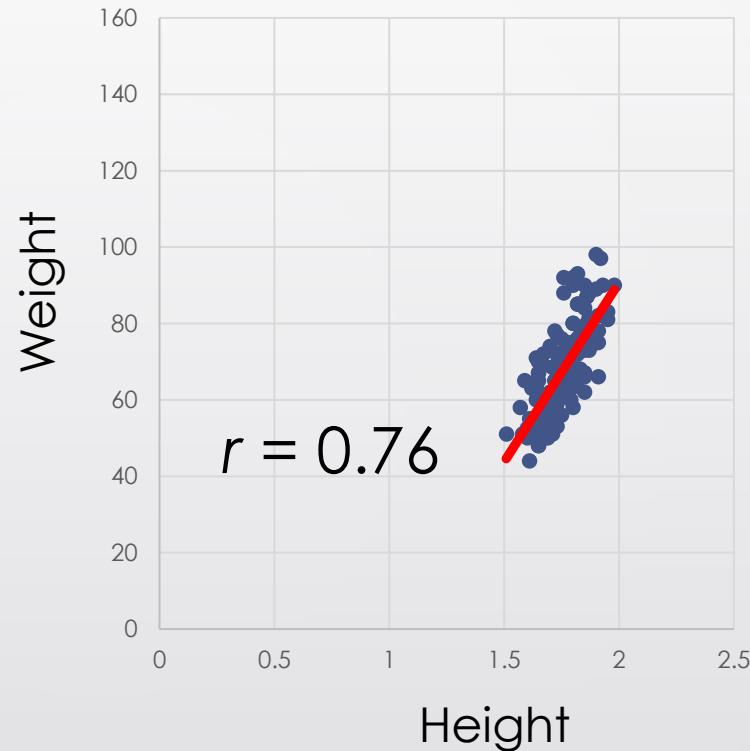
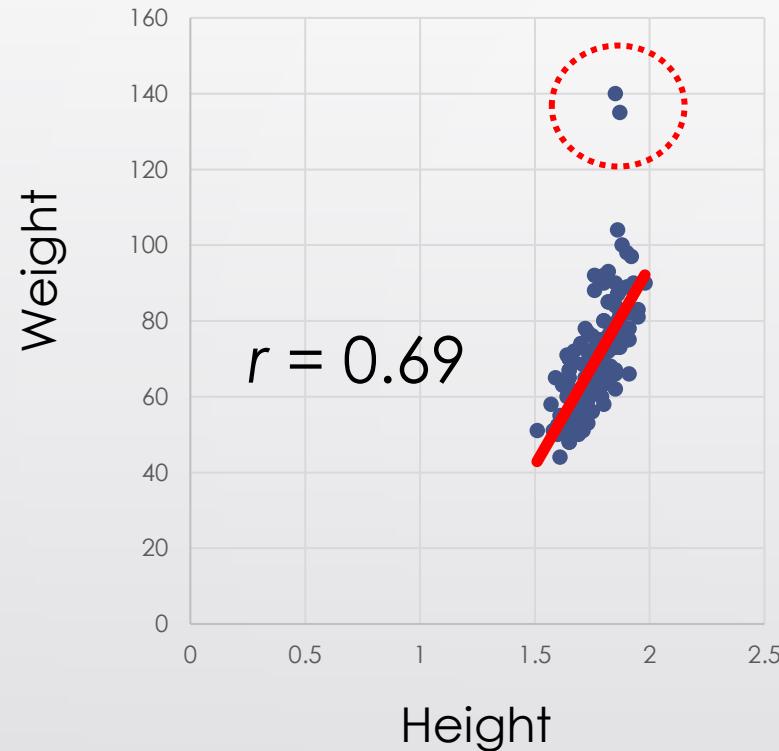
$$\sigma = \sqrt{\frac{1}{N} \sum_1^N (y_n - y'_n)^2}$$

RSSを最小化 Minimize RSS

過学習のサイン Signs of Overfitting



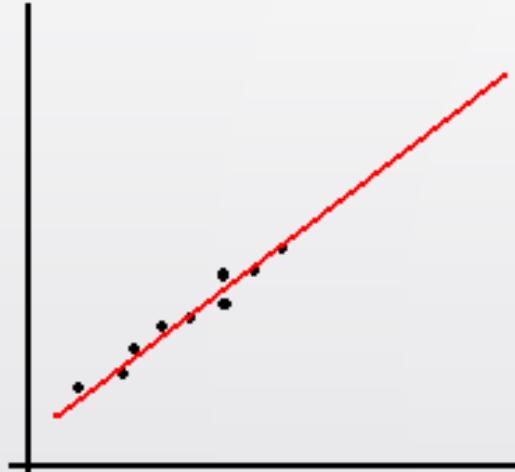
外れ値 Outlier



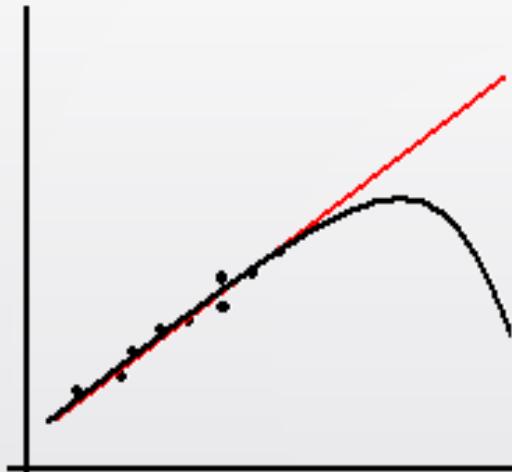
結果に影響を与える可能性がある外れ値の存在をチェックする必要がある
Data set should be checked for the existence of outliers that can influence the results



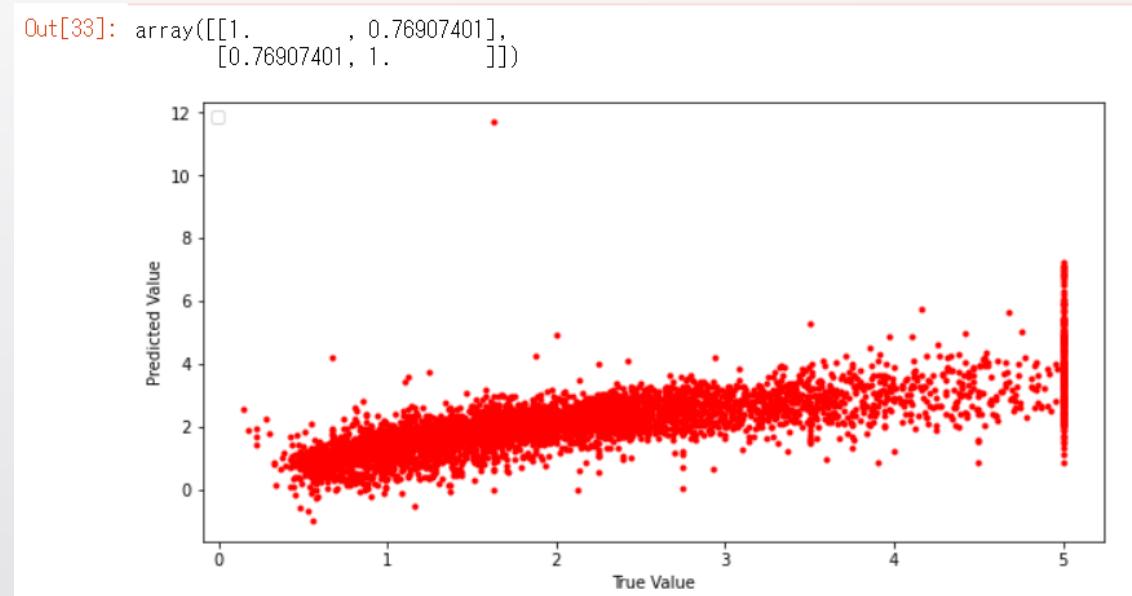
外挿の危険性 Peril of Extrapolation



(a) Beware of
extrapolation
past the end of
the data.



(b) Extrapolated
line is red,
actual response
curve is black.

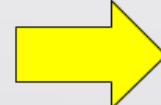


カテゴリ変数の扱い方

Treatment of Categorical variable

ダミー変数 Dummy Variable 男 ⇒ 1, 女 ⇒ 2

One-hot Encoding



Color	Red	Yellow
Red	1	0
Red	1	0
Yellow	0	1
Green	0	0
Yellow	0	0

<https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/notebook>



データマイニング

Data Mining

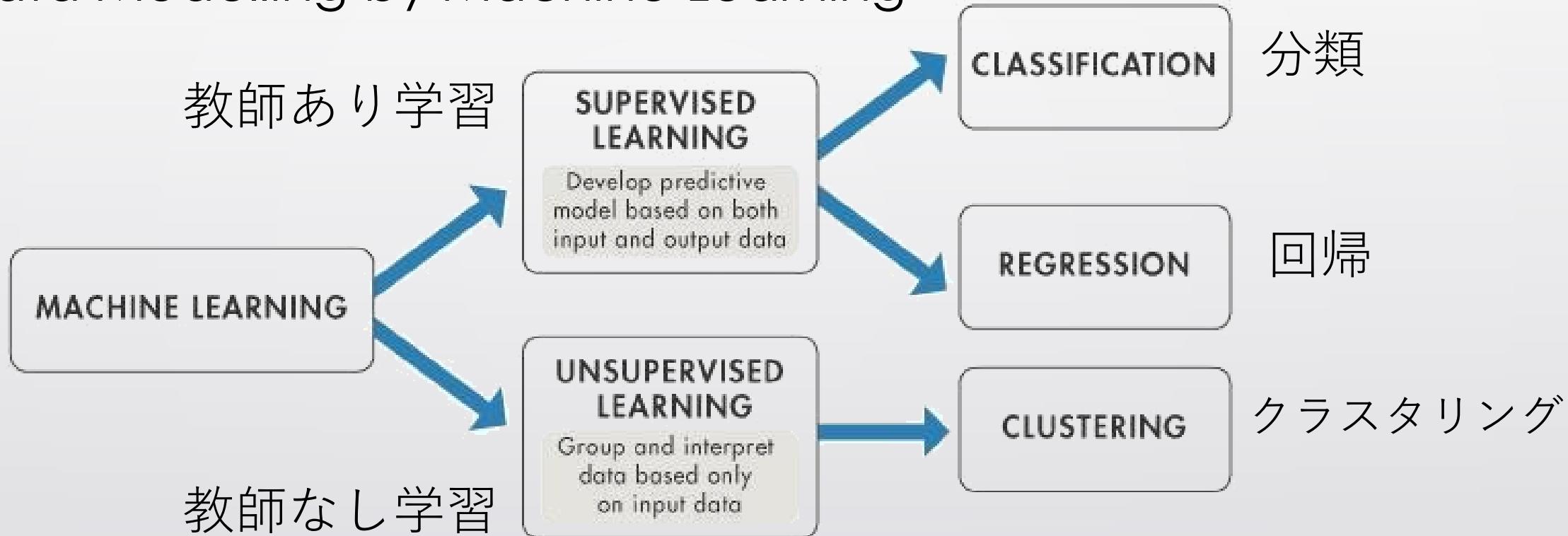
6: 分類① Classification

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

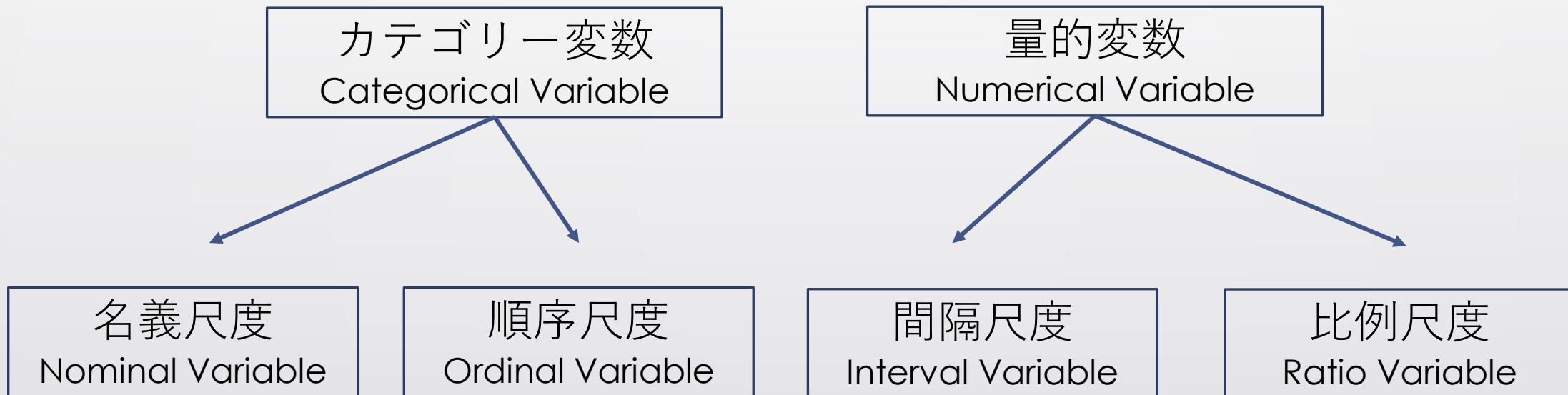
機械学習によるモデル化

Data Modelling by Machine Learning



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

変数の種類 Types of Variables





変数の種類 Types of Variables

名義尺度

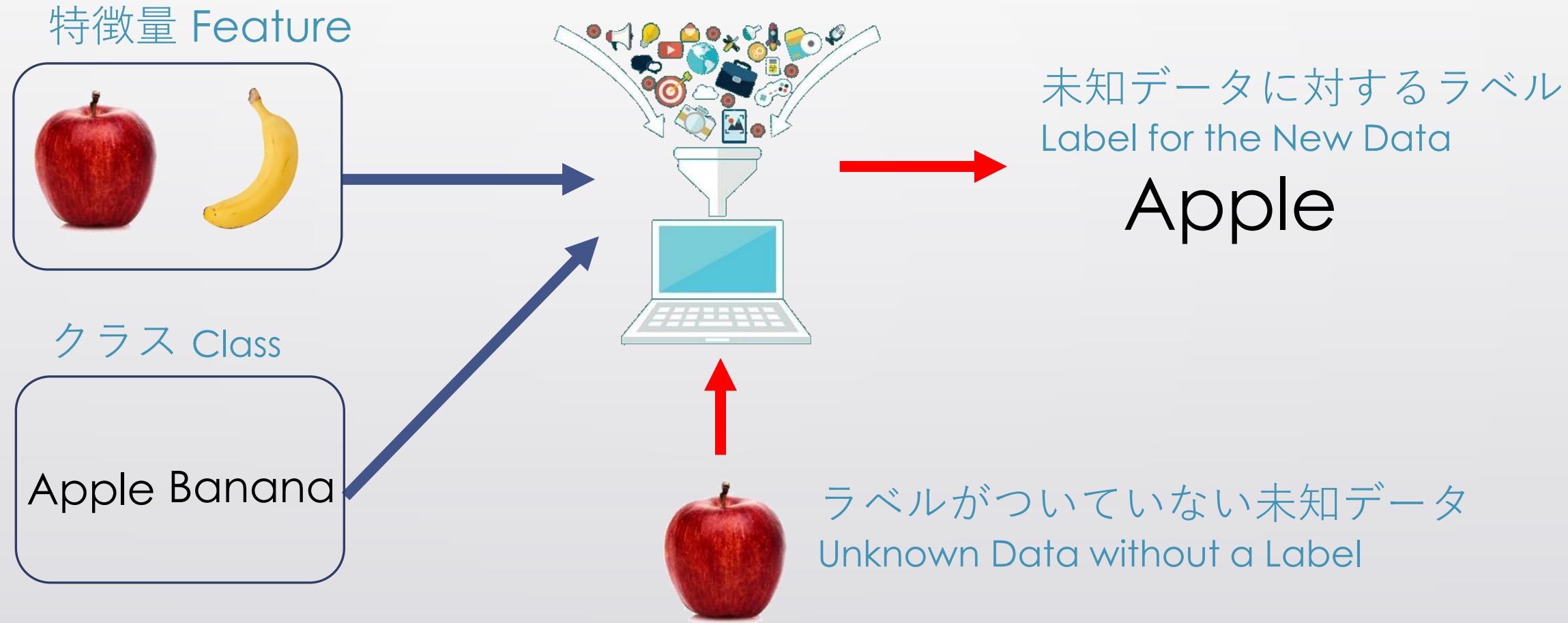
Nominal Variable

あるカテゴリーを、別のカテゴリーと区別するために用いられる、数値自体には意味がない変数

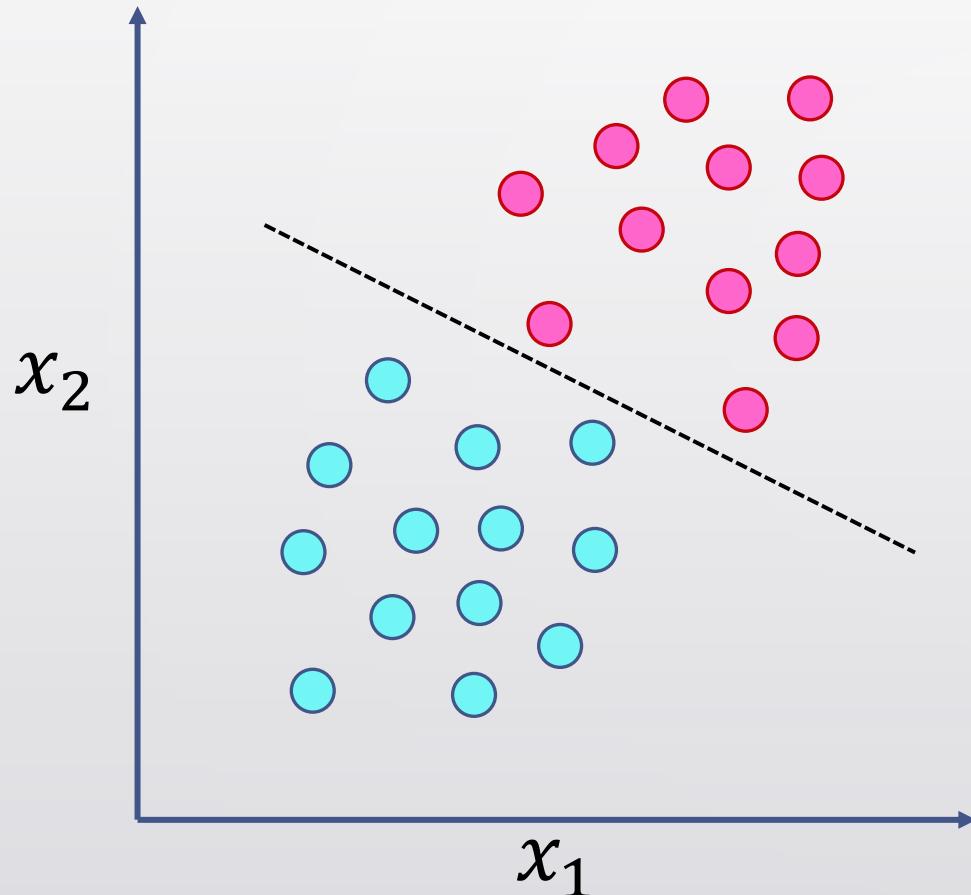
Variables, whose number has no numerical value, often used to discriminate multiple categories

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

教師あり学習 Supervised Learning



線型判別分析 Linear Discriminant Analysis (LDA)

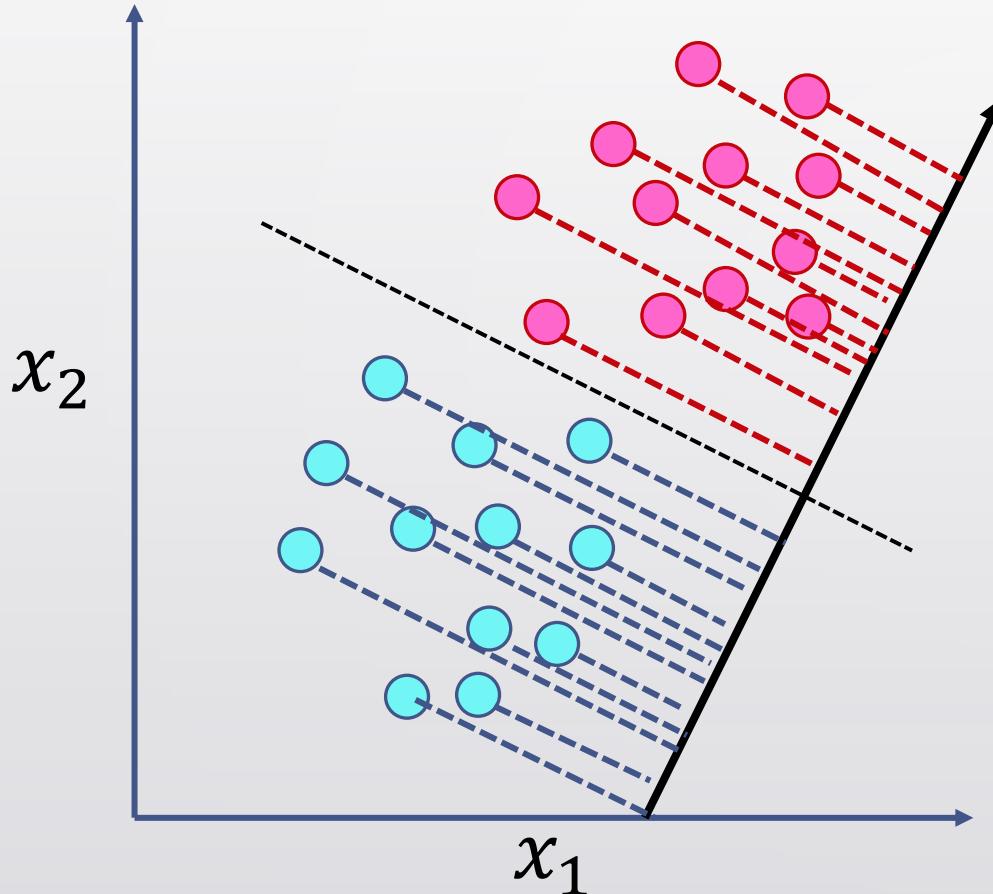


クラス1とクラス2を分離できる決定
境界の引き方を見つける

Find how to draw a decision boundary that
separates Class 1 and Class 2

決定境界
Decision Boundary

線型判別分析 Linear Discriminant Analysis (LDA)

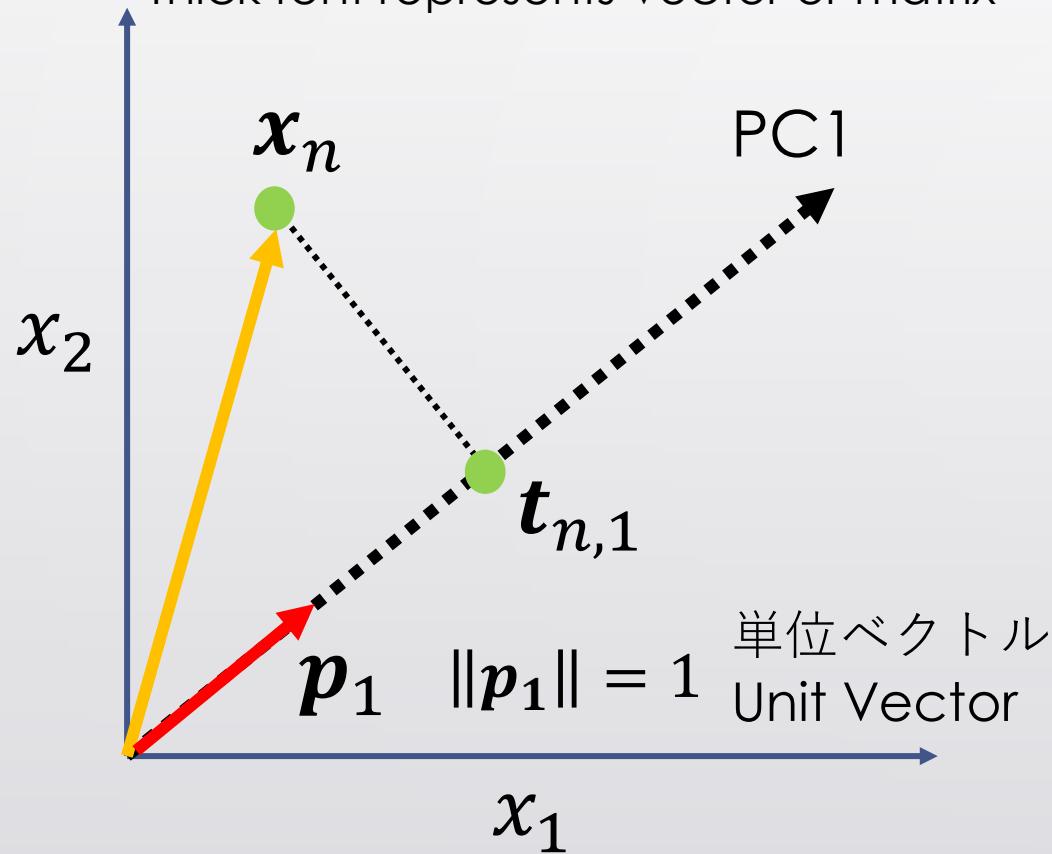


決定境界に直交する軸へのデータの射影を計算する
Consider the projection of data onto axis orthogonal to the decision boundary

第1主成分の計算 Computation of PC1

太字はベクトルか行列

Thick font represents vector or matrix



変数を標準化しておく

Normalize the variables

観測データ x_n の第1主成分軸方向への

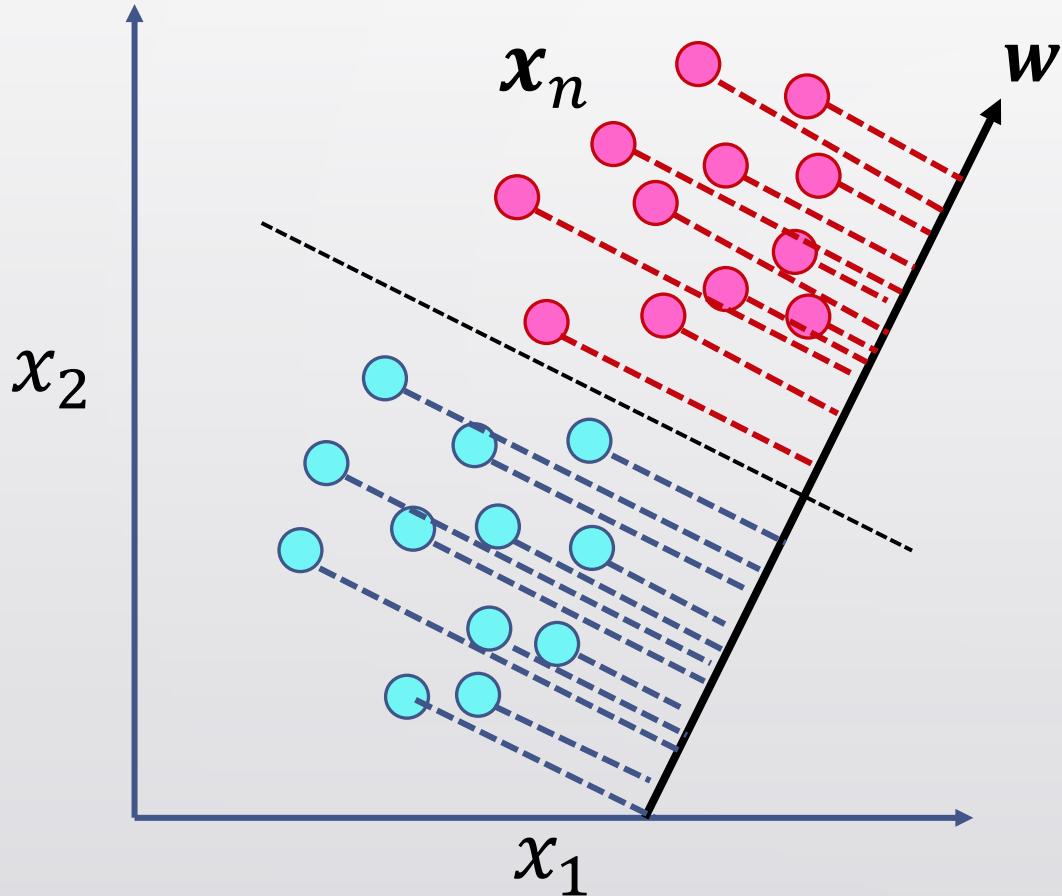
射影 $t_{n,1}$ を計算する

Compute the projection $t_{n,1}$ of the observed data x_n onto the first PC axis

$t_{n,1}$ は x_n と p_1 の内積

$t_{n,1}$ is dot(inner) product of x_n and p_1

線型判別分析 Linear Discriminant Analysis (LDA)

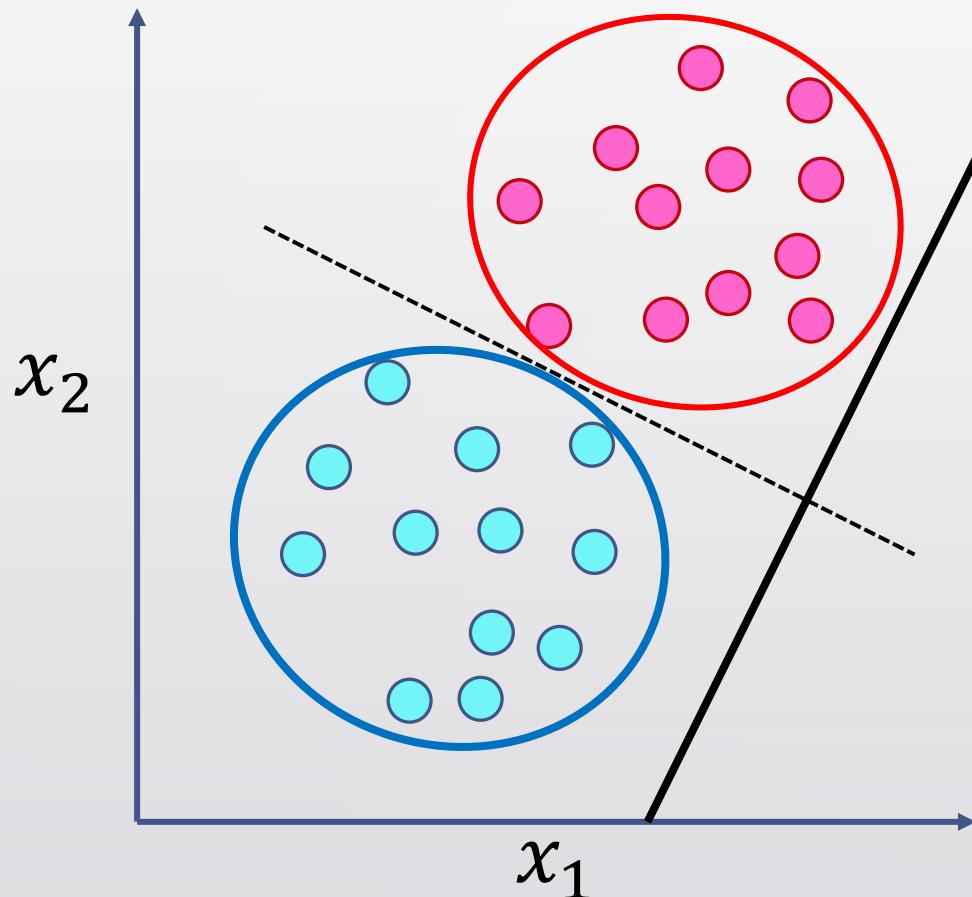


$$x_n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix} \quad \|w\| = 1$$

y_n は x_n の軸 w への射影
 y_n is projection of x_n onto axis w

$$y_n = w^T x_n$$

線型判別分析 Linear Discriminant Analysis (LDA)

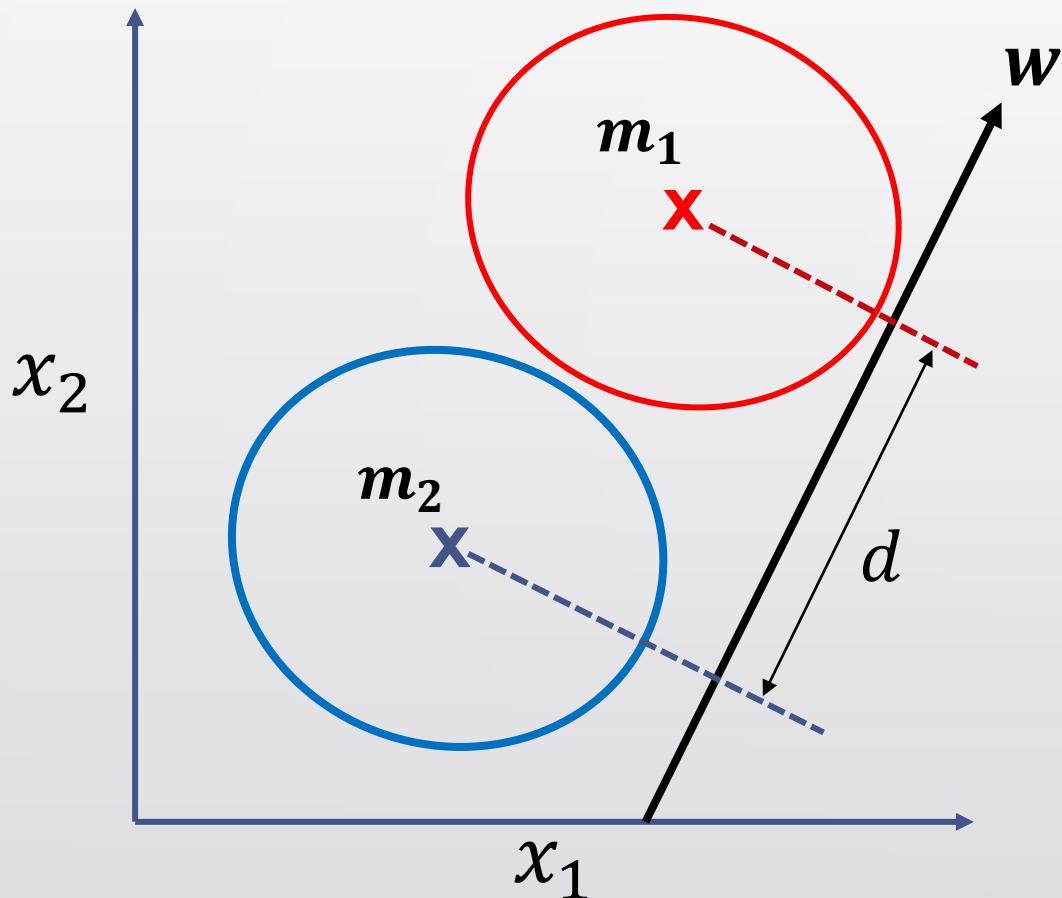


よい決定境界は、下の二つの条件を満たす

A good decision boundary meets the two conditions below

1. 1. クラスの中心が離れている
Centers of the two classes are distant from each other
2. 各クラスのクラス内分散が小さい
Within-class variance of each class is small

線型判別分析 Linear Discriminant Analysis (LDA)



1.2 クラスの中心が離れている
Centers of the two classes are distant from each other

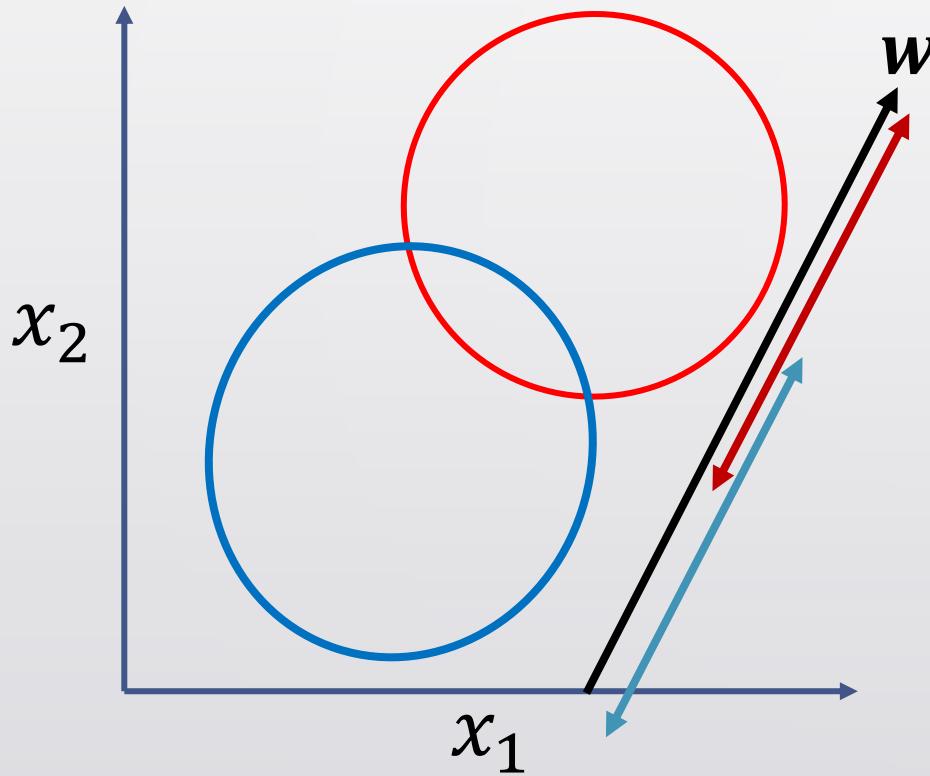
$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{x_k \in C_1} x_k \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{x_k \in C_2} x_k$$

$$d = \mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2 = \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)$$

線型判別分析 Linear Discriminant Analysis (LDA)

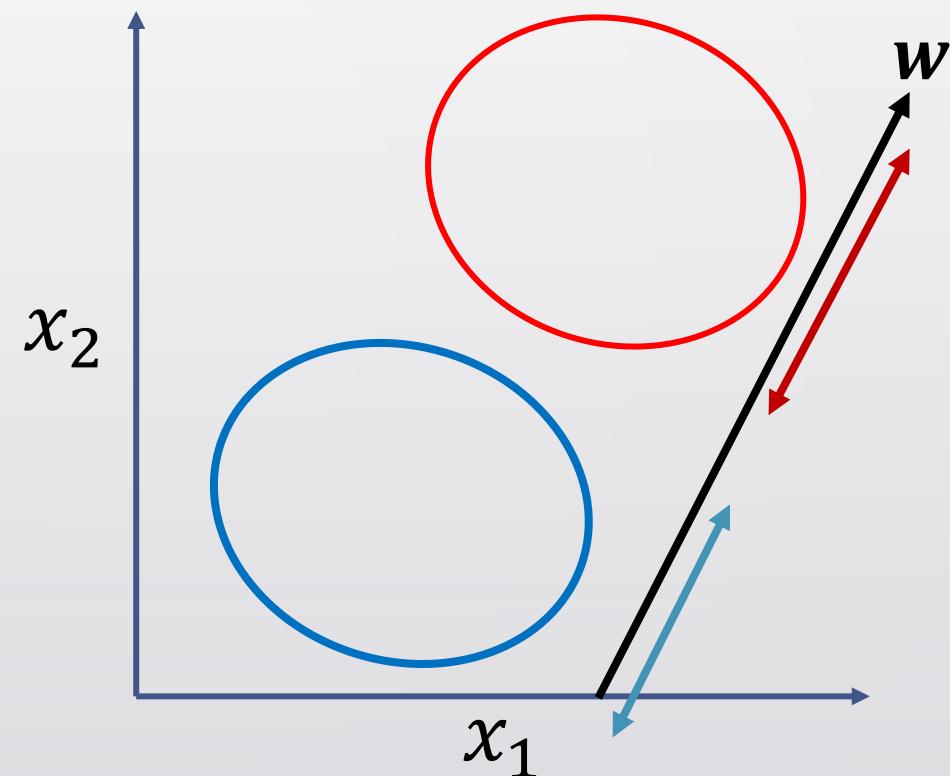
クラス内分散 大

Large within-class variance



クラス内分散 小

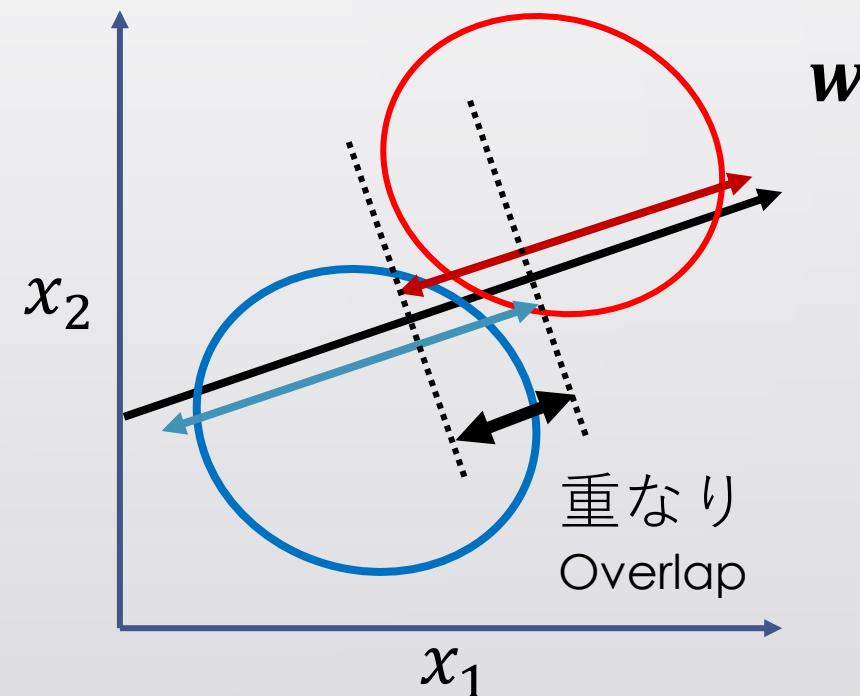
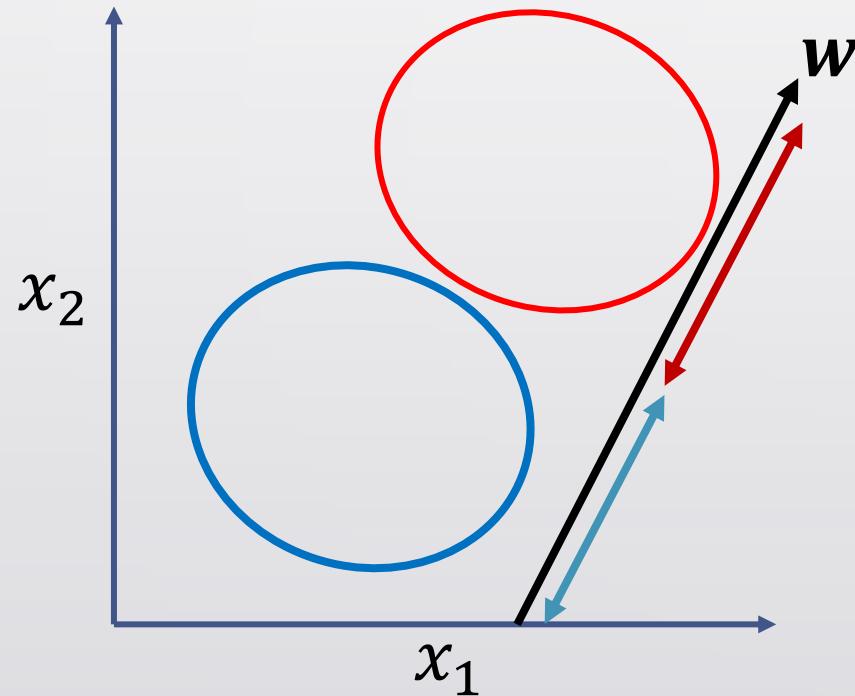
Small within-class variance



線型判別分析 Linear Discriminant Analysis (LDA)

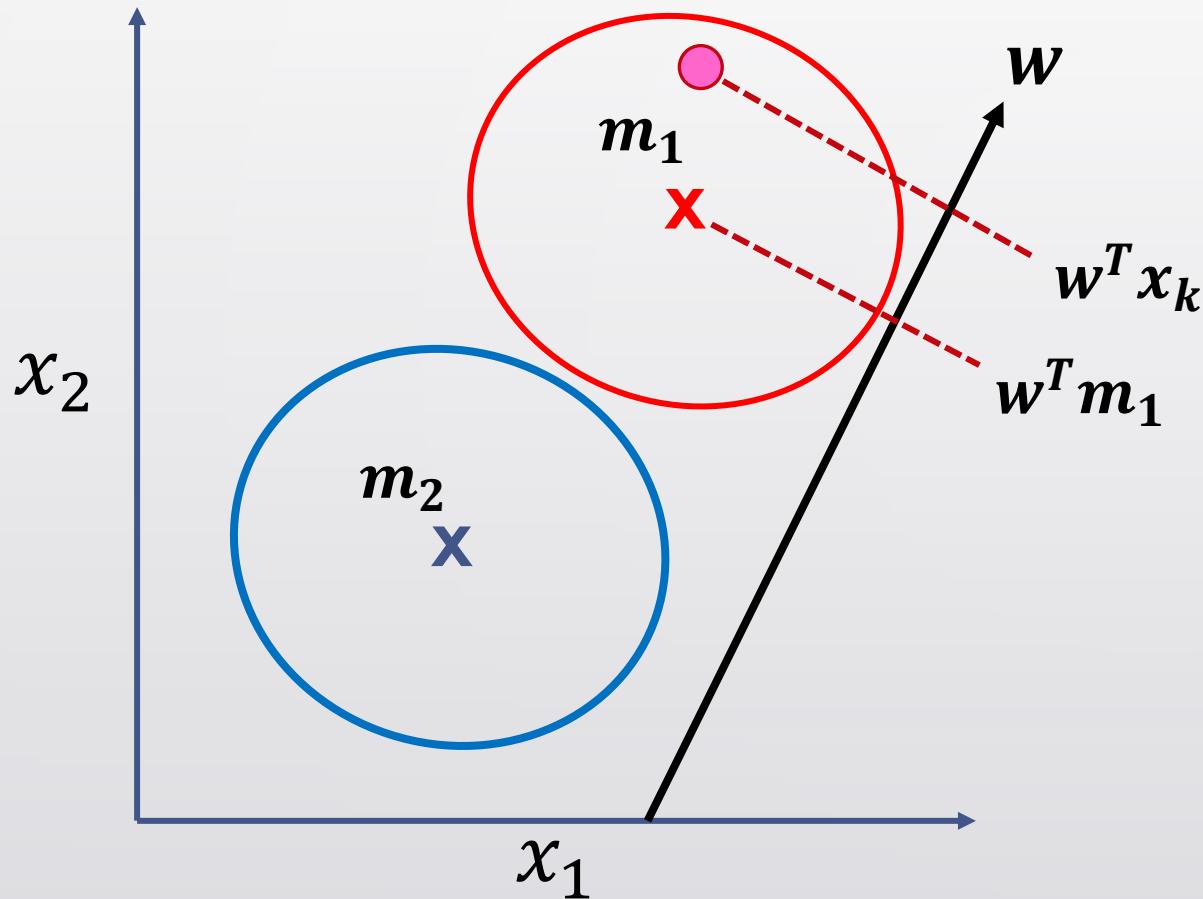
射影のクラス内分散は軸 w の向きにより変化する

Within-class variance of projection is dependent on the direction of axis w





線型判別分析 Linear Discriminant Analysis (LDA)



$$s_1^2 = \frac{1}{N_1} \sum_{x_k \in C_1} (w^T x_k - w^T m_1)^2$$

$$s_2^2 = \frac{1}{N_2} \sum_{x_k \in C_2} (w^T x_k - w^T m_2)^2$$

$$S^2 = s_1^2 + s_2^2$$

線型判別分析 Linear Discriminant Analysis (LDA)

1.2 クラスの中心が離れている Centers of the two classes are distant from each other

→ d を最大化する Maximize d

$$d = \mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2 = \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)$$

2.各クラスのクラス内分散が小さい Within-class variance of each class is small

→ s^2 を最小化する Minimize s^2

$$s^2 = s_1^2 + s_2^2 \quad s_j^2 = \frac{1}{N_j} \sum_{x_k \in C_j} (\mathbf{w}^T \mathbf{x}_k - \mathbf{w}^T \mathbf{m}_j)^2$$

線型判別分析 Linear Discriminant Analysis (LDA)

1.2 クラスの中心が離れている Centers of the two classes are distant from each other

→ d を最大化する → d^2 を最大化する

$$d^2 = \{\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\}^2 = \{\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\} \{\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\}^T$$

$$= \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} = \mathbf{w}^T \mathbf{S}_B \mathbf{w}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

線型判別分析 Linear Discriminant Analysis (LDA)

2. 各クラスのクラス内分散が小さい Within-class variance of each class is small

→ s^2 を最小化する Minimize s^2

$$s^2 = s_1^2 + s_2^2 = \mathbf{w}^T \mathbf{S}_w \mathbf{w}$$

$$\mathbf{S}_w = \sum_{x_k \in C_1} (x_k - \mathbf{m}_1)(x_k - \mathbf{m}_1)^T + \sum_{x_k \in C_2} (x_k - \mathbf{m}_2)(x_k - \mathbf{m}_2)^T$$

線型判別分析 Linear Discriminant Analysis (LDA)

1.2 クラスの中心が離れている Centers of the two classes are distant from each other

2.各クラスのクラス内分散が小さい Within-class variance of each class is small

→ $J(w)$ を最大化する Maximize $J(w)$

$$J(w) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

線型判別分析 Linear Discriminant Analysis (LDA)

$$J(w) \text{を最大化する} \quad \text{Maximize } J(w) \quad J(w) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

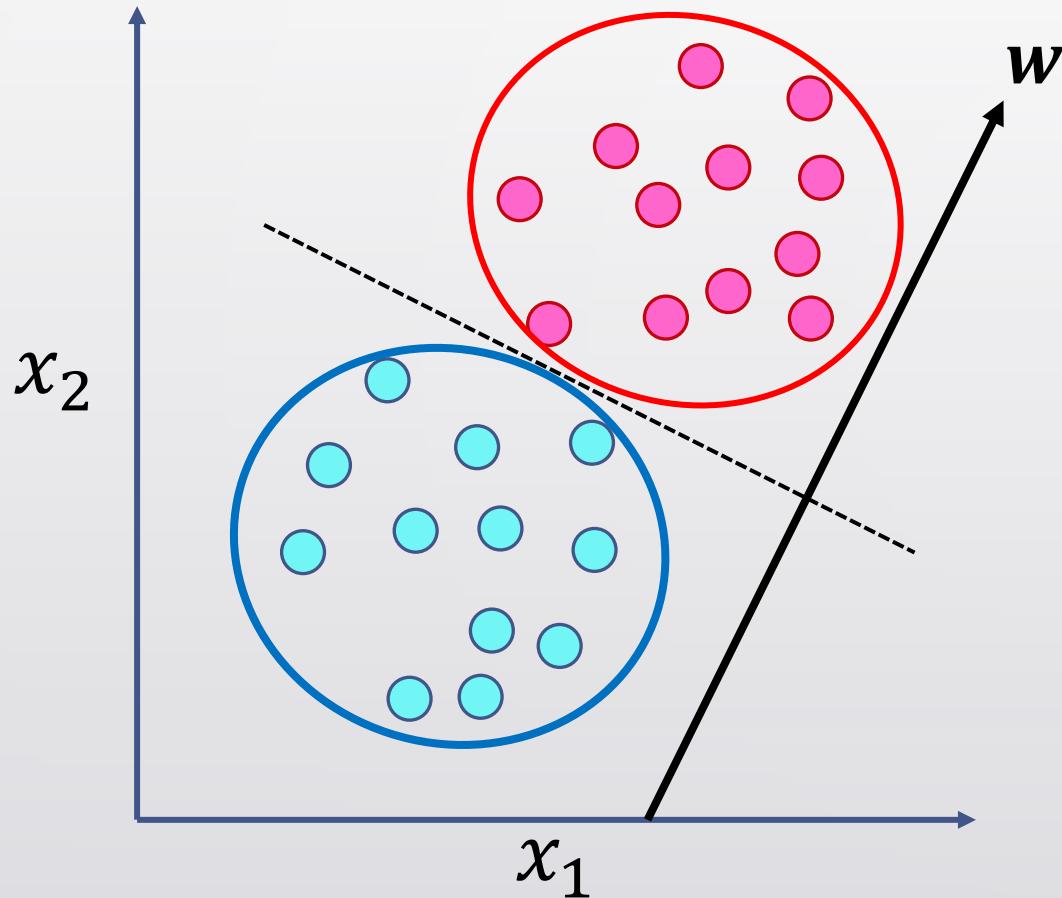
$J(w)$ を最大化する \mathbf{w} は下の固有方程式を満たす

\mathbf{w} that maximizes $J(w)$ satisfies the eigen equation below

$$\mathbf{S}_w^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w} \quad \mathbf{S}_B \mathbf{w} = \mathbf{S}_w \mathbf{S}_w^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$$

$$J(w) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} = \frac{\lambda \mathbf{w}^T \mathbf{S}_w \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} = \lambda$$

線型判別分析 Linear Discriminant Analysis (LDA)



$$S_w^{-1} S_B w = \lambda w$$

$$J(w) = \frac{w^T S_B w}{w^T S_w w} = \frac{\lambda w^T S_w w}{w^T S_w w} = \lambda$$

軸 w は最大の固有値に対応する固有ベクトルと並行

Axis w is in parallel with eigen vector corresponding to the largest eigen value

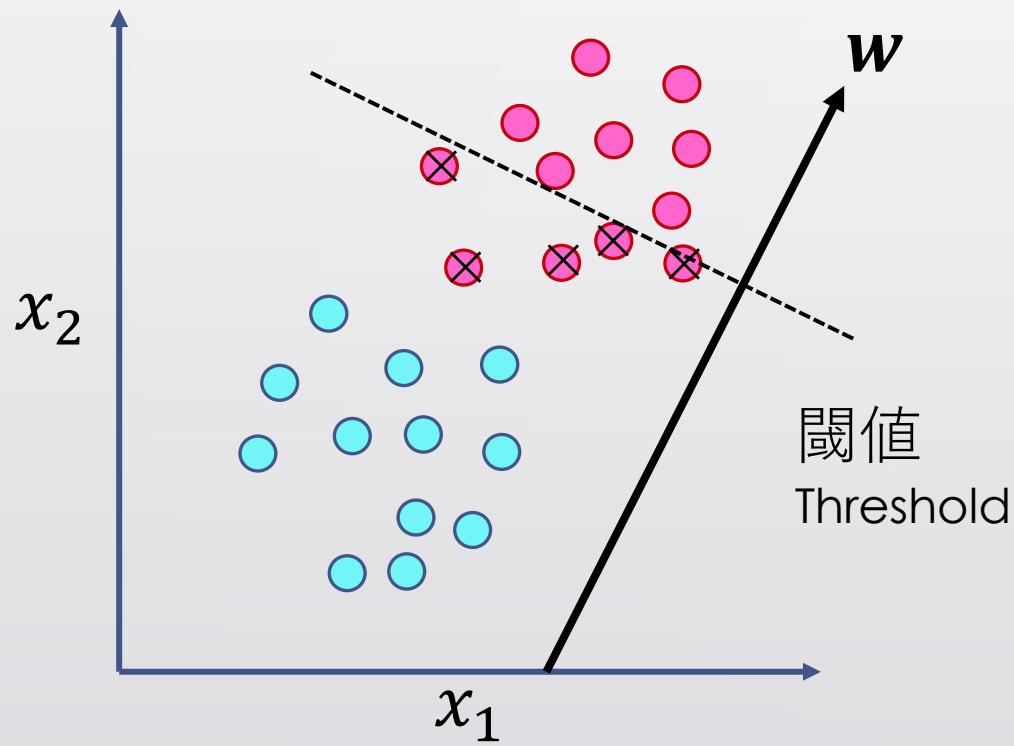
→ 決定境界の向きが決まる

Orientation of decision boundary is determined

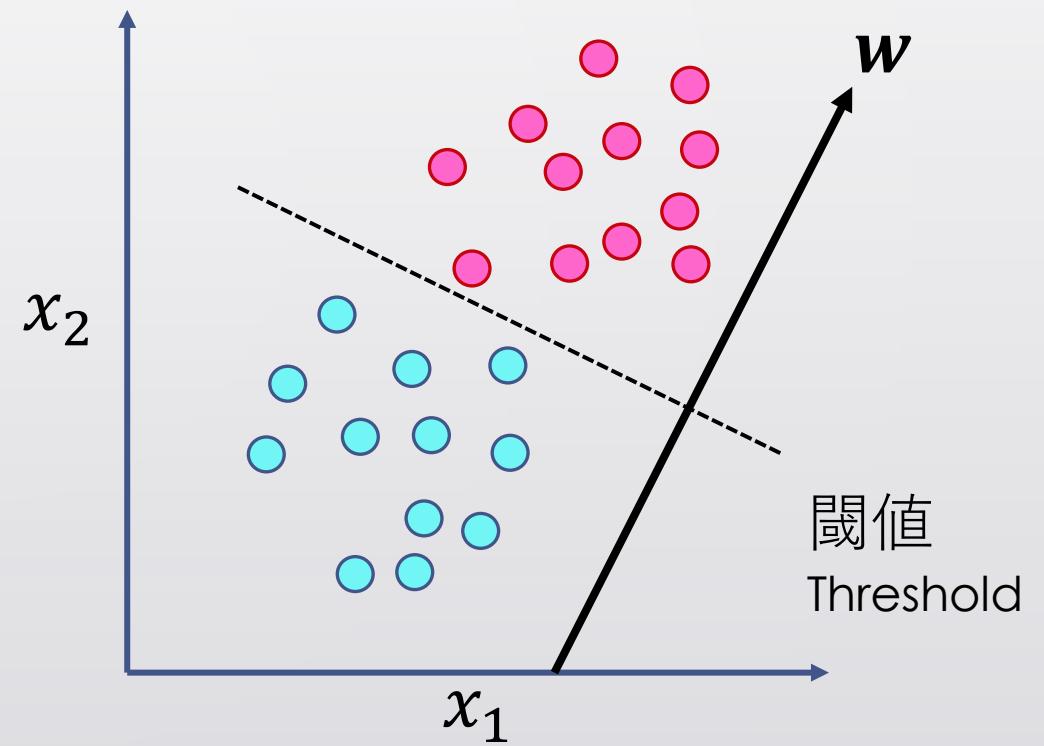
///

閾値をどう決定するか？ How should we determine the threshold?

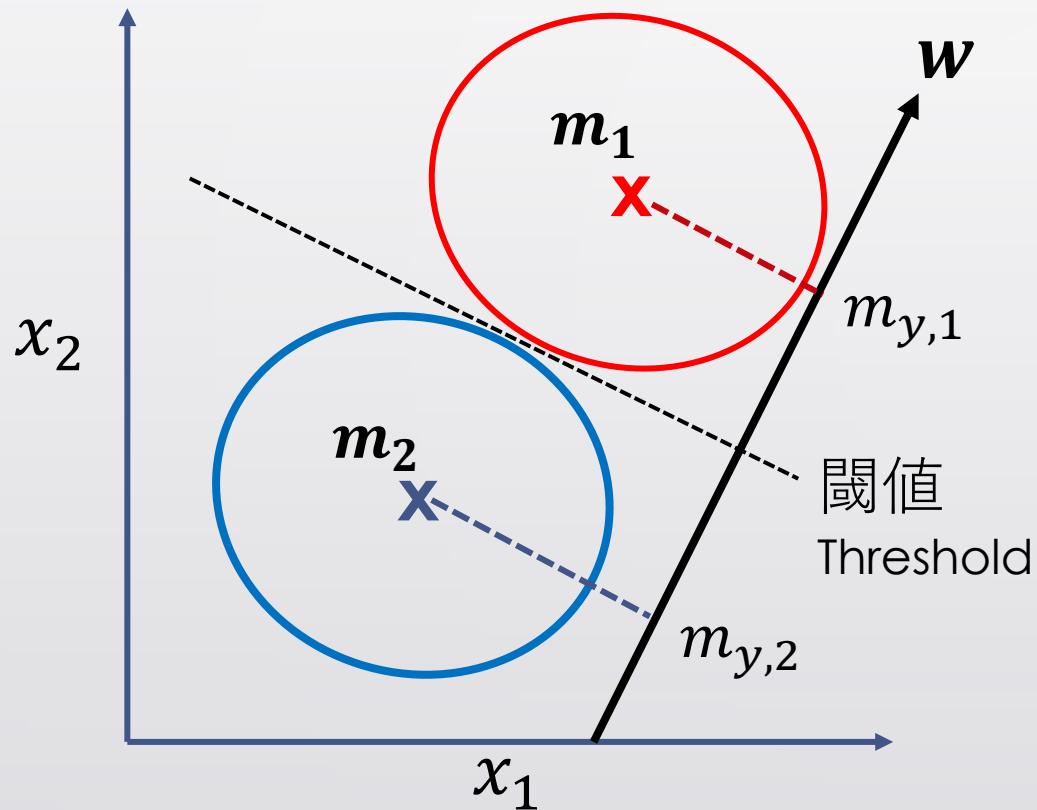
不適切な閾値 Inappropriate threshold



適切な閾値 Appropriate threshold



閾値をどう決定するか？ How should we determine the threshold?



各クラスの中心の射影の加重平均を
閾値にする

Adopt as threshold value the weighted mean of
projection of center of each class

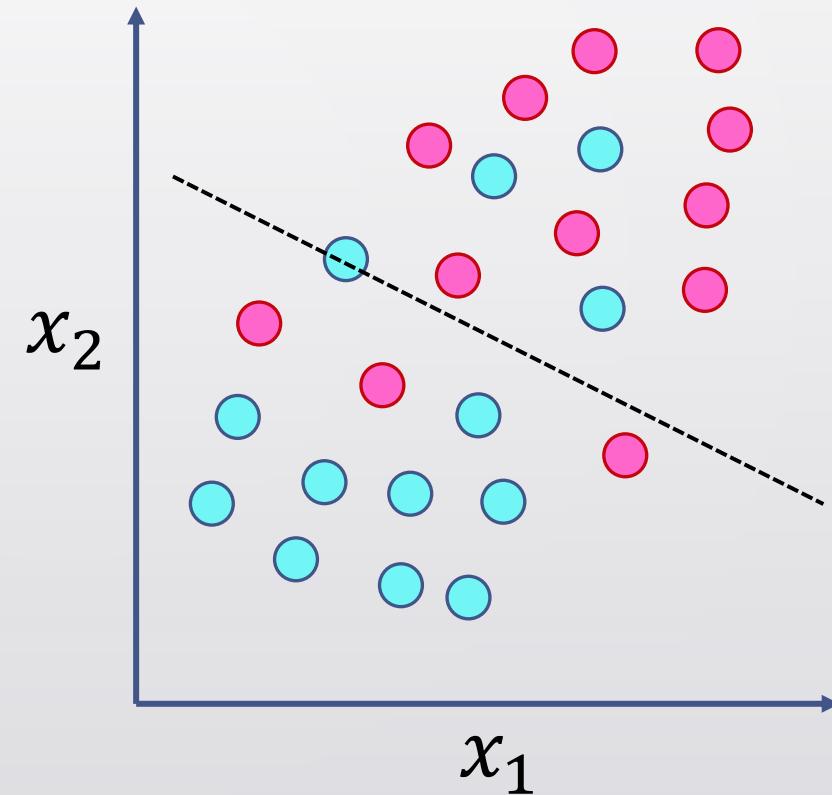
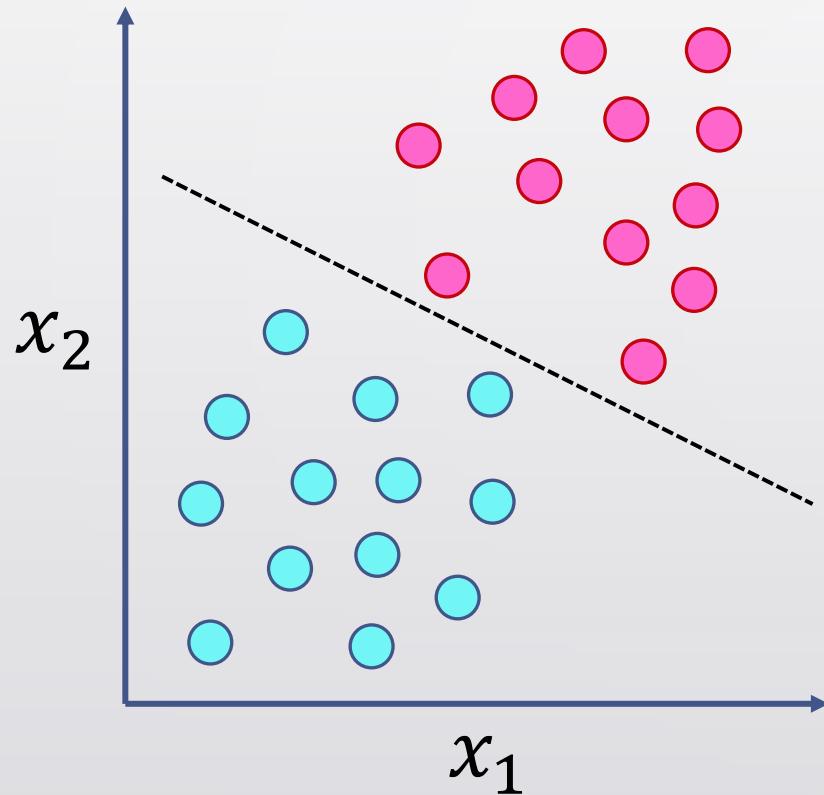
$$Threshold = \frac{N_1 s_{y,1}^2 m_{y,1} + N_2 s_{y,2}^2 m_{y,2}}{N_1 s_{y,1}^2 + N_2 s_{y,2}^2}$$

$s_{y,j}^2$: クラスjのデータの射影の分散

$m_{y,j}$: クラスjの重心の射影

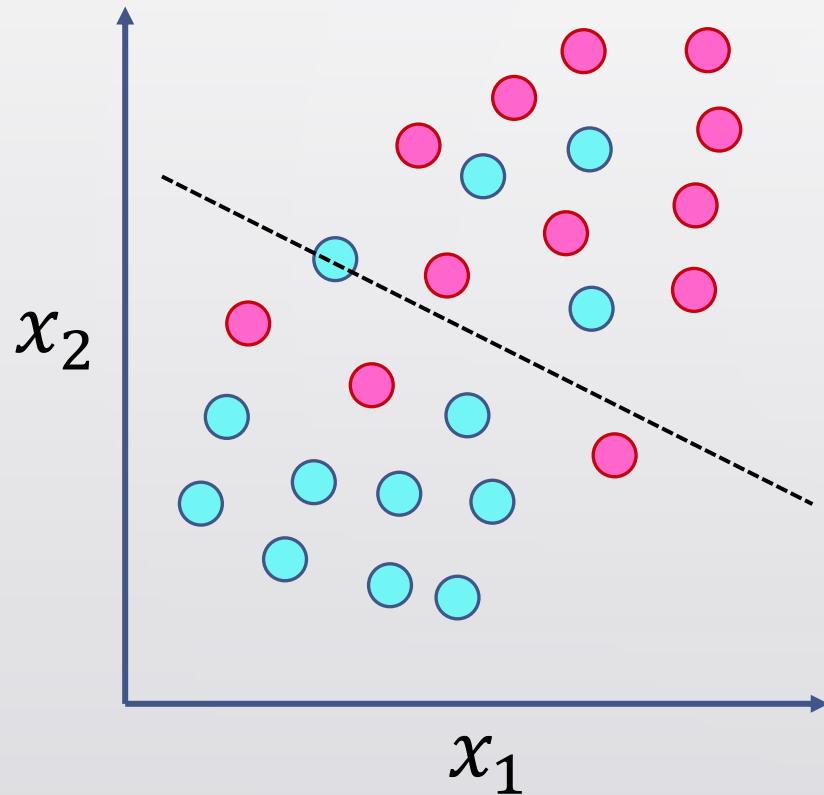
“線型分離可能”とは？

What does “Linearly Separable” Mean?



“線型分離可能”とは？

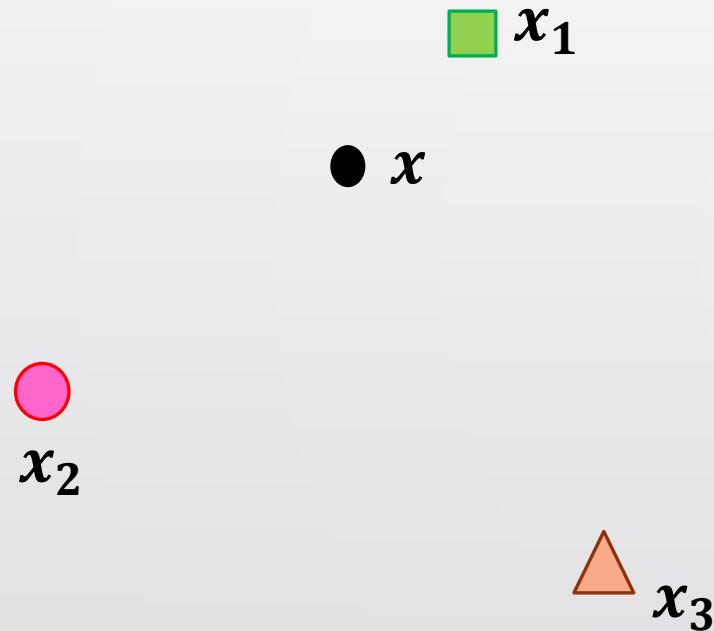
What does “Linearly Separable” Mean?



線型分離不可能な問題には、LDA
がうまく機能しない

LDA does not work well for linearly
inseparable problems

最近傍法 Nearest Neighbor Method



データ x は最も近くにある鋳型データ x_j と同じクラスに属するとみなす

Data x is judged to belong to the same class as template data x_j

最近傍法 Nearest Neighbor Method

クラス C_i の鋳型と x の最小距離

Shortest distance between x and templates of class C_i

$$\operatorname{argmin}_i \left\{ \min_j d(x, x_j^i) \right\} \text{ if } \min_{i,j} d(x, x_j^i) < t$$

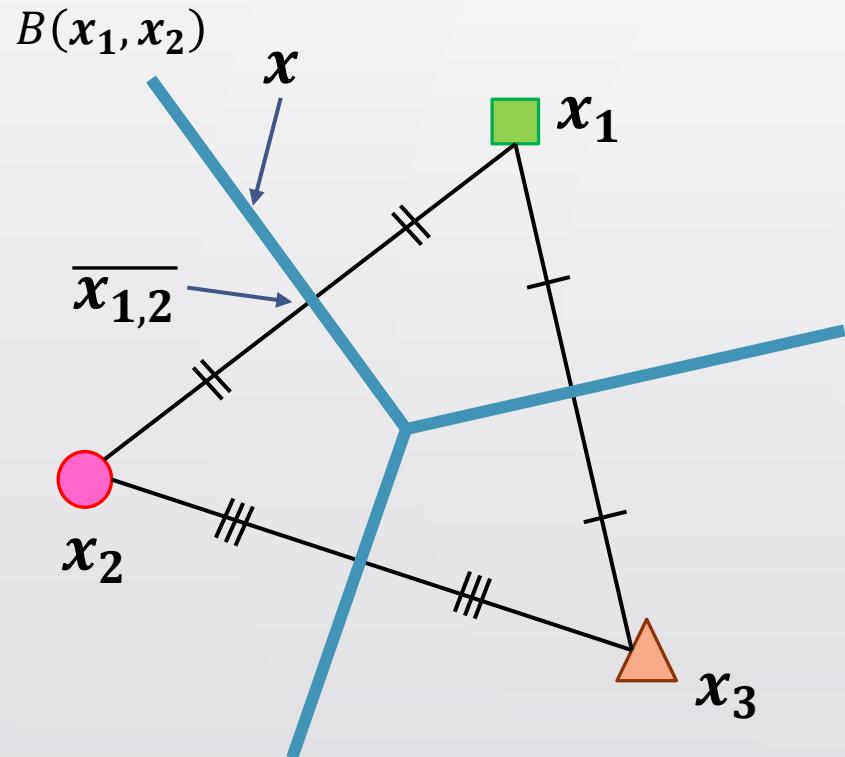
x と最も近い鋳型が属するクラスの識別番号を返す

Returns the identifier of the class to which the template closest to x

$$\text{Reject if } \min_{i,j} d(x, x_j^i) \geq t$$



ボロノイ境界 Voronoi Boundary

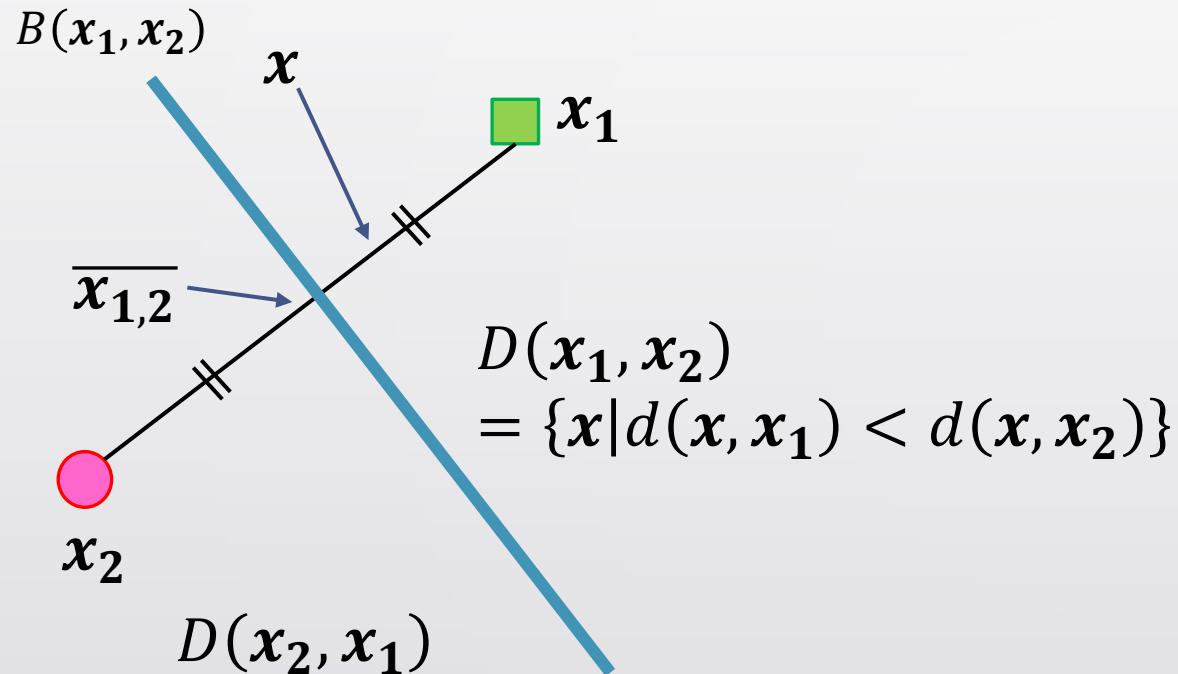


一対の鋳型から等距離にある点の集合
Set of points equidistant from a pair of templates

$$B(x_j, x_j) = \{x | d(x, x_i) = d(x, x_j)\}$$

$$(x - \overline{x_{i,j}}) \cdot (x_i - x_j) = 0$$

ボロノイ領域 Voronoi Region

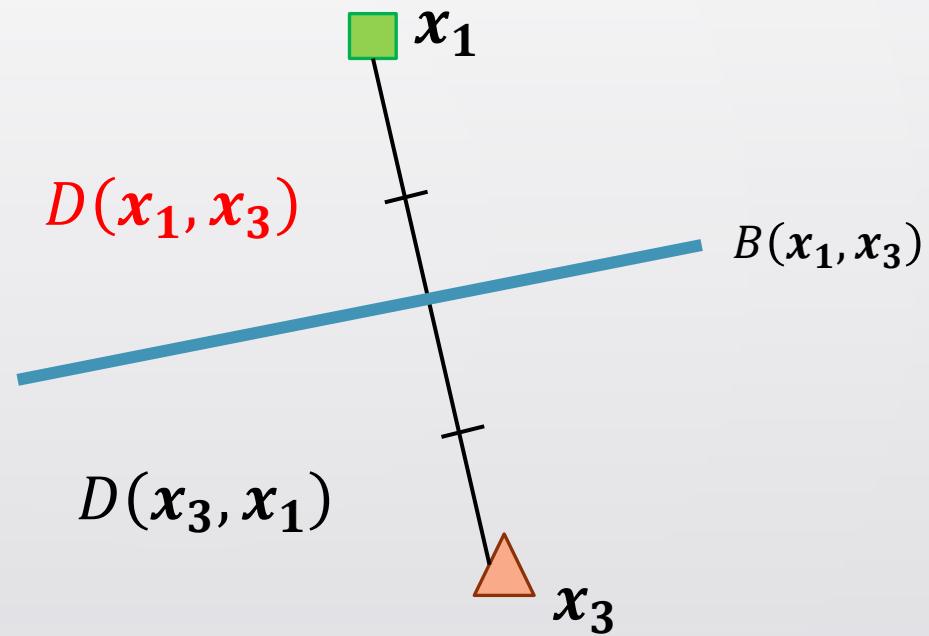
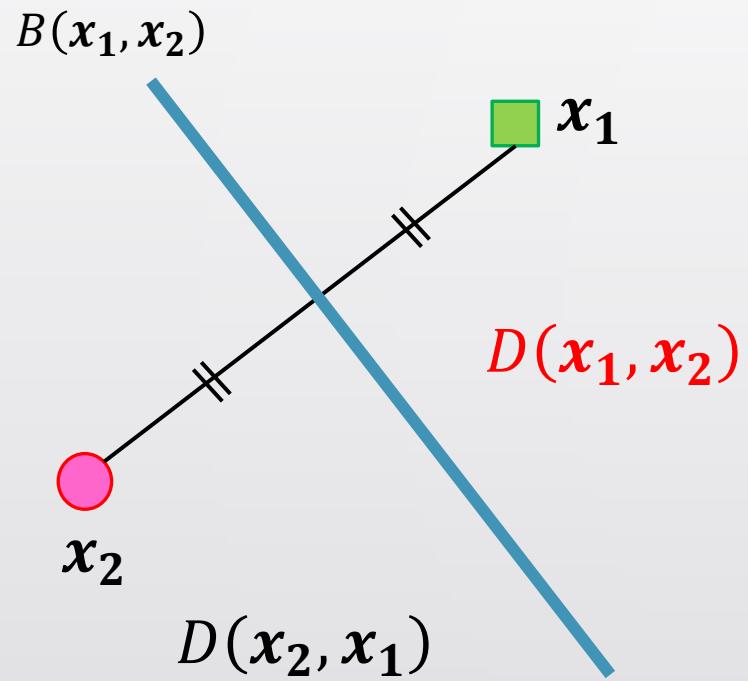


$D(x_i, x_j)$:

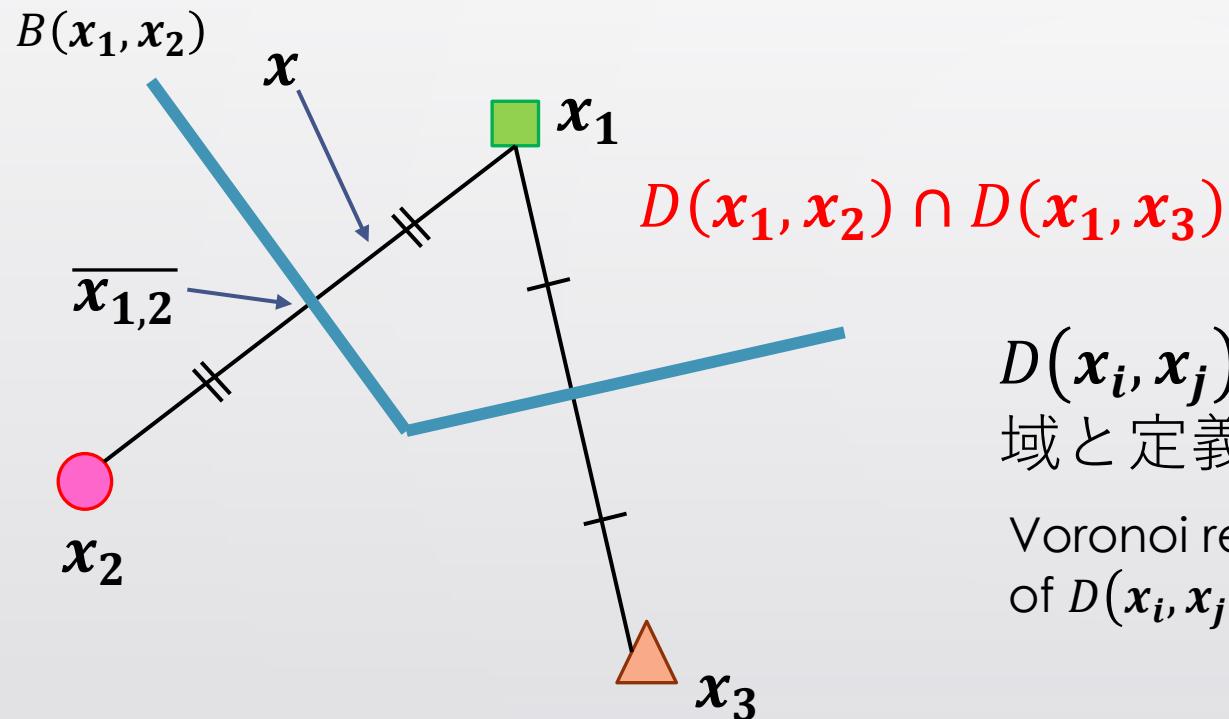
点 x_j より点 x_i に距離が近い点の集合
Set of points closer to x_i than x_j



ボロノイ領域 Voronoi Region



ボロノイ領域 Voronoi Region



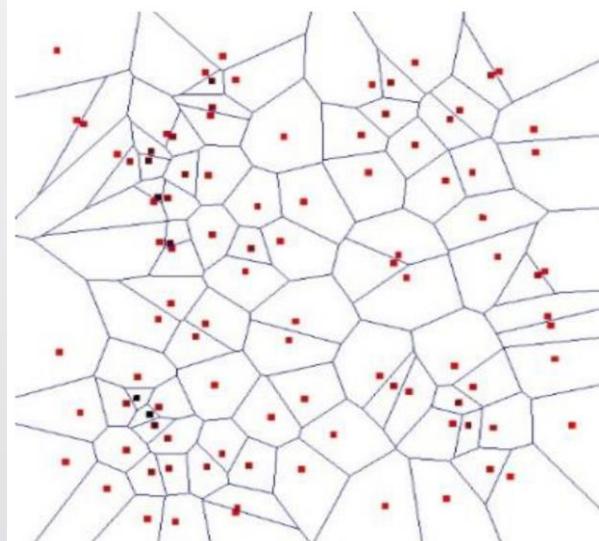
$D(x_i, x_j)(j \neq i)$ の積集合を x_i のボロノイ領域と定義する

Voronoi region of x_i is defined as set intersection of $D(x_i, x_j)(j \neq i)$

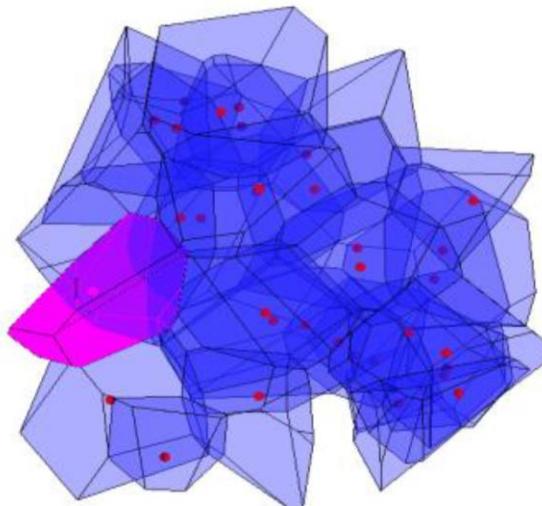
ボロノイ図 Voronoi Diagram

最近傍法の決定境界はボロノイ図を描く

Voronoi diagram shows configuration of decision boundaries in nearest neighbor method



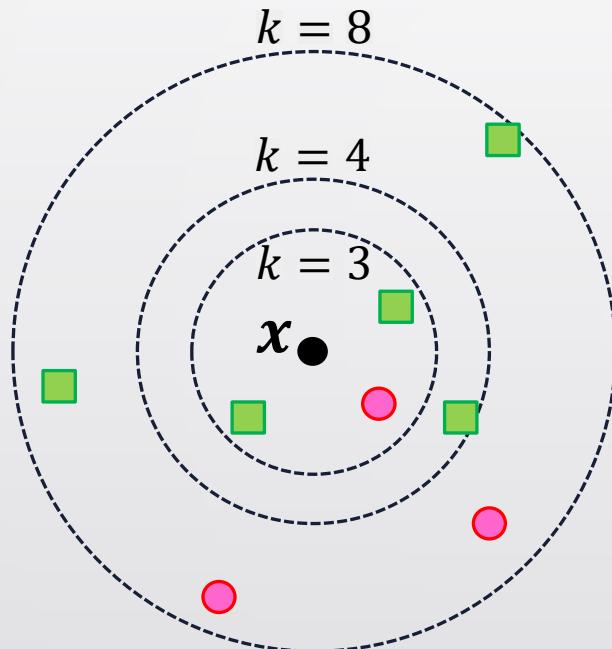
(a)



(b)

<https://www.mdpi.com/2220-9964/4/3/1480>

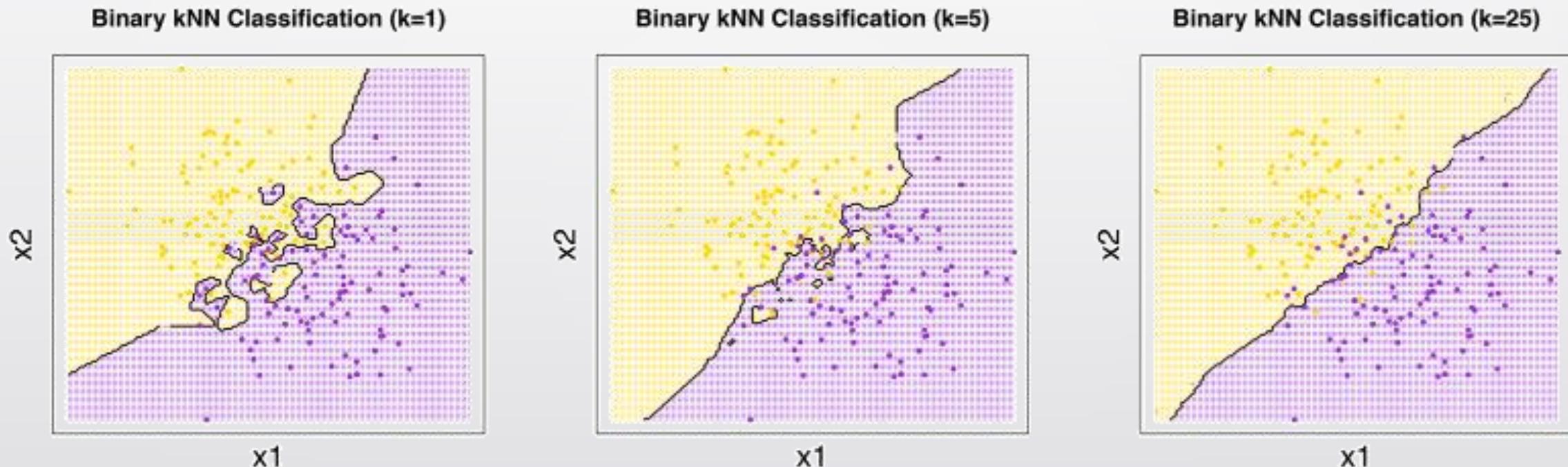
k 最近傍法 k Nearest Neighbor Method



データ x のクラスを最近傍にある k 個のデータの多数決投票により決定する

Class of data x is determined by majority voting of k data points closest to x

k最近傍法 k Nearest Neighbor Method





データマイニング

Data Mining

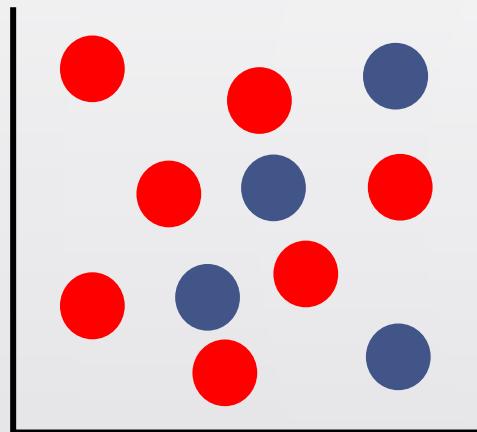
7: 分類② Classification

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

ベイズの定理 Bayes Theorem

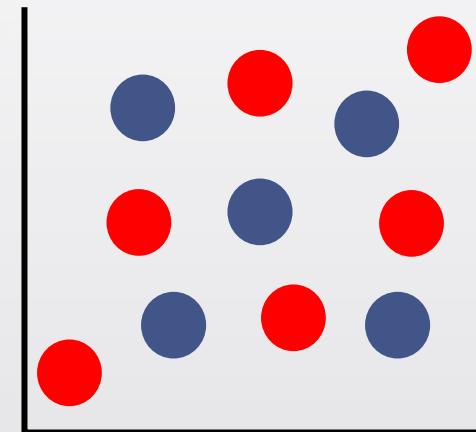
箱1 Box 1



赤玉 Red Ball: 70

青玉 Blue Ball: 30

箱2 Box 2



赤玉 Red Ball: 40

青玉 Blue Ball: 60

どちらかの箱から、1個ずつ玉をとりだし、その色を確認する

Take one ball out from either one of the boxes, and check its color

ベイズの定理 Bayes Theorem

- 1] H_1 : 玉を箱1から取り出した Ball was taken out from box 1
 - 2] H_2 : 玉を箱2から取り出した Ball was taken out from box 2
-
- D_1 : 取り出した玉が赤色だった Color of the ball taken out was red
 - D_2 : 取り出した玉が青色だった Color of the ball taken out was blue



条件付き確率 Conditional Probability

$P(A|B)$: 事象Bが起こっているという条件の下で、事象Aが起こる確率
Probability of event A under the condition that event B occurs

cf. $P(A \cap B)$: 同時確率 Joint Probability

事象AとBが共に起こる確率 Probability that both event A and B occur

$P(H_1|D_1)$: 取り出した玉の色が赤の時、玉を取り出した箱が箱1である確率
Probability that you took out a ball from box 1 when the ball taken out from a box was red.

ベイズの定理 Bayes Theorem

取り出した玉の色が赤の時、玉を取り出した箱が箱1である確率

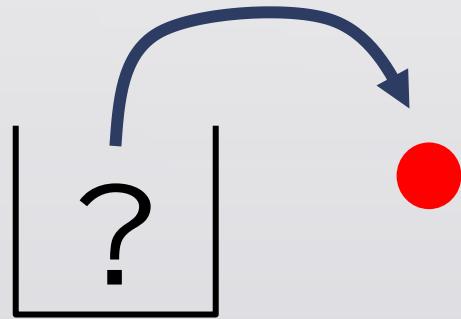
Probability that you took out a ball from box 1 when the ball taken out from a box was red.

$$P(H_1 \cap D_1) = P(H_1|D_1)P(D_1)$$

$$P(H_1|D_1)P(D_1) = P(D_1|H_1)P(H_1)$$

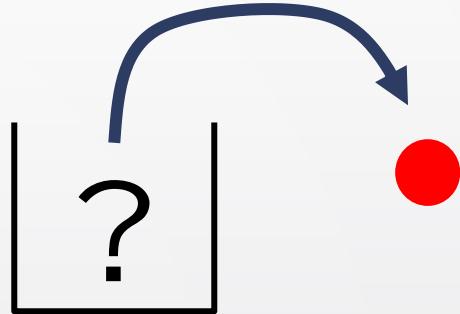
$$P(H_1 \cap D_1) = P(D_1|H_1)P(H_1)$$

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1)}$$



ベイズの定理 Bayes Theorem

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1)}$$



	箱1 Box 1	箱2 Box 2
赤玉 Red Ball	70	40
青玉 Blue Ball	30	60

$$P(D_1|H_1) = \frac{70}{100} \quad P(D_1) = \frac{70 + 40}{(70 + 30) + (40 + 60)}$$

$$P(H_1) = \frac{70 + 30}{(70 + 30) + (40 + 60)}$$

$$P(H_1|D_1) = \frac{70}{100} \times \frac{100}{200} \times \frac{200}{110} = 63.7\%$$

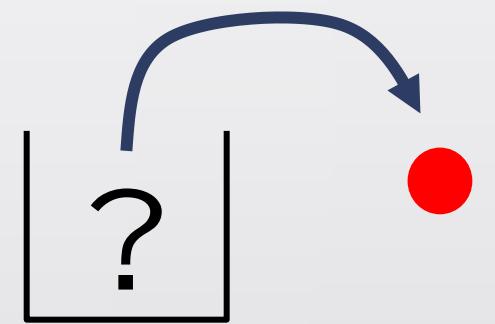


ベイズの定理 Bayes Theorem

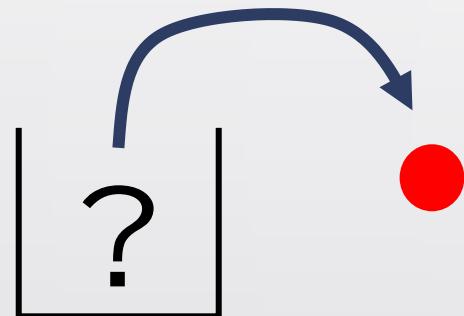
$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1)}$$

$$\begin{aligned} P(D_1) &= P(D_1 \cap H_1) + P(D_1 \cap H_2) = \sum P(D_1 \cap H_k) \\ &= \sum P(D_1 \cap H_k) = \sum P(D_1|H_k)P(H_k) \end{aligned}$$

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{\sum P(D_1|H_k)P(H_k)}$$



事前確率と事後確率 Prior Probability and Posterior Probability



事前確率 Prior Probability

データを観察する”前に”推定した、箱が箱1である確率

Probability that the box is box 1 estimated “**before**” observing the data

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{\sum P(D_1|H_k)P(H_k)}$$

事後確率 Posterior Probability

データを観察した”後に”推定した、箱が箱1である確率

Probability that the box is box 1 estimated “**after**” observing the data

ベイズの定理 Bayes Theorem

箱から取り出した玉が赤色だった The color of ball taken out from a box was red

箱1から玉を取り出した確率

Probability that the box was box 1

箱2から玉を取り出した確率

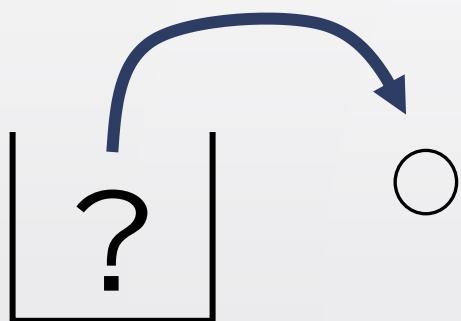
Probability that the box was box 2

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{\sum P(D_1|H_k)P(H_k)}$$

$$P(H_2|D_1) = \frac{P(D_1|H_2)P(H_2)}{\sum P(D_1|H_k)P(H_k)}$$

$$\frac{P(H_1|D_i)}{P(H_2|D_i)} = \frac{P(D_i|H_1)P(H_1)}{P(D_i|H_2)P(H_2)} \quad i = 1, 2$$

ベイズ更新 Bayesian Updating



箱からボールを取り出し、その色を確認する。

Take out one ball from a box and check its color

この操作を3回繰り返す。 Repeat this procedure three times



	箱1 Box 1	箱2 Box 2
赤玉 Red Ball	7	4
青玉 Blue Ball	3	6

箱が箱1である事後確率はどう変化するか？

How does the posterior probability that the box is box 1 change?

ベイズ更新 Bayesian Updating

	H_1	H_2	
	箱1 Box 1	箱2 Box 2	
D_1	赤玉 Red Ball	7	4
D_2	青玉 Blue Ball	3	6

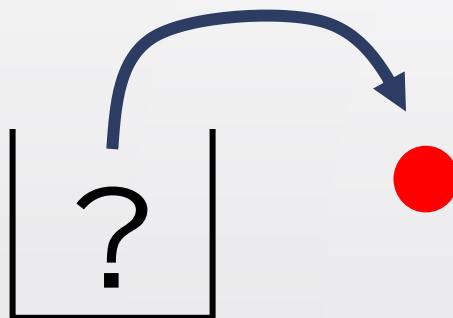
$$P(D_1|H_1) = \frac{7}{10}$$

$$P(D_2|H_1) = \frac{3}{10}$$

$$P(D_1|H_2) = \frac{4}{10}$$

$$P(D_2|H_2) = \frac{6}{10}$$

ベイズ更新-1回目 Bayesian Updating-1st round



$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{\sum P(D_1|H_k)P(H_k)} \quad P(H_2|D_1) = \frac{P(D_1|H_2)P(H_2)}{\sum P(D_1|H_k)P(H_k)}$$

玉を取り出す前は、どちらの箱か手がかりがない

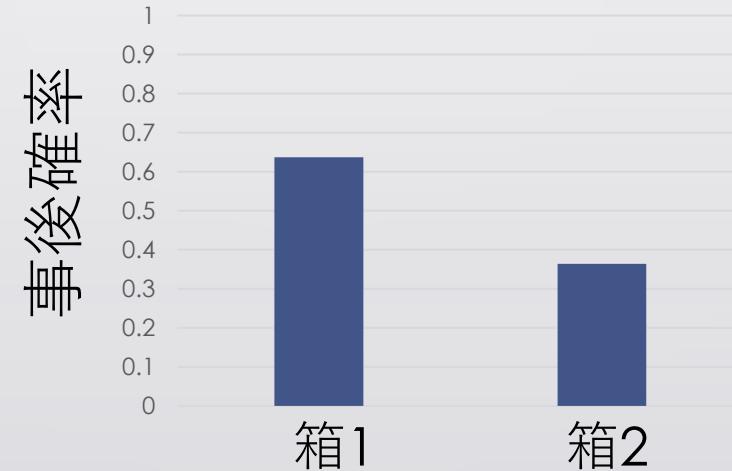
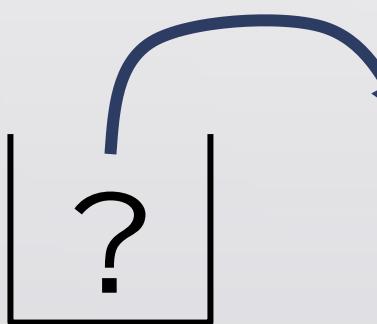
Before taking out a ball, we have no clue as to which box it is taken out from

$$P(H_1) = \frac{1}{2} \quad P(H_2) = \frac{1}{2}$$

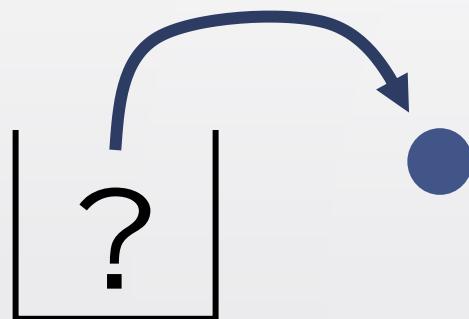
ベイズ更新-1回目 Bayesian Updating-1st round

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1|H_1)P(H_1) + P(D_1|H_2)P(H_2)} = \frac{7 \times 1}{7 \times 1 + 4 \times 1} = \frac{7}{11}$$

$$P(H_2|D_1) = \frac{P(D_1|H_2)P(H_2)}{P(D_1|H_1)P(H_1) + P(D_1|H_2)P(H_2)} = \frac{4}{11}$$



ベイズ更新-2回目 Bayesian Updating-2nd round



$$P(H_1|D_2) = \frac{P(D_2|H_1)P(H_1)}{\sum P(D_2|H_k)P(H_k)} \quad P(H_2|D_2) = \frac{P(D_2|H_2)P(H_2)}{\sum P(D_2|H_k)P(H_k)}$$

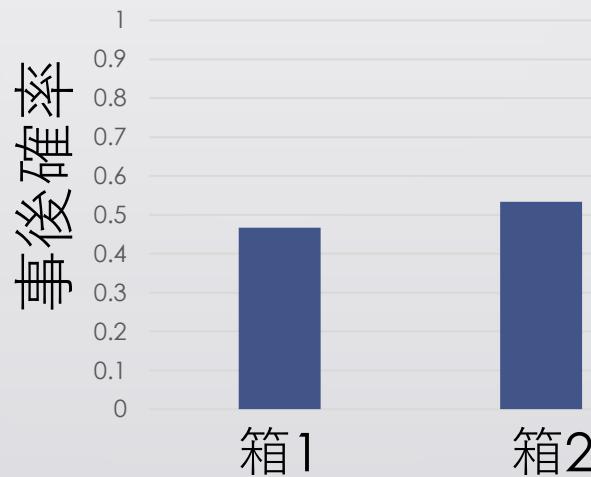
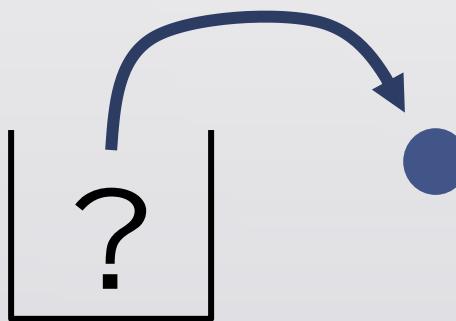
1回目のベイズ更新で計算した事後確率を、事前確率として用いる

Use as Prior Probability the posterior probabilities calculated in the first round of Bayesian updating

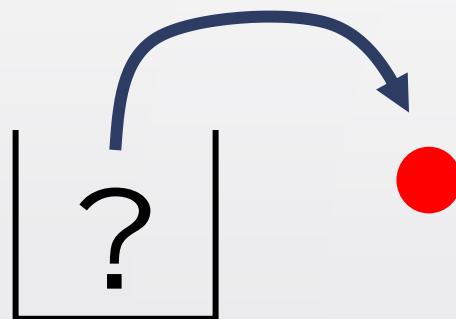
$$P(H_1) = \frac{7}{11} \quad P(H_2) = \frac{4}{11}$$

ベイズ更新-2回目 Bayesian Updating-2nd round

$$P(H_1|D_2) = \frac{P(D_2|H_1)P(H_1)}{P(D_2|H_1)P(H_1) + P(D_2|H_2)P(H_2)} = \frac{\frac{3}{10} \times \frac{7}{11}}{\frac{3}{10} \times \frac{7}{11} + \frac{6}{10} \times \frac{4}{11}}$$
$$= \frac{21}{21 + 24} = \frac{21}{45} \quad P(H_2|D_2) = \frac{24}{45}$$



ベイズ更新-3回目 Bayesian Updating-3rd round



$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{\sum P(D_1|H_k)P(H_k)} \quad P(H_2|D_1) = \frac{P(D_1|H_2)P(H_2)}{\sum P(D_1|H_k)P(H_k)}$$

2回目のベイズ更新で計算した事後確率を、事前確率として用いる

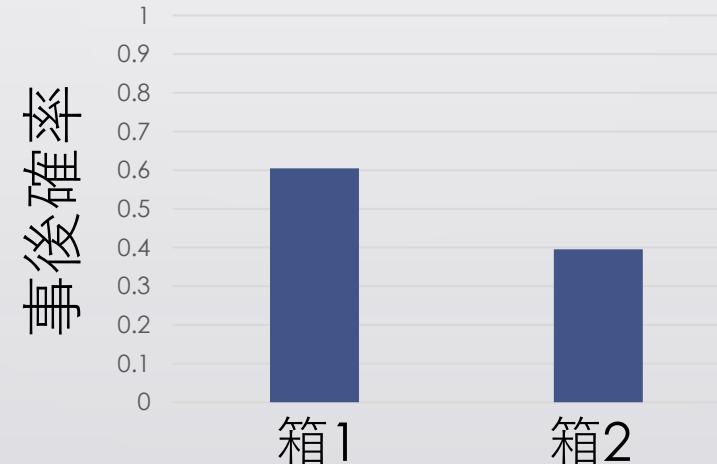
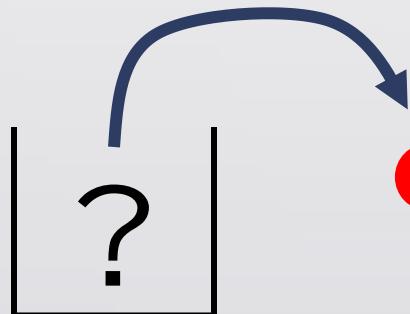
Use as Prior Probability the posterior probabilities calculated in the second round of Bayesian updating

$$P(H_1) = \frac{21}{45} \quad P(H_2) = \frac{24}{45}$$

ベイズ更新-3回目 Bayesian Updating-3rd round

$$P(H_1|D_1) = \frac{P(D_1|H_1)P(H_1)}{P(D_1|H_1)P(H_1) + P(D_1|H_2)P(H_2)} = \frac{21 \times 7}{21 \times 7 + 24 \times 4} = \frac{147}{243}$$

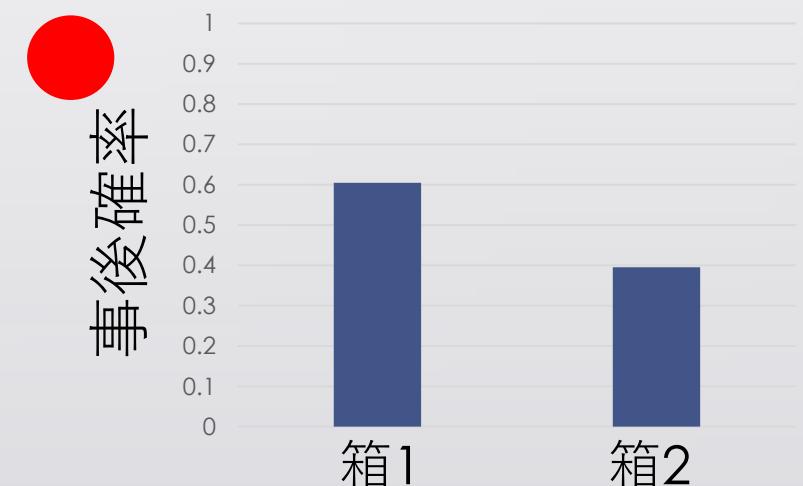
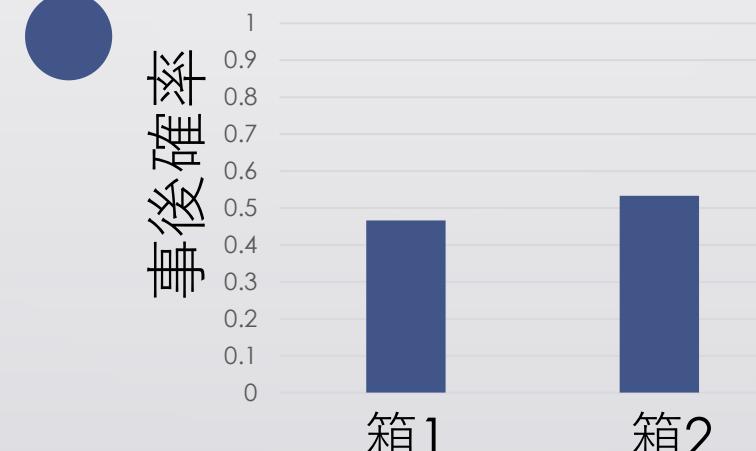
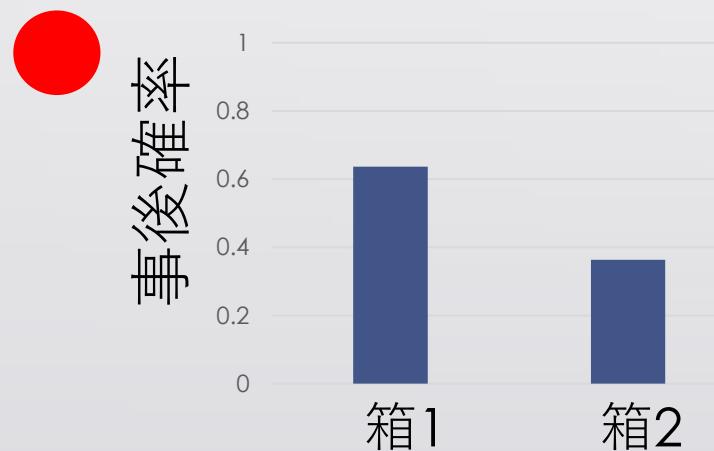
$$P(H_2|D_2) = \frac{96}{243}$$





ベイズ更新 Bayesian Updating

	箱1 Box 1	箱2 Box 2
赤玉 Red Ball	7	4
青玉 Blue Ball	3	6



スパムメールの分類 Classification of spam mail

メールに“秘密”“技術”“大当たり”という3つの単語が含まれていた。

An e-mail contains three words, “Secret”, “Technology” and “Jackpot”

このメールはスパムメールだろうか？

Is this e-mail a spam mail?

spam



<https://pc-yougo.com/spam-mail/>

ナイーブベイズ Naïve Bayes

スパムメールかどうかの事後確率をベイズ更新で計算する



?

“秘密”
Secret

“技術”
Technology

“大当たり”
Jackpot

ナイーブベイズ Naïve Bayes

スパムメール Spam Mail	普通のメール Authentic Mail
60%	40%

$$P_1(H_s) = 0.6$$

$$P_1(H_a) = 0.4$$

	スパムメール Spam Mail	普通のメール Authentic Mail
単語1 Word 1 “秘密” Secret	60%	40%
単語2 Word 2 “技術” Technology	20%	80%
単語3 Word 3 “大当たり” Jackpot	90%	10%

$$P(W_1|H_s) = 0.6, P(W_1|H_a) = 0.4$$

$$P(W_2|H_s) = 0.2, P(W_2|H_a) = 0.8$$

$$P(W_3|H_s) = 0.9, P(W_3|H_a) = 0.1$$

ナイーブベイズ Naïve Bayes

メールの文章に、単語 1 “秘密”が含まれていた

Contents in an e-mail contained word 1 “Secret”

そのメールがスパムである事後確率

Posterior probability that the e-mail is
a spam

$$P_1(H_s|W_1) = \frac{P(W_1|H_s)P_1(H_s)}{P_1(W_1)}$$

そのメールが普通のメールである事後確率

Posterior probability that the e-mail is
authentic one

$$P_1(H_a|W_1) = \frac{P(W_1|H_a)P_1(H_a)}{P_1(W_1)}$$

ナイーブベイズ Naïve Bayes

$$P_1(H_s|W_1) = \frac{P(W_1|H_s)P_1(H_s)}{P_1(W_1)} \quad P_1(H_a|W_1) = \frac{P(W_1|H_a)P_1(H_a)}{P_1(W_1)}$$

$P_1(W_1)$: 受信したメールの文章に単語 1 “秘密”が含まれる確率

Probability that contents of a received e-mail contain word 1 “secret”

両式から $P_1(W_1)$ を消去する Remove $P_1(W_1)$ from both equations

$$\frac{P_1(H_s|W_1)}{P_1(H_a|W_1)} = \frac{P(W_1|H_s)P_1(H_s)}{P(W_1|H_a)P_1(H_a)}$$

ナイーブベイズ Naïve Bayes

メールの文章に、単語 2 “技術”が含まれていた
Contents in an e-mail contained word 2 “Technology”

$$P_1(H_s|W_1) = \frac{P(W_1|H_s)P_1(H_s)}{P_1(W_1)}$$

$$\frac{P_2(H_s|W_2)}{P_2(H_a|W_2)} = \frac{P(W_2|H_s)P_2(H_s)}{P(W_2|H_a)P_2(H_a)}$$

$$P_1(H_a|W_1) = \frac{P(W_1|H_a)P_1(H_a)}{P_1(W_1)}$$

ナナイーブベイズ Naïve Bayes

メールの文章に、単語 2 “技術”が含まれていた
Contents in an e-mail contained word 2 “Technology”

$$\frac{P_2(H_s|W_2)}{P_2(H_a|W_2)} = \frac{P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_2|H_a)P(W_1|H_a)P_1(H_a)}$$

メールの文章に、単語 3 “大当たり”が含まれていた
Contents in an e-mail contained word 3 “Jackpot”

$$\frac{P_3(H_s|W_3)}{P_3(H_a|W_3)} = \frac{P(W_3|H_s)P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_3|H_a)P(W_2|H_a)P(W_1|H_a)P_1(H_a)}$$

ナイーブベイズ Naïve Bayes

メールに“秘密”“技術”“大当たり”という 3 つの単語が含まれていた。

An e-mail contains three words, “Secret”, “Technology” and “Jackpot”

$$\frac{P_3(H_s|W_3)}{P_3(H_a|W_3)} = \frac{P(W_3|H_s)P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_3|H_a)P(W_2|H_a)P(W_1|H_a)P_1(H_a)} = \frac{628 \times 10^{-4}}{128 \times 10^{-4}}$$

$$Probability\ of\ being\ a\ spam = \frac{628}{628 + 128} = 0.83$$

ナイーブベイズ Naïve Bayes

メールに $\{W_1, W_2, \dots, W_n\}$ という n 個の単語が含まれていた。

An e-mail contains n words, $\{W_1, W_2, \dots, W_n\}$

$$P(H_s | W_1, W_2, \dots, W_n) = \frac{P(W_1, W_2, \dots, W_n | H_s)P(H_s)}{P(W_1, W_2, \dots, W_n)}$$

$$\begin{aligned} P(W_1, W_2, \dots, W_n | H_s)P(H_s) &= P(W_1, W_2, \dots, W_{n-1} | W_n, H_s)P(W_n | H_s)P(H_s) \\ &= P(W_1, W_2, \dots, W_{n-2} | W_{n-1}, W_n, H_s)P(W_{n-1} | W_n, H_s)P(W_n | H_s)P(H_s) \\ &= P(W_1, W_2, \dots, W_{n-3} | W_{n-2}, W_{n-1}, W_n, H_s)P(W_{n-2} | W_{n-1}, W_n, H_s)P(W_{n-1} | W_n, H_s)P(W_n | H_s)P(H_s) \\ &\dots \end{aligned}$$

ナイーブベイズ Naïve Bayes

$$P(W_1, W_2, \dots, W_n | H_s) P(H_s)$$

$$= P(W_1, W_2, \dots, W_{n-3} | W_{n-2}, W_{n-1}, W_n, H_s) P(W_{n-2} | W_{n-1}, W_n, H_s) P(W_{n-1} | W_n, H_s) P(W_n | H_s) P(H_s)$$

スパムメールに単語 W_j が出現する条件付確率は、他の単語 $\{W_1, W_2, \dots, W_{j-1}, W_{j+1}, \dots, W_n\}$ とは独立であると仮定すると

Under the assumption that the conditional probability of occurrence of the word W_j in a spam-mail is independent of the occurrence of the other words $\{W_1, W_2, \dots, W_{j-1}, W_{j+1}, \dots, W_n\}$

$$P(W_j | W_{j+1} \dots W_n, H_s) = P(W_j | H_s)$$

ナイーブベイズ Naïve Bayes

“単純化した”仮定 Naïve Assumption

スパムメールに単語 W_j が出現する条件付確率は、他の単語 $\{W_1, W_2, \dots, W_{j-1}, W_{j+1}, \dots, W_n\}$ とは独立であると仮定すると

Under the assumption that the conditional probability of occurrence of the word W_j in a spam-mail is independent of the occurrence of the other words $\{W_1, W_2, \dots, W_{j-1}, W_{j+1}, \dots, W_n\}$

$$P(W_1, W_2, \dots, W_n | H_s) P(H_s)$$

$$= P(W_1 | H_s) \cdots P(W_n | H_s) P(H_s) = P(H_s) \prod_{j=1}^n P(W_j | H_s)$$

ナイーブベイズ Naïve Bayes

メールに $\{W_1, W_2, \dots, W_n\}$ という n 個の単語が含まれていた。
An e-mail contains n words, $\{W_1, W_2, \dots, W_n\}$

“単純化した”仮定をおくと Under the naïve assumption

$$P(H_s|W_1, W_2, \dots, W_n) = \frac{P(H_s) \prod_{j=1}^n P(W_j|H_s)}{P(W_1, W_2, \dots, W_n)} \quad P(H_a|W_1, W_2, \dots, W_n) = \frac{P(H_a) \prod_{j=1}^n P(W_j|H_a)}{P(W_1, W_2, \dots, W_n)}$$

$$\frac{P(H_s|W_1, W_2, \dots, W_n)}{P(H_a|W_1, W_2, \dots, W_n)} = \frac{P(H_s) \prod_{j=1}^n P(W_j|H_s)}{P(H_a) \prod_{j=1}^n P(W_j|H_a)}$$

ナイーブベイズ Naïve Bayes

メールに“秘密”“技術”“大当たり”という3つの単語が含まれていた。

An e-mail contains three words, “Secret”, “Technology” and “Jackpot”

$$\frac{P_3(H_s|W_3)}{P_3(H_a|W_3)} = \frac{P(W_3|H_s)P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_3|H_a)P(W_2|H_a)P(W_1|H_a)P_1(H_a)} = \frac{628 \times 10^{-4}}{128 \times 10^{-4}}$$

$$Probability\ of\ being\ a\ spam = \frac{628}{628 + 128} = 0.83$$

スパムメールと判定するかどうかは閾値による

It depends on the threshold whether the e-mail is judged to be a spam or not

閾値と偽陽性 Threshold and False Positives

正解 Answer

判定
Judgment

	スパムメール Spam Mail	普通のメール Authentic Mail
スパムメール Spam Mail	真陽性 True Positive	偽陽性 False Positive
普通のメール Authentic Mail	偽陰性 False Negative	真陰性 True Negative

スパムと判定する閾値を下げる → 偽陽性率が上がる
Lowering threshold for judging to be a spam Higher false positive rate



データマイニング

Data Mining

8: 分類③ Classification

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

ロジスティック回帰 Logistic Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M$$

回帰をクラス分類に適用する

Sold within 1 year or not = $\beta_1 MedInc + \beta_2 HouseAge + \dots \beta_8 Longitude$

Yes: 1, No: 0

ダミー変数 Dummy Variable

House Value = $\beta_1 MedInc + \beta_2 HouseAge + \dots \beta_8 Longitude$

量的変数 Quantitative Variable

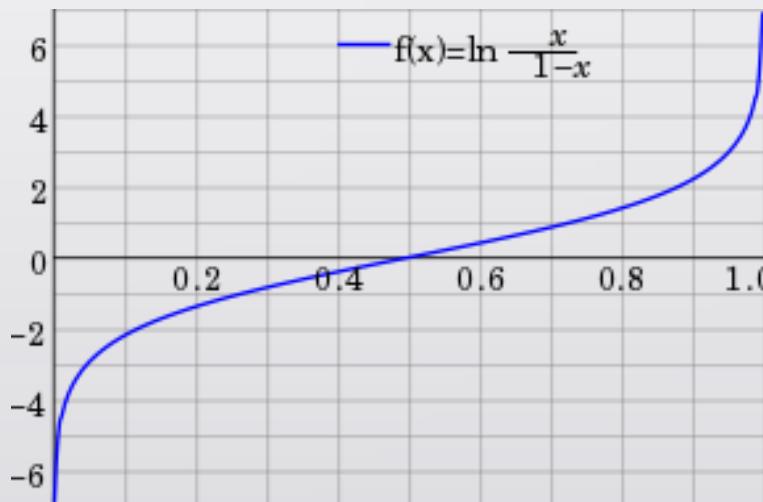
オッズとロジット Odds and Logit

$$Odds = \frac{P(Y = 1)}{1 - P(Y = 1)}$$

ある事象 ($Y = 1$) が起こる確率と、
起こらない確率の比

Ratio between the probability that an event ($Y = 1$)
occurs and the probability that it does not occur

$$Logit(P) = \log \frac{P(Y = 1)}{1 - P(Y = 1)}$$



[https://en.wikipedia.org/
wiki/Logit](https://en.wikipedia.org/wiki/Logit)



ロジスティック回帰 Logistic Regression

$$\log \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M$$

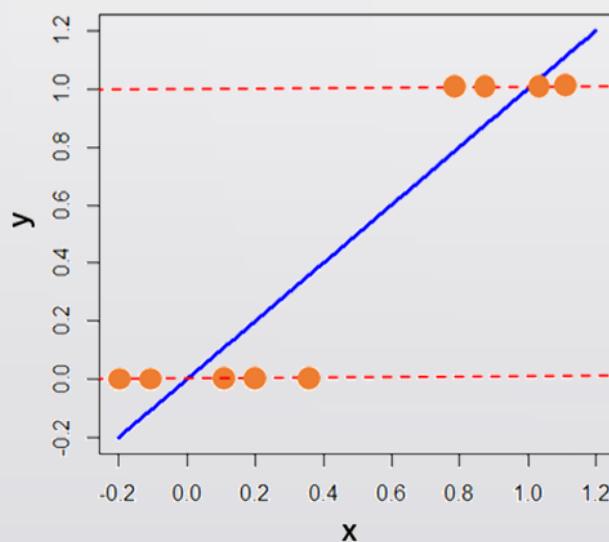
$$\frac{P}{1 - P} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M}$$

$$P = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M)}}$$

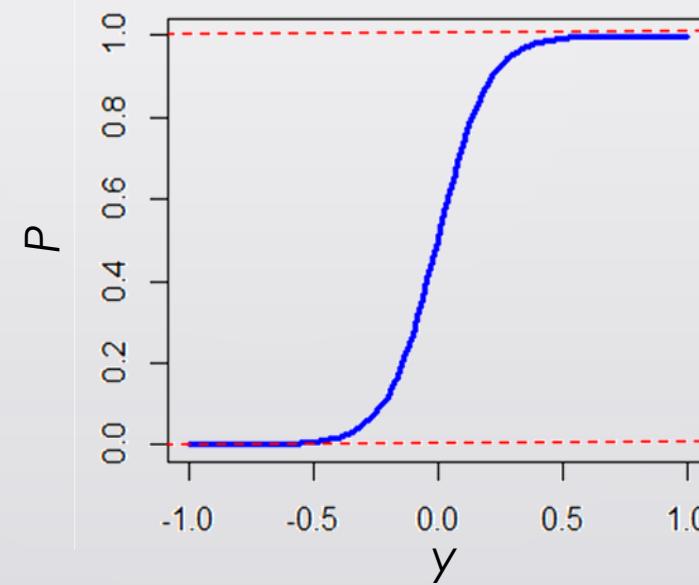
ロジスティック回帰 Logistic Regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M \quad P = \frac{1}{1 + e^{-y}}$$

変換前 Before Conversion



変換後 Before Conversion



● ターゲット変数
Target Variable

<https://bellcurve.jp/statistics/course/26934.html>



最尤推定法 Maximum Likelihood Estimation

与えられたデータが観測される確率が最大になるよう回帰係数 β_k を決定する

Determine regression coefficient β_k so as to maximize the probability that given data is observed

$$Y_i = \begin{cases} 1 & P_i = p(Y_i = 1) \quad i\text{番目のデータ } Y_i \text{ が } 1 \text{ である確率} \\ 0 & \text{Probability that } i\text{-th data } Y_i \text{ is } 1 \end{cases}$$

$$L = \prod_{i=1}^N P_i^{Y_i} (1 - P_i)^{1-Y_i} \quad \begin{array}{l} L \text{を最大化する} \\ \text{Maximize } L \end{array}$$



最尤推定法 Maximum Likelihood Estimation

$$L = \prod_{i=1}^N P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

Lを最大化する
Maximize L

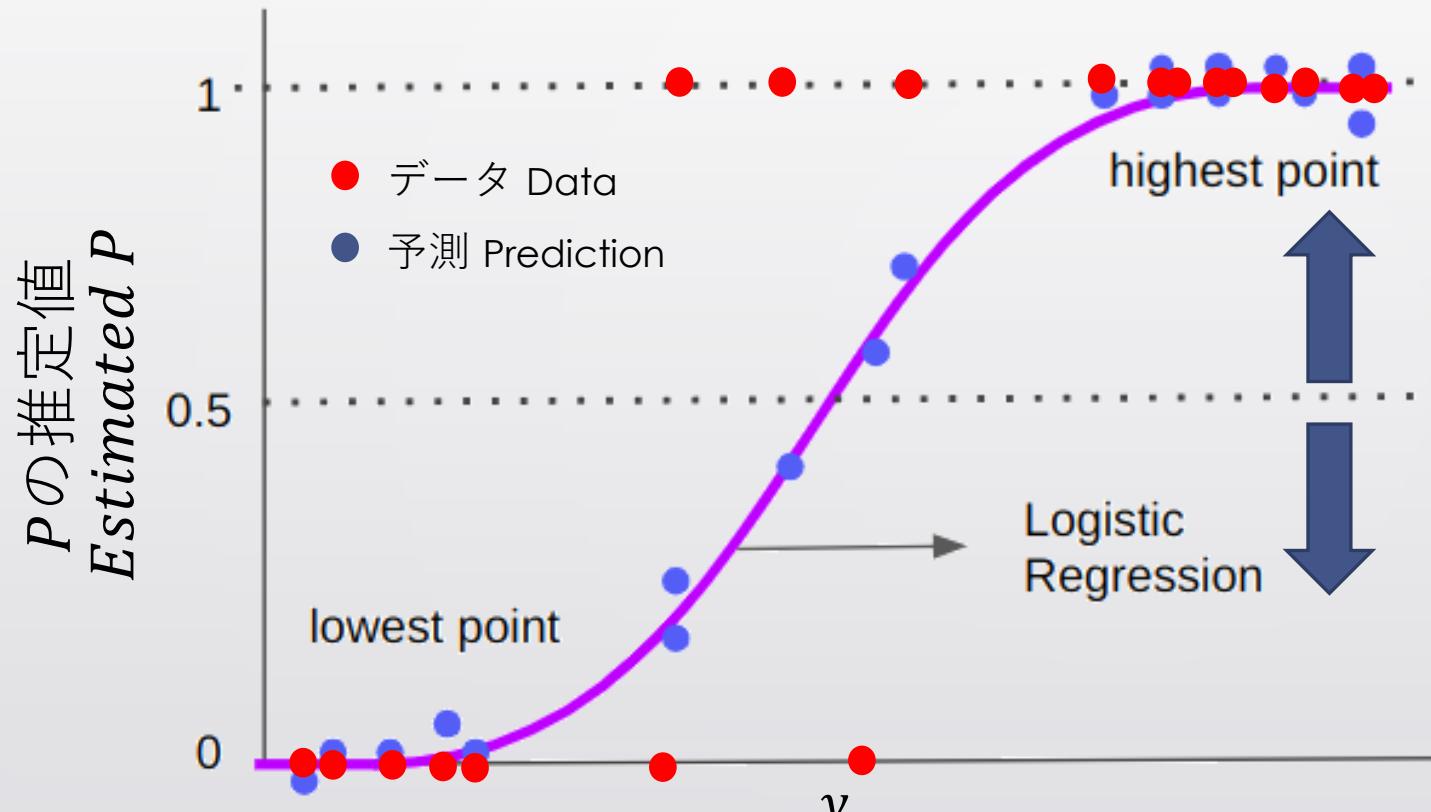
$$\underline{\log(L)} = \sum_{i=1}^N \{\log(P_i)Y_i + \log(1 - P_i)(1 - Y_i)\}$$

対数尤度関数

Log-likelihood function

ニュートン・ラフソン法で回帰係数をもとめる

閾値の設定 Setting Threshold



$P \geq$ 閾値ならばデータがクラス 1 に分類される

Data is classified into Class 1 if
 $P \geq Threshold$

threshold

どのように閾値を設定すればいいか？

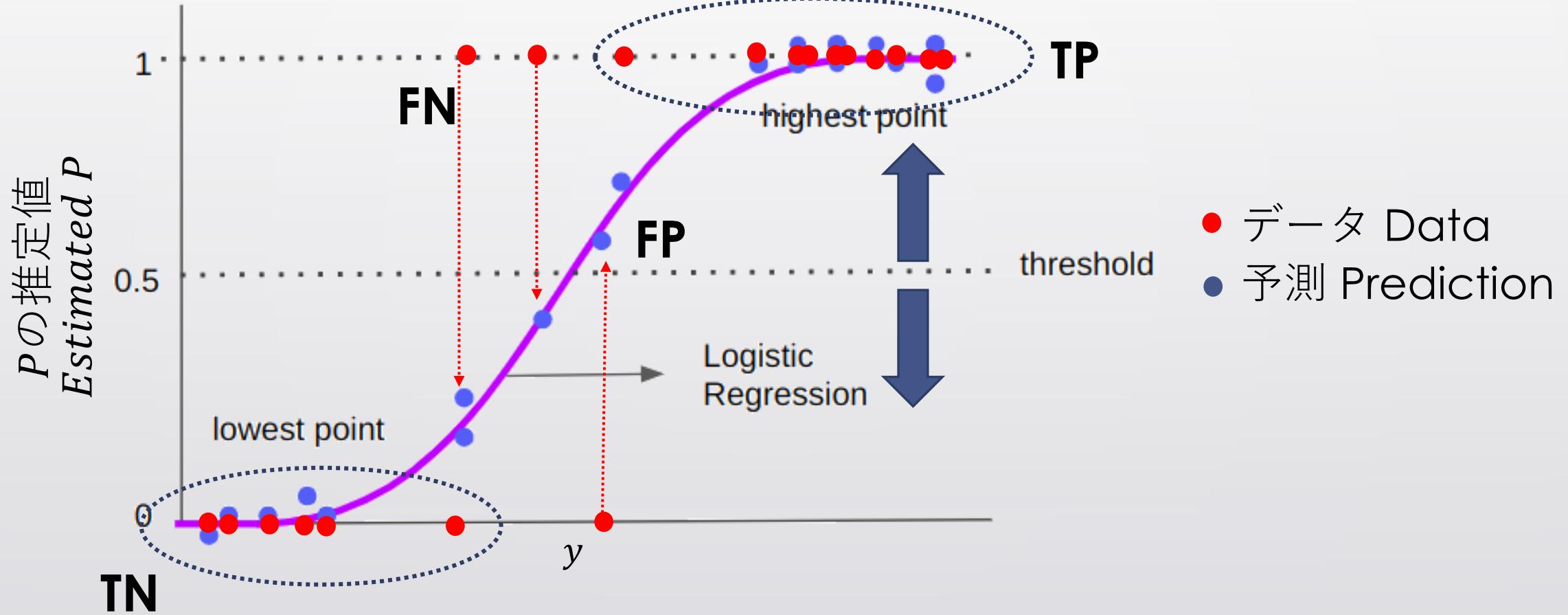
How should we set the threshold?

<https://stackabuse.com/definitive-guide-to-logistic-regression-in-python/>

混同行列 Confusion Matrix

		正解 Answer	
		クラス1 Class 1	クラス0 Class 0
分類 Classification	クラス1 Class 1	真陽性 (TP) True Positive (TP)	偽陽性 (FP) False Positive (FP)
	クラス0 Class 0	偽陰性 (FN) False Negative (FN)	真陰性 (TN) True Negative (TN)

ロジスティック回帰と混同行列





分類性能の評価 Evaluation of Classification Performance

正答率

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Value = [1, 1, 1, 0, 1, 1, 1, 1, 0, 1]

Prediction = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

$$Accuracy = 0.8$$



分類性能の評価 Evaluation of Classification Performance

感度 再現率 TP
 $Sensitivity = Recall = \frac{TP}{TP + FN}$

陽性データが正しく陽性と判定される確率

The probability that positive data is correctly classified as “positive”

適合率 陽性的中率 TP
 $Precision = Positive Predictive Value = \frac{TP}{TP + FP}$

陽性と判定されたデータが実際に陽性である確率

The probability that data classified as “positive” is truly positive



分類性能の評価 Evaluation of Classification Performance

特異度

$$Specificity = \frac{TN}{TN + FP}$$

陰性データを正しく陰性と判定する確率

The probability that negative data is correctly classified as “negative”

$$1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} = \begin{matrix} \text{偽陽性率} \\ \text{False Positive Rate} \end{matrix}$$



分類性能の評価 Evaluation of Classification Performance

$$Sensitivity = Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP}$$

再現率と適合率の間にはトレードオフがある

There is a trade-off between recall and precision

$$F1 = \text{再現率と適合度の調和平均} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = \frac{2Recall \times Precision}{Recall + Precision}$$

Harmonic mean of recall and precision

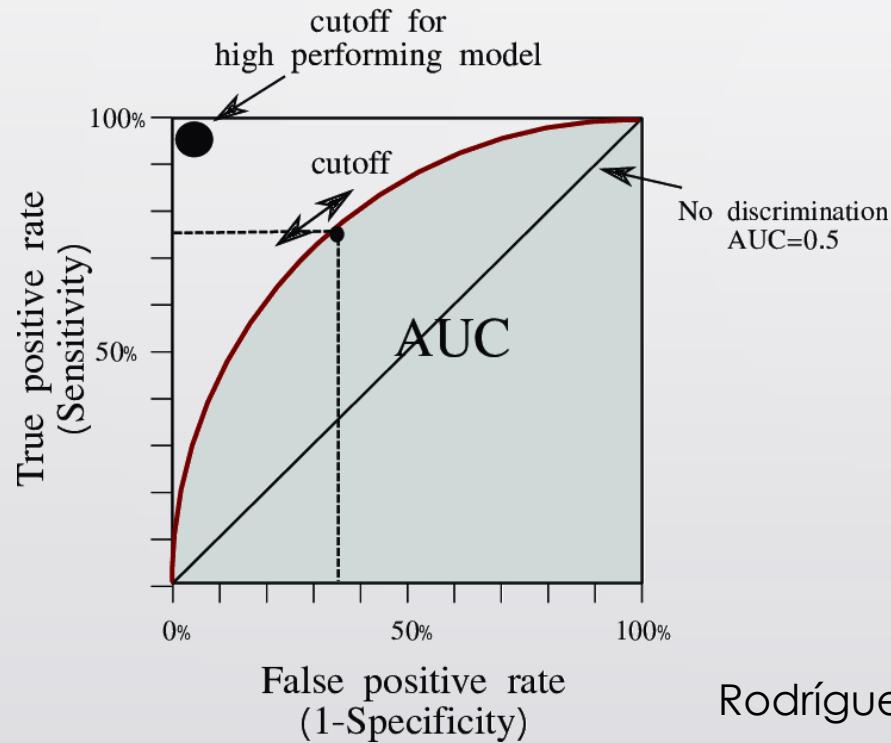
F1値は再現率と適合度のバランスを反映する

F1-value represents balance between recall and precision

ROC曲線 ROC(Receiver-Operator Characteristics) Curve

良い分類器は、感度が高く偽陽性率が低い

A good classifier has high sensitivity and low false positive rate

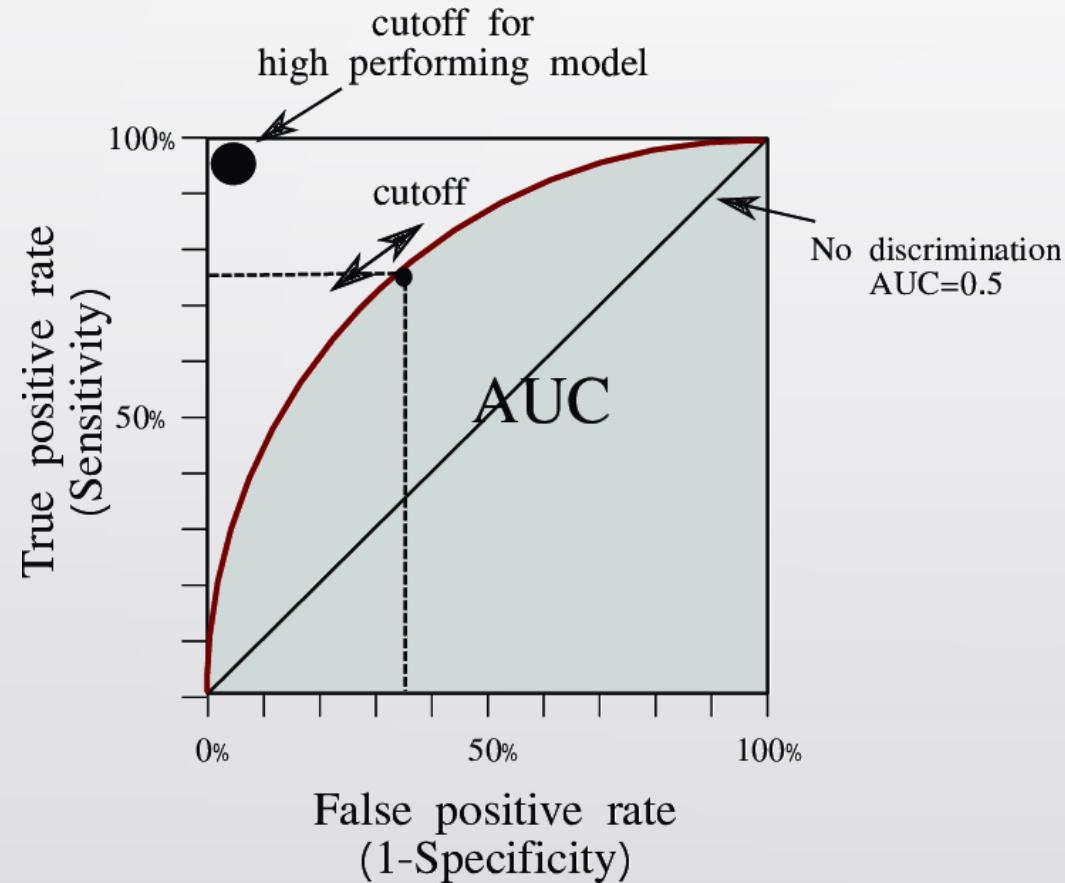


ROC曲線は、異なる閾値における感度・偽陽性率を表す

ROC represents relationship between sensitivity and false positive rate under varying threshold

Rodríguez-Hernández et al, 2021

ROC曲線 ROC(Receiver-Operator Characteristics) Curve



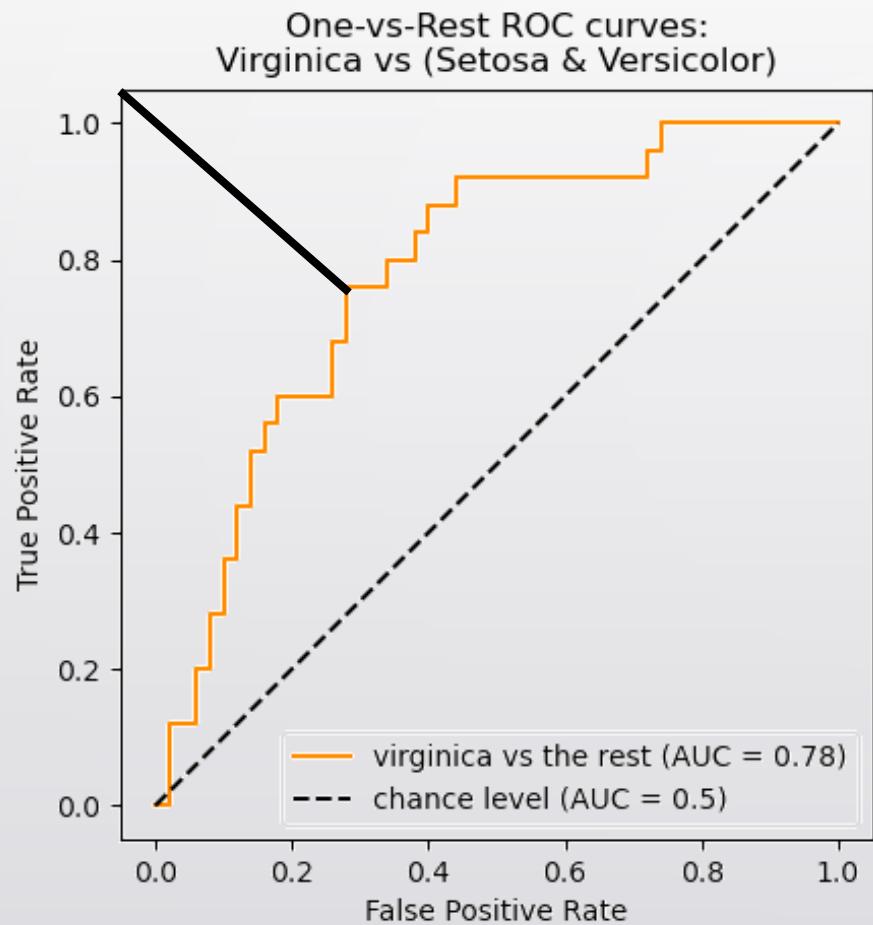
AUC: Area Under Curve

AUCが大きいほど、分類器の性能が良い

Larger AUC indicates better performance of classifier

AUC	
0.9 - 1.0	High accuracy
0.9 - 0.7	Moderate accuracy
0.5 - 0.7	Low accuracy

カットオフの決定方法 How to determine “Cut-Off”

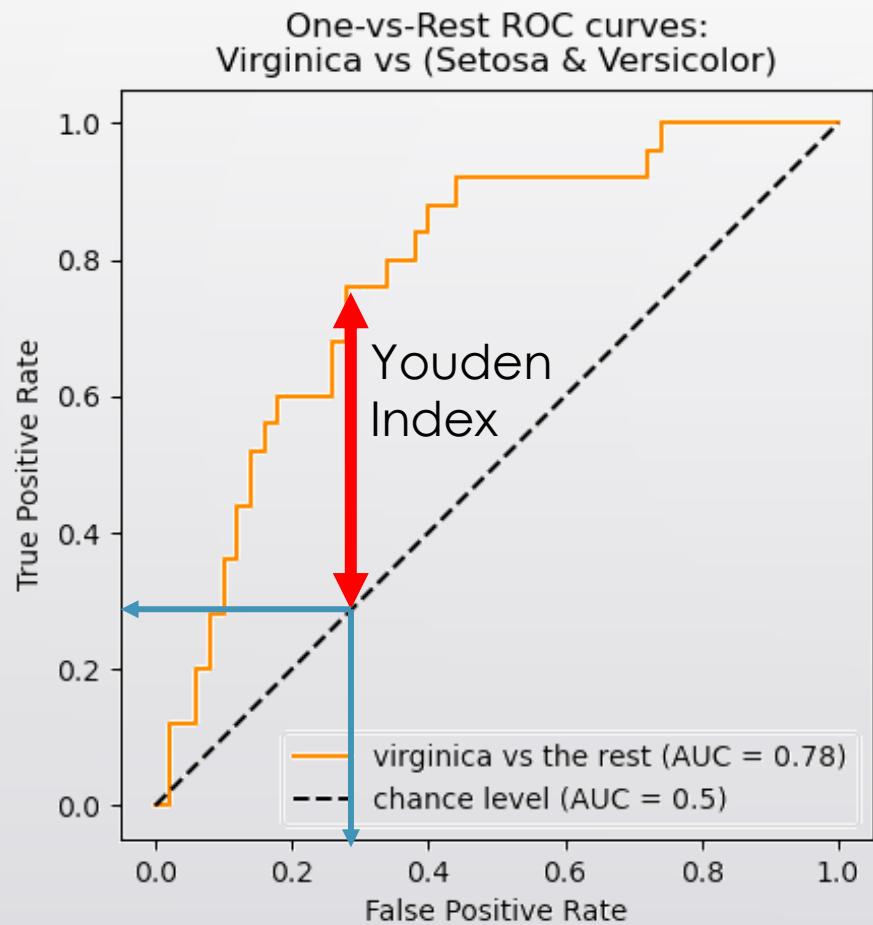


最適な性能との距離が最小になる閾値

The threshold at which distance from the optimal performance is minimized

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

カットオフの決定方法 How to determine “Cut-Off”

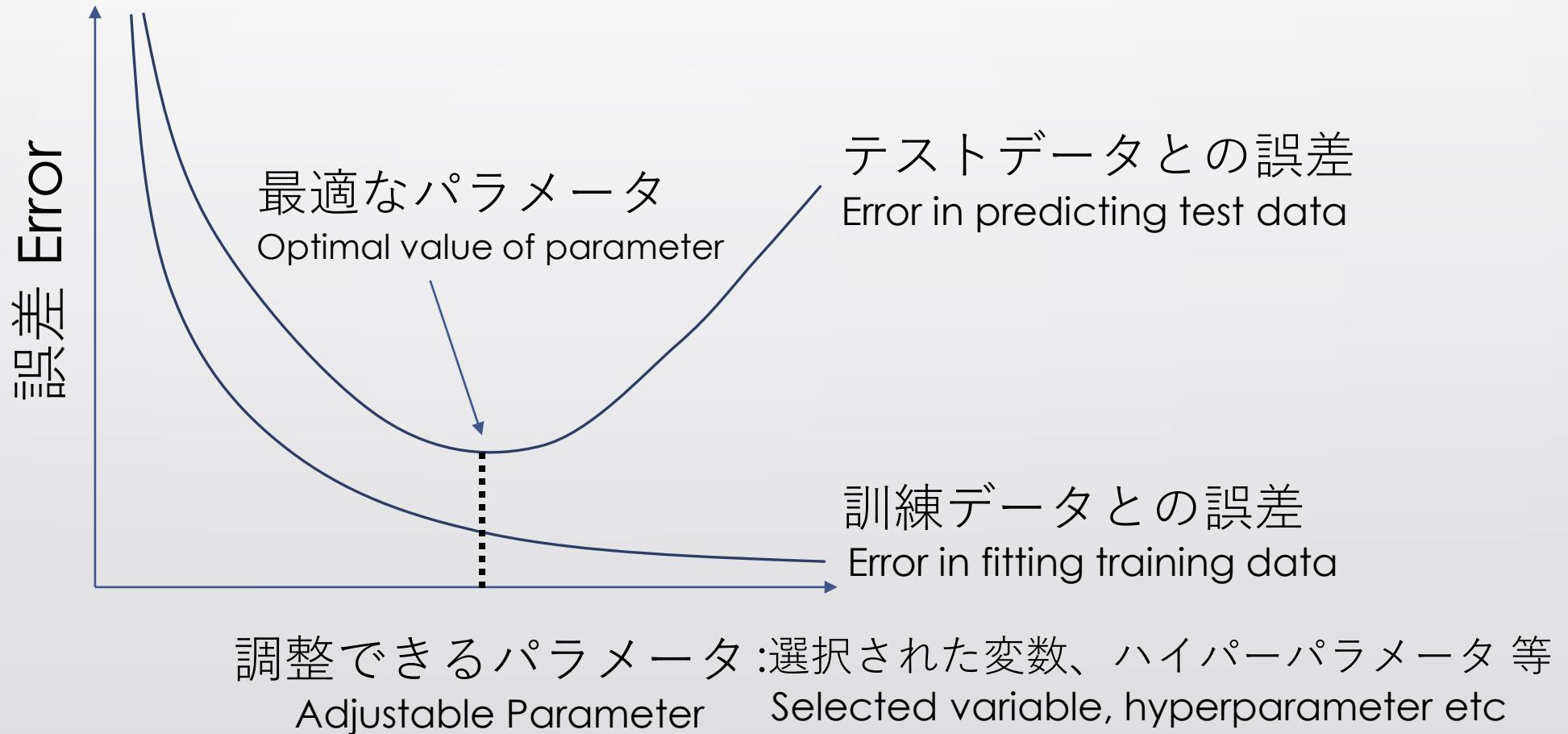


Youden Indexが最大になる閾値
The threshold at which Youden Index is maximized

$$\begin{aligned} \text{Youden Index} &= \text{Sensitivity} - \text{False Positive Rate} \\ &= \text{Sensitivity} - (1 - \text{Specificity}) \\ &= \text{Sensitivity} + \text{Specificity} - 1 \end{aligned}$$

https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

過学習のサイン Signs of Overfitting





交差検証 Cross validation

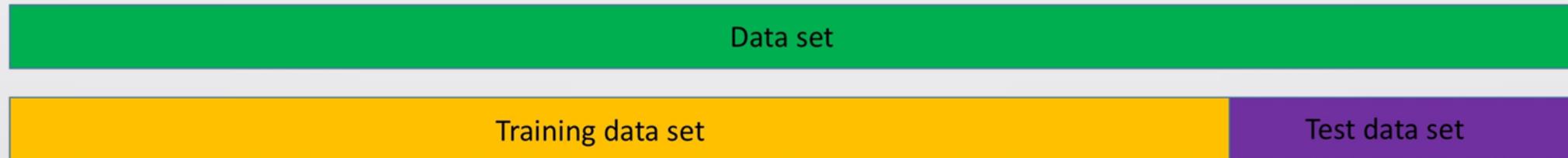
1. データを学習(訓練)データとテストデータに分割する
Splitting data into training and test data
2. 学習(訓練)データを使って分類モデルを作る
Create classification model based on training data
3. 分類モデルの予測性能をテストデータで検証する
Evaluate prediction performance of classification model using test data

ホールドアウト法 Hold-out Method

データを一定の比率で学習データとテストデータに分割し性能検証を行う

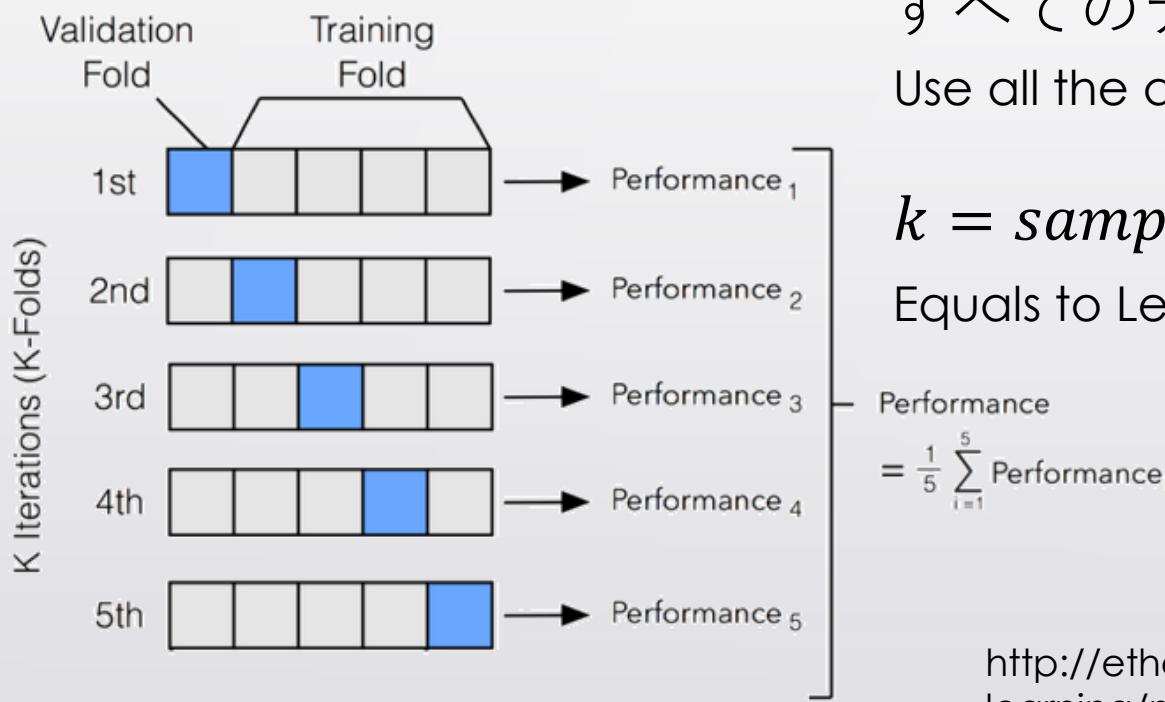
Evaluate classification performance by dividing the dataset to training/test data with certain proportion

Hold-out



<https://qiita.com/ZESSU/items/8aaad3cdfeae35fa0820>

k -分割交差検証 k -fold cross validation



すべてのデータを学習/テストデータとして使用する
Use all the data as training/test data

$k = \text{sample size}$ の時がLeave-one-out交差検証
Equals to Leave-one-out cross validation when $k = \text{sample size}$

$$\text{Performance} = \frac{1}{5} \sum_{i=1}^5 \text{Performance}_i$$

http://ethen8181.github.io/machine-learning/model_selection/model_selection.html



ロジスティック回帰 Logistic Regression

$$\log \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M$$

$$\frac{P}{1 - P} = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M}$$

$$P = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots \beta_M x_M)}}$$



多クラス分類への拡張 Extension to Multiclass Classification

データが複数に属する可能性がある場合

When a data can belong to multiple classes simultaneously

各クラスごとに回帰モデルを作る

Estimate regression model for each class

$$\log \frac{P_{C_k}}{1 - P_{C_k}} = \beta_{1,C_k} x_1 + \beta_{2,C_k} x_2 + \beta_{3,C_k} x_3 \dots \beta_{M,C_k} x_M$$

P_{C_k} : データがクラス C_k に属する確率

The probability that data belongs to class C_k

多クラス分類への拡張 Extension to Multiclass Classification

データが一つのクラスのみに属する場合

When a data can be classified into only one class

ソフトマックス関数で各データが観測される確率を計算する

Compute the probability that each data is observed by softmax function

$$f_{C_k} = \beta_{1,C_k} x_1 + \beta_{2,C_k} x_2 + \beta_{3,C_k} x_3 \dots \beta_{M,C_k} x_M$$

$$P_{C_k} = \frac{\exp(f_{C_k})}{\sum_1^K \exp(f_{C_i})}$$

K: クラスの総数
Total number of classes



データマイニング

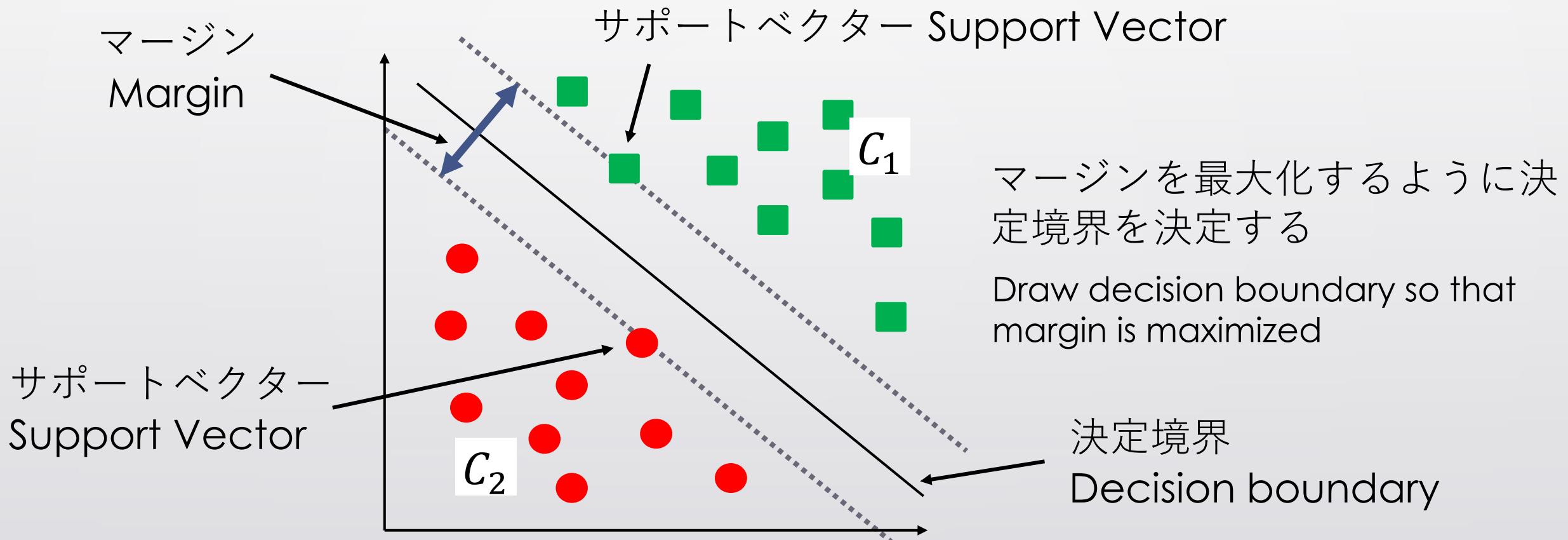
Data Mining

9: 分類④ Classification

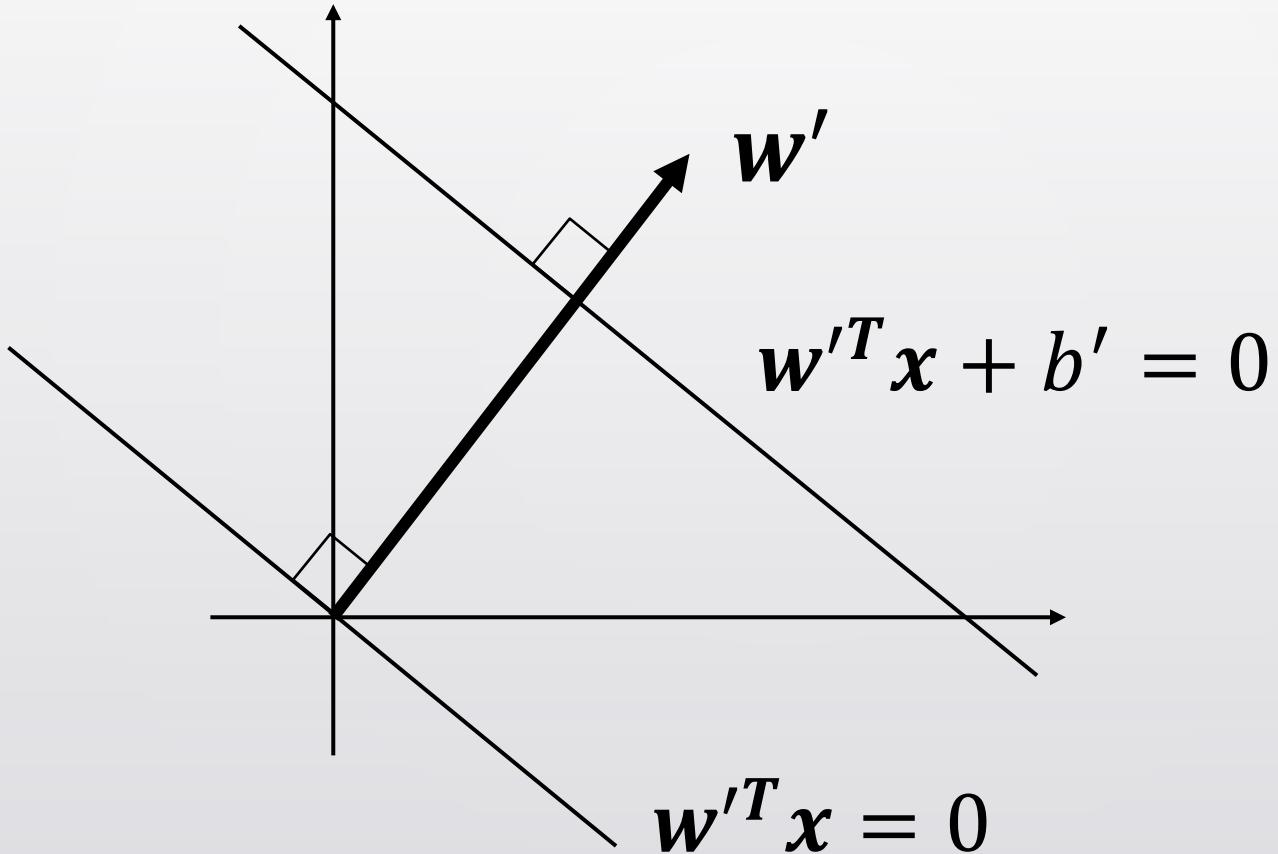
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

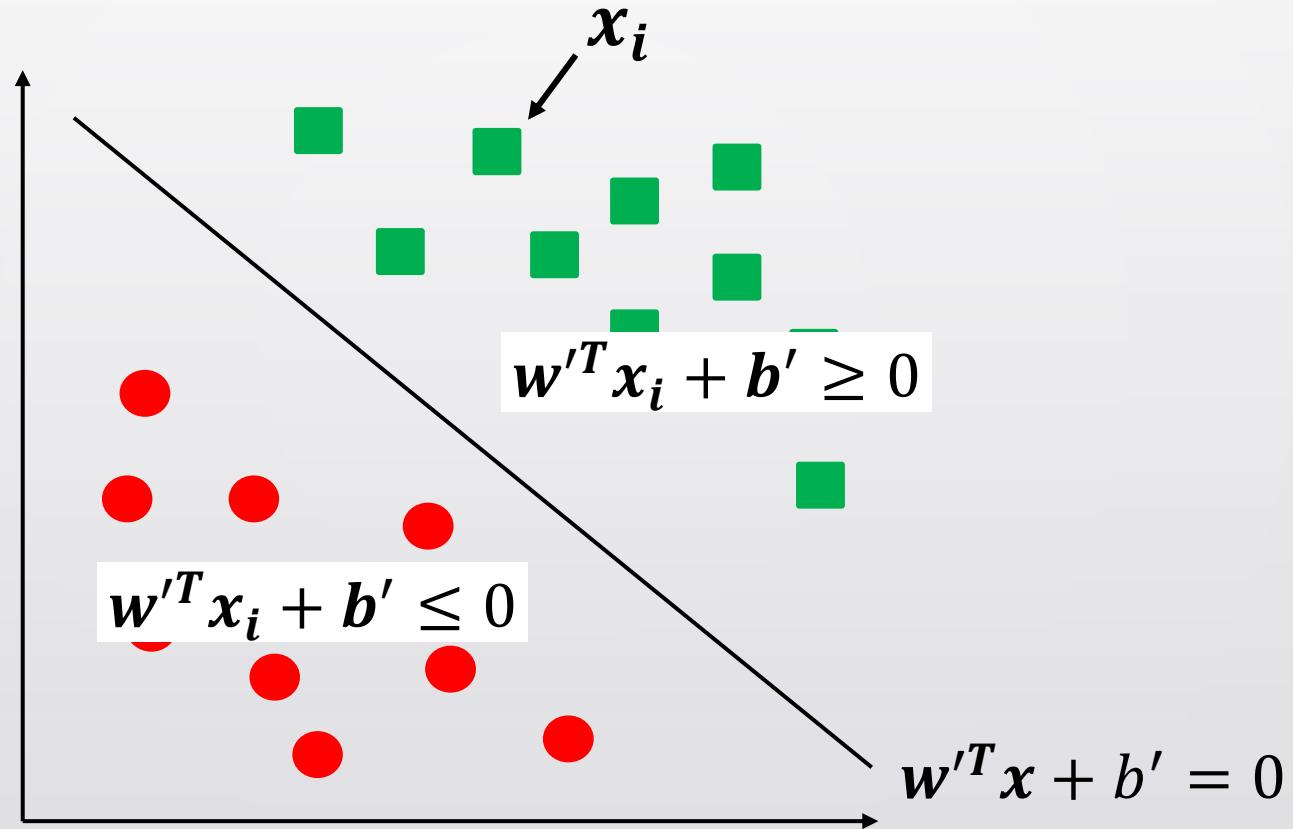
サポートベクターマシン Support Vector Machine



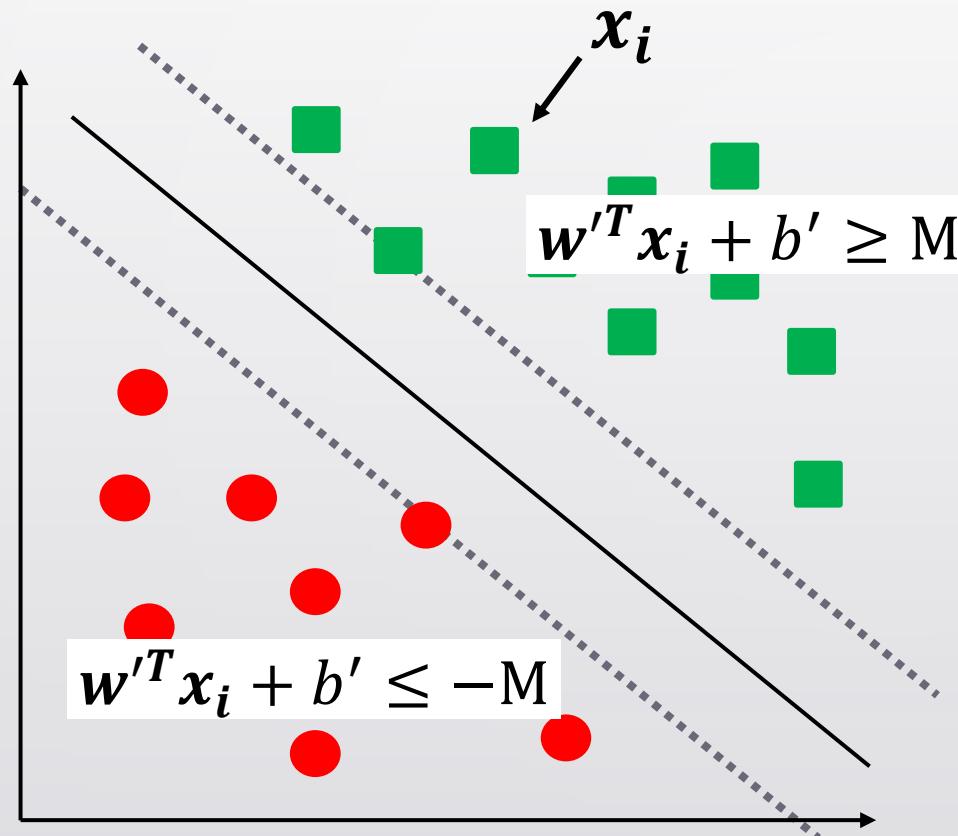
平面の方程式 Equation of a plane



サポートベクターマシン Support Vector Machine



サポートベクターマシン Support Vector Machine



$x_i \in C_1$ の場合 $w'^T x_i + b' \geq M$

In case of $x_i \in C_1$

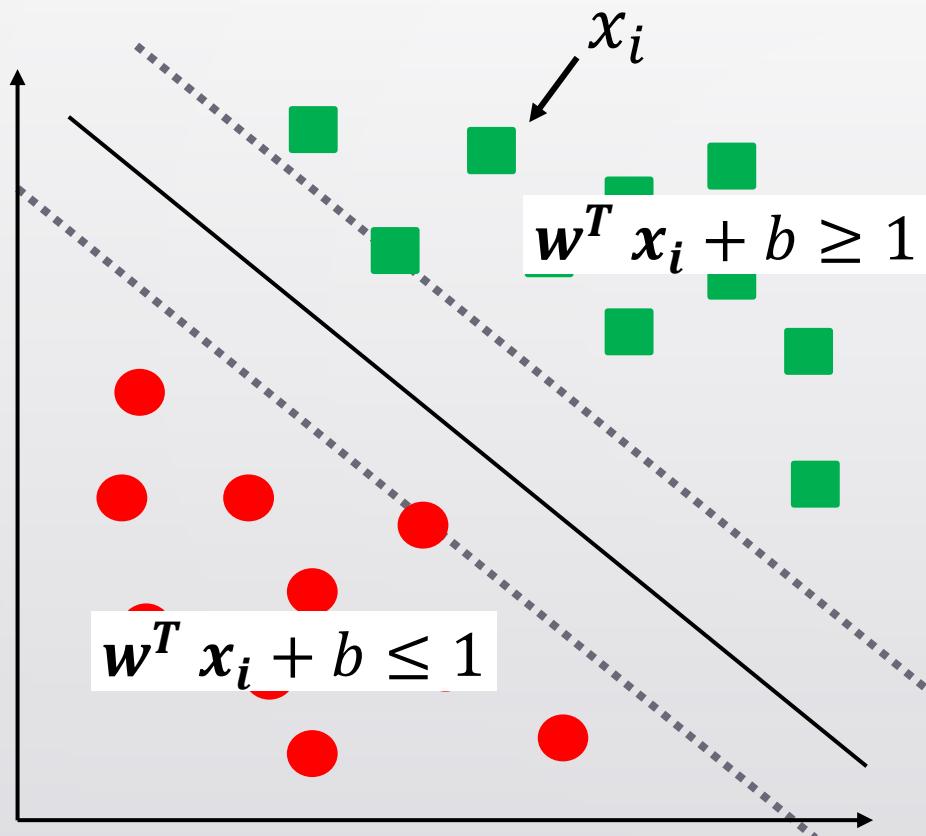
$x_i \in C_2$ の場合 $w'^T x_i + b' \leq -M$

In case of $x_i \in C_2$

すべての x_i に対して For all x_i

$$|w'^T x_i + b'| \geq M$$

サポートベクターマシン Support Vector Machine



すべての x_i に対して As for all x_i

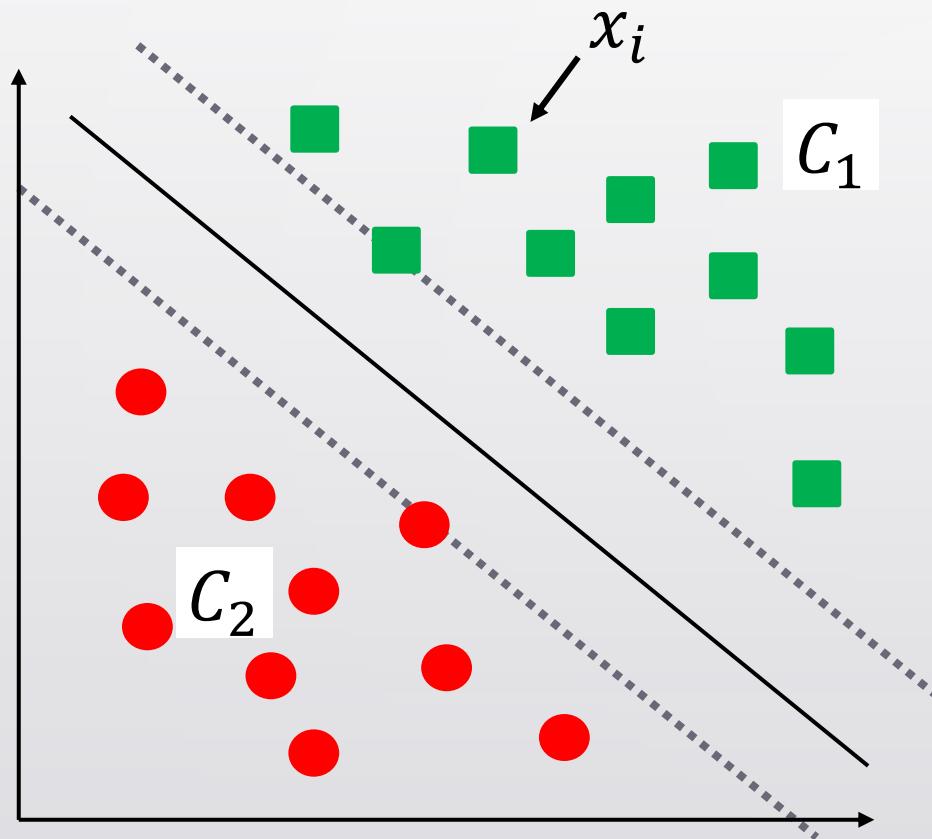
$$|w'^T x_i + b'| \geq M$$



$$|w^T x_i + b| \geq 1$$

$$w = \frac{1}{M} w' \quad b = \frac{1}{M} b'$$

サポートベクターマシン Support Vector Machine



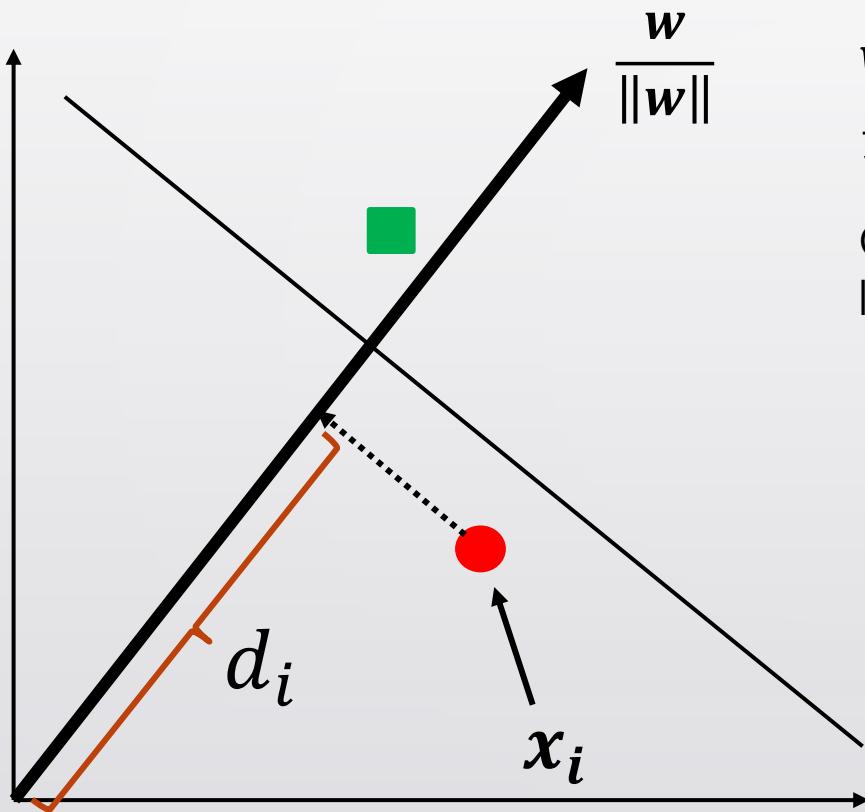
$x_i \in C_1$ の場合 $\mathbf{w}^T \mathbf{x}_i + b \geq 1$

In case of $x_i \in C_1$

$x_i \in C_2$ の場合 $\mathbf{w}^T \mathbf{x}_i + b \leq -1$

In case of $x_i \in C_2$

サポートベクターマシン Support Vector Machine



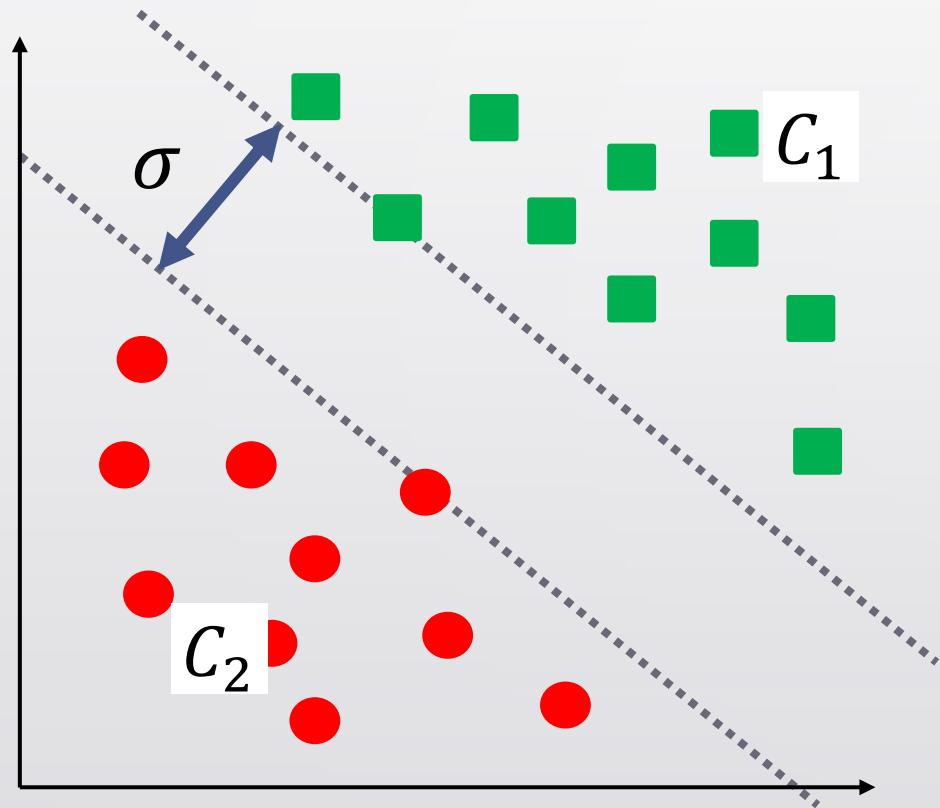
\mathbf{w} と平行な向きに対する \mathbf{x}_i の射影の長さ d_i を計算する

Calculate the length d_i of projection of \mathbf{x}_i onto the line parallel to \mathbf{w}

$$d_i = \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|}$$

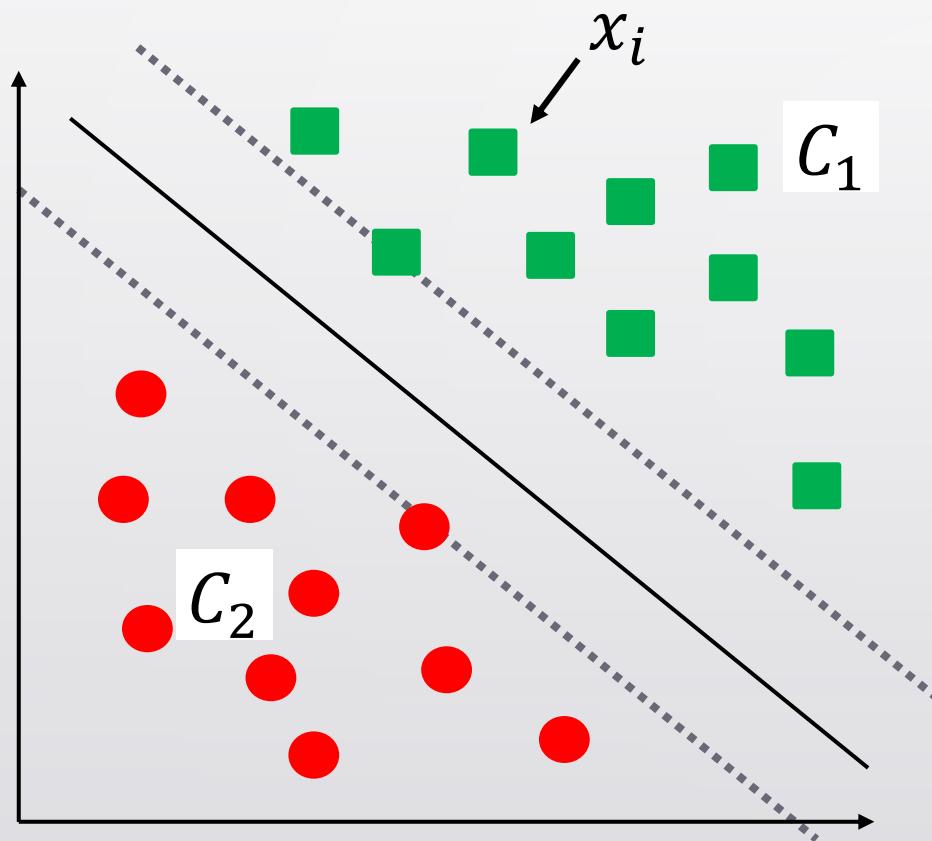


サポートベクターマシン Support Vector Machine



$$\begin{aligned}\sigma &= \min_{x_i \in C_1} d_i - \max_{x_i \in C_2} d_i \\ &= \min_{x_i \in C_1} \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|} - \max_{x_i \in C_2} \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|}\end{aligned}$$

サポートベクターマシン Support Vector Machine



$x_i \in C_1$ の場合 $w^T x_i + b \geq 1$

In case of $x_i \in C_1$

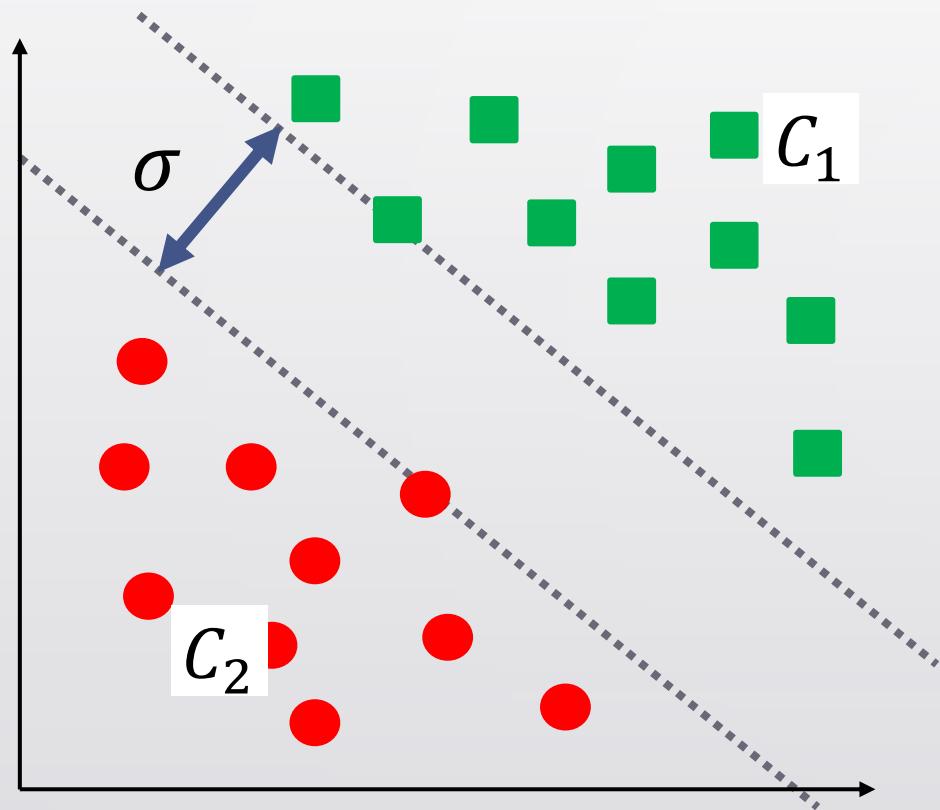
なので Then $\min_{x_i \in C_1} \frac{w^T x_i}{\|w\|} = \frac{1 - b}{\|w\|}$

$x_i \in C_2$ の場合 $w^T x_i + b \leq -1$

In case of $x_i \in C_2$

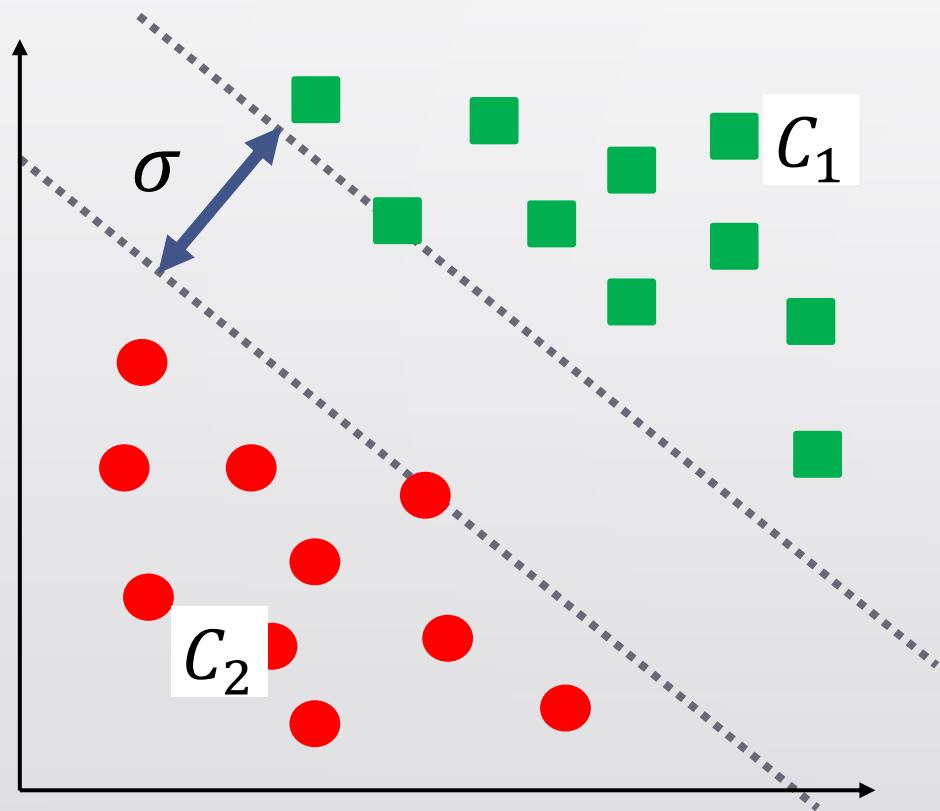
なので Then $\max_{x_i \in C_2} \frac{w^T x_i}{\|w\|} = \frac{-1 - b}{\|w\|}$

サポートベクターマシン Support Vector Machine



$$\begin{aligned}\sigma &= \min_{x_i \in C_1} \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|} - \max_{x_i \in C_2} \frac{\mathbf{w}^T \mathbf{x}_i}{\|\mathbf{w}\|} \\ &= \frac{1 - b}{\|\mathbf{w}\|} - \frac{-1 - b}{\|\mathbf{w}\|} \\ &= \frac{2}{\|\mathbf{w}\|}\end{aligned}$$

サポートベクターマシン Support Vector Machine



$|\mathbf{w}^T \mathbf{x}_i + b| \geq 1$ という制約の下で

Under the constraint that $|\mathbf{w}^T \mathbf{x}_i + b| \geq 1$

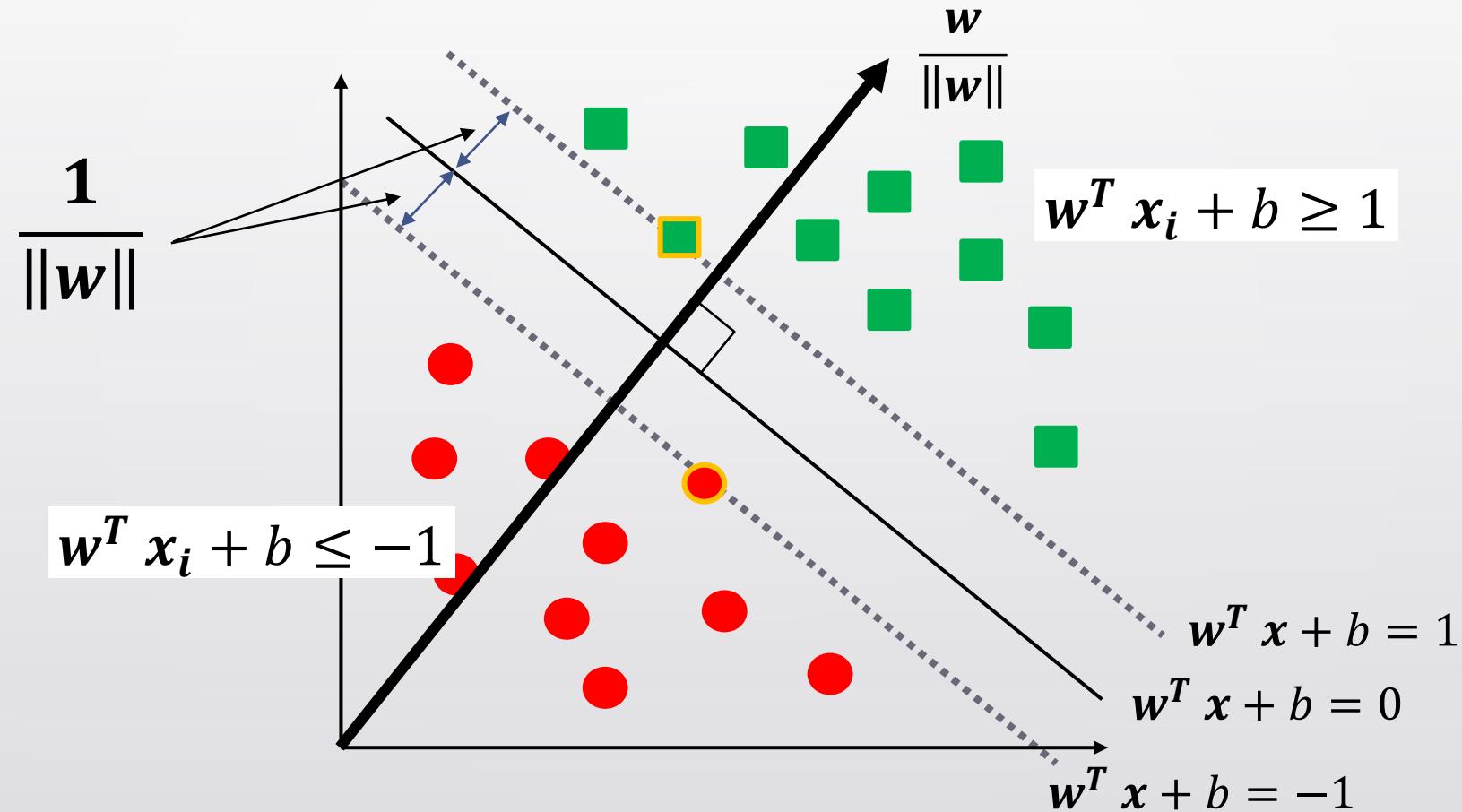
$\sigma = \frac{2}{\|\mathbf{w}\|}$ を最大化する

Maximize $\sigma = \frac{2}{\|\mathbf{w}\|}$

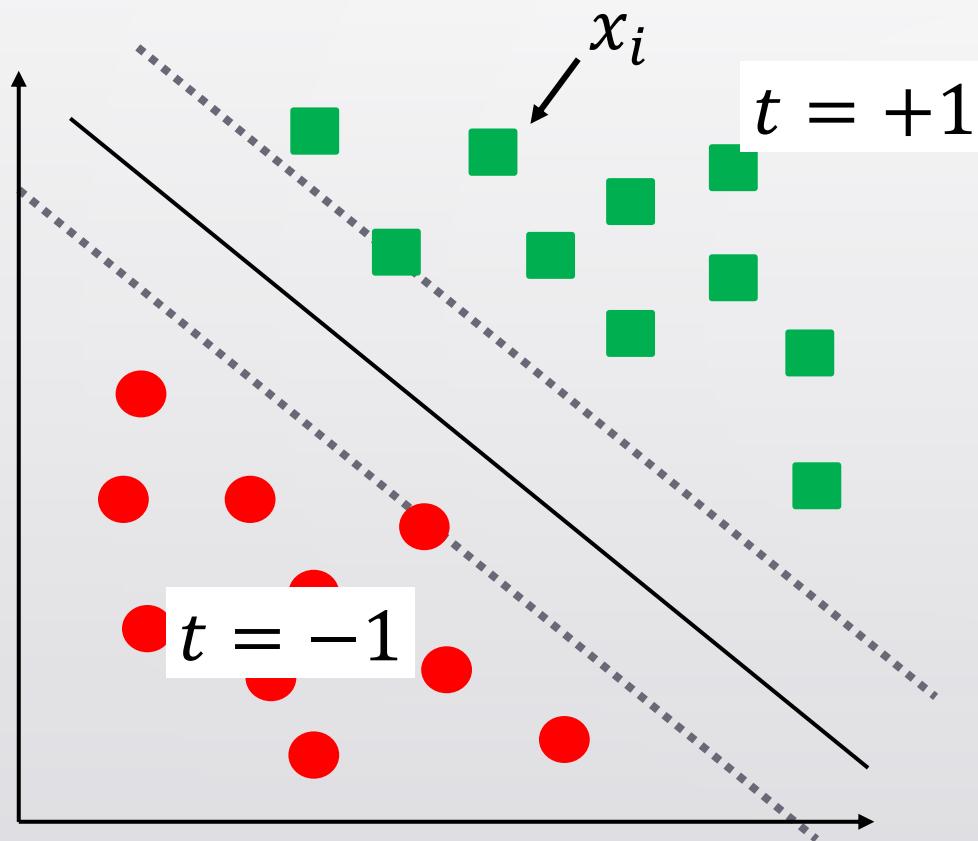
$\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w}$ を最小化する

Minimize $\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w}$

サポートベクターマシン Support Vector Machine



サポートベクターマシン Support Vector Machine

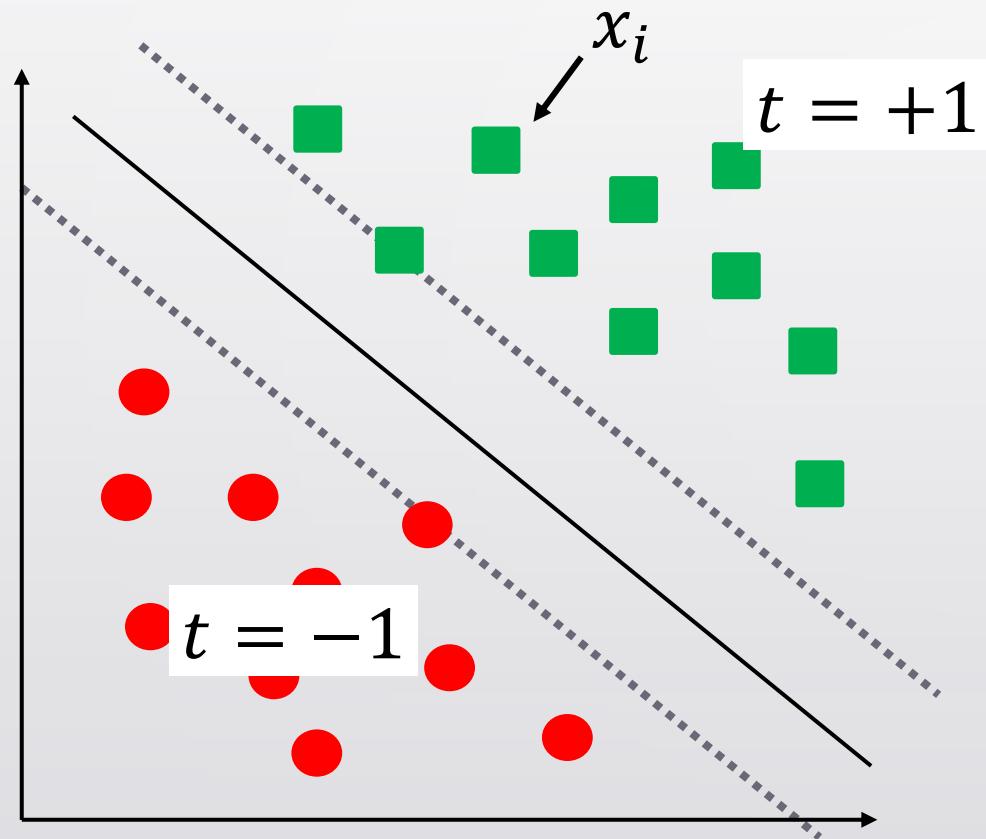


t_i : x_i のクラスを表す変数

A variable representing the class of x_i

$$t_i = \{-1, +1\}$$

サポートベクターマシン Support Vector Machine



$x_i \in C_1$ の場合 $w^T x_i + b \geq 1$

In case of $x_i \in C_1$

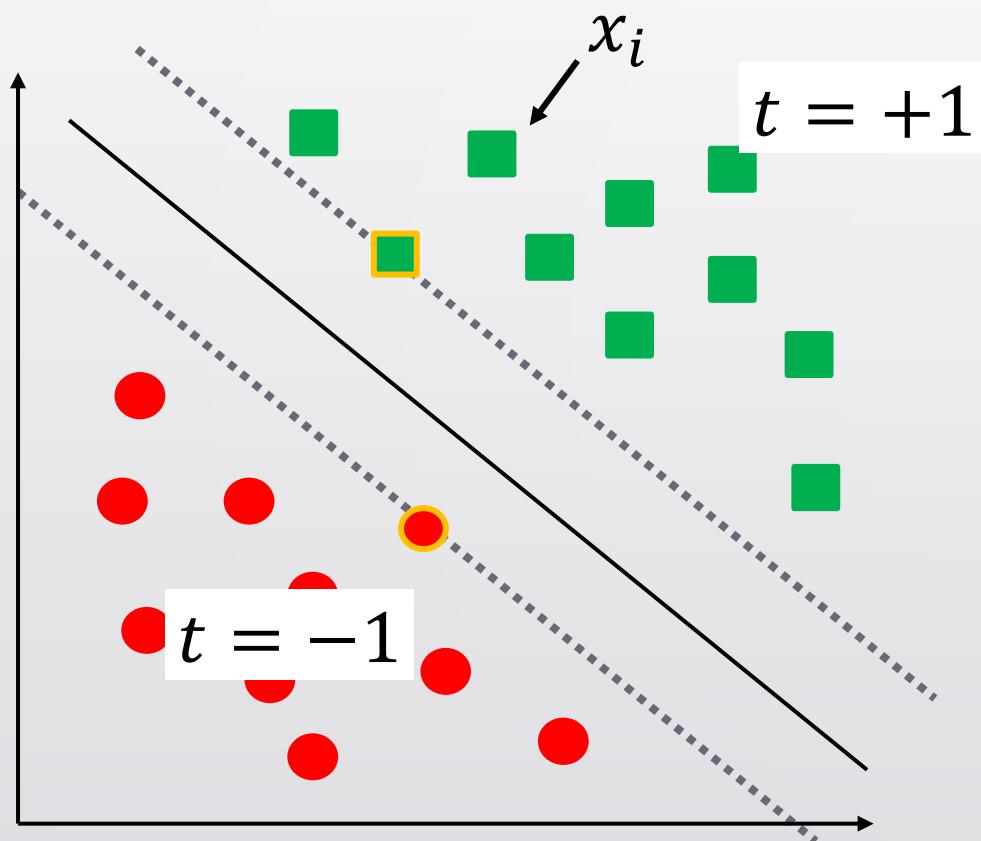
$x_i \in C_2$ の場合 $w^T x_i + b \leq -1$

In case of $x_i \in C_2$



$$t_i(w^T x_i + b) \geq 1$$

ハードマージンSVM Hard Margin SVM



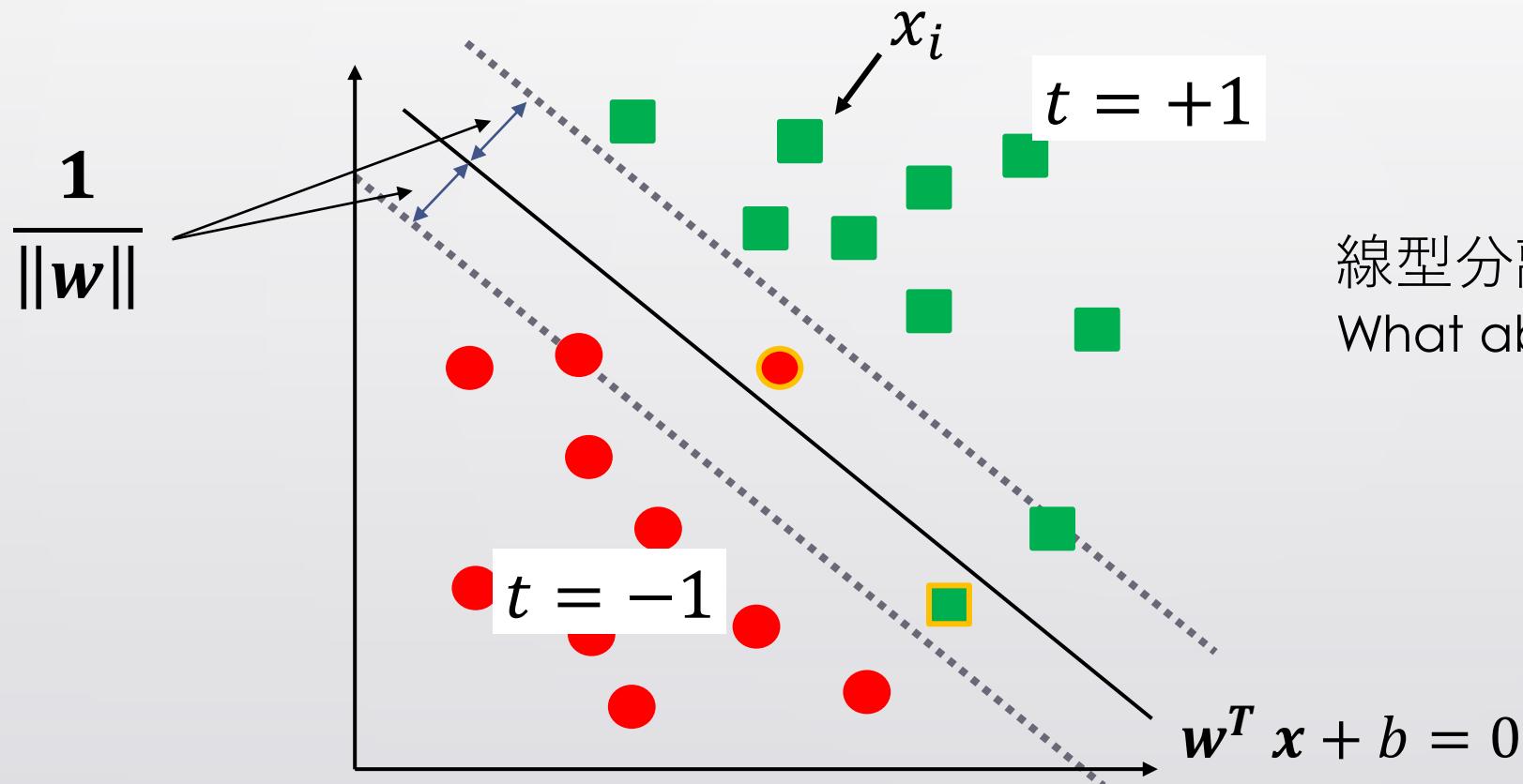
ハードマージンSVMは線型分離可能な問題に適用

Hard margin SVM is applicable to linearly separable data

$$\min_{w, b} \frac{1}{2} \|w\|^2$$

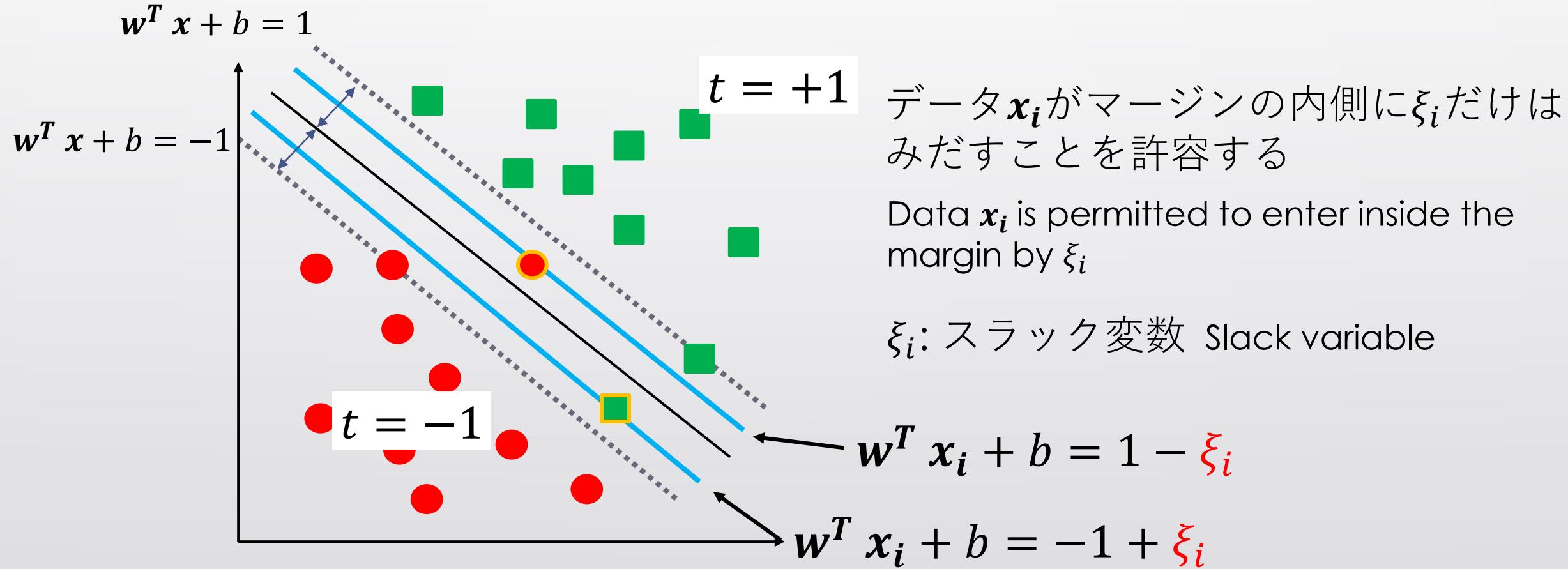
$$\text{subject to } t_i(w^T x_i + b) \geq 1$$

ソフトマージンSVM Soft Margin SVM

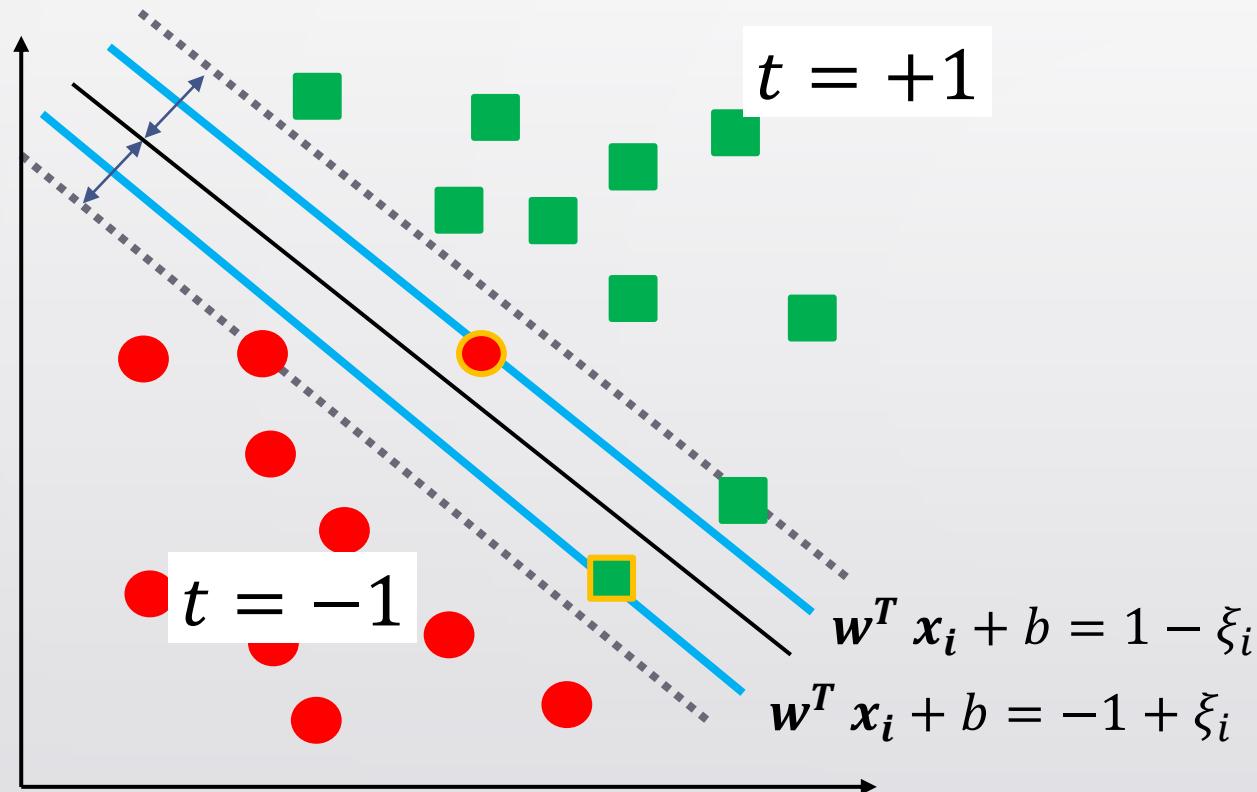


線型分離不可能な場合は?
What about linearly inseparable case?

ソフトマージンSVM Soft Margin SVM



サポートベクターマシン Support Vector Machine



$x_i \in C_1$ の場合 $w^T x_i + b \geq 1 - \xi_i$

In case of $x_i \in C_1$

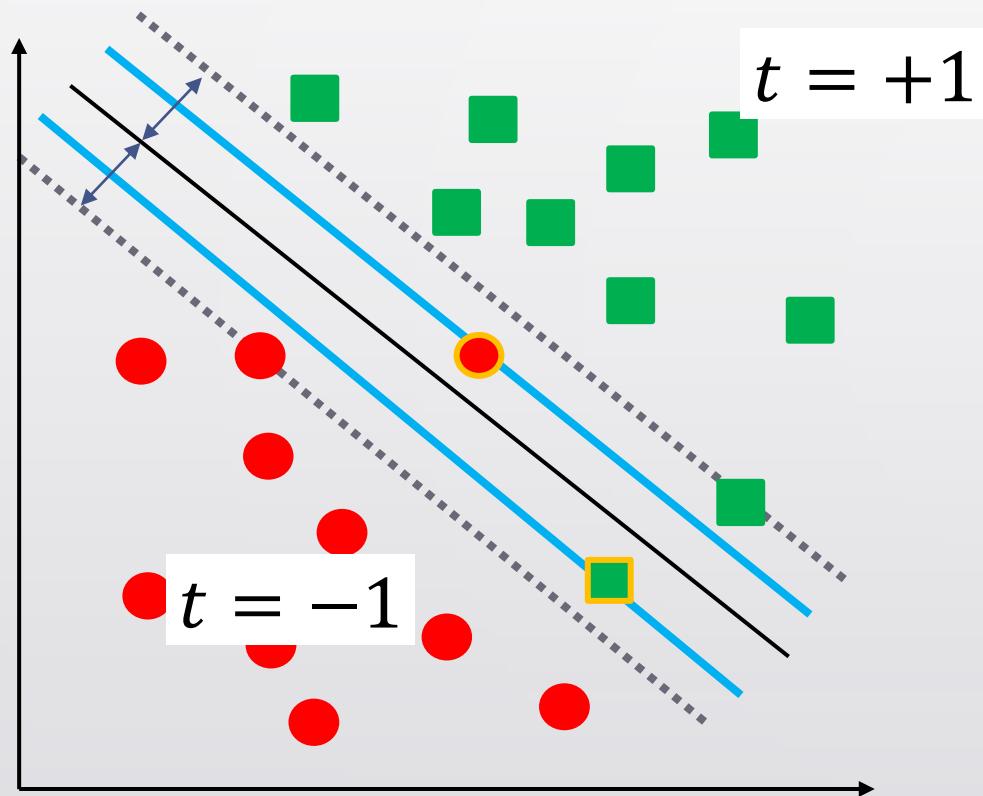
$x_i \in C_2$ の場合 $w^T x_i + b \leq -1 + \xi_i$

In case of $x_i \in C_2$



$$t_i(w^T x_i + b) \geq 1 - \xi_i$$

ソフトSVM Soft Margin SVM



ソフトマージンSVMは線型分離不可能な問題に適用

Soft margin SVM is applicable to linearly inseparable data

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to } t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \end{aligned}$$

ハードマージンSVM Hard Margin SVM

<主問題> Primal Problem

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^n \alpha_i (1 - t_i(\mathbf{w}^T \mathbf{x}_i + b))$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \frac{\partial L}{\partial b} = 0 \quad \frac{\partial L}{\partial \alpha} = 0 \quad \begin{array}{l} \text{KKT条件を考慮} \\ \text{Take KKT condition into account} \end{array}$$

ハードマージンSVM Hard Margin SVM

<双対問題> Dual Problem

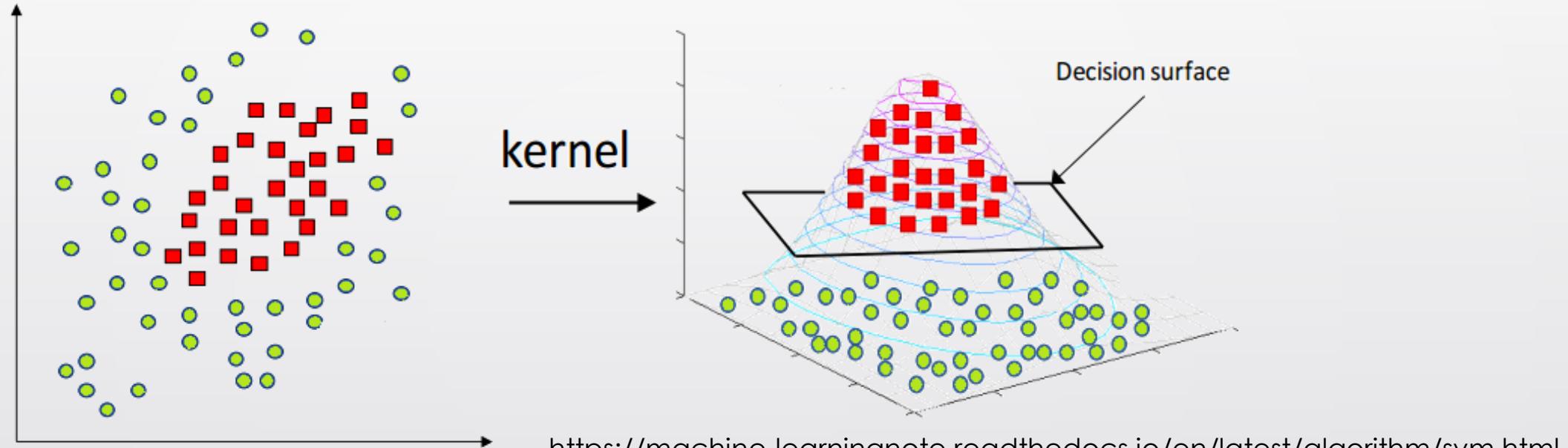
$$\max_{\alpha} L(\alpha) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^n \alpha_i$$

$$\text{subject to } \sum_{i=1}^n \alpha_i t_i = 0, \alpha_i \geq 0$$

α を求めるとき、それを用いて \mathbf{w} と b の最適解が分かる

Optimal values of \mathbf{w} and b are obtained based on α meeting the constraints above

カーネル法 Kernel Methods



データを高次元空間に写像することで、線型分離不可能な問題を線型分離可能にする

Transforming linearly inseparable problem to linearly separable one by mapping data to higher-dimensional space

基底関数 Basis Function

基底関数 φ は m 次元データ \mathbf{x}_i を p ($p > m$) 次元データ $\varphi(\mathbf{x}_i)$ に変換する

Basis function φ transforms m -dimensional data \mathbf{x}_i to p -dimensional data $\varphi(\mathbf{x}_i)$

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$$

$$\varphi(\mathbf{x}_i) = (\varphi_1(\mathbf{x}_i), \varphi_2(\mathbf{x}_i), \dots, \varphi_p(\mathbf{x}_i))$$

例 Example

$$\mathbf{x}_i = (x_{i1}, x_{i2})$$

$$\varphi(\mathbf{x}_i) = (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2c}x_{i1}, \sqrt{2c}x_{i2}, c)$$

基底関数 Basis Function

<双対問題>

$$\max_{\alpha} L(\alpha) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \varphi(x_i)^T \varphi(x_j) + \sum_{i=1}^n \alpha_i$$

$$subject\ to \sum_{i=1}^n \alpha_i t_i = 0, \alpha_i \geq 0$$

α を求めるとき、それを用いて w と b の最適解が分かる

Optimal values of w and b are obtained based on α meeting the constraints above

カーネル関数 Kernel Function

$$\max_{\alpha} L(\alpha) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \varphi(x_i)^T \varphi(x_j) + \sum_{i=1}^n \alpha_i$$

$\varphi(x_i)^T \varphi(x_j)$ の計算はコストが大きい

A lot of resource is required to compute $\varphi(x_i)^T \varphi(x_j)$

$$\underline{k(x_i, x_j)} = \varphi(x_i)^T \varphi(x_j)$$

カーネル関数 Kernel Function

カーネル関数は、基底関数 φ による写像の内積と一致する

Kernel function equals to the dot product of mappings by basis function φ

多項式カーネル Polynomial Kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \gamma (\mathbf{x}_i^T \mathbf{x}_j + c)^d$$

例 Example

$$\mathbf{x}_i = (x_{i1}, x_{i2}) \quad \gamma = 1, d = 2$$

$$\varphi(\mathbf{x}_i) = (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2c}x_{i1}, \sqrt{2c}x_{i2}, c)$$

$$\begin{aligned} k(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i^T \mathbf{x}_j + c)^2 \\ &= (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2c}x_{i1}, \sqrt{2c}x_{i2}, c)^T (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2c}x_{i1}, \sqrt{2c}x_{i2}, c) \\ &= \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) \end{aligned}$$

カーネルトリック Kernel Trick

基底関数をカーネル関数に置き換えることで、少ない計算コストで双対問題を解ける

Dual problem can be solved with reduced computational cost by replacing basis function with kernel function

$$\begin{aligned} \max_{\alpha} L(\alpha) &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j) + \sum_{i=1}^n \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j k(\mathbf{x}_i, \mathbf{x}_j) + \sum_{i=1}^n \alpha_i = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j (\mathbf{x}_i^T \mathbf{x}_j + c)^2 + \sum_{i=1}^n \alpha_i \end{aligned}$$



その他のカーネル Other Kernels

- linear

$$k(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2$$

- polynomial

$$k(\mathbf{x}_1, \mathbf{x}_2) = (\gamma \mathbf{x}_1 \cdot \mathbf{x}_2 + c)^d$$

- Gaussian or radial basis

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|^2)$$

- sigmoid

$$k(\mathbf{x}_1, \mathbf{x}_2) = \tanh(\gamma \mathbf{x}_1 \cdot \mathbf{x}_2 + c)$$

<https://www.analyticsvidhya.com/blog/2021/07/svm-support-vector-machine-algorithm/>



データマイニング

Data Mining

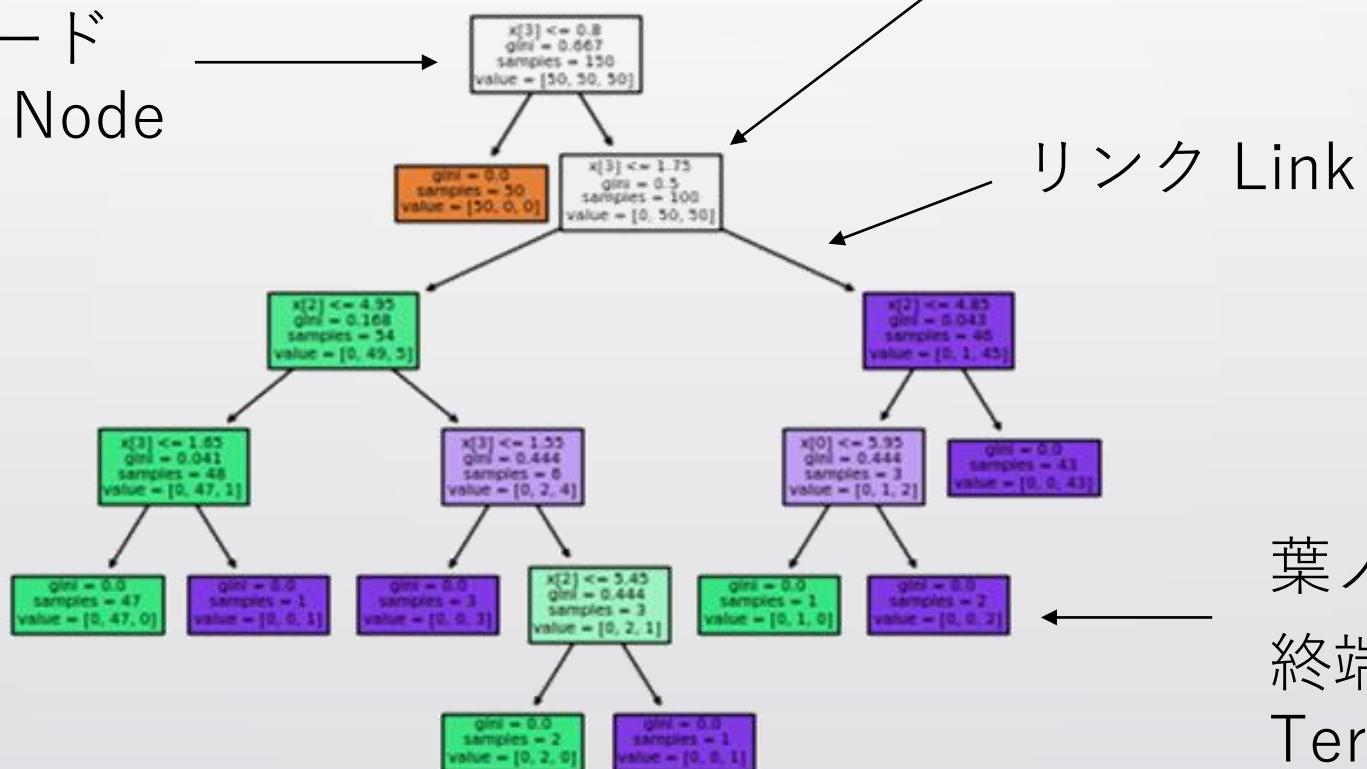
10: 分類⑤ Classification

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

決定木 Decision Tree

根ノード
Root Node



ノード Node

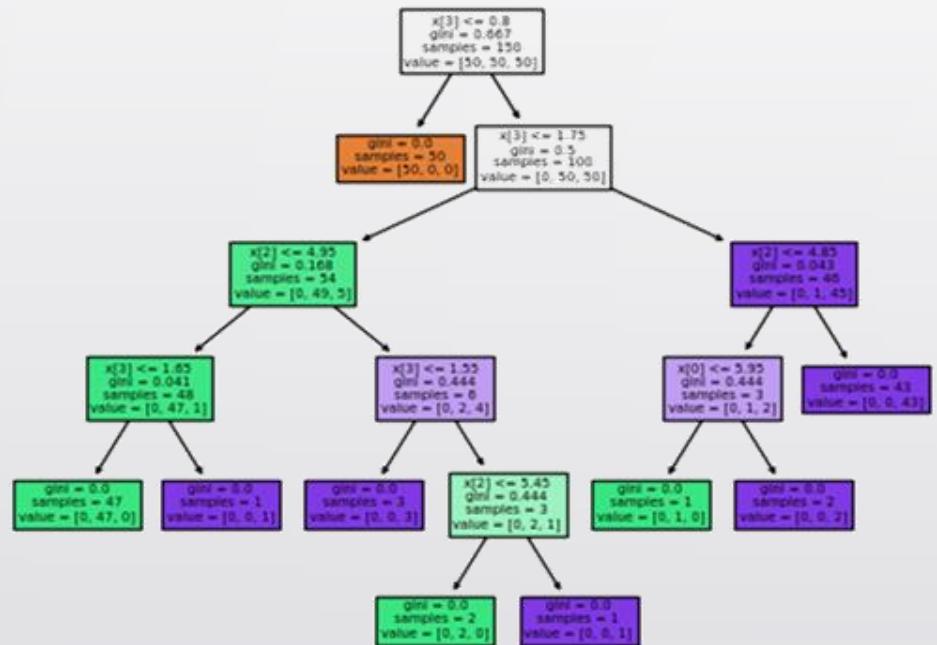
リンク Link

葉ノード Leaf Node

終端ノード

Terminal Node

分割統治法 Divide and Conquer Induction



出来るだけ誤りなくデータを分類できる2分割基準でデータを分類する

Classify data by dichotomous criteria that minimizes false classification rate

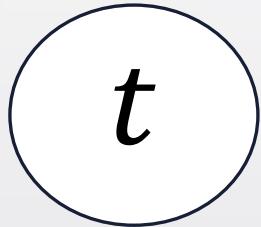


この手続きを繰り返すことで決定木を成長させる

Grow decision tree by repeating this procedure



ノードのクラスの決定 Decision of the Class of a Node



$\{(x_i, t_i)\}$:学習データ x_i とクラスラベル t_i の集合

C_j :クラス j

N_j :クラス j に属するデータの数

$N(t)$:ノード t に属するデータの数

$N_j(t)$:ノード t に属するクラス j のデータの数



ベイズの定理 Bayes Theorem

$$P(H \cap C) = P(H|C)P(C)$$

$$P(H \cap C) = P(C|H)P(H)$$

$$P(C|H)P(H) = P(H|C)P(C)$$

$$P(C|H) = \frac{P(H|C)P(C)}{P(H)}$$

ノードのクラスの決定 Decision of the Class of a Node

ノード t に属するデータが、クラス C_j に属する事後確率を最大化するクラス C_j を、ノード t のクラスとする

Designate class C_j as the class of node t so that posterior probability of data in node t belonging to class C_j is maximized

$$P(C_j|t) = \frac{P(t|C_j)P(C_j)}{P(t)}$$

ノード t のクラス = $\operatorname{argmax}_j P(C_j|t)$
Class of node t

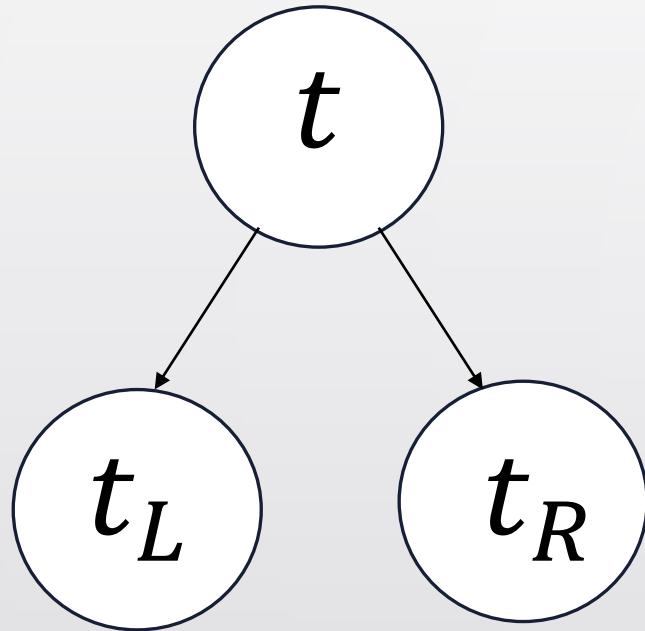
ノードのクラスの決定 Decision of the Class of a Node

$$P(t|C_j)P(C_j) = \frac{N_j(t)}{N_j} \times \frac{N_j}{N} = \frac{N_j(t)}{N} \quad P(t) = \frac{N(t)}{N}$$

$$P(C_j|t) = \frac{P(t|C_j)P(C_j)}{P(t)} = \frac{N_j(t)}{N} \times \frac{N}{N(t)} = \frac{N_j(t)}{N(t)}$$

ノード t の クラス = $\operatorname{argmax}_j P(C_j|t) = \operatorname{argmax}_j \frac{N_j(t)}{N(t)}$
Class of node t

ノードの不純度 Impurity of a Node



$$\Delta \text{Impurity} = \text{Impurity}_{\text{before}} - \text{Impurity}_{\text{after}}$$

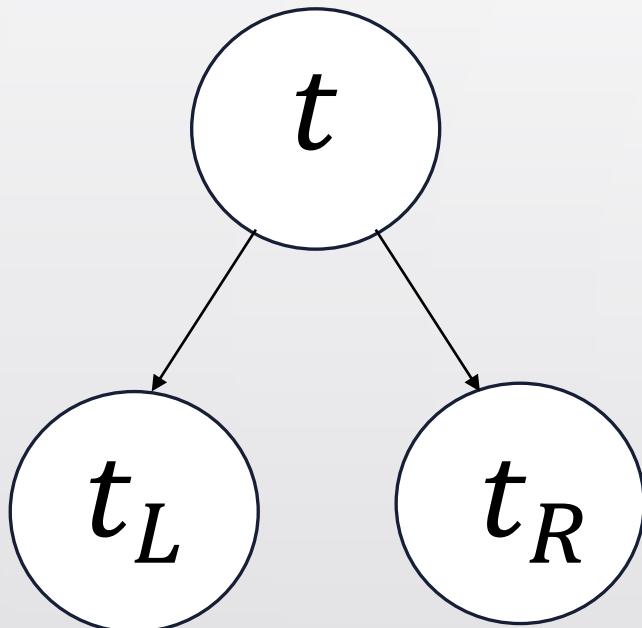
$$\text{Impurity}_{\text{before}} = \text{Impurity}(t)$$

$$\text{Impurity}_{\text{after}} = P(t_L)\text{Impurity}(t_L) + P(t_R)\text{Impurity}(t_R)$$

分割前後の不純度の減少が最大になるようにデータを
2分割する

Divide data so as to maximize the decrease of impurity
 $\Delta \text{Impurity}$ after division

ジニ係数 Gini Index

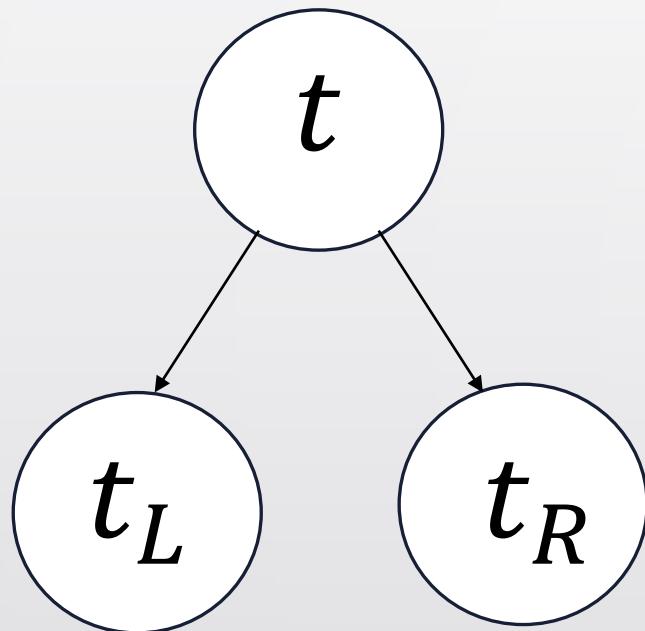


$$Gini\ Index = 1 - \sum_1^K P(C_j|t)^2$$

$P(C_j|t)$ のクラス間の違いが二乗により強調される
Inter-class difference in $P(C_j|t)$ is pronounced when squared



エントロピー Entropy

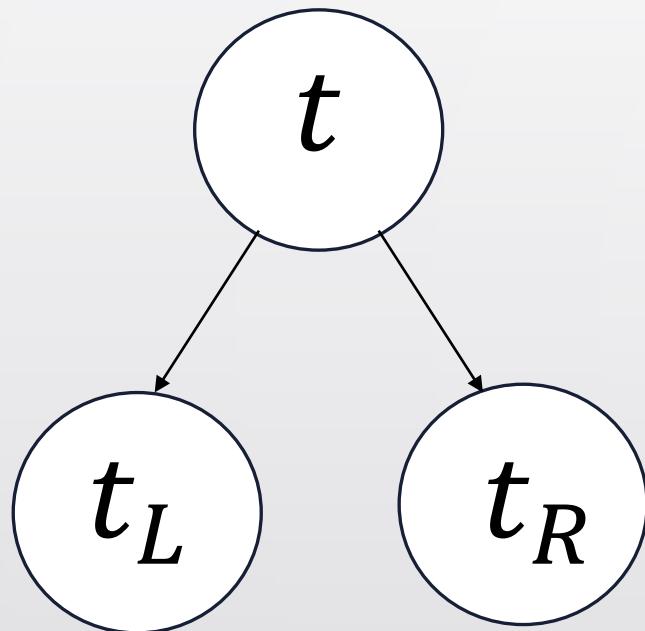


ノード t に含まれるクラスがばらついている
Class of data belonging to node t is not uniform



ノード t の不純度が高いと $P(C_j|t)$ が小さくなる
 $P(C_j|t)$ gets small when impurity of node t is large

エントロピー Entropy

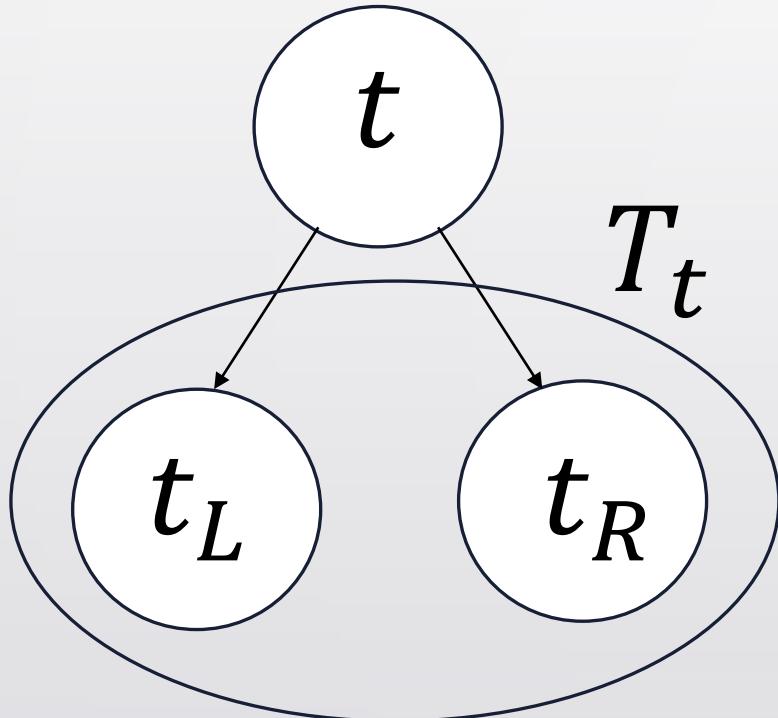


$$Entropy = - \int p \log(p) = E \left[\log \frac{1}{p} \right]$$

確率が低い事象が起こると大きくなる
Increases when an event with low-probability occurs

$$Impurity = - \sum_1^K P(C_j|t) \log P(C_j|t)$$

木の剪定 Pruning Decision Tree

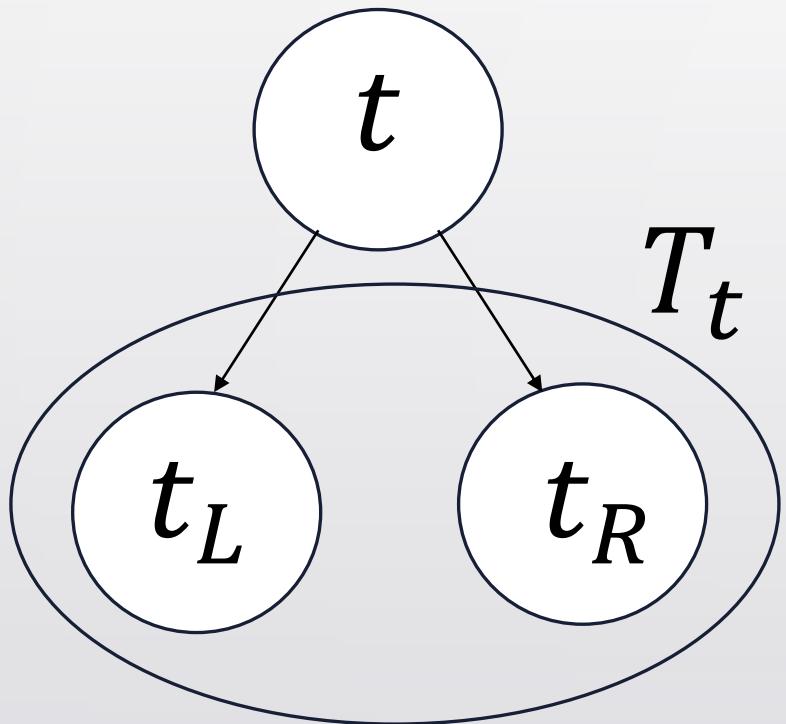


ノード t の分岐 Branch of node t

ノード t を終端ノードとすべきかどうか
= ノード t の分岐 T_t を削除すべきかどうか

The problem of whether node t should be determined as a terminal node equals to whether branch T_t of node t should be removed or not.

木のコスト Cost of Tree



ノード t の分岐 Branch of node t

$$R_\alpha(t) = \underline{R(t)} + \alpha$$

ノード t における誤り率 Error rate at node t

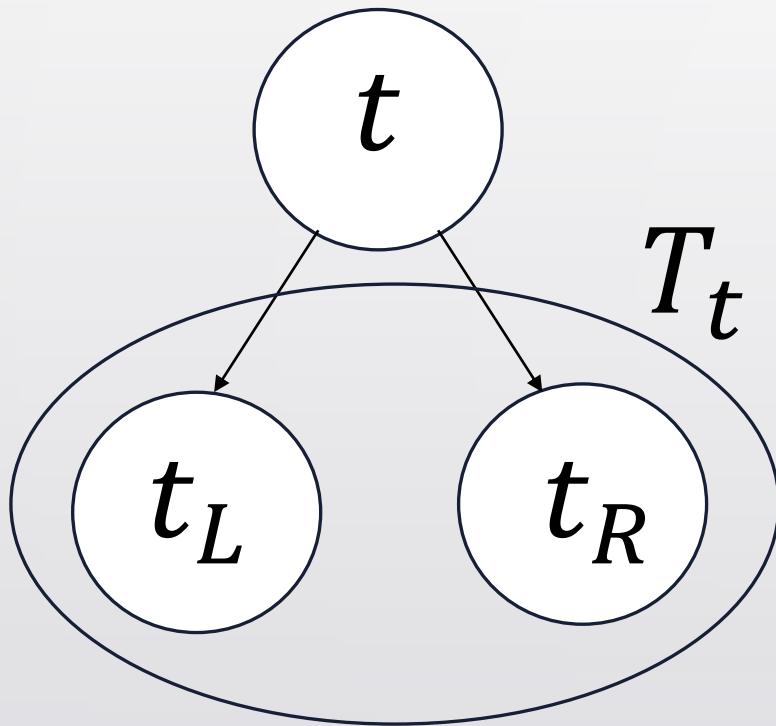
$$R_\alpha(T) = \sum_{t \in T} R_\alpha(t) = R(T) + \alpha |\tilde{T}|$$

$|\tilde{T}|$: 終端ノードの数 Number of terminal node

小さい木で高い正答率を達成したい

Aims to achieve high accuracy by decision tree with smaller size

木のコスト Cost of Tree



ノード t の分岐 Branch of node t

$$R_\alpha(t) = R(t) + \alpha$$

$$R_\alpha(T_t) = R(T_t) + \underline{\alpha |\tilde{T}_t|}$$

T_t は複数のノードを含むことに注意

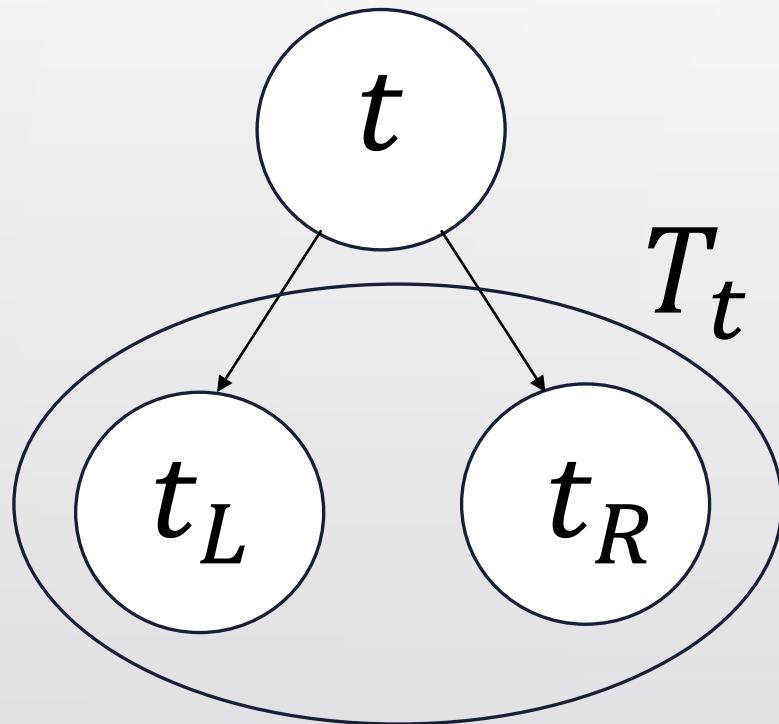
Note that T_t includes multiple nodes

$R_\alpha(T_t) \geq R_\alpha(t)$ なら T_t を削除する

Remove T_t if $R_\alpha(T_t) \geq R_\alpha(t)$



木のコスト Cost of Tree



ノード t の分岐 Branch of node t

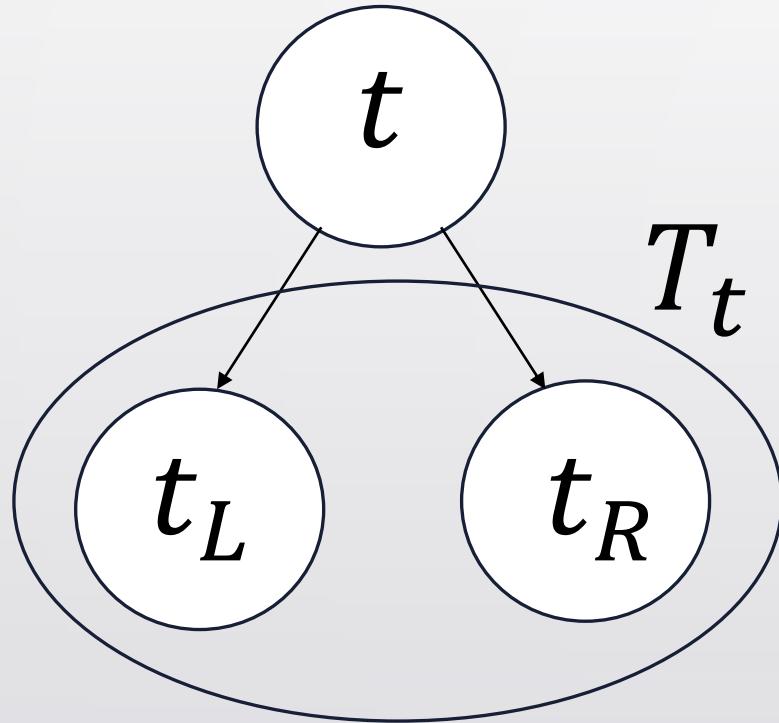
$$R(T_t) + \alpha |\tilde{T}_t| \geq R(t) + \alpha$$



$$\frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1} \leq \alpha \text{ なら } T_t \text{ を削除する}$$

Remove T_t if $\frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1} \leq \alpha$

木のコスト Cost of Tree



ノード t の分岐 Branch of node t

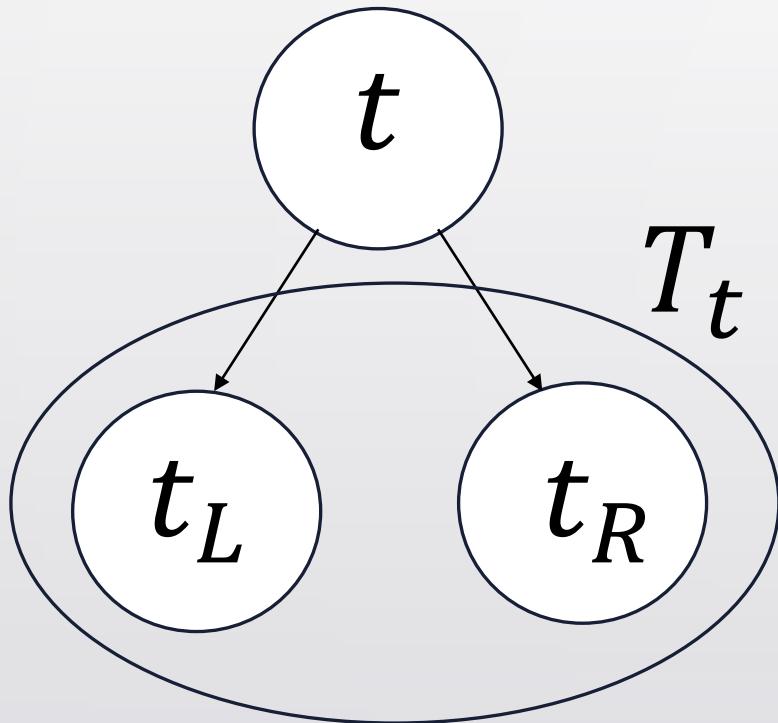
$$\frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1} \leq \alpha \text{ なら } T_t \text{ を削除する}$$



$$g(t) = \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}$$

$g(t)$ は剪定するかどうかの判断基準を与える
 $g(t)$ gives a criteria to decide whether to prune branch

剪定のアルゴリズム Algorithm of Pruning



ノード t の分岐 Branch of node t

- ・ 全ての内部ノードについて $g(t)$ を計算する
Compute $g(t)$ for all the internal nodes
- ・ $g(t)$ が最小のノードの分岐を削除する
Remove branch of the node with minimum $g(t)$
- ・ この手続きを繰り返す
Repeat the procedure above until certain criteria is met

アンサンブル学習 Ensemble Learning

ノーフリーランチ定理 No Free Lunch Theorem

あらゆる問題に対して最適な分類器は存在しない

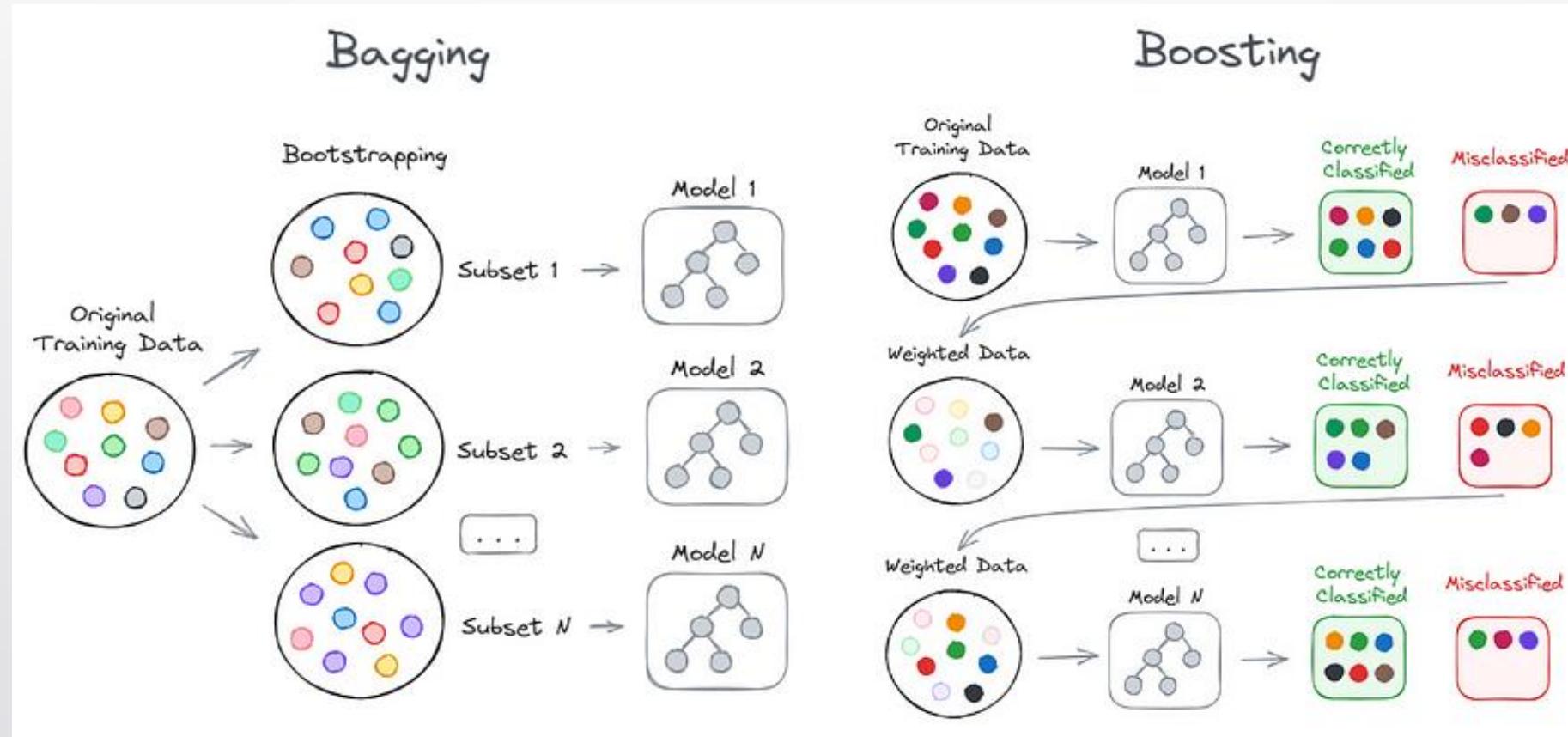
There is no classifier that shows best performance for every classification problem



複数の弱識別器を組み合わせることで、精度のよい分類器を作る

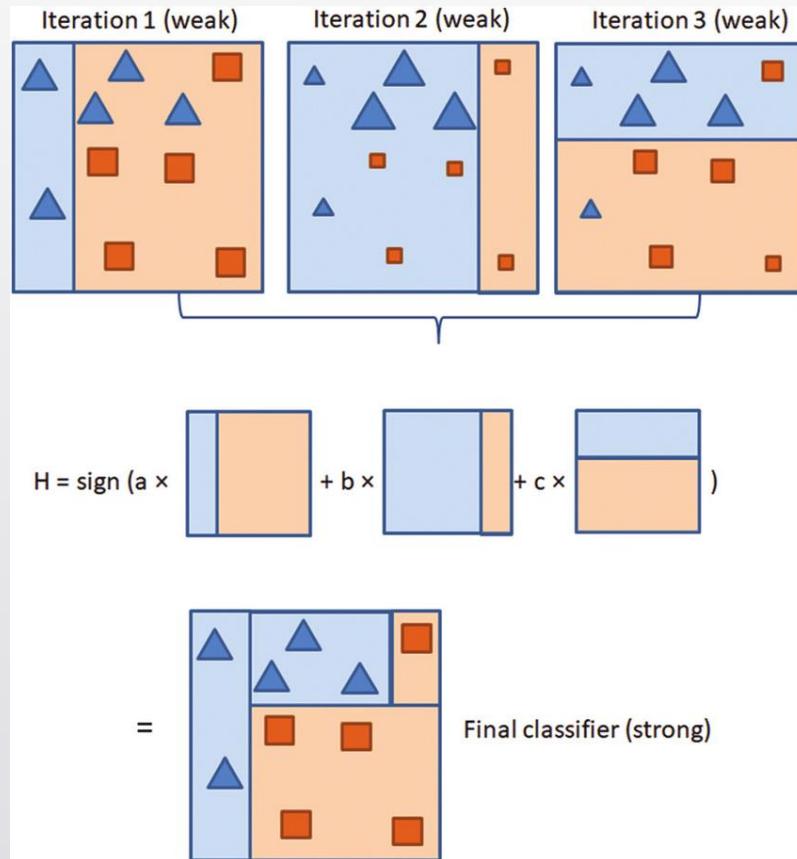
Create superior classifier by combining multiple weak classifiers

バギングとブースティング Bagging and Boosting



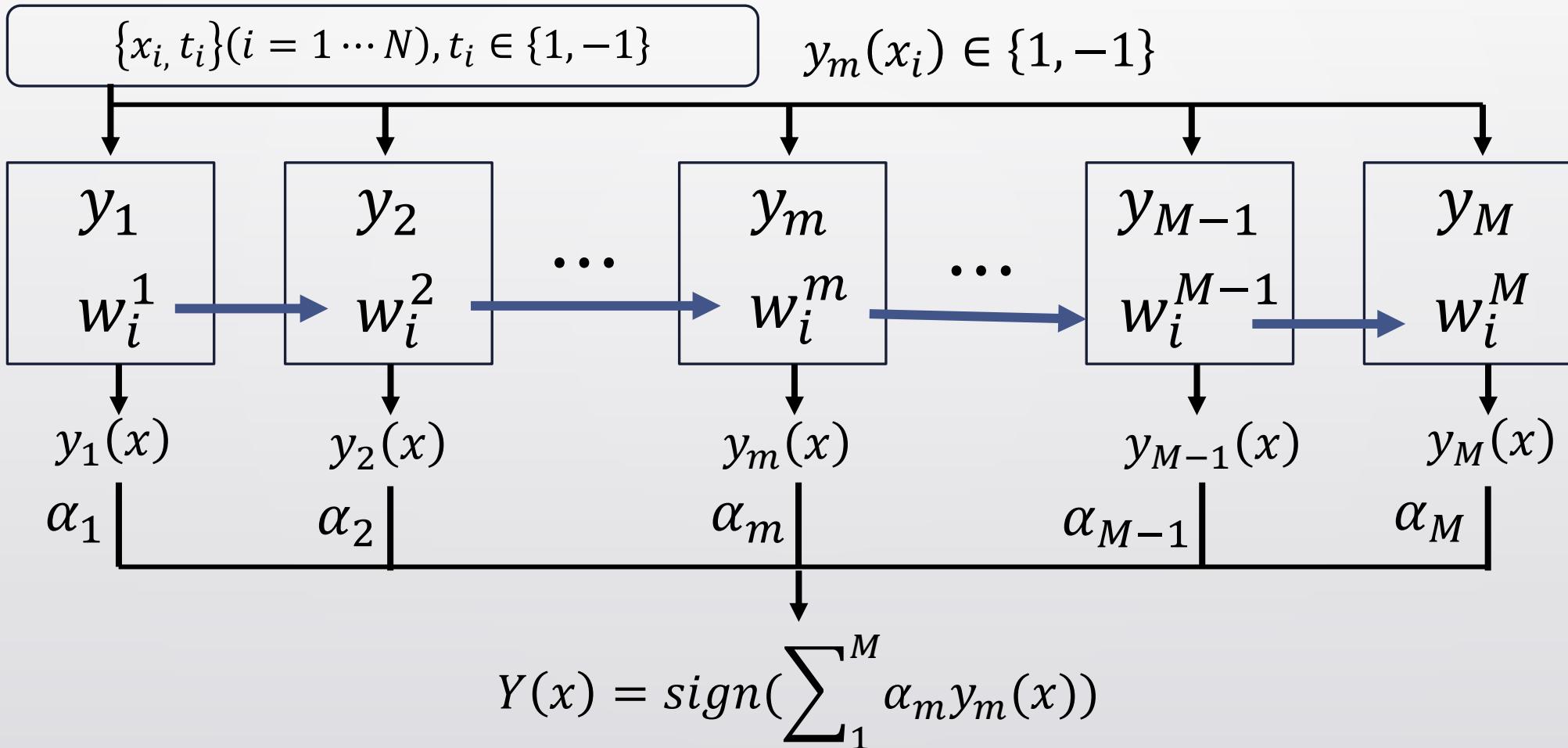
<https://pub.towardsai.net/bagging-vs-boosting-the-power-of-ensemble-methods-in-machine-learning-6404e33524e6>

アダブースト Ada(ptive)Boost(ing)



- ・ ブースティングアルゴリズムの一つ
One of boosting algorithms
- ・ t 個めの弱学習器が誤識別したデータの重みを大きくして、 $t + 1$ 個めの弱学習器をトレーニングする
Train $t + 1$ -th weak learner by giving large weight to data points that t -th weak learner misclassified

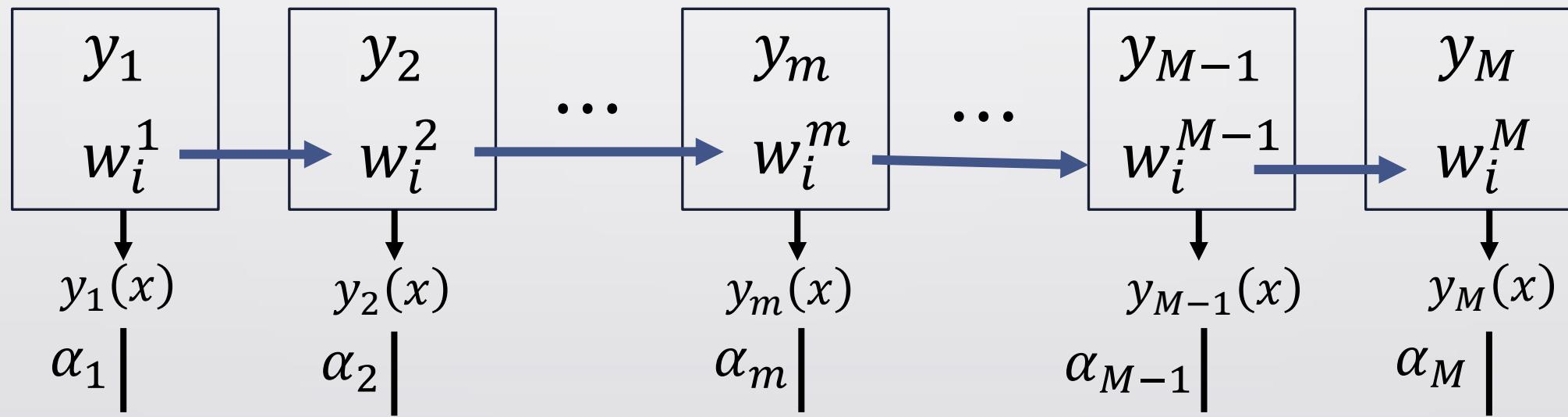
アダブースト AdaBoost



アダブースト AdaBoost

弱識別器 y_m を逐次的に訓練する

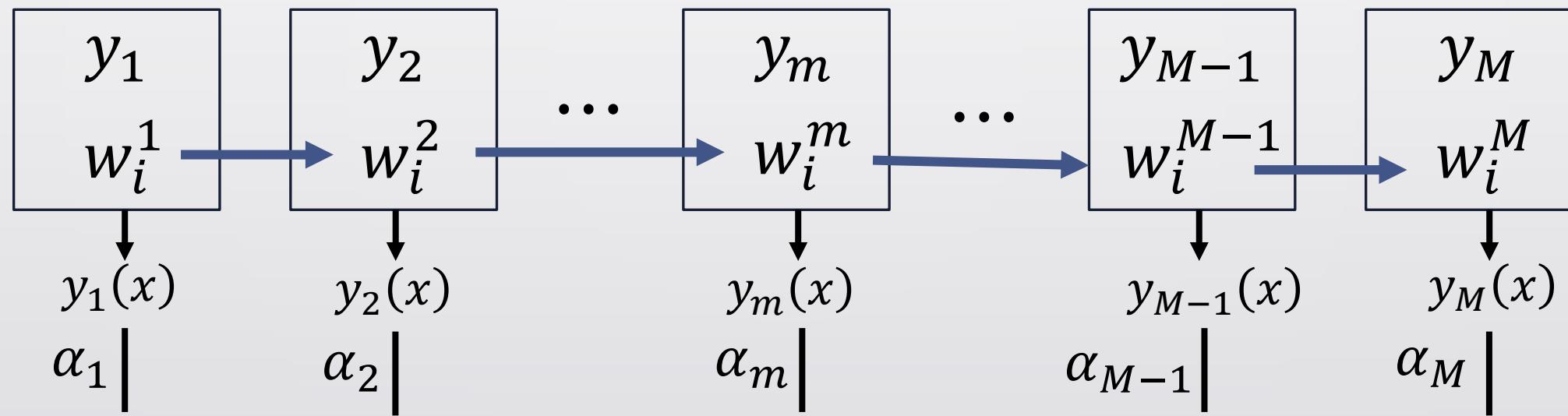
Train weak learners y_m sequentially



アダブースト AdaBoost

データの重み w_i^m と弱学習器の重み α_m が更新される

Update weight of data w_i^m and weight of weak learner α_m



アダブースト AdaBoost

E_m が最小になるよう弱識別器 y_m を訓練する

Train weak learners y_m so that E_m is minimized

$$E_m = \frac{\sum_1^N w_i^m I(y_m(x_i))}{\sum_1^N w_i^m} \quad I(y_m(x_i)) = \begin{cases} 1 & (y_m(x_i) \neq t_i) \\ 0 & (y_m(x_i) = t_i) \end{cases}$$

アダブースト AdaBoost

データの重み w_i^m と弱学習器の重み α_m が更新される

Update weight of data w_i^m and weight of weak learner α_m

$$\alpha_m = \ln \left(\frac{1}{E_m} - 1 \right) \geq 0$$

$$w_i^{m+1} = \begin{cases} w_i^m \exp(\alpha_m) & (y_m(x_i) \neq t_i) \\ w_i^m & (y_m(x_i) = t_i) \end{cases}$$

アダブースト AdaBoost

最終的な出力は、弱学習器の出力の重み付き和によって決まる

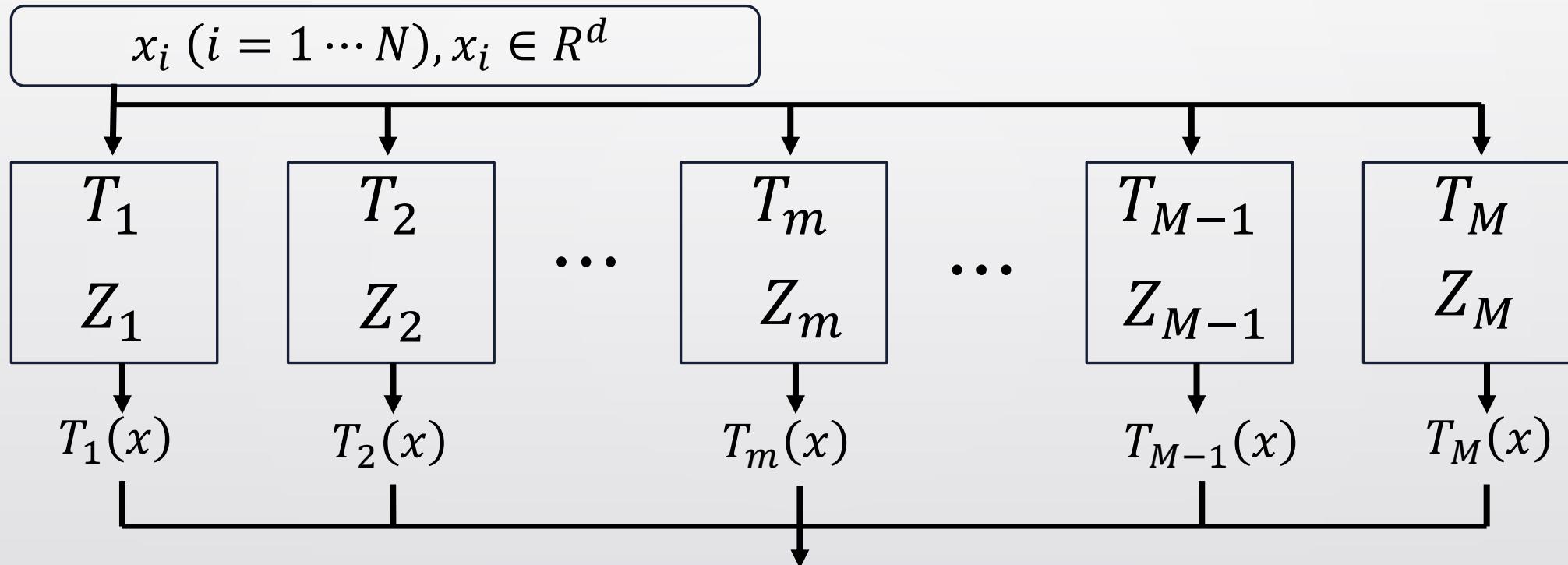
Output of AdaBoost classifier is determined by weighted sum of output of each weak classifier

$$y_1(x) \quad y_2(x) \quad \dots \quad y_m(x) \quad \dots \quad y_{M-1}(x) \quad y_M(x)$$
$$\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_m \quad \dots \quad \alpha_{M-1} \quad \alpha_M$$

A diagram illustrating the AdaBoost process. At the top, five outputs are labeled: $y_1(x)$, $y_2(x)$, $y_m(x)$, $y_{M-1}(x)$, and $y_M(x)$. Below them, their corresponding weights α_1 , α_2 , α_m , α_{M-1} , and α_M are shown. A horizontal line connects the outputs, and a vertical arrow points downwards from the line to the equation below.

$$Y(x) = sign\left(\sum_1^M \alpha_m y_m(x)\right)$$

ランダムフォレスト Random Forest



多数決投票 Majority Voting

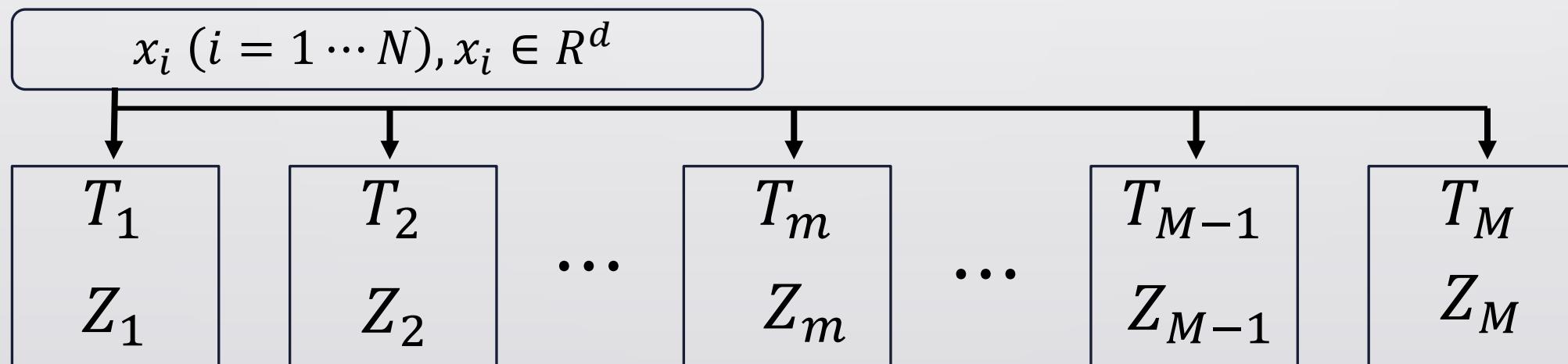
ランダムフォレスト Random Forest

データからブートストラップサンプル Z_m を抽出する

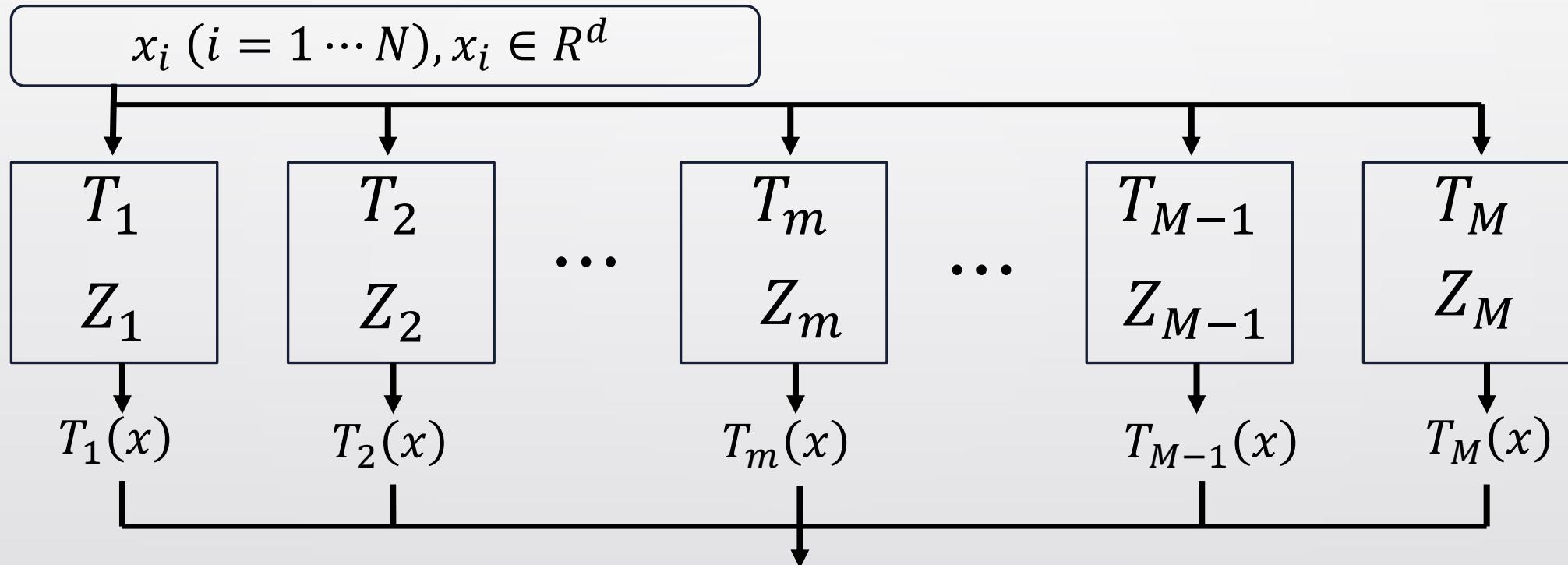
Extract bootstrapped samples Z_m from dataset

ランダムに選択した d' 次元の特徴量を使って決定木 T_m を構成する

Grow decision tree T_m based on randomly-selected d' -dimensional features



ランダムフォレスト Random Forest



多数決投票 Majority Voting



データマイニング

Data Mining

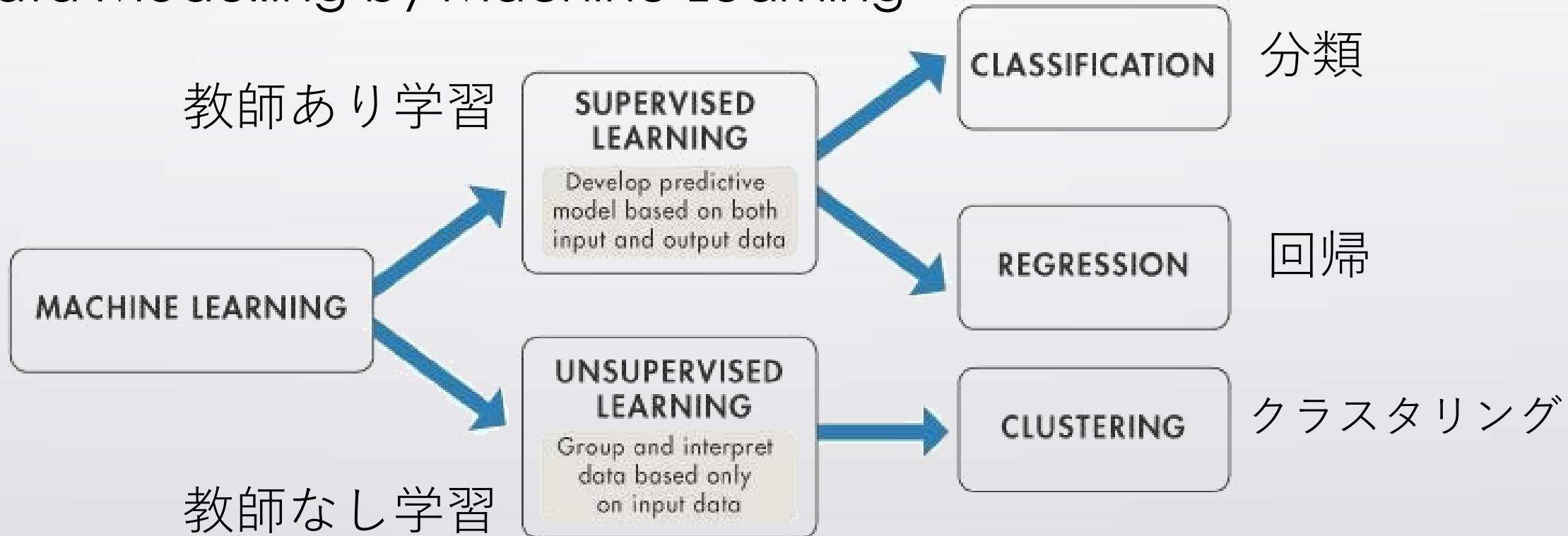
11: クラスタリング① Clustering

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

機械学習によるモデル化

Data Modelling by Machine Learning



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

クラスタリングの種類 Types of Clustering

- 非階層的クラスタリング
Non-Hierarchical Clustering
- 階層的クラスタリング
Hierarchical Clustering
- モデル・ベース・クラスタリング
Model-Based Clustering

データの統計的分布についての仮定をおく
Make presumptions about statistical distribution of data



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

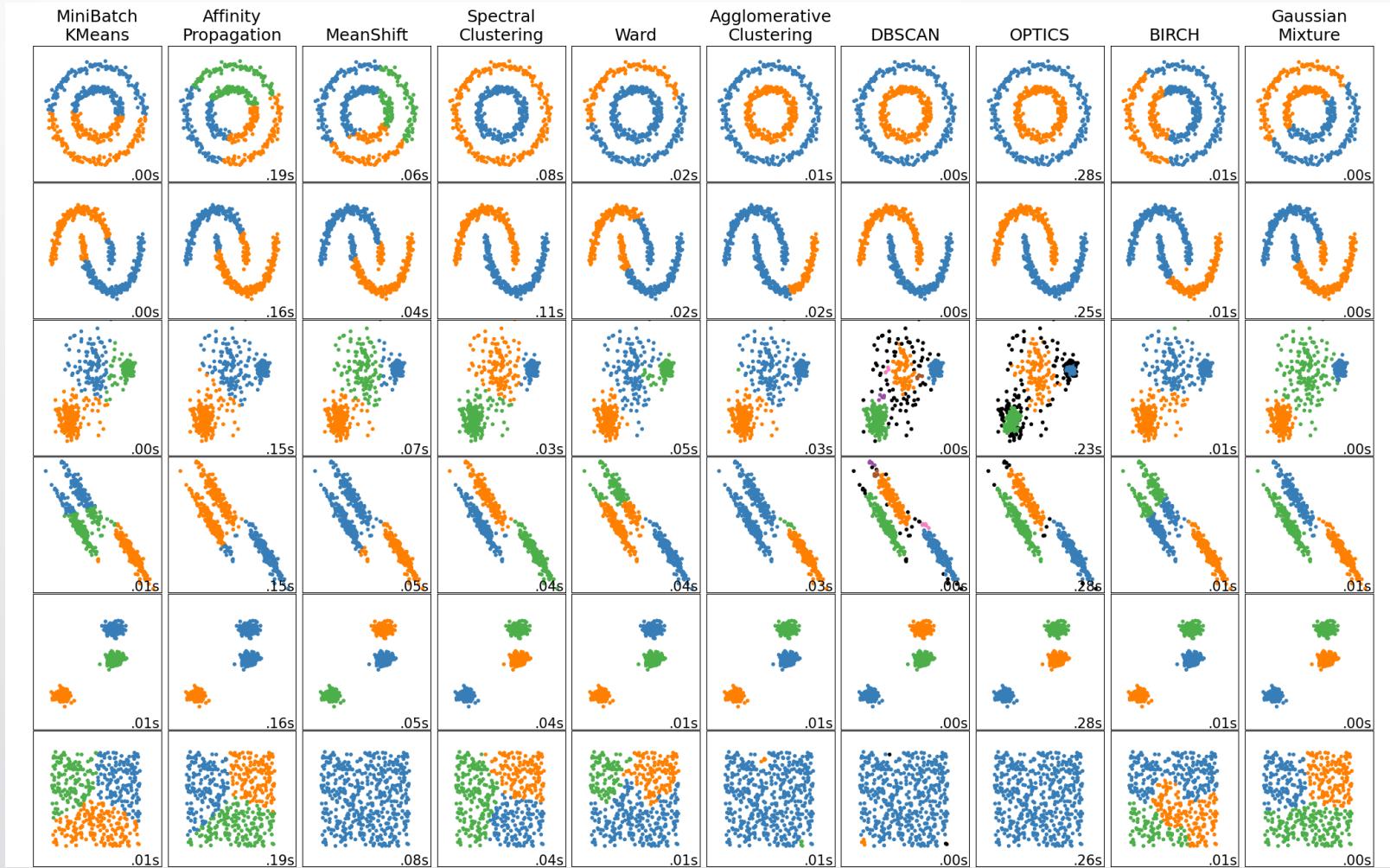
クラスタリング

Final result of clustering
depends on

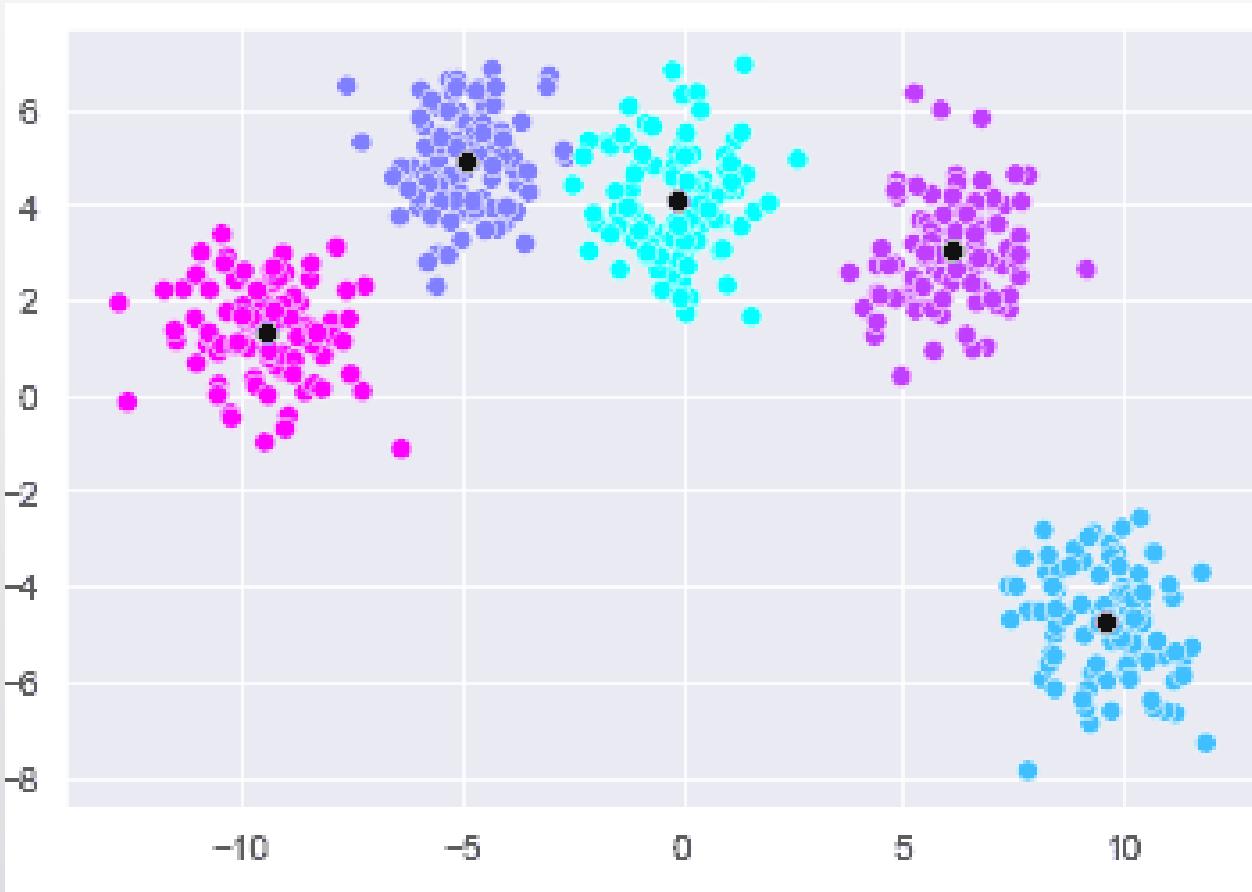
Type of algorithm
アルゴリズムの種類

Parameter Setting
パラメータ設定

<https://scikit-learn.org/stable/modules/clustering.html>



K平均クラスタリング k-means clustering



クラスターの数を指定しなくては
いけない

You have to specify the number of
clusters, k .

K平均クラスタリング *k-means clustering*

非階層的クラスタリングの代表的なアルゴリズム

Representative algorithm of non-hierarchical clustering

各クラスターの中心とデータとの距離に基いてクラスタリングを行う

Clustering based on distance between data point and center of each cluster

予めクラスターの数を指定する必要がある

It is necessary to specify the number of clusters beforehand

K平均クラスタリング *k*-means clustering

$$D = \{x_1, x_2, \dots, x_N\}, x_i \in R^d$$

d次元データがN個ある There are N d-dimensional data points

μ_k : *k*番目のクラスターの代表ベクトル
Representative vector of *k*-th cluster

$M(\mu_k)$: μ_k のボロノイ領域
Voronoi region of representative vector μ_k

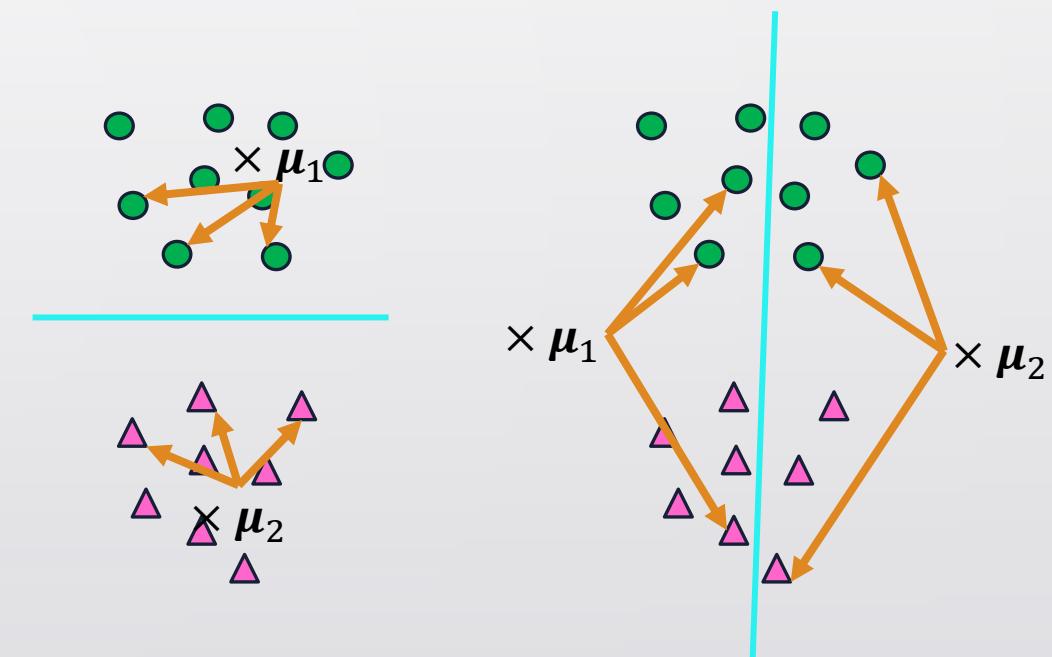
K平均クラスタリング *k*-means clustering

$$q_{i,k} = \begin{cases} 1 & (x_i \in M(\mu_k) \text{ の場合}) \quad \text{In case of } x_i \in M(\mu_k) \\ 0 & \end{cases}$$

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^N \sum_{k=1}^K q_{i,k} \|x_i - \mu_k\|^2$$

$J(q_{i,k}, \mu_k)$ を最小化する $q_{i,k}$ と μ_k を求める

Find $q_{i,k}$ and μ_k that minimize $J(q_{i,k}, \mu_k)$



K平均クラスタリング *k*-means clustering

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^N \sum_{k=1}^K q_{i,k} \|x_i - \mu_k\|^2$$

$$\frac{\partial J}{\partial \mu_k} = 0 \quad -2 \sum_{i=1}^N q_{i,k} (x_i - \mu_k) = 0$$

$$\mu_k = \frac{\sum_{i=1}^N q_{i,k} x_i}{\sum_{i=1}^N q_{i,k}}$$

μ_k と $q_{i,k}$ を同時に最適化するにはどうすればいいか？

How can we optimize μ_k and $q_{i,k}$ simultaneously?

K平均クラスタリング *k*-means clustering

$$J(q_{i,k}, \mu_k) = \sum_{i=1}^N \sum_{k=1}^K q_{i,k} \|x_i - \mu_k\|^2$$

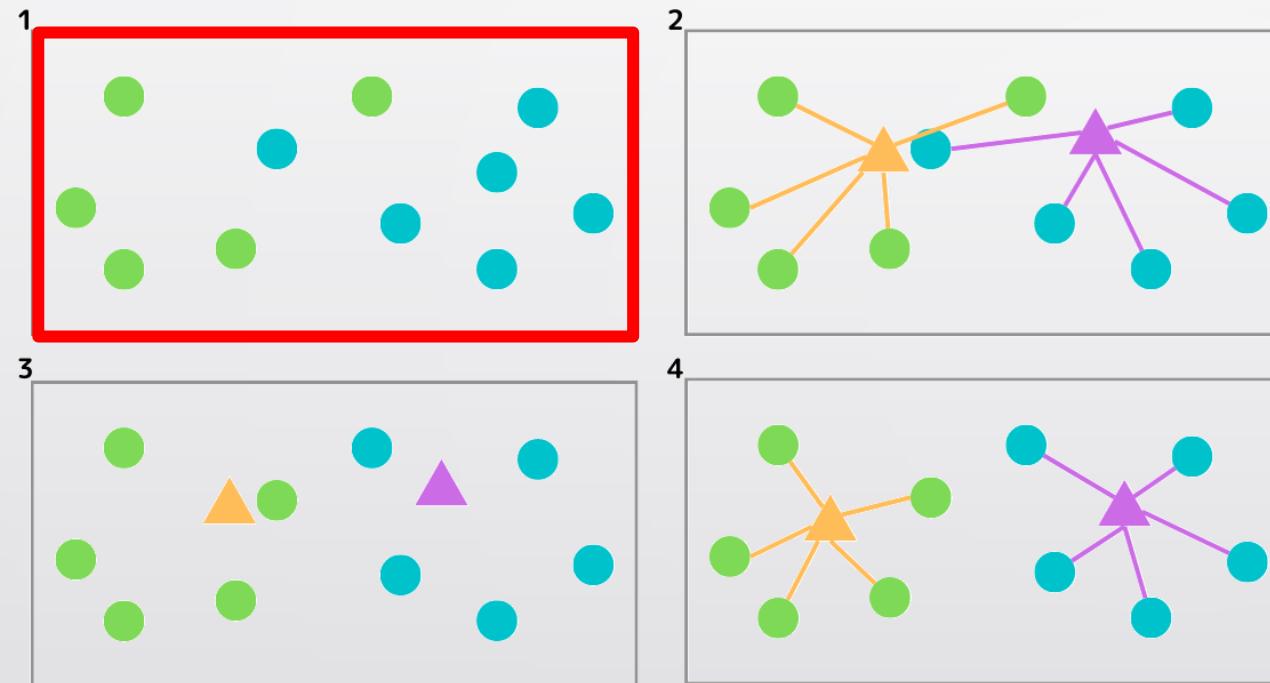
$$\frac{\partial J}{\partial \mu_k} = 0 \quad -2 \sum_{i=1}^N q_{i,k} (x_i - \mu_k) = 0$$

$$\mu_k = \frac{\sum_{i=1}^N q_{i,k} x_i}{\sum_{i=1}^N q_{i,k}}$$

μ_k と $q_{i,k}$ を同時に最適化するにはどうすればいいか？

How can we optimize μ_k and $q_{i,k}$ simultaneously?

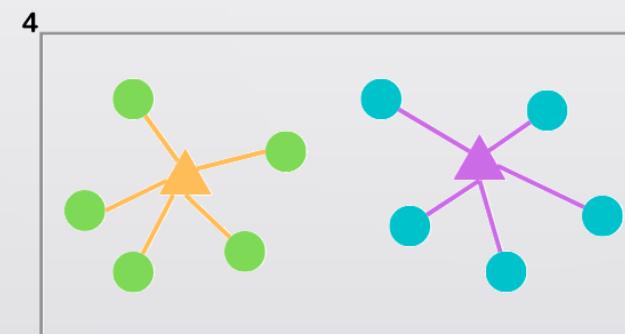
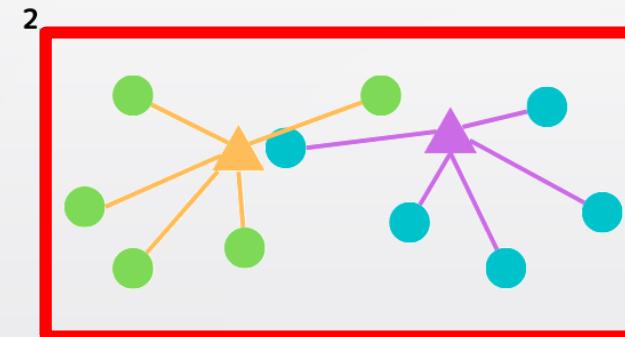
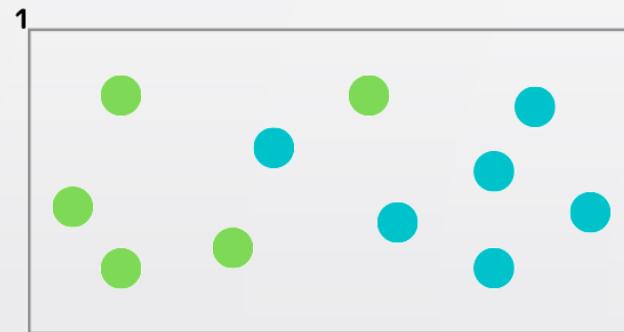
K平均クラスタリング k-means clustering



1. データをランダムにクラスターに割り当てる

Randomly assign data to one of the clusters

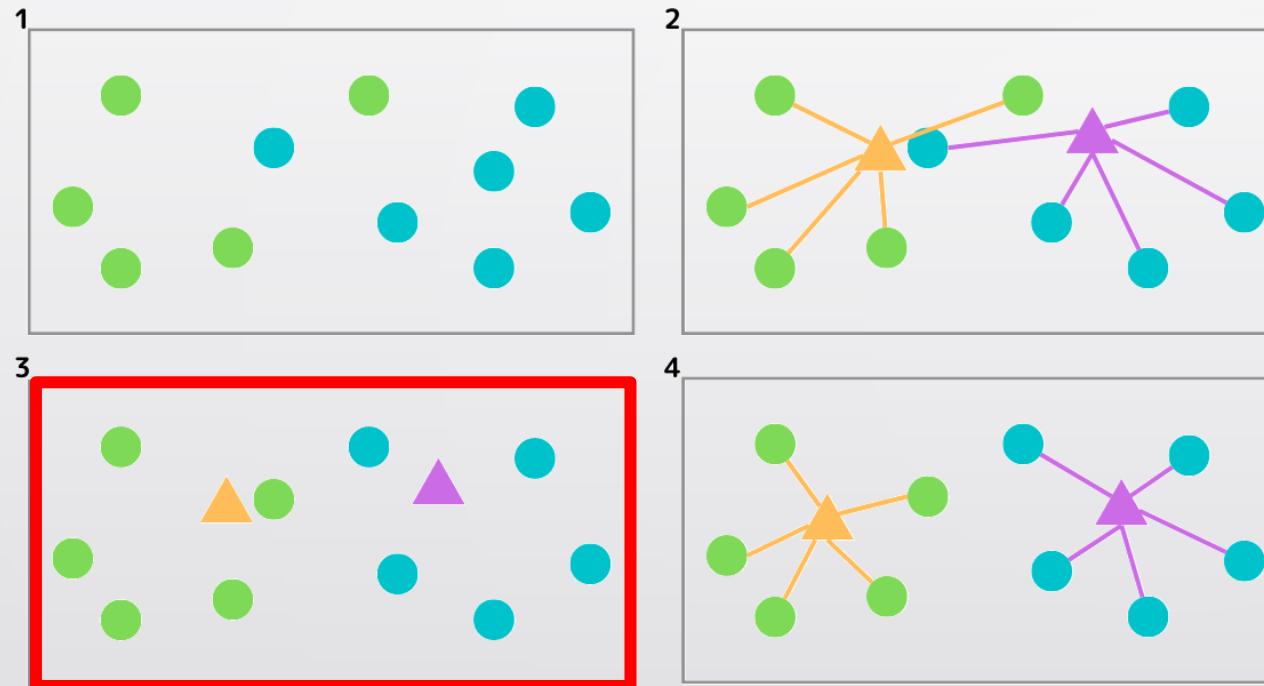
K平均クラスタリング k-means clustering



3. 各クラスター中心とデータとの距離を計算する

Compute the distance of data from center of each cluster

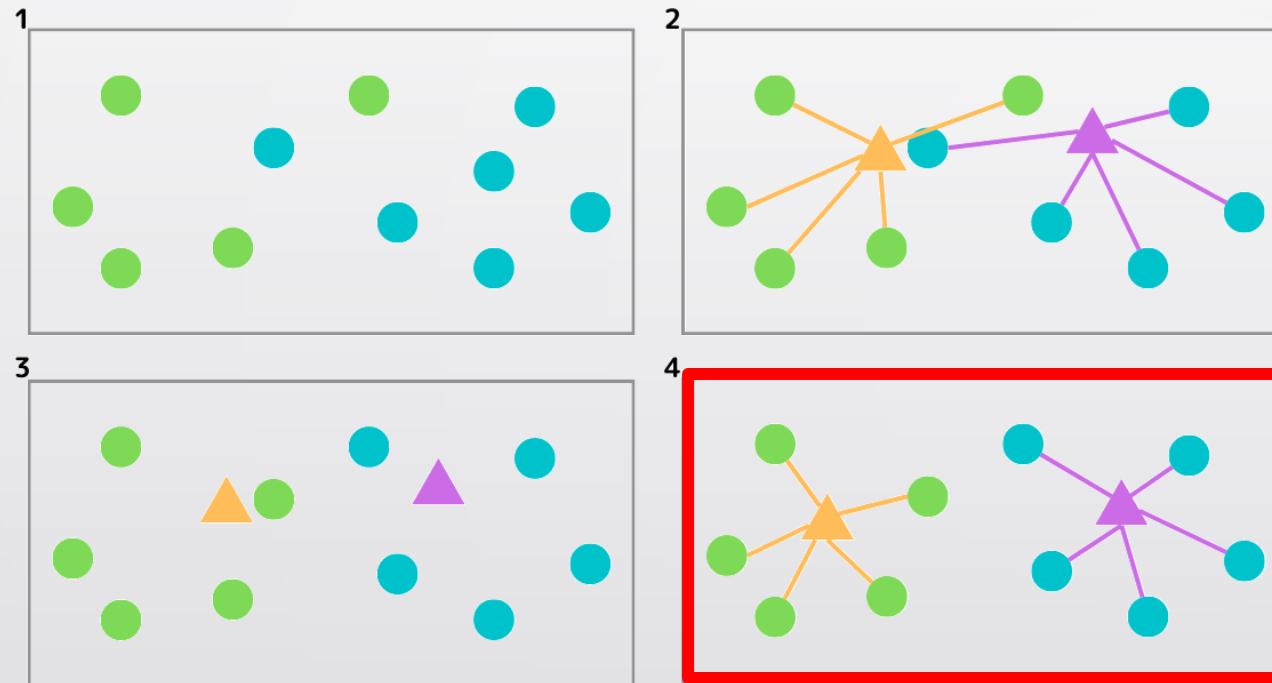
K平均クラスタリング k-means clustering



3. データと最も中心からの距離が近い
クラスターに割り当てる

Assign data point to the cluster with
smallest distance

K平均クラスタリング k-means clustering



4. データと最も中心からの距離が近い
クラスターに割り当てる

Assign data point to the cluster with
smallest distance

K平均クラスタリング *k*-means clustering

1. μ_k を固定し以下の方法で $q_{i,k}$ を決定する

Fix μ_k and determine $q_{i,k}$ following the rule below

$$q_{i,k} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|x_i - \mu_j\|^2 \\ 0 & \end{cases}$$

x_i を重心 μ_j との距離が一番近いクラスタに割り当てる

Assign x_i to the cluster whose centroid μ_j is closest to x_i

2. μ_k を最適化する

Optimize μ_k

$$\mu_k = \frac{\sum_{i=1}^N q_{i,k} x_i}{\sum_{i=1}^N q_{i,k}}$$

K平均クラスタリング K-means clustering

1-2.の手続きを収束するまで繰り返す Repeat Step 1-4 until the result converges

クラスター内の誤差平方和が閾値以下になることが収束条件である

Convergence Criterion is usually that squared-sum within cluster SSE_k becomes smaller than threshold

$$SSE_k = \sum \|x_i - \mu_k\|^2 \quad \sum SSE_k \leq Threshold$$

距離は、通常、ユークリッド距離を計算する

Usually, Euclidian distance is computed as the index of distance between data point and cluster center

クラスター数の決定法: エルボー法

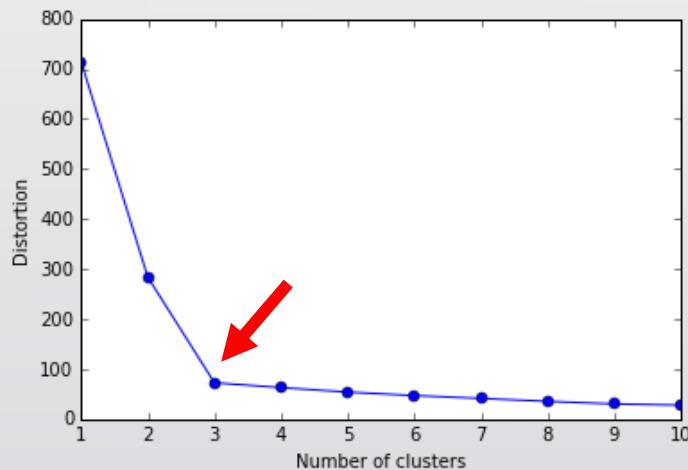
How to determine the number of Clusters: Elbow Method

- 様々なクラスター数でクラスタリングを行いSSEを計算する

Compute SSE after clustering with varying number of clusters

- SSEをプロットし、SSEの減少が平たんになるクラスター数を見つける

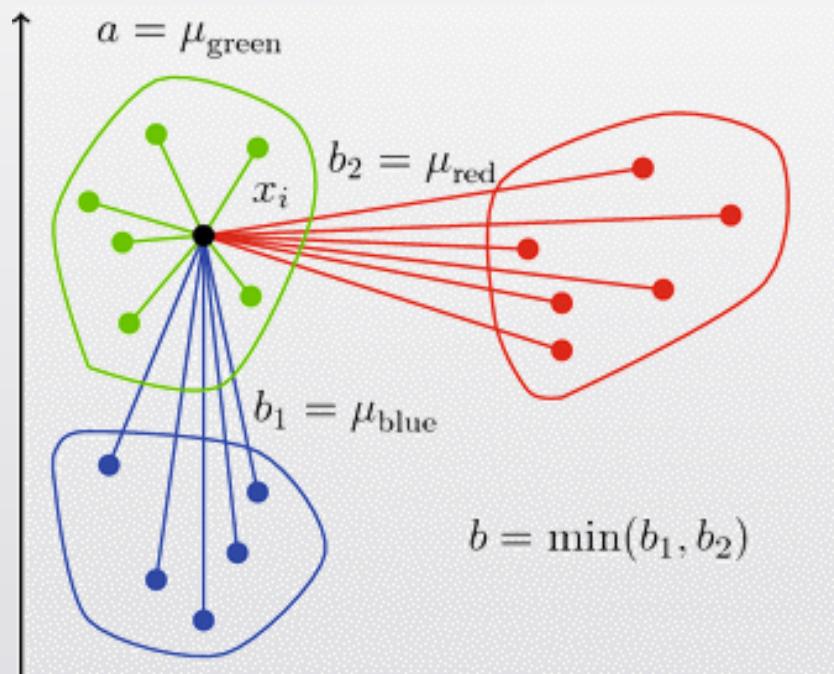
Plot SSEs and find the cluster number at which decrease of SSE reaches plateau



<https://qiita.com/deaikei/items/11a10fde5bb47a2cf2c2>

クラスター数の決定法: シルエット係数

How to determine the number of Clusters: Silhouette Coefficient



Usha Narra et al, 2016

a: クラスター内の他のデータとの距離の平均
a: Mean distance from other data points within cluster

b: 最も近いクラスターのデータとの距離の平均
a: Mean distance from data points in nearest cluster

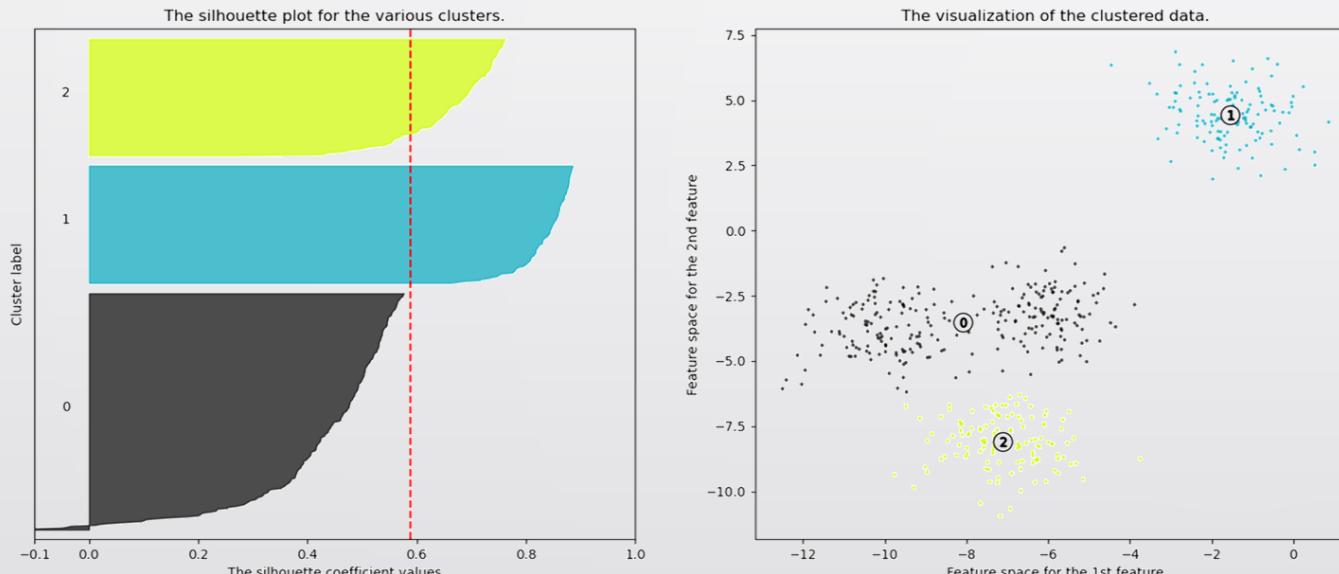
$$\textit{Silhouette Coefficient} = \frac{b - a}{\max(b, a)}$$

[−1, 1]の範囲で変動する
Ranges within [−1, 1]

クラスター数の決定法: シルエット係数

How to determine the number of Clusters: Silhouette Coefficient

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



全てのデータのシルエット係数をプロットしている
Silhouette coefficient of every data point is plotted

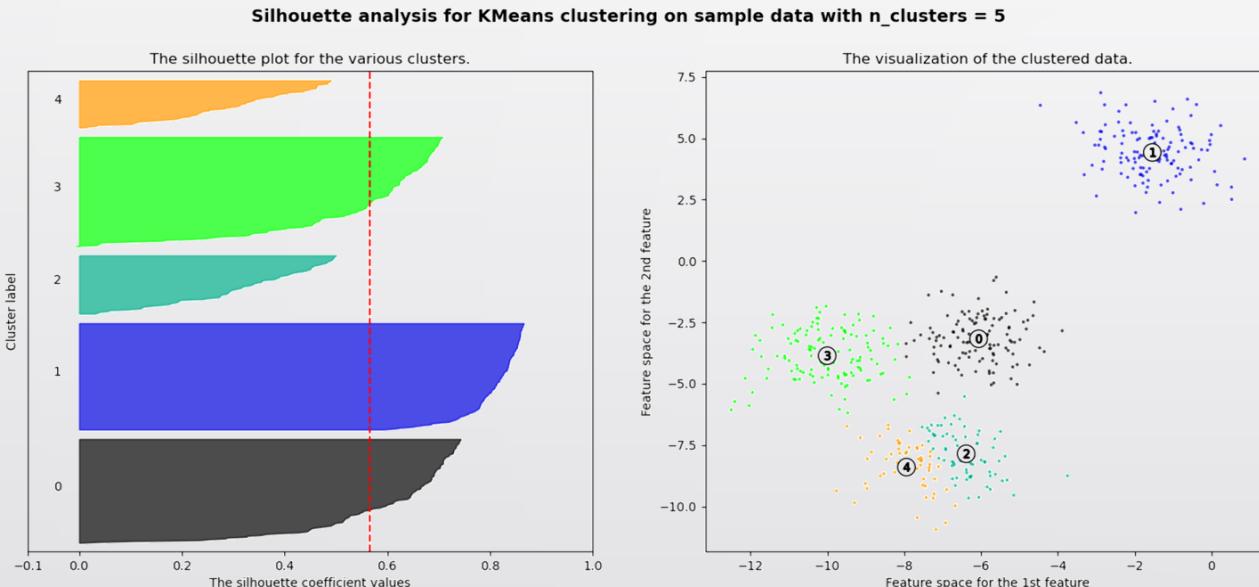
クラスター数が不適切な時
When number of clusters is not appropriate

シルエット係数が小さなクラスターがある
There are clusters with small silhouette coefficient

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

クラスター数の決定法: シルエット係数

How to determine the number of Clusters: Silhouette Coefficient



全てのデータのシルエット係数をプロットしている
Silhouette coefficient of every data point is plotted

クラスター数が不適切な時
When number of clusters is not appropriate

シルエット係数が小さなクラスターがある
There are clusters with small silhouette coefficient

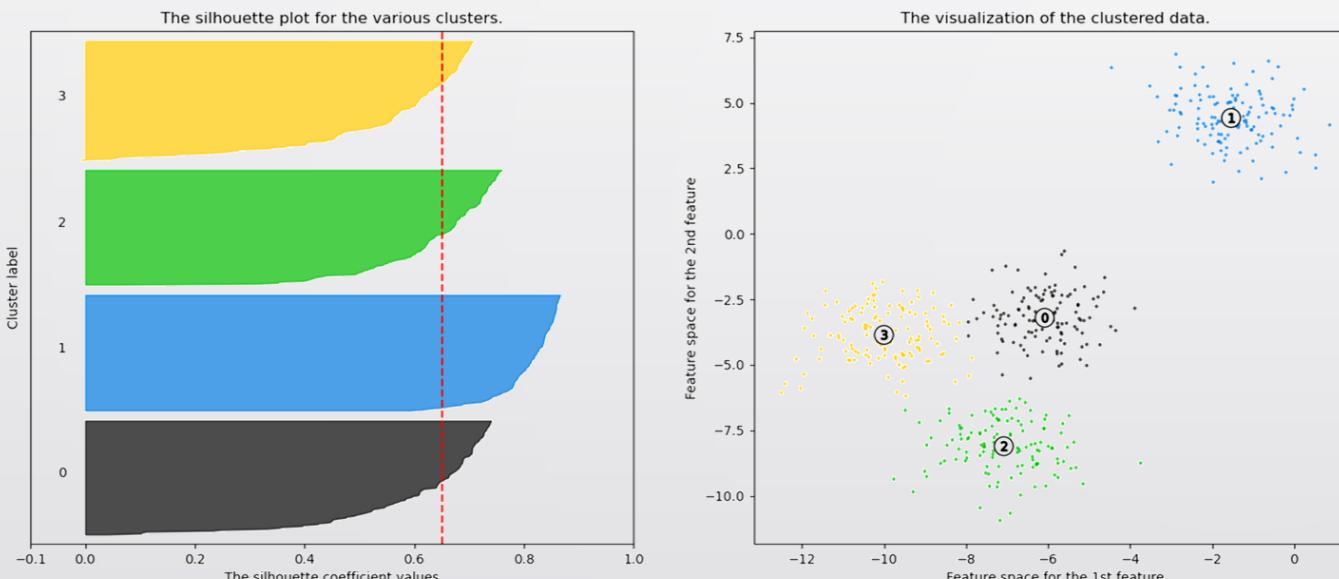
クラスターの大きさが不均一
Size of cluster is inhomogenous

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

クラスター数の決定法: シルエット係数

How to determine the number of Clusters: Silhouette Coefficient

Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



全てのデータのシルエット係数をプロットしている

Silhouette coefficient of every data point is plotted

クラスター数が適切な時

When number of clusters is not appropriate

全てのクラスターのシルエット係数
が大きい

Silhouette coefficient of every cluster is
large enough

クラスターの大きさが均一

Size of cluster is homogenous

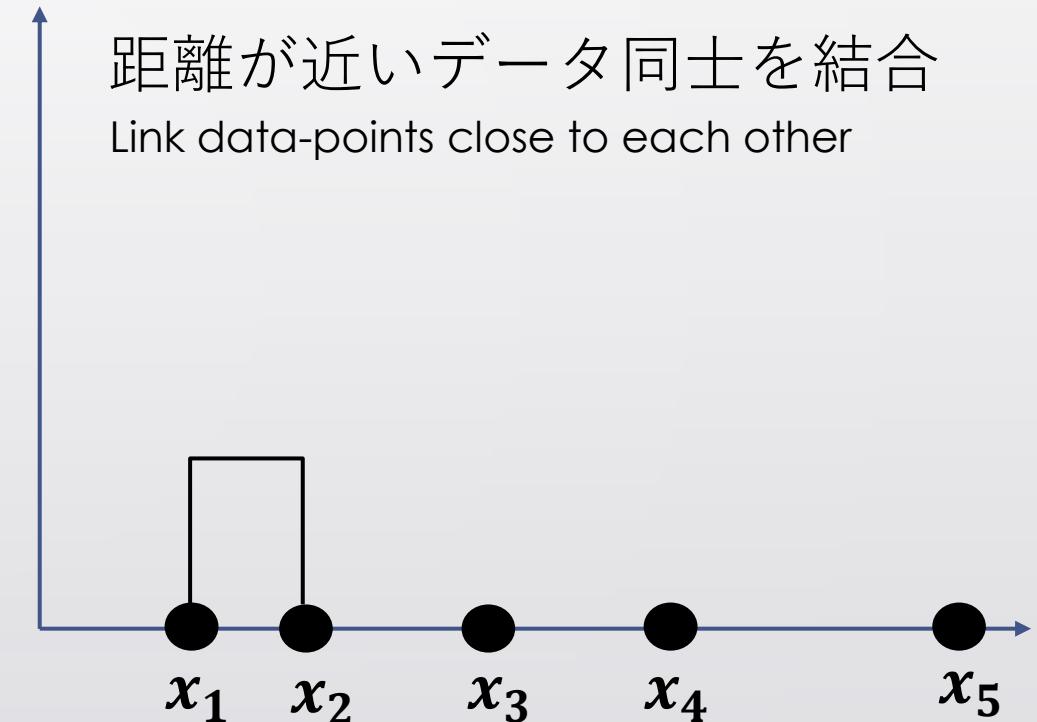
https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html



凝聚性階層的クラスタリング Agglomerative Hierarchical Clustering

	x_1	x_2	x_3	x_4	x_5
x_1	0				
x_2	1	0			
x_3	4	3	0		
x_4	6	5	2	0	
x_5	10	9	6	4	0

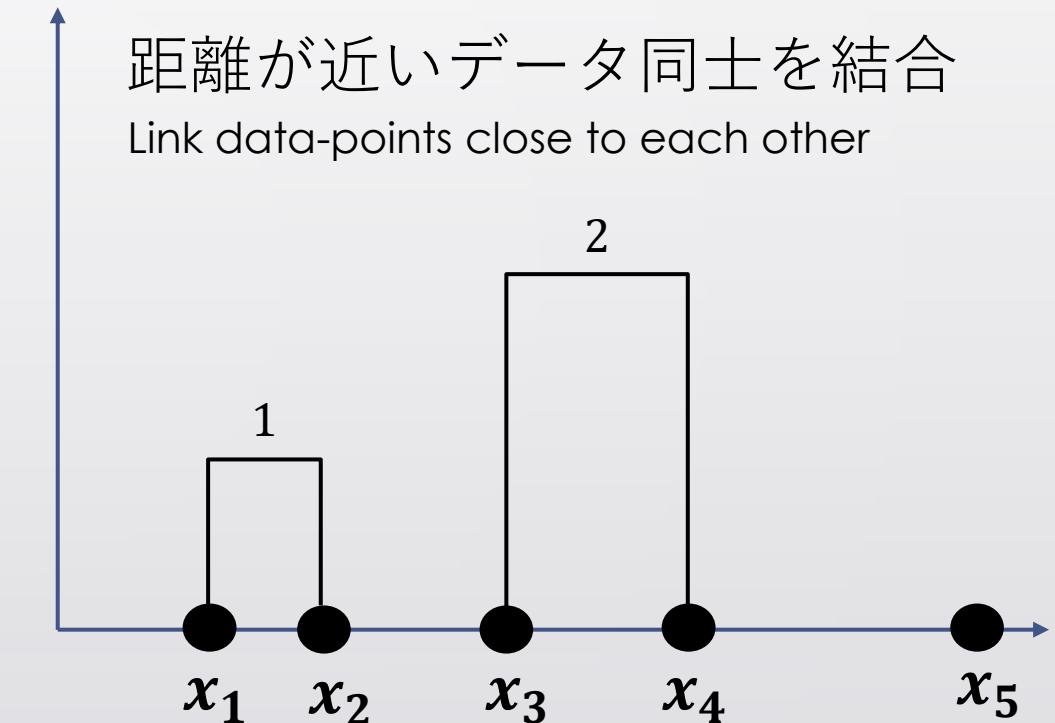
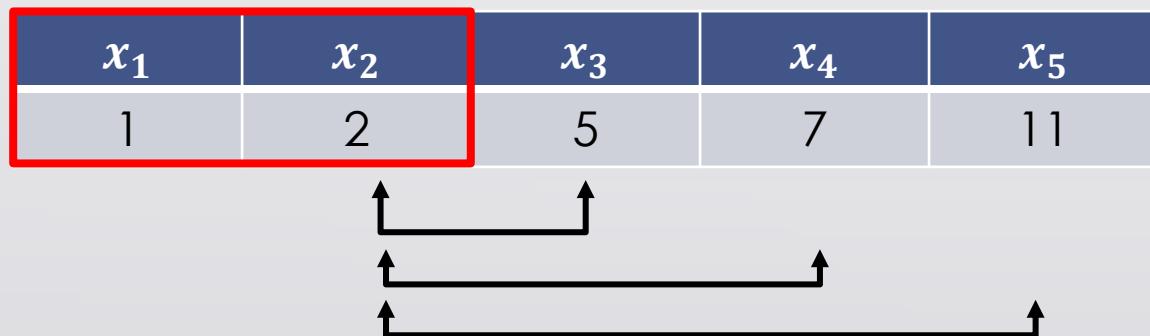
x_1	x_2	x_3	x_4	x_5
1	2	5	7	11





凝聚性階層的クラスタリング Agglomerative Hierarchical Clustering

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0			
x_3	3	0		
x_4	5	2	0	
x_5	9	6	4	0

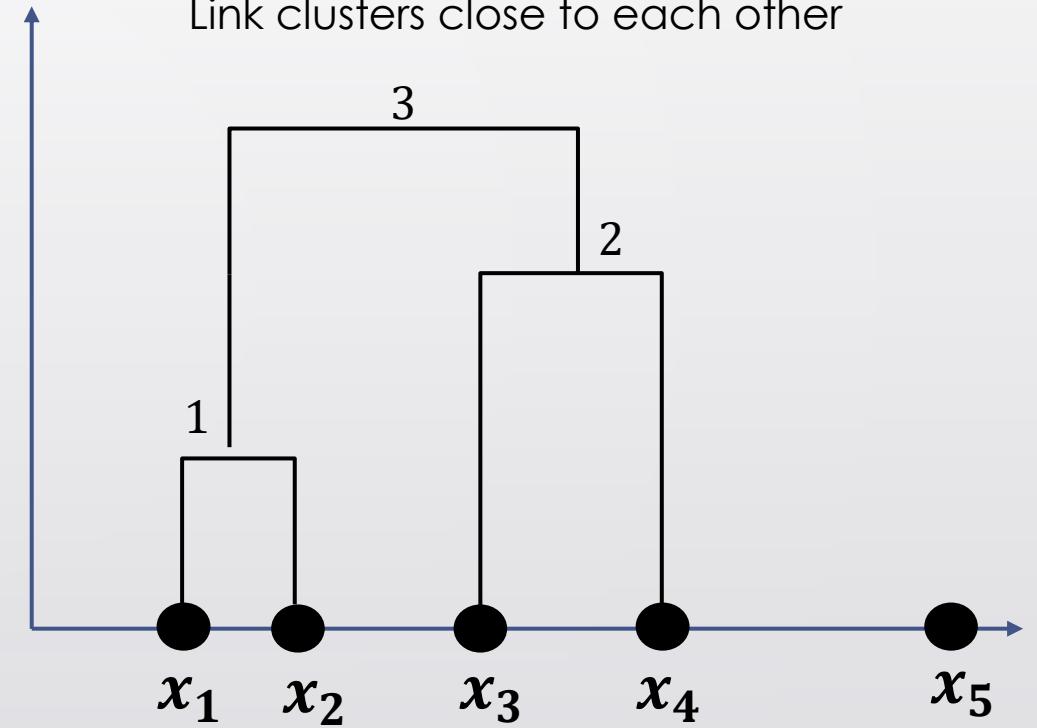
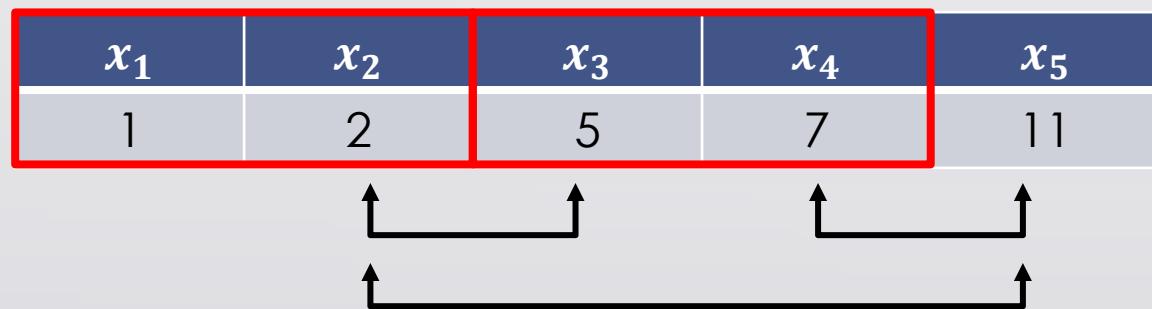


凝聚性階層的クラスタリング

Agglomerative Hierarchical Clustering

距離が近いクラスター同士を結合
Link clusters close to each other

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0		
$\{x_3, x_4\}$	3	0	
x_5	9	4	0

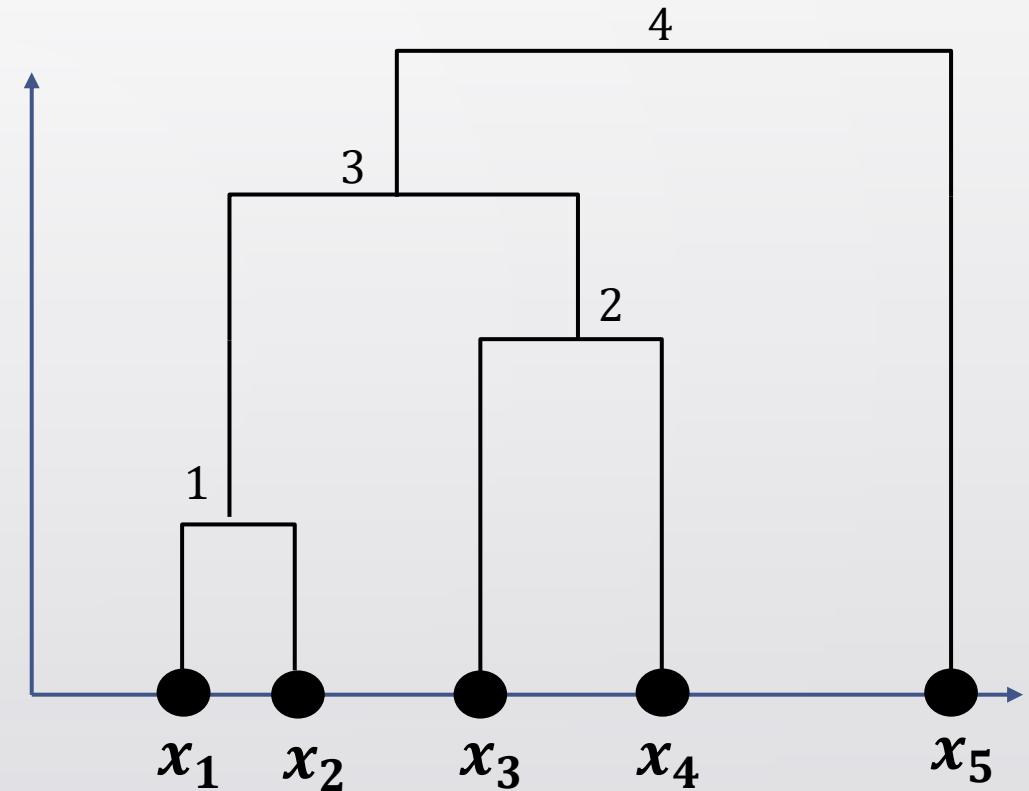
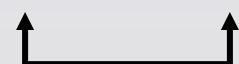




凝聚性階層的クラスタリング Agglomerative Hierarchical Clustering

	$\{x_1, x_2, x_3, x_4\}$	x_5
$\{x_1, x_2, x_3, x_4\}$	0	
x_5	4	0

x_1	x_2	x_3	x_4	x_5
1	2	5	7	11



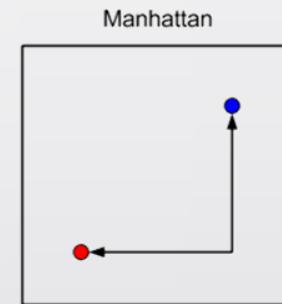


データ間の距離 Distance between Data-Points

ミンコフスキ距離 Minkowski Distance $Minkowski\ Distance = \left(\sum |x_{i,k} - x_{j,k}|^p \right)^{\frac{1}{q}}$

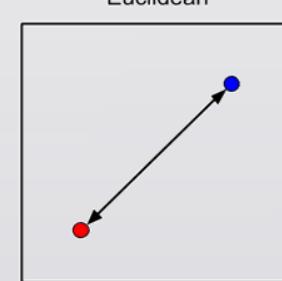
$p = 1, q = 1$ の時 In case of $p = 1, q = 1$

マンハッタン距離 Manhattan Distance = $\sum |x_{i,k} - x_{j,k}|$



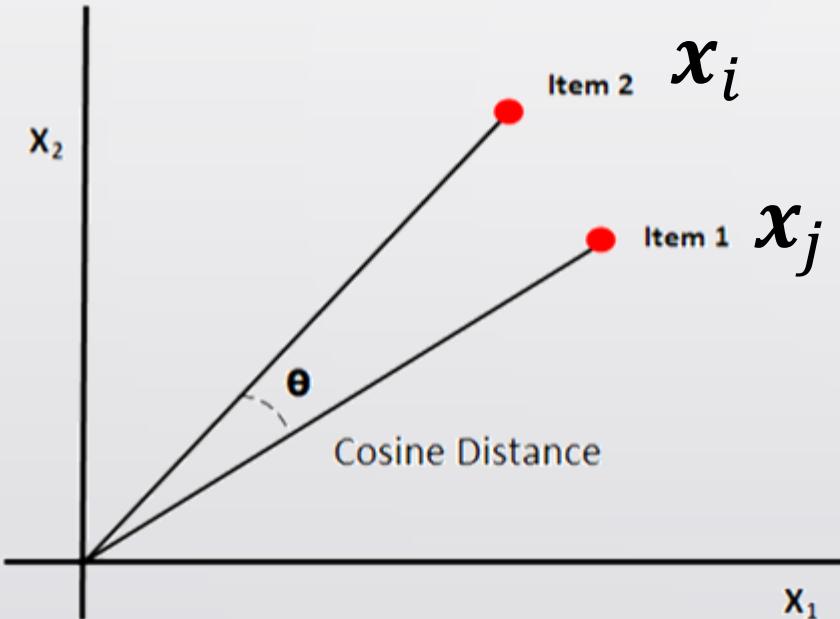
$p = 2, q = 2$ の時 In case of $p = 2, q = 2$

ユークリッド距離 Euclidean Distance = $\sqrt{\sum (x_{i,k} - x_{j,k})^2}$



データ間の距離 Distance between Data-Points

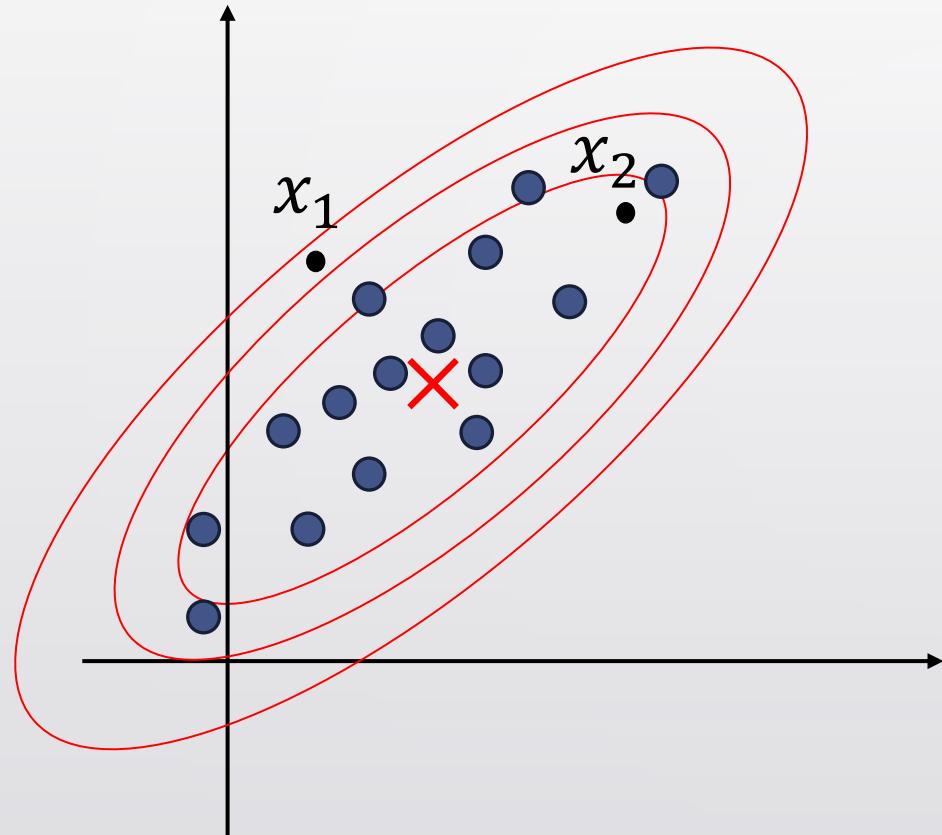
コサイン類似度 Cosine Similarity



$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\text{dot}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|} = \frac{\sum x_{i,k} \times x_{j,k}}{\sum x_{i,k}^2 \times \sum x_{j,k}^2}$$

<https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml>

マハラノビス距離 Mahalanobis Distance



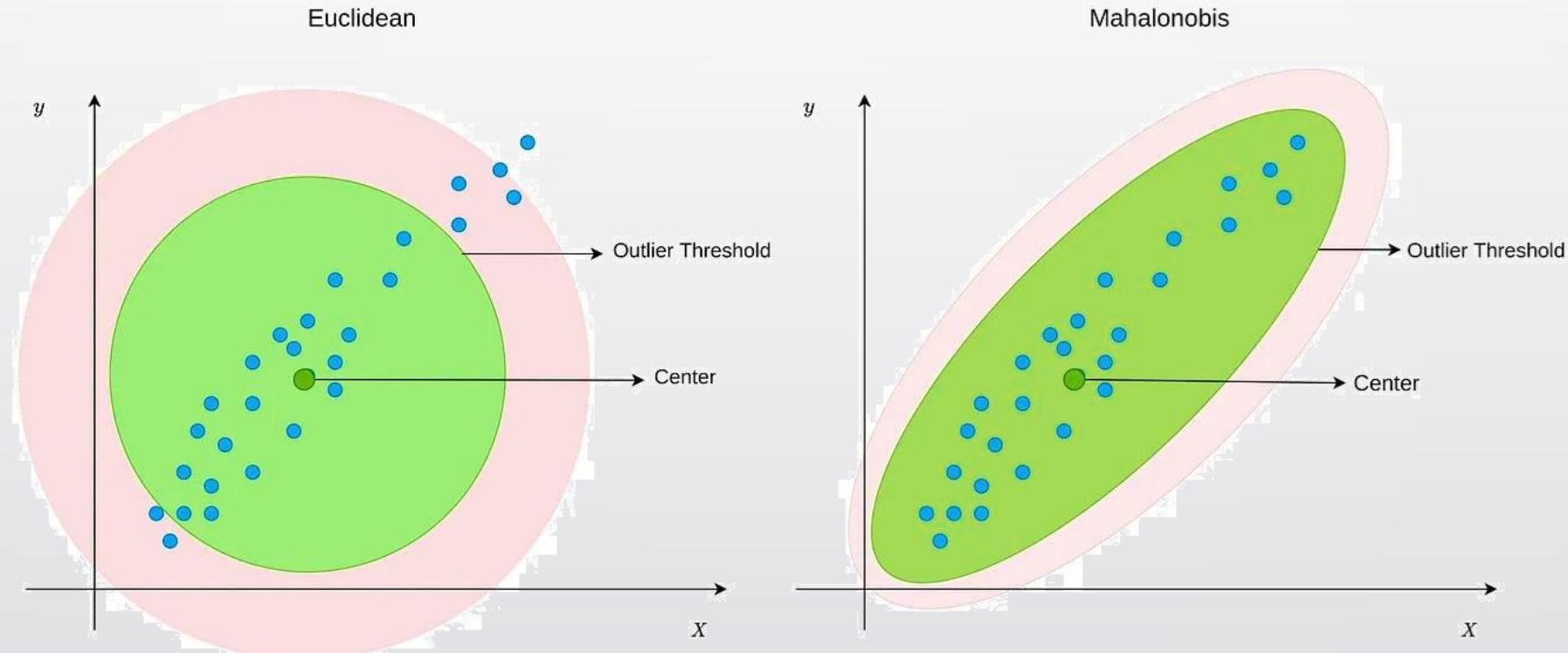
$$D_M(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\Sigma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T}$$

Σ : 分散共分散行列
Variance-Covariance Matrix

分布の形状を考慮した分布の中心からの距離の指標

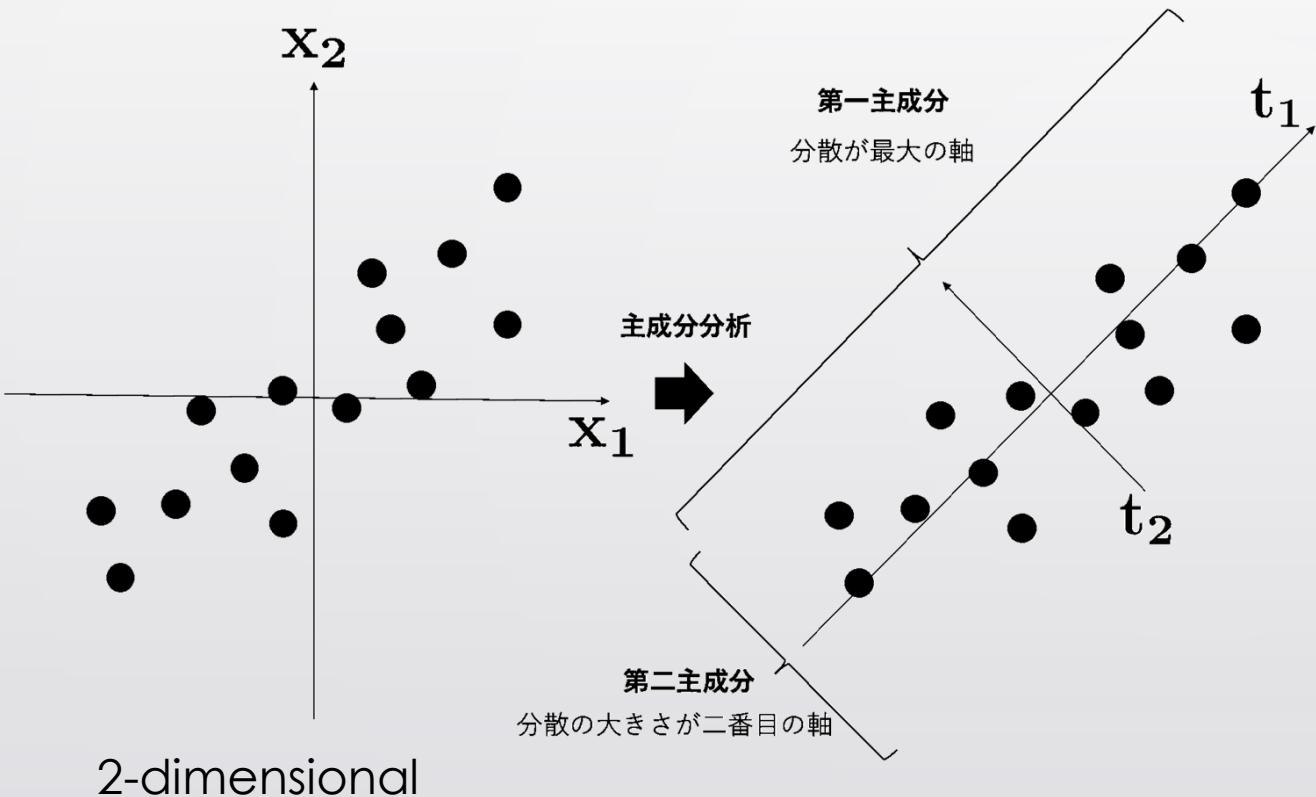
Measure of distance from distribution center
adjusted by the shape of data distribution

マハラノビス距離 Mahalanobis Distance



<https://bob3.hatenablog.com/entry/2023/04/22/113540>

主成分 Principal Components



第1主成分軸は、データの分散が最大化される方向を向いている

The first PC axis is oriented in the direction along which variance of projected data is maximized

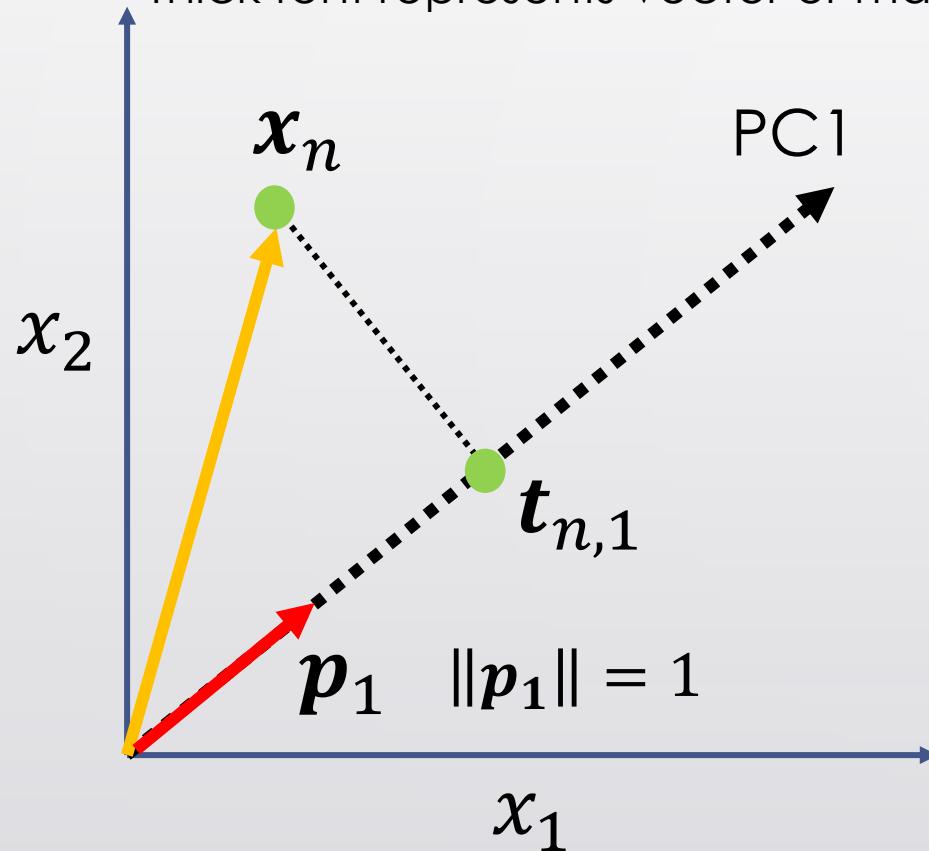
第 j 主成分軸は、データの分散が j 番目の大きさになる方向を向いている

The j -th PC axis is oriented in the direction along which projected data has j -th largest variance

第1主成分の計算 Computation of PC1

太字はベクトルか行列

Thick font represents vector or matrix



変数を中心化しておく Center the variables

観測データ x_n の第1主成分軸方向への
射影 $t_{n,1}$ を計算する

Compute the projection $t_{n,1}$ of the observed
data x_n onto the first PC axis

$t_{n,1}$ は x_n と p_1 の内積

$t_{n,1}$ is dot(inner) product of x_n and p_1

第1主成分の計算 Computation of PC1

$V\mathbf{p}_1 = \lambda\mathbf{p}_1$ \mathbf{p}_1 は V の固有ベクトルである \mathbf{p}_1 is eigenvector of V

$V = \frac{1}{N} \mathbf{X}^T \mathbf{X}$ \mathbf{t}_1 の分散は λ に一致する Variance of \mathbf{t}_1 equals to λ

V とは何か？ What is V ？

$$V = \frac{1}{N} \mathbf{X}^T \mathbf{X} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

分散共分散行列 Variance-Covariance Matrix

V は X の分散共分散行列である V is variance-covariance matrix of X

$$V = \frac{1}{N} X^T X = \frac{1}{N} \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,M} & x_{2,M} & \dots & x_{N,M} \end{bmatrix} \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \dots & \sigma_{1,M}^2 \\ \sigma_{2,1}^2 & \ddots & & \vdots \\ \vdots & & & \sigma_{M-1,M}^2 \\ \sigma_{M,1}^2 & \sigma_{M,2}^2 & \dots & \sigma_{M,M}^2 \end{bmatrix} \quad \sigma_{i,j}^2 = \frac{1}{N} \sum_{k=1}^N x_{k,i} x_{k,j}$$

分散共分散行列の対角化

Diagonalization of Variance-Covariance Matrix

$$\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M)$$

$$V\mathbf{p}_i = \lambda_i \mathbf{p}_i \quad \|\mathbf{p}_i\| = 1$$

対称行列の異なる固有値に対する固有ベクトルは直交するので

Since eigenvectors of symmetric matrix corresponding to different eigen values are orthogonal

$$\mathbf{p}_i \mathbf{p}_j^T = \begin{cases} 1 & (\lambda_i = \lambda_j) \\ 0 & (\lambda_i \neq \lambda_j) \end{cases}$$

分散共分散行列の対角化

Diagonalization of Variance-Covariance Matrix

$$V\mathbf{P} = V(\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M) = (\lambda_1 \mathbf{p}_1, \lambda_2 \mathbf{p}_2, \dots, \lambda_M \mathbf{p}_M)$$

$$\mathbf{p}_i^T V \mathbf{P} = (\lambda_1 \mathbf{p}_i^T \mathbf{p}_1, \lambda_2 \mathbf{p}_i^T \mathbf{p}_2, \dots, \lambda_M \mathbf{p}_i^T \mathbf{p}_M) = (0, 0, \dots, \lambda_i, \dots, 0)$$

$$\mathbf{P}^T V \mathbf{P} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_M \end{bmatrix}$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$\mathbf{x}_1 = (x_{11}, x_{12}) \quad \mathbf{x}_2 = (x_{21}, x_{22})$$



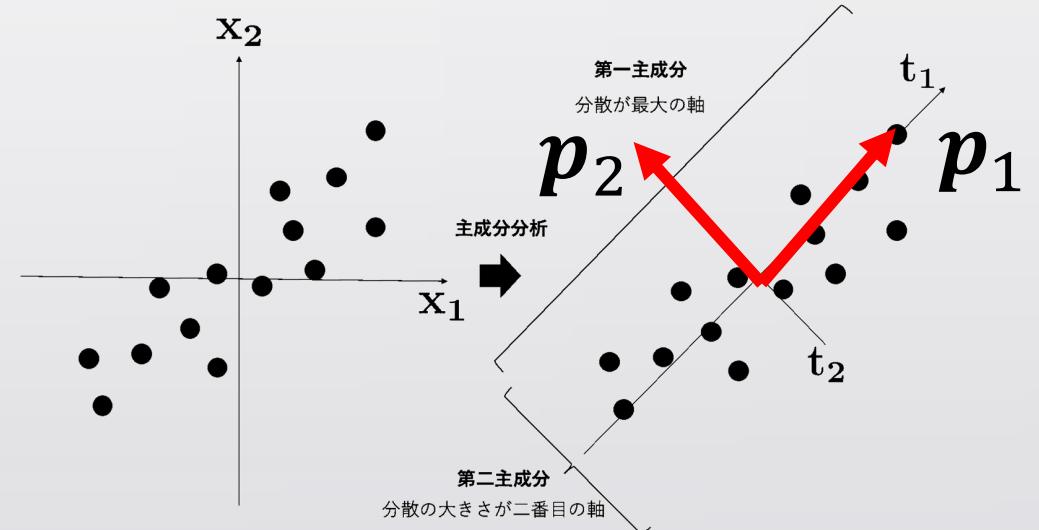
\mathbf{x}_i の \mathbf{p}_k への射影を計算 Project \mathbf{x}_i onto \mathbf{p}_k

$$\mathbf{u}_1 = (u_{11}, u_{12}) = \mathbf{x}_1(\mathbf{p}_1, \mathbf{p}_2)$$

$$\mathbf{u}_2 = (u_{21}, u_{22}) = \mathbf{x}_2(\mathbf{p}_1, \mathbf{p}_2)$$

$(\mathbf{p}_1, \mathbf{p}_2)$ を基底とする座標系では \mathbf{x}_1 は \mathbf{u}_1 と表現される

\mathbf{x}_1 is expressed as \mathbf{u}_1 in a coordinate system defined by basis vectors of $(\mathbf{p}_1, \mathbf{p}_2)$





マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$\mathbf{u}_1 = (u_{11}, u_{12}) = \mathbf{x}_1(\mathbf{p}_1, \mathbf{p}_2) \quad \mathbf{u}_2 = (u_{21}, u_{22}) = \mathbf{x}_2(\mathbf{p}_1, \mathbf{p}_2)$$

$$\begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \mathbf{p}_1 & \mathbf{x}_1 \mathbf{p}_2 \\ \mathbf{x}_2 \mathbf{p}_1 & \mathbf{x}_2 \mathbf{p}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} (\mathbf{p}_1, \mathbf{p}_2)$$

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,M} \\ u_{2,1} & u_{2,2} & \dots & u_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N,1} & u_{N,2} & \dots & u_{N,M} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} (\mathbf{p}_1, \mathbf{p}_2 \dots \mathbf{p}_M) = X P$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \dots & u_{1,M} \\ u_{2,1} & u_{2,2} & \dots & u_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N,1} & u_{N,2} & \dots & u_{N,M} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix} (\mathbf{p}_1, \mathbf{p}_2 \dots \mathbf{p}_M) = X P$$

U の 分 散 共 分 散 行 列 は Variance-covariance matrix of U is

$$\Sigma = \frac{1}{N} \mathbf{U}^T \mathbf{U} = \frac{1}{N} (X P)^T X P = \frac{1}{N} P^T X^T X P = P^T V P = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \lambda_M \end{bmatrix}$$

マハラノビス距離の意味 Meaning of Mahalanobis Distance

固有ベクトルで線型変換した後、マハラノビス距離を計算する

Calculate Mahalanobis distance after linear transformation by eigen vectors

$$D_M(\mathbf{u}_i) = \sqrt{(\mathbf{u}_i - \bar{\mathbf{u}})\Sigma^{-1}(\mathbf{u}_i - \bar{\mathbf{u}})^T}$$

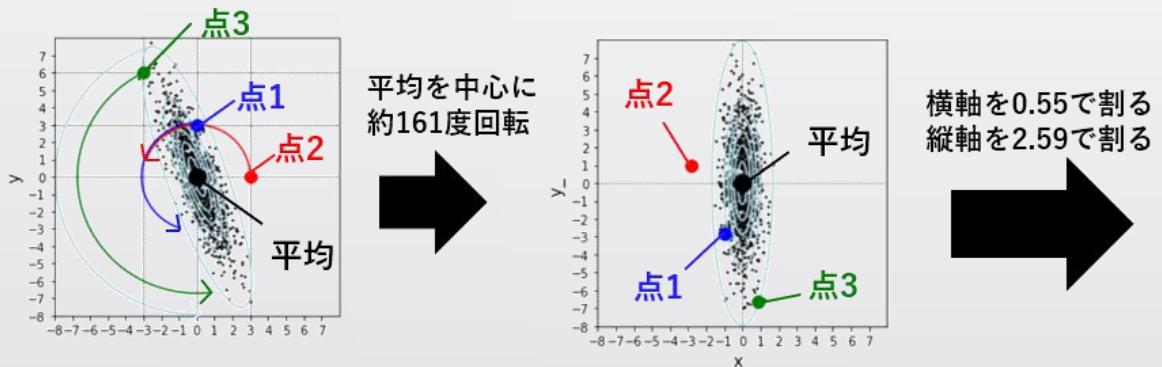
λ_i はデータの \mathbf{p}_i 方向の分散と一致

λ_i equals to variance of data along the direction of \mathbf{p}_i

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\lambda_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\lambda_M} \end{bmatrix}$$

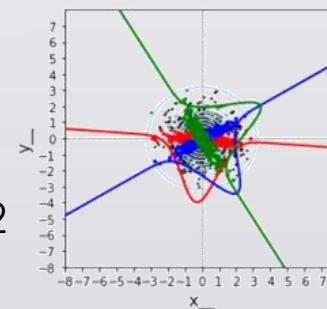
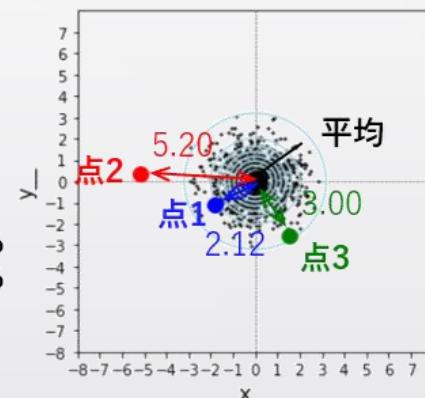
マハラノビス距離の意味 Meaning of Mahalanobis Distance

$$D_M(\mathbf{u}_i) = \sqrt{(\mathbf{u}_i - \bar{\mathbf{u}})\Sigma^{-1}(\mathbf{u}_i - \bar{\mathbf{u}})^T}$$



固有ベクトルで線型変換
Linear transformation by eigen vectors

<https://qiita.com/yutera12/items/db425fafce2d87a25a1f>



すべての方向の分散を一致させた後、ユークリッド距離を計算

Compute Euclidian distance after equalizing variance across all the directions

Jaccard 系数 Jaccard Coefficient

集合の類似度の指標 Measure of similarity between two sets

ベクトル間の類似度の指標としても用いることが出来る

Can be used as a measure of similarity between vectors

$$x = \{x_1, x_2 \dots x_n\} \quad y = \{y_1, y_2 \dots y_n\}$$

$$Jaccard(x, y) = \frac{|x \cap y|}{|x \cup y|} \quad Jaccard(x, y) = \frac{xy^T}{xx^T + yy^T - xy^T}$$

Dice 系数 Dice Coefficient

集合の類似度の指標 Measure of similarity between two sets

ベクトル間の類似度の指標としても用いることが出来る

Can be used as a measure of similarity between vectors

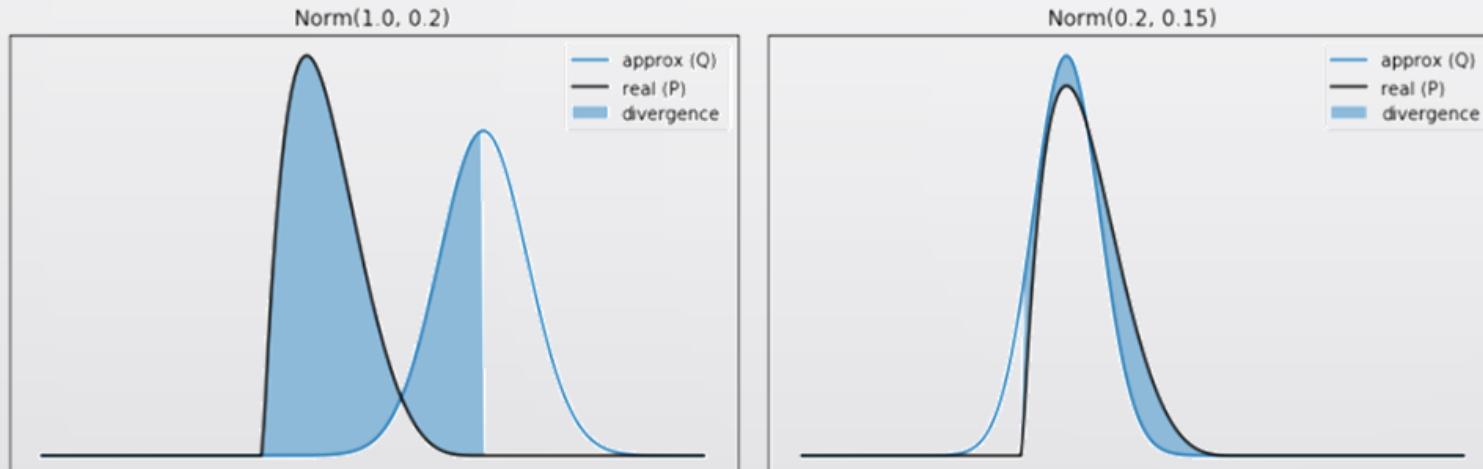
$$x = \{x_1, x_2 \dots x_n\} \quad y = \{y_1, y_2 \dots y_n\}$$

$$Dice(x, y) = \frac{2|x \cap y|}{|x| + |y|}$$

$$Dice(x, y) = \frac{2xy^T}{\|x\| + \|y\|}$$

KLダイバージェンス Kullback-Leibler Divergence

分布同士の類似度の評価指標 Measure of similarity between two distributions



$$KL(p(x)|q(x)) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

<https://jessicastringham.net/2018/12/27/KL-Divergence/>



データマイニング

Data Mining

12: クラスタリング② Clustering

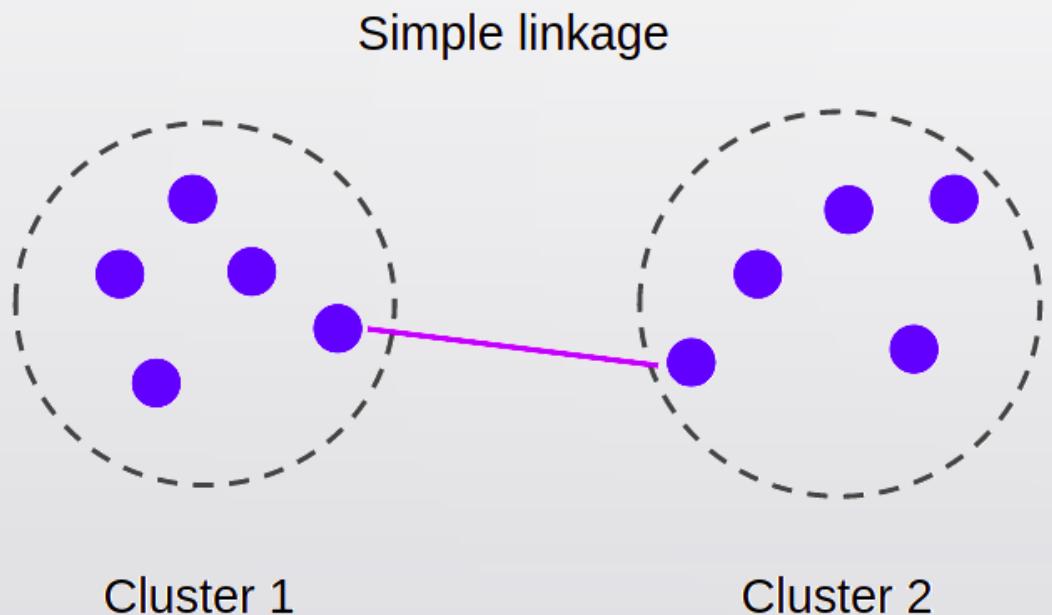
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology



クラスター間の距離 Distance between Clusters

単リンク法 Single Linkage



$$D(A, B) = \min_{x \in A, y \in B} d(x, y)$$

各クラスターのデータの内、最も近いデータ間の距離を、クラスター間の距離とする

Distance between clusters is defined as the distance between their closest members

<https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/>

单リンク法 Single Linkage

- ・大きなクラスターが形成されやすい Large cluster is likely to be formed
- ・近いデータ同士が別のクラスターに含まれてしまう連鎖効果が起きやすい
Neighboring data points tend to be included in separate clusters (chain effect)

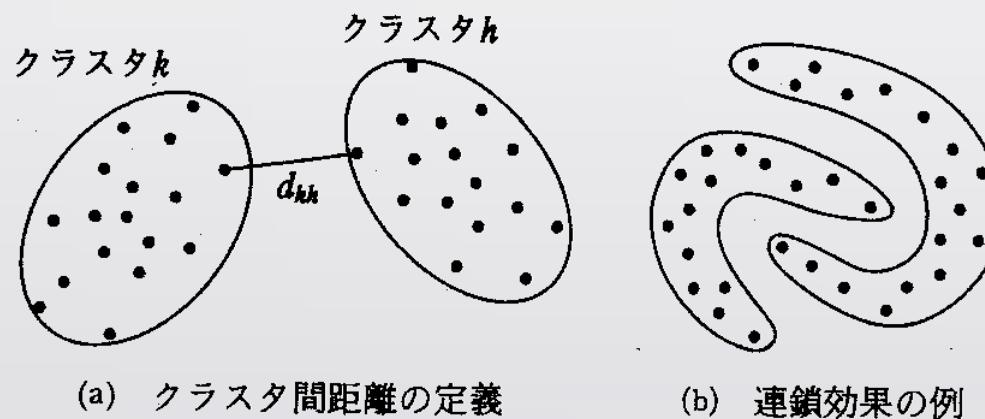
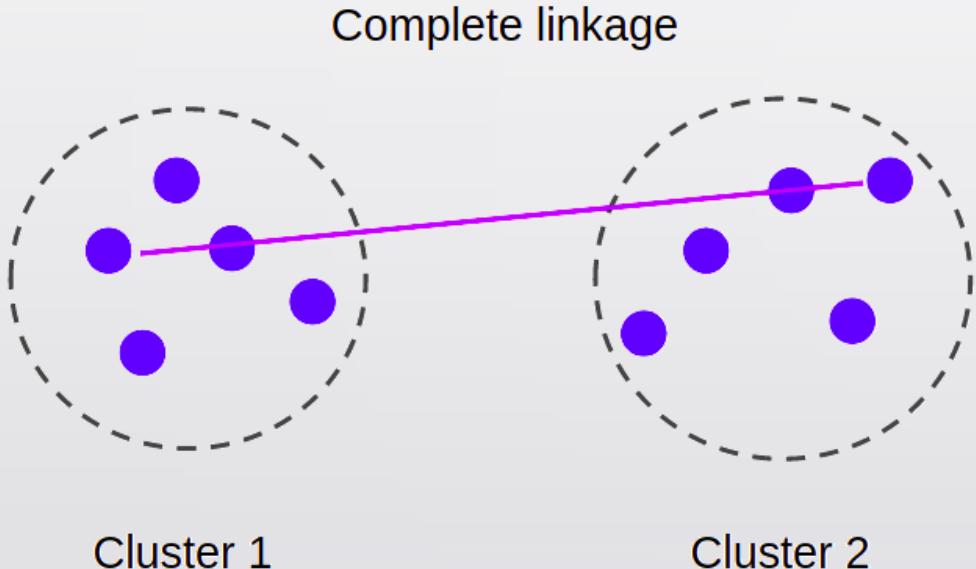


図 1.2.5 最短距離法におけるクラスタ間距離の定義と連鎖効果 <https://www.is.kochi-u.ac.jp/kyoko/edu/image/c.html>

クラスター間の距離 Distance between Clusters

完全リンク法 Complete Linkage



$$D(A, B) = \max_{x \in A, y \in B} d(x, y)$$

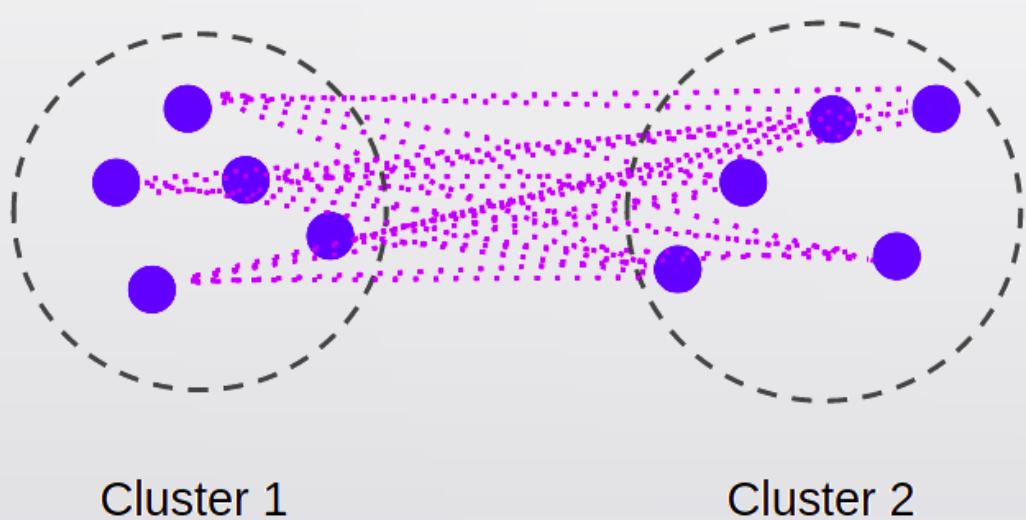
各クラスターのデータの内、最も遠いデータ間の距離を、クラスター間の距離とする

Distance between clusters is defined as the distance between their farthest members

クラスター間の距離 Distance between Clusters

平均リンク法 Average Linkage

Average linkage



$$D(A, B) = \frac{1}{N_A N_B} \sum_{x \in A} \sum_{y \in B} d(x, y)$$

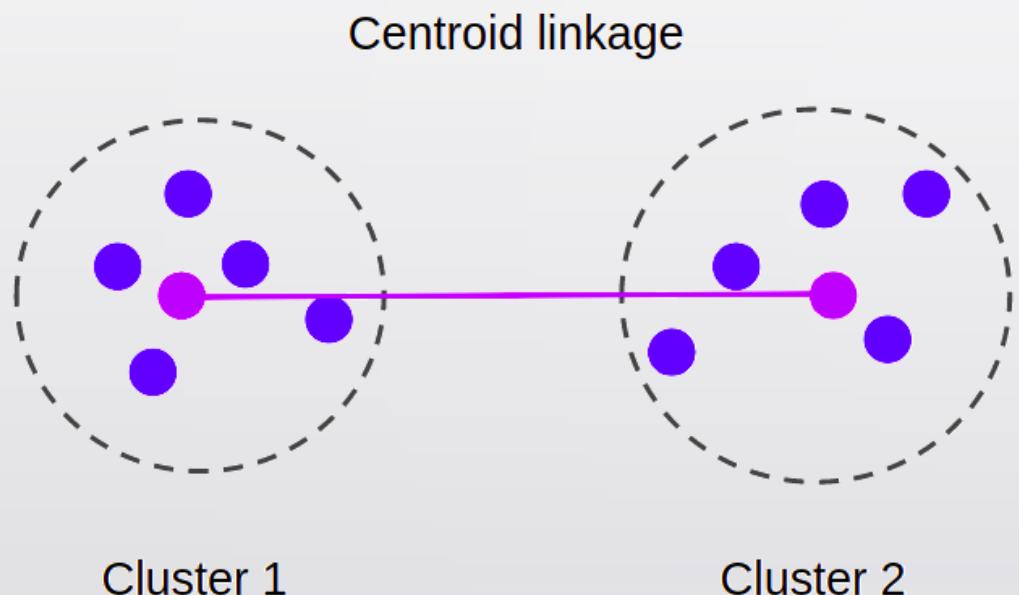
各クラスターのすべてのデータペアの平均

Distance between clusters is defined as average distance of all the between-cluster data pairs



クラスター間の距離 Distance between Clusters

中心リンク法 Centroid Linkage



$$D(A, B) = d(\mu_A, \mu_B)$$

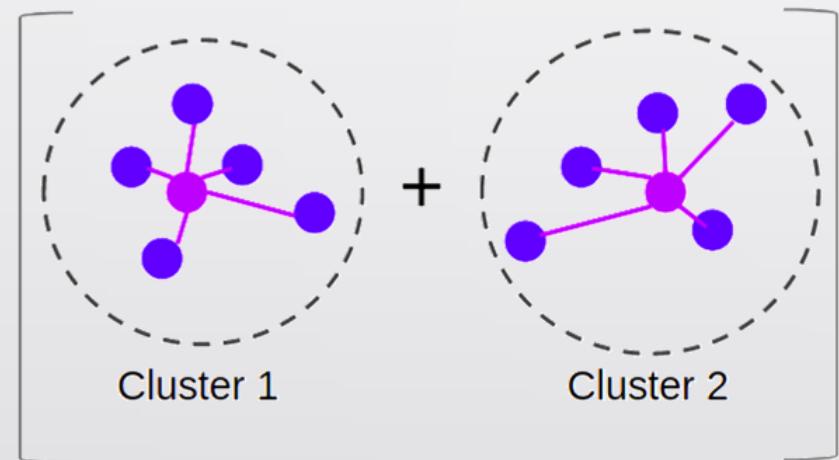
各クラスターの中心間の距離

Distance between clusters is defined as the distance between cluster centers

ウォード法 Ward Linkage

Δ が最小になるようなクラスター同士を結合する

Link clusters with minimum Δ



クラスター内SSEの合計を計算する

Compute sum of intra-cluster sum of squared error (SSE)

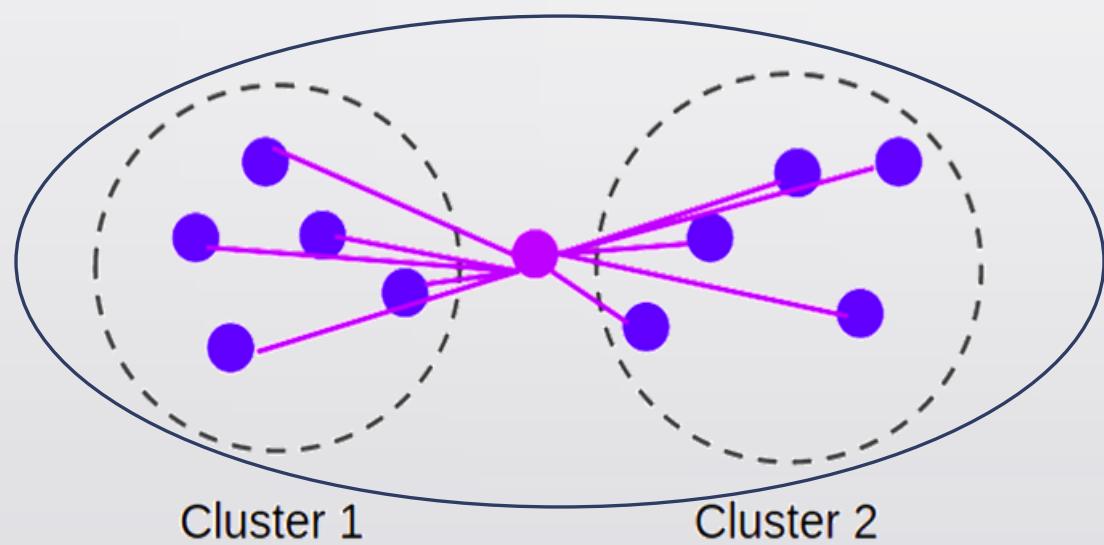
$$\sum_{x \in A} d(x, \mu_A)^2 + \sum_{y \in B} d(y, \mu_B)^2$$



ウォード法 Ward Linkage

Δ が最小になるようなクラスター同士を結合する

Link clusters with minimum Δ



クラスターを結合した時のSSEを計算する

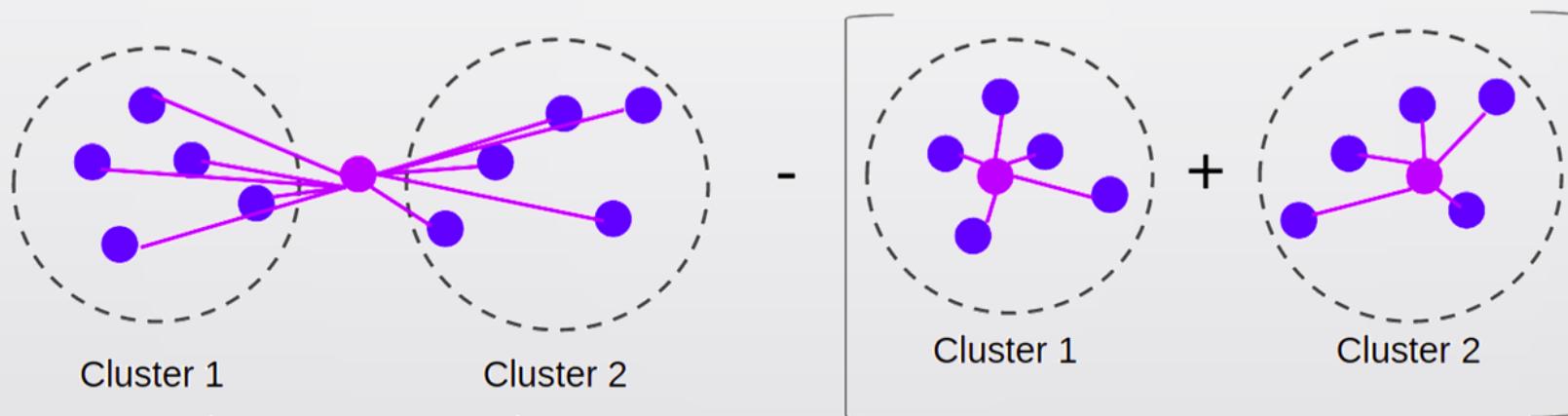
Compute within-cluster sum of squared error (SSE)
when the two clusters are joined to form single
cluster

$$\sum_{x \in AB} d(x, \mu_{AB})^2$$

ウォード法 Ward Linkage

Δ が最小になるようなクラスター同士を結合する

Link clusters with minimum Δ

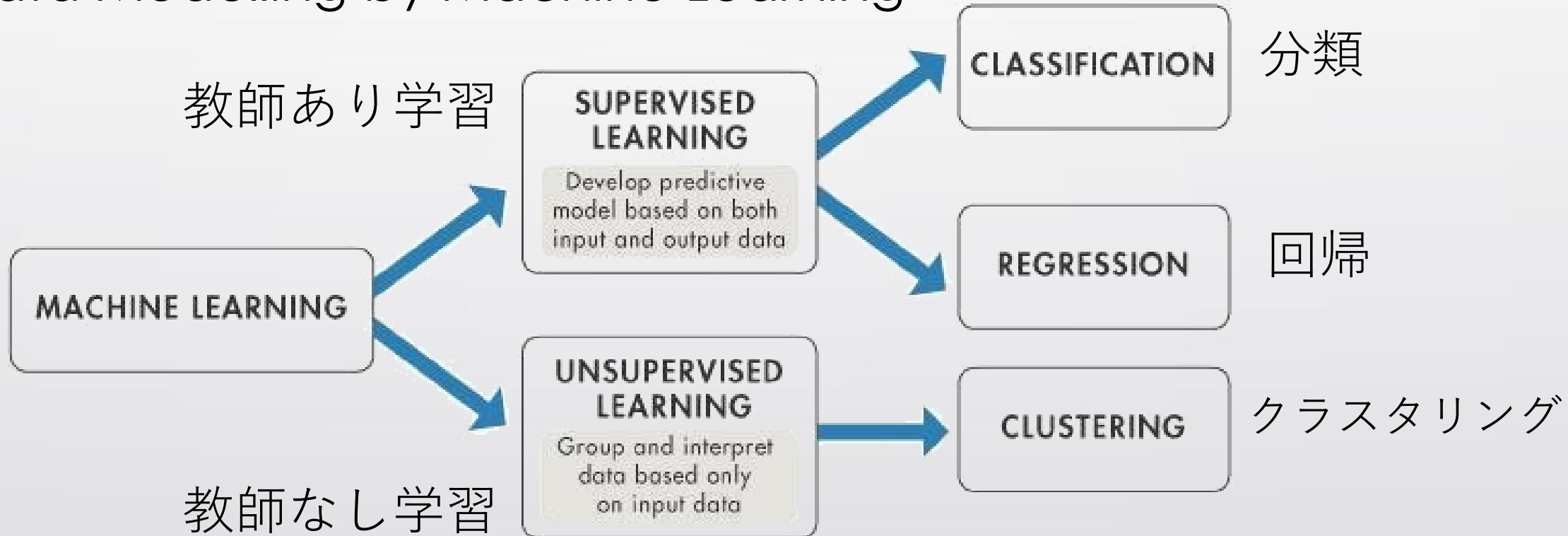


$$\Delta = SSE \text{ after linkage} - SSE \text{ before linkage}$$

Δ を情報ロスと呼ぶ Δ is referred to as “Information loss”

機械学習によるモデル化

Data Modelling by Machine Learning



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

クラスタリングの種類 Types of Clustering

- 非階層的クラスタリング
Non-Hierarchical Clustering
- 階層的クラスタリング
Hierarchical Clustering
- モデル・ベース・クラスタリング
Model-Based Clustering

データの統計的分布についての仮定をおく
Make presumptions about statistical distribution of data



ソフトクラスタリング Soft Clustering

ハードクラスタリング Hard Clustering

各データは一つのクラスターにしか所属できない

Each data belongs to single cluster

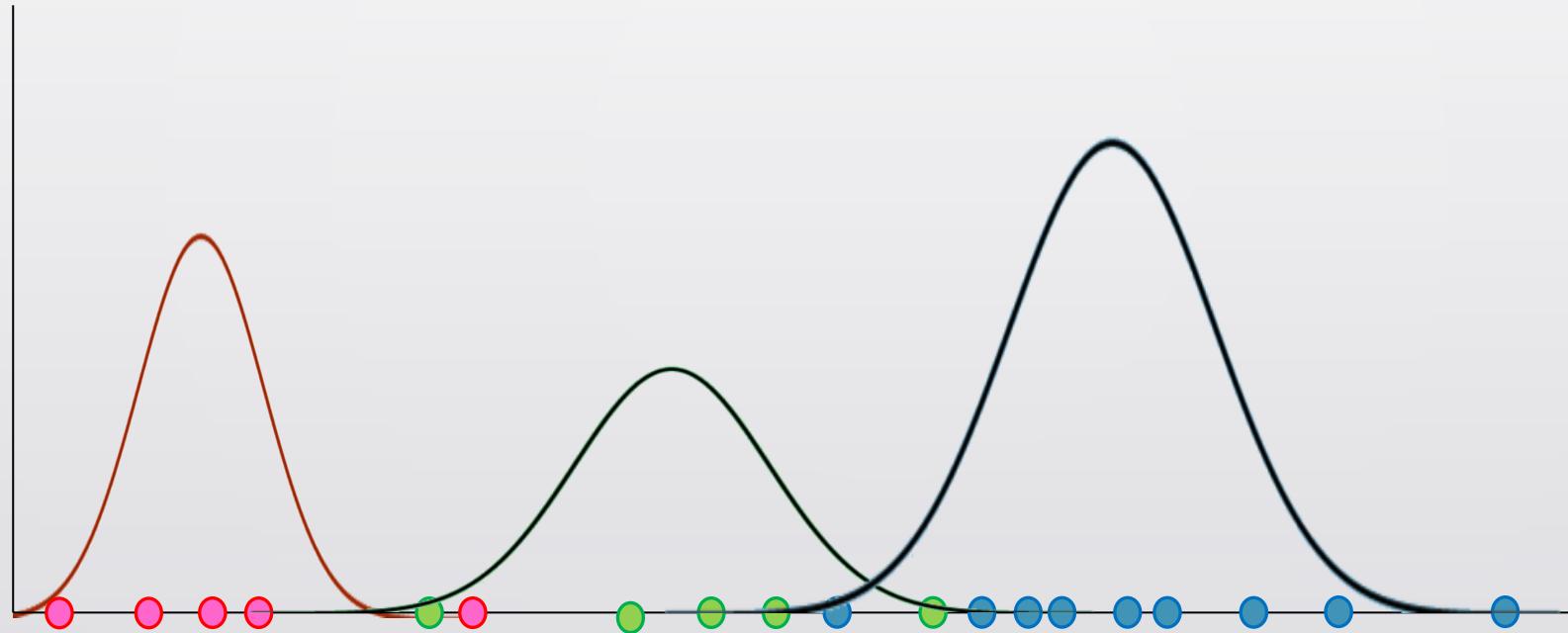
ソフトクラスタリング Soft Clustering

各データが複数のクラスに所属しうる

Each data can belong to multiple classes

混合ガウス分布モデル Gaussian Mixture Model

観測されたデータが複数のガウス分布の重ね合わせから生成されたと仮定する
Assume that observations are generated by multiple overlapping Gaussian distributions



多次元ガウス分布 Multidimensional Gaussian Distribution

正規分布を多次元に拡張した分布

Probability density distribution obtained by extending normal distribution to multi-dimensional space

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

$\boldsymbol{\Sigma}$: 分散共分散行列 Variance-covariance matrix

$|\boldsymbol{\Sigma}|$: 分散共分散行列の行列式 Determinant of variance-covariance matrix

$\boldsymbol{\Sigma}^{-1}$: 分散共分散行列の逆行列 Inverse matrix of variance-covariance matrix

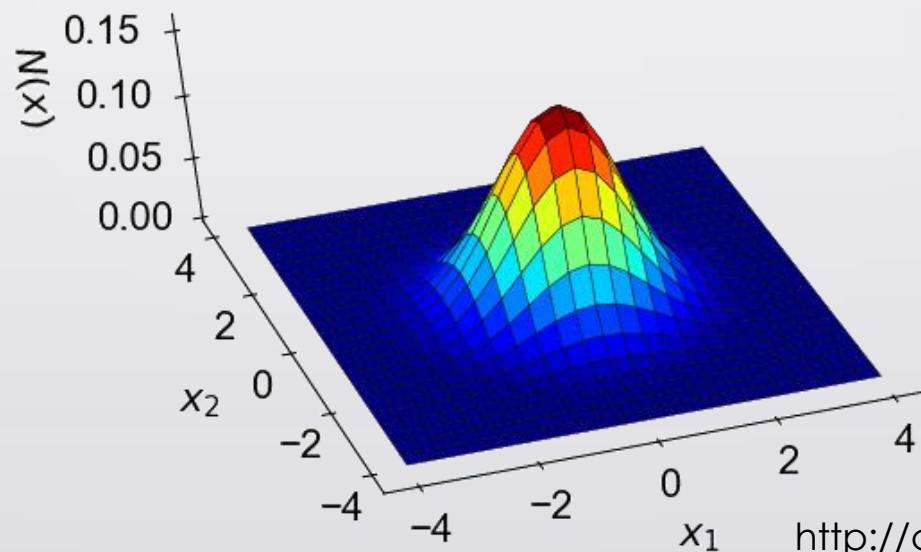
多次元ガウス分布 Multidimensional Gaussian Distribution

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

$$\mathbf{x} = (x_1, x_2)$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{pmatrix}$$

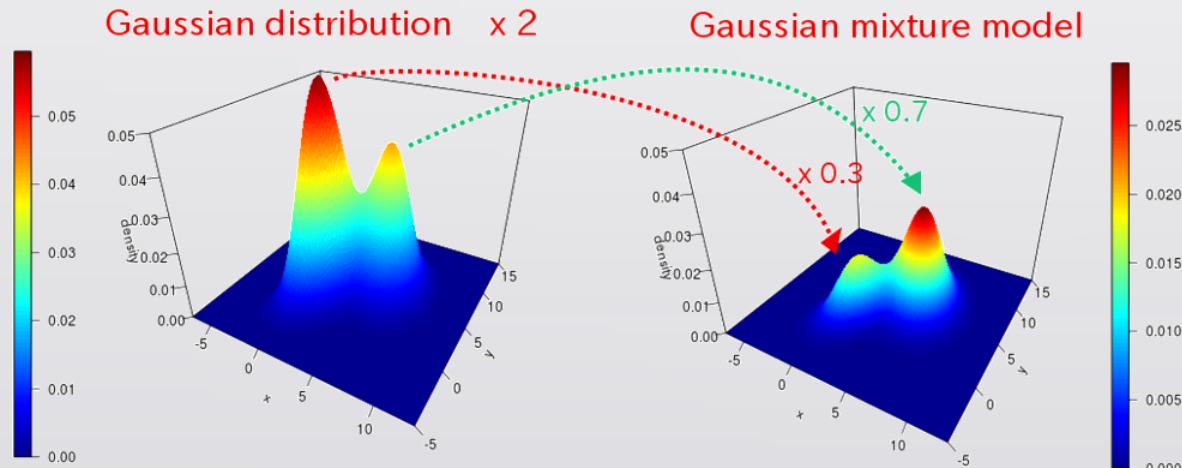


混合ガウス分布 Gaussian Mixture Distribution

M 個の正規分布の重ね合わせにより確率分布を表現する

Represent probability distribution as weighted mixture of M normal distributions

$$p(x) = \sum_{m=1}^M \pi_m N(x|\mu_m, \sigma_m) \quad 0 \leq \pi_m \leq 1 \quad \sum_{m=1}^M \pi_m = 1 \quad \pi_m : \text{混合比 Mixing Ratio}$$



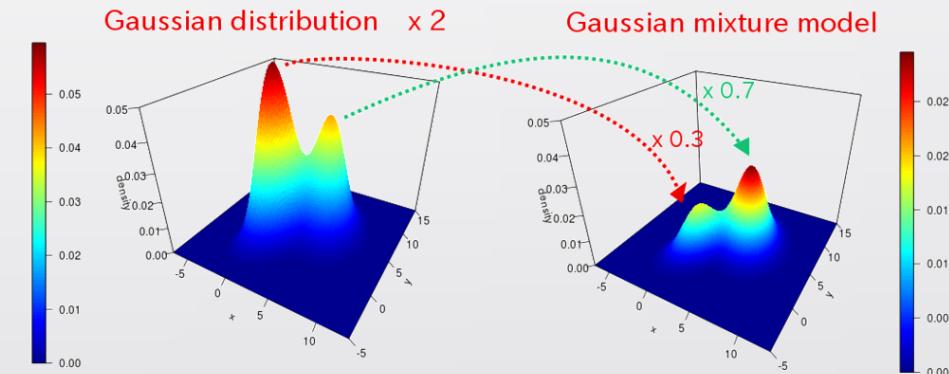
<https://work-in-progress.hatenablog.com/entry/2018/11/08/224826>

潜在変数 Latent Variable (隠れ変数 Hidden Variable)

観測データからは直接得ることが出来ない情報

Information that cannot be obtained directly from observations

$$p(x) = \sum_{m=1} \pi_m N(x|\mu_m, \sigma_m)$$

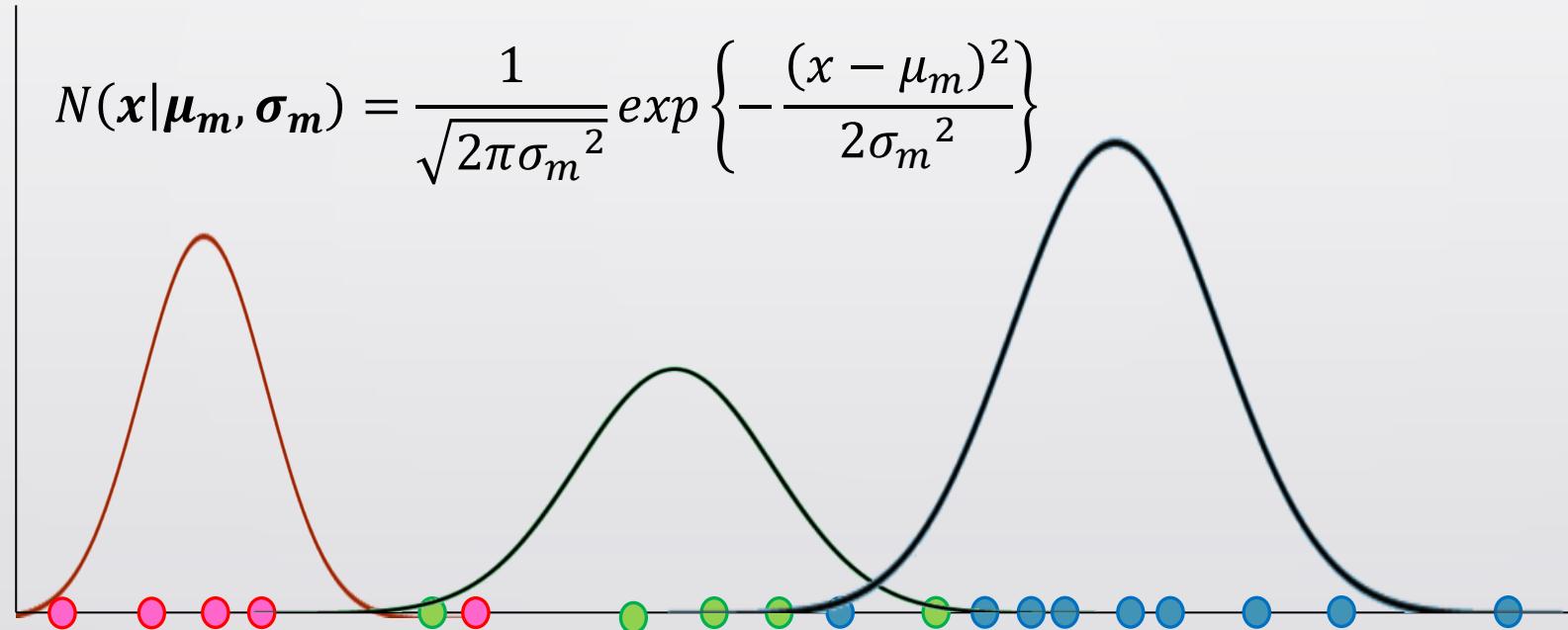


潜在変数を使って、どの分布からデータが生成されたかを表現する

Latent variable represents from which gaussian distribution a data is generated

一次元混合ガウスモデル 1-Dimensional GMM

観測されたデータが複数のガウス分布の重ね合わせから生成されたと仮定する
Assume that observations are generated by multiple overlapping Gaussian distributions





一次元混合ガウスモデル 1-Dimensional GMM

π_m : 混合比 Mixing Ratio $\sum_{m=1}^M \pi_m = 1 \quad 0 \leq \pi_m \leq 1$

\mathbf{z} : 潜在変数 Latent Variables

$$z_m \in \{0, 1\} \quad \mathbf{z} = (z_1, z_2, z_3, z_4, \dots, z_{M-1}, z_M)$$

$$\sum_{m=1}^M z_m = 1 \quad ex) \mathbf{z} = (0, 0, 0, 1, \dots, 0, 0, 0)$$



一次元混合ガウスモデル 1-Dimensional GMM

$$N(x|\mu_m, \sigma_m) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left\{-\frac{(x - \mu_m)^2}{2\sigma_m^2}\right\}$$

$$p(x) = \sum_{m=1}^M p(x|z_m = 1)p(z_m = 1) = \sum_{m=1}^M N(x|\mu_m, \sigma_m)\pi_m$$



一次元混合ガウスモデル 1-Dimensional GMM

$$p(z_m = 1|x) = \frac{p(x|z_m = 1)p(z_m = 1)}{p(x)} = \frac{p(x|z_m = 1)p(z_m = 1)}{\sum_{m=1}^M p(x|z_m = 1)\pi_m}$$



観測された x が分布 m から生成された事後確率

Posterior probability that observation x is generated by distribution m

潜在変数の期待値は、その事後確率と一致する

Expected value of latent variable corresponds to its posterior probability

$$E[z_m] = p(z_m = 1|x)$$



完全データ Complete Data

$$X = \{x_1, x_2 \cdots x_N\} \quad Z = \{z_1, z_2 \cdots z_N\} \quad z_n = (z_{n1}, z_{n2} \cdots z_{nM})$$

$$Y = \{X, Z\} = \{x_1, x_2 \cdots x_N, z_1, z_2 \cdots z_N\}$$

$$p(x_n, z_{nm} = 1 | \mu, \sigma, \pi)$$

$$= p(x_n | z_{nm} = 1) p(z_{nm} = 1 | \mu, \sigma, \pi)$$

$$= N(x_n | \mu_m, \sigma_m) \pi_m$$

MLEによるパラメータ推定 Parameter Estimation by MLE

$$p(x_n, z_{nm} = 1 | \mu, \sigma, \pi) = N(x_n | \mu_m, \sigma_m) \pi_m$$

$$p(Y | \mu, \sigma, \pi) = \prod_{n=1}^N \prod_{m=1}^M [N(x_n | \mu_m, \sigma_m) \pi_m]^{z_{nm}}$$

$$\mu, \sigma, \pi = \operatorname{argmax}_{\mu, \sigma, \pi} p(Y | \mu, \sigma, \pi)$$

観測される**X**と対応する**Z**の同時確率を最大化するパラメータセットを求める

Search for parameter set that maximizes observations **X** and corresponding **Z**



MLEによるパラメータ推定 Parameter Estimation by MLE

$$\log p(Y | \mu, \sigma, \pi) = \sum_{n=1}^N \sum_{m=1}^M z_{nm} \log N(x_n | \mu_m, \sigma_m) + \sum_{n=1}^N \sum_{m=1}^M z_{nm} \log \pi_m$$

潜在変数は直接的に観察できない

Latent variables are not directly observable



潜在変数の期待値 = 事後確率で置き換えてパラメータ推定

Estimate parameters by replacing latent variables with their expected values



Q関数 Q function

$$\log p(Y | \mu, \sigma, \pi) = \sum_{n=1}^N \sum_{m=1}^M z_{nm} \log N(x_n | \mu_m, \sigma_m) + \sum_{n=1}^N \sum_{m=1}^M z_{nm} \log \pi_m$$

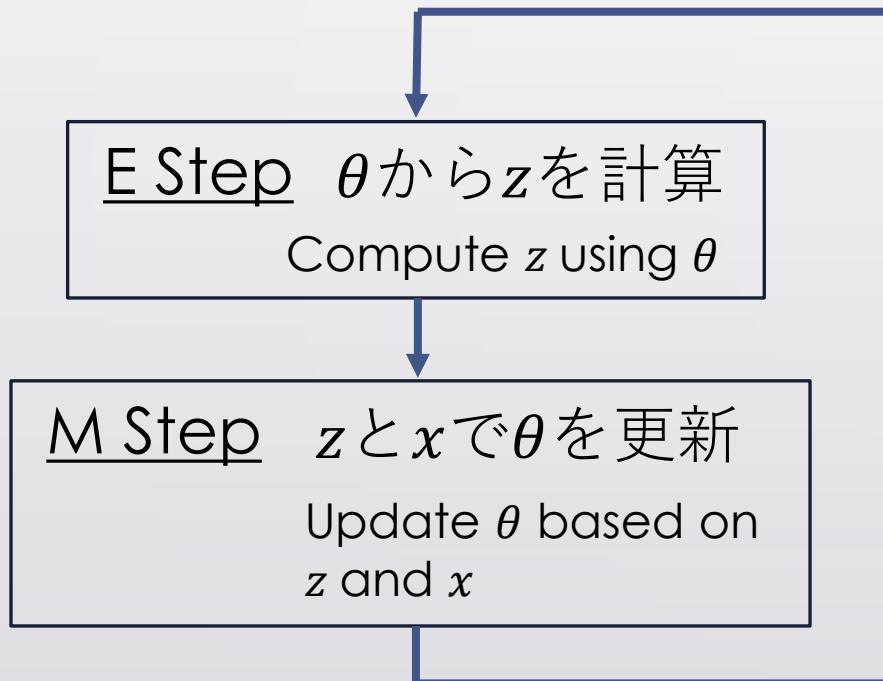


$$Q = \sum_{n=1}^N \sum_{m=1}^M E[z_{nm}] \log N(x_n | \mu_m, \sigma_m) + \sum_{n=1}^N \sum_{m=1}^M E[z_{nm}] \log \pi_m$$

EM アルゴリズム Expectation-Maximizing Algorithm

潜在変数を含むモデルの代表的なパラメータ推定法

Algorithm for parameter estimation of models including latent variables



x : 観測 Observations

θ : 確率密度関数のパラメータセット
Parameter set of probability distribution functions

z : 潜在変数 Latent Variables

E-ステップ Expectation-Step

現在のパラメータセット $\theta^{(t)}$ を用いて z の期待値を求める

Compute expected value of z based on current parameter set $\theta^{(t)}$

$$\boldsymbol{\theta}^{(t)} = \{\boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\pi}^{(t)}\}$$

$$\mathbf{z} = (z_1, z_2 \dots z_m \dots z_M)$$

$$E[z_m] = \frac{N(x | \mu_m^{(t)}, \sigma_m^{(t)}) \pi_m^{(t)}}{\sum_{m=1}^M N(x | \mu_m^{(t)}, \sigma_m^{(t)}) \pi_m^{(t)}}$$



M-ステップ Maximize-Step

$$p(x, z | \theta^{(t)}) = \prod_{n=1}^N \prod_{m=1}^M p(x|z_n, \theta^{(t)}) p(z_m | \theta^{(t)})$$

$$= \prod_{n=1}^N \prod_{m=1}^M [p(x_n | z_{n,m}, \theta^{(t)}) p(z_m | \theta^{(t)})]^{z_{n,m}}$$

$$= \prod_{n=1}^N \prod_{m=1}^M [N(x_n | \mu_m^{(t)}, \sigma_m^{(t)}) \pi_m^{(t)}]^{z_{n,m}}$$

M-ステップ Maximation-Step

Q関数を最大化するようパラメータセット $\theta^{(t)}$ を更新

Update parameter set $\theta^{(t)}$ so that Q function is maximized

$$p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}^{(t)}) = \prod_{n=1}^N \prod_{m=1}^M [N(x_n | \mu_m^{(t)}, \sigma_m^{(t)}) \pi_m^{(t)}]^{z_{n,m}}$$

$$Q(\boldsymbol{\theta}^{(t)}) = \sum_{n=1}^N \sum_{m=1}^M E[z_{n,m}] \left[\log(\pi_m^{(t)}) + \log(N(x_n | \mu_m^{(t)}, \sigma_m^{(t)})) \right]$$



M-ステップ Maximizeation-Step

$$Q(\boldsymbol{\theta}^{(t)}) = \sum_{n=1}^N \sum_{m=1}^M E[z_{n,m}] \left[\log(\pi_m^{(t)}) + \log \left(N(x_n | \mu_m^{(t)}, \sigma_m^{(t)}) \right) \right]$$

$$E[z_m] = p(z_m = 1 | x) = \frac{p(x | z_m = 1)p(z_m = 1)}{\sum_{m=1}^M p(x | z_m = 1)\pi_m}$$

$$\frac{\partial Q}{\partial \pi_m} = 0 \quad \frac{\partial Q}{\partial \mu_m} = 0 \quad \frac{\partial Q}{\partial \sigma_m} = 0$$

EMアルゴリズム Expectation-Maximizing Algorithm

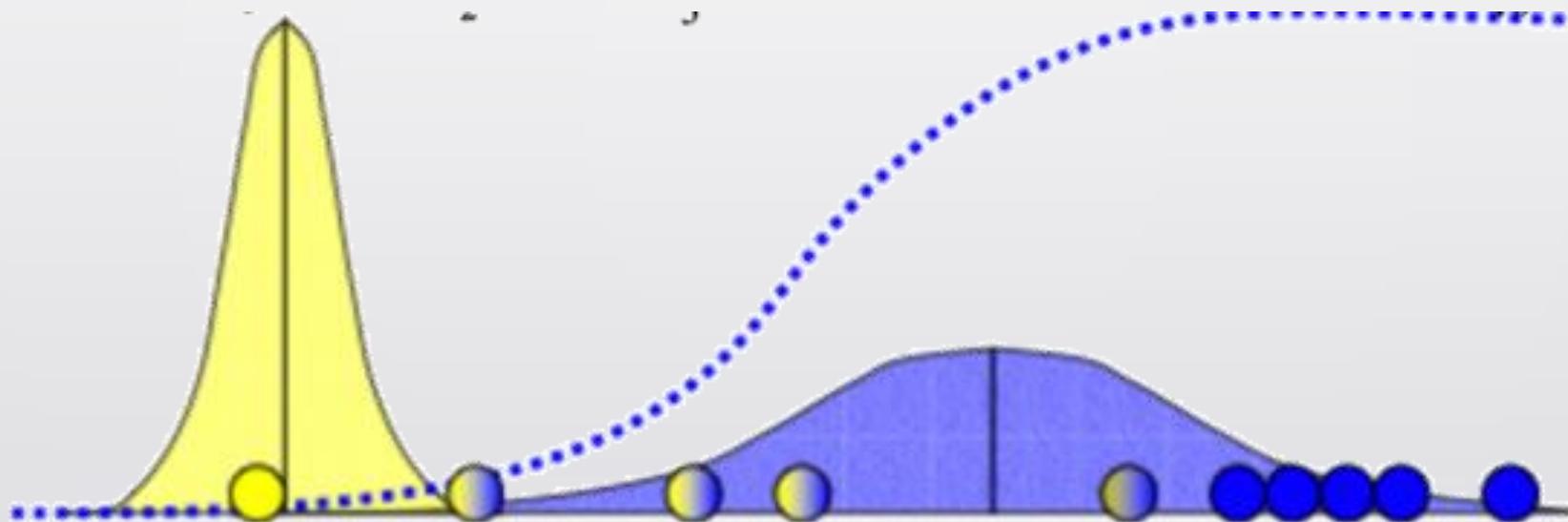
1. θ を初期化する Initialize θ



<https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf>

EMアルゴリズム Expectation-Maximizing Algorithm

2. E-ステップ：現在のパラメータセット $\theta^{(t)}$ を用いて z の期待値を求める
データの色が $E[z_{n,m}]$ を表す

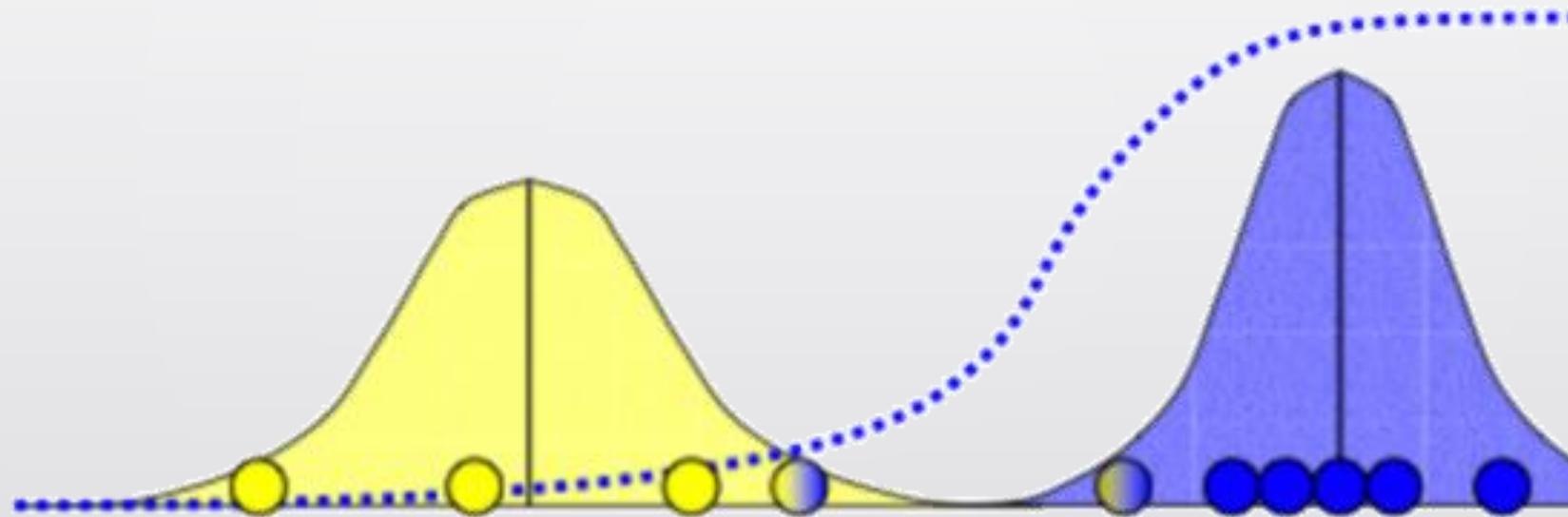


<https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf>



EMアルゴリズム Expectation-Maximizing Algorithm

3. M-ステップ：パラメータセット $\theta^{(t)}$ を更新する Update θ



<https://courses.cs.washington.edu/courses/cse416/22sp/lectures/12/12.pdf>

クラスタリングの評価 Evaluation of Clustering Results

	C^1	C^2
a	90	10
b	20	80

C^k : k 番目のクラスター k -th cluster

$n_{m,k}$: C^k のうち m 番目のクラス C_m に属するデータの数
Number of data belonging to class C_m within C^k

$|C^k|$: クラスター C^k に属するデータの数
Number of data belonging to cluster C^k

純度 Purity

局所的純度 Local Purity

$$Purity = \frac{\max_m n_{m,k}}{|C^k|}$$

最大多数派のクラスのデータがクラスターに占める割合

Proportion of data of majority class

大域的純度 Global Purity

$$Purity = \frac{\sum_k \max_m n_{m,k}}{\sum_k |C^k|}$$

	C^1	C^2
a	90	10
b	20	80

逆純度 Inverse Purity

クラスターの純度は 2 つのテーブルで同じ

Purity of the clusters is the same across the tables below

	C^1	C^2
a	90	10
b	20	80

	C^1	C^2
a	90	40
b	20	5

$$Purity = \frac{\max_m n_{m,k}}{|C^k|}$$

逆純度 Inverse Purity

$M_k = \operatorname{argmax}_m n_{m,k}$ クラスター C^k において最大多数派のクラス
Majority class within cluster C_k

$|C_{M_k}|$: クラス M_k に属するデータの総数
Total number of data belonging to class M_k

$$|C_{M_k}| = \sum_{k=1}^K n_{M_k,k}$$

逆純度 Inverse Purity

$$Inverse\ Purity = \frac{1}{N} \sum_{k=1}^K \frac{\max_m n_{m,k}}{|C_{M_k}|} |C^k|$$

$$M_1 = a$$

	C^1	C^2
a	90	10
b	20	80

$$M_2 = b$$

	C^1	C^2
a	90	10
b	20	80

逆純度 Inverse Purity

$$Inverse\ Purity = \frac{1}{N} \sum_{k=1}^K \frac{\max_m n_{m,k}}{|C_{M_k}|} |C^k|$$

$$M_1 = a$$

	C^1	C^2
a	90	40
b	20	5

$$M_2 = a$$

	C^1	C^2
a	90	40
b	20	5



F值 F-value

$$Purity = \frac{\sum_k max_m n_{m,k}}{\sum_k |C^k|} \quad Inverse\ Purity = \frac{1}{N} \sum_{k=1}^K \frac{max_m n_{m,k}}{|C_{M_k}|} |C^k|$$

$$F = \frac{1}{\frac{1}{2} \left(\frac{1}{Purity} + \frac{1}{Inverse\ Purity} \right)} = \frac{2\ Purity \cdot Inverse\ Purity}{Purity + Inverse\ Purity}$$



データマイニング

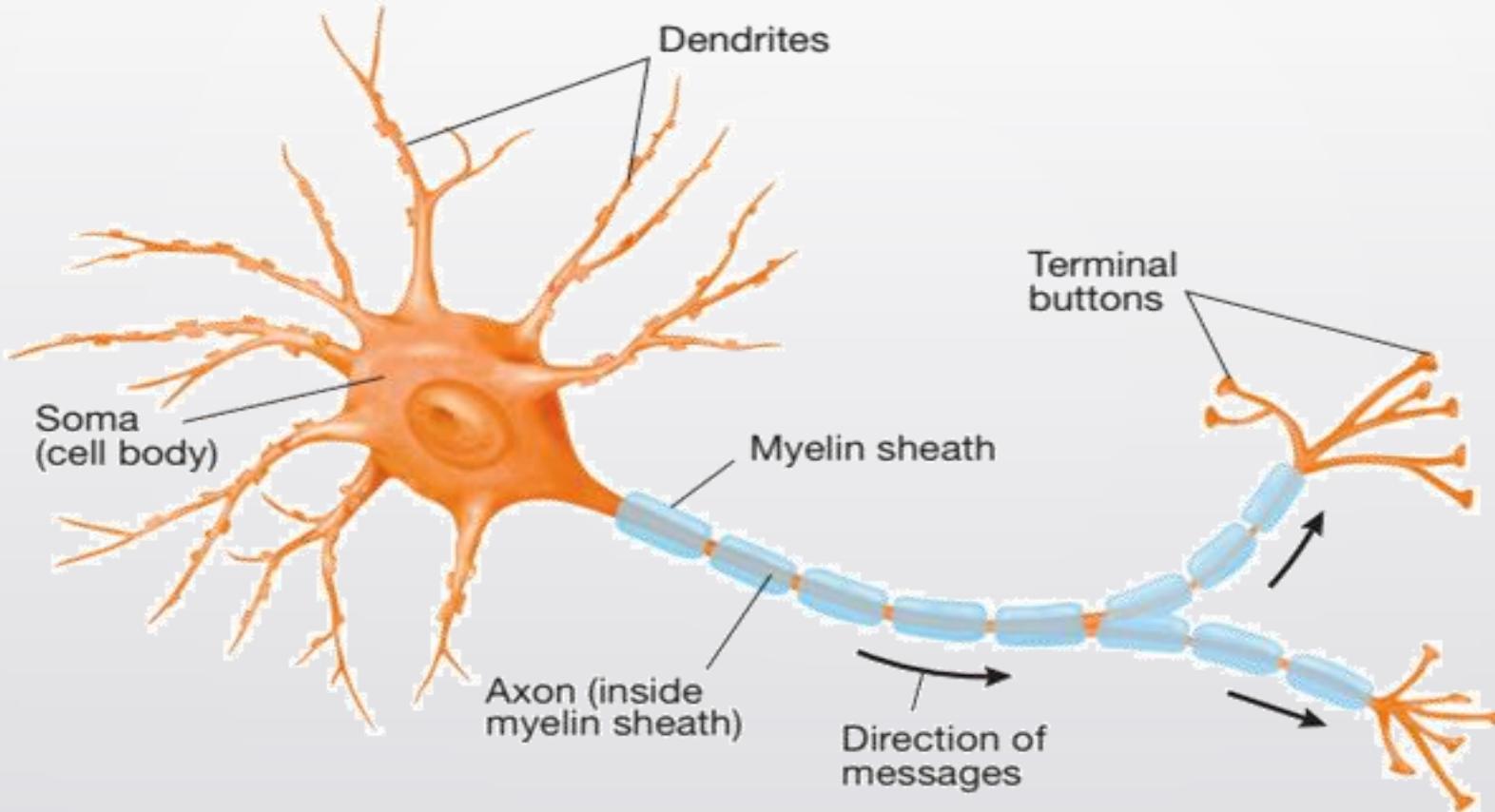
Data Mining

13: ニューラルネットワーク① Neural Network

土居 裕和 Hirokazu Doi

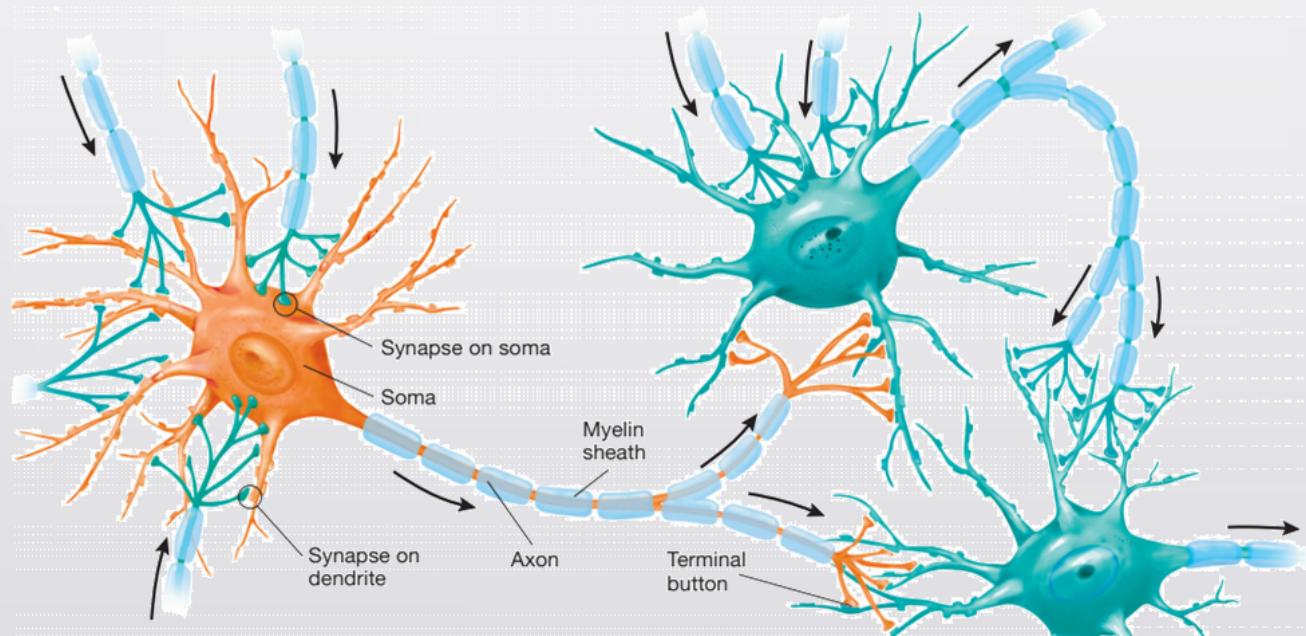
長岡技術科学大学 Nagaoka University of Technology

神經細胞 Neuron

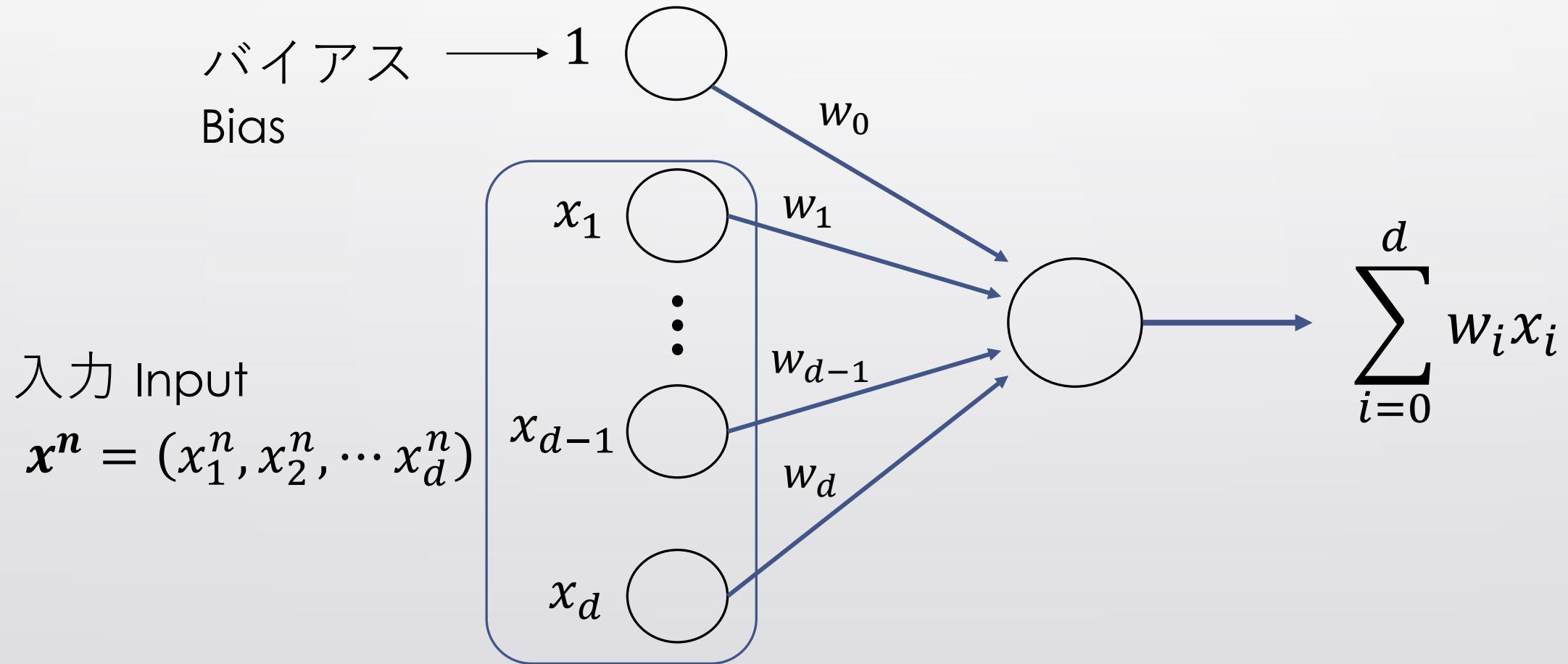


神経細胞の興奮 Neuronal Excitation

神経系の活動 = 神経細胞が電気活動を発生させ、神経細胞間で
伝えていくこと
Inter-neuronal transmission of electrical activity

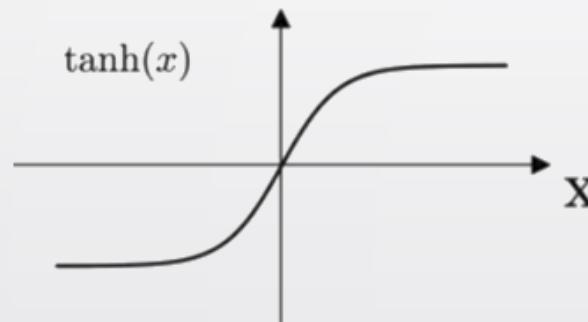


パーセプトロン Perception

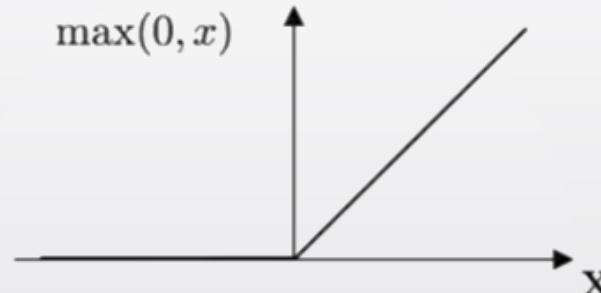


活性化関数 Activation Function

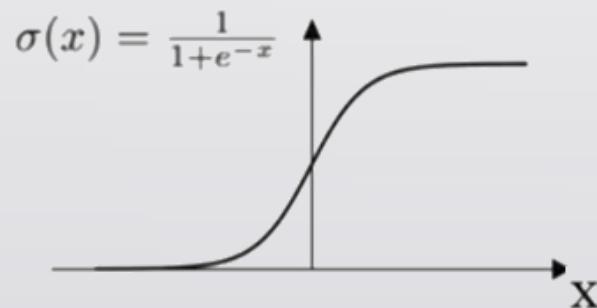
Tanh



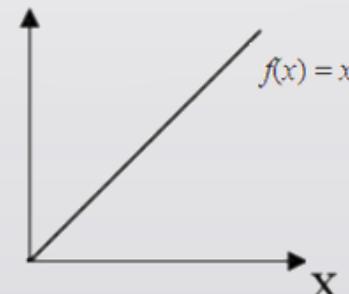
ReLU



Sigmoid

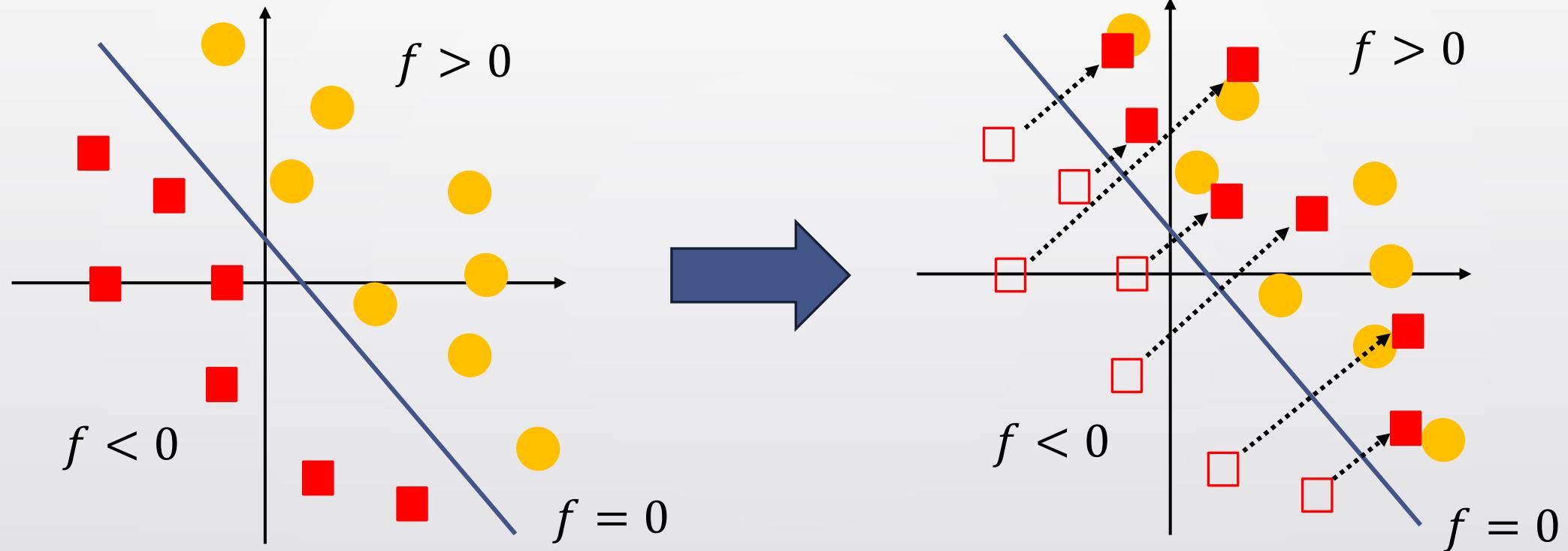


Linear



<https://machine-learning.paperspace.com/wiki/activation-function>

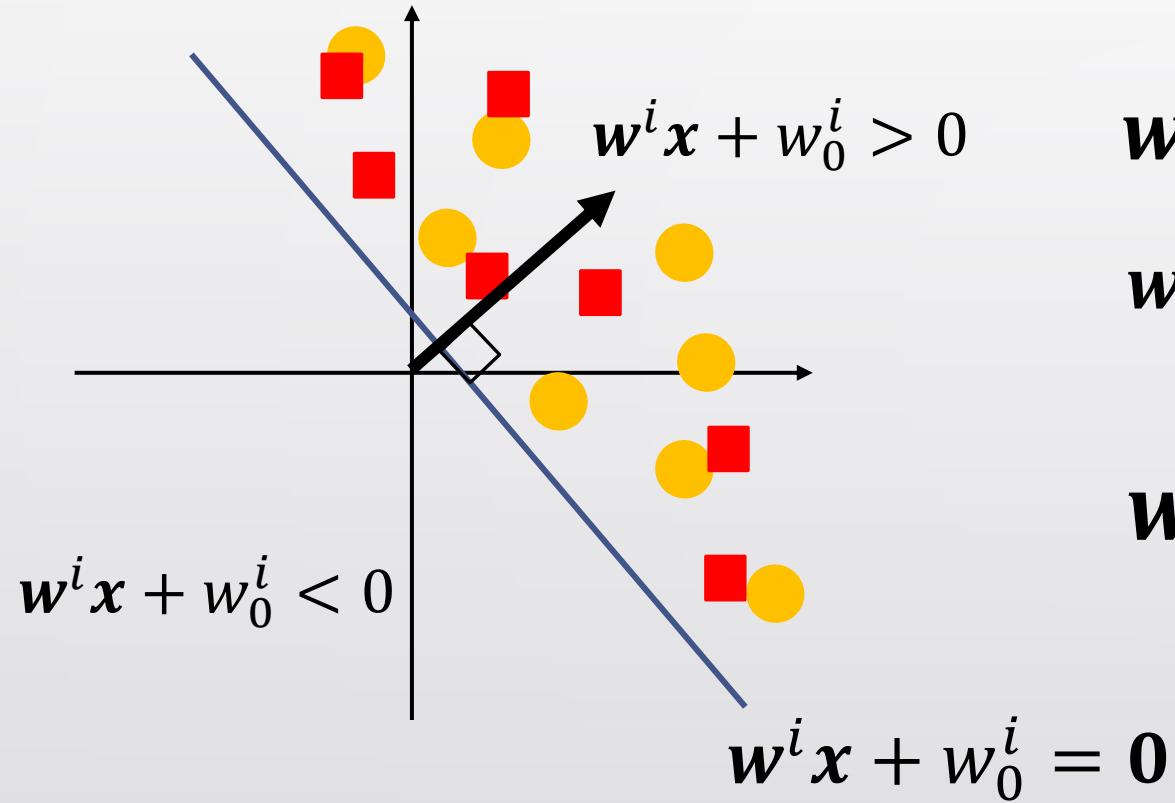
パーセプトロン学習則 Learning Rule of Perceptron



線型分離可能なデータは、符号反転により、すべて $f > 0$ の領域にくる

If linearly separable, all the data can be moved to the region $f > 0$ by sign inversion

パーセプトロン学習則 Learning Rule of Perceptron

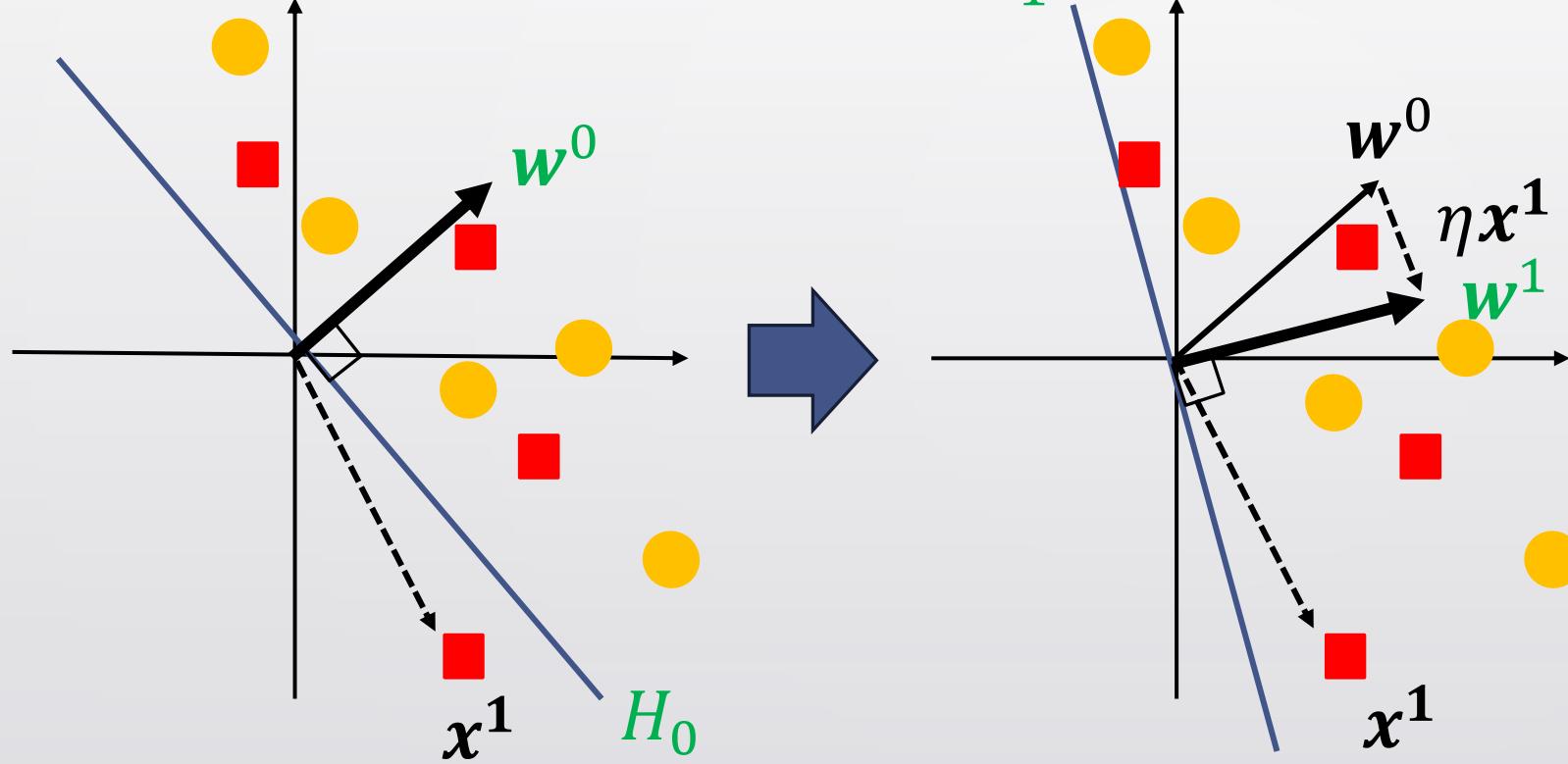


$$w^i, x^n \in R^d$$

$$w^i = (w_1^i, w_2^i, \dots, w_d^i)$$

$$w^{i+1} = \begin{cases} w^i + \eta x^n & (f(x^n) < 0) \\ w^i & (f(x^n) > 0) \end{cases}$$

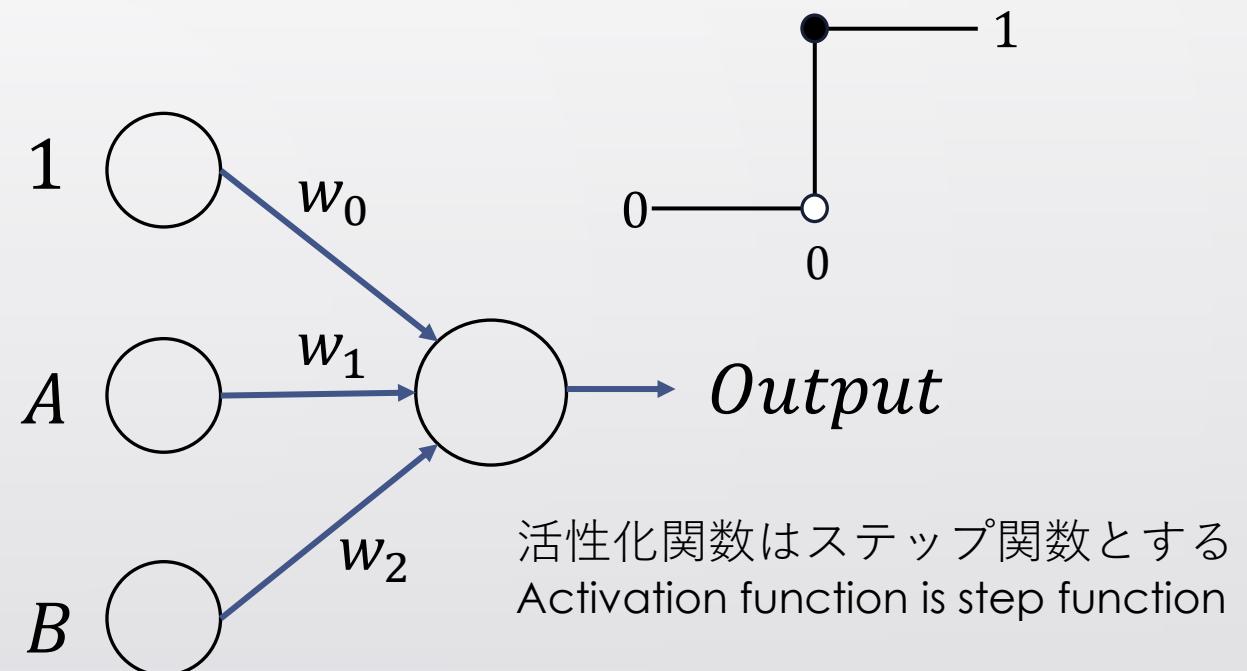
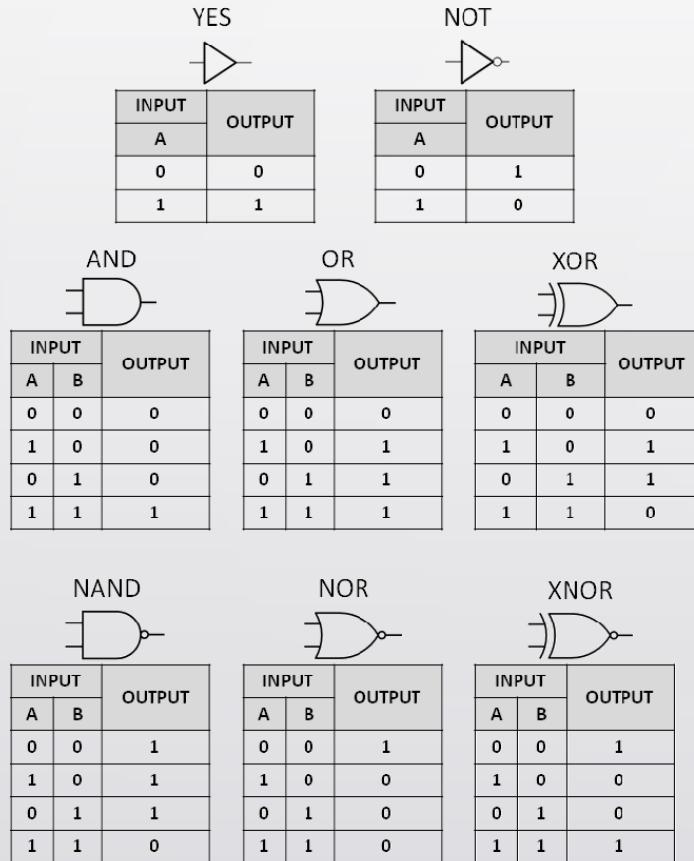
パーセプトロン学習則 Learning Rule of Perceptron



超平面を回転させることで、識別性能が向上する
Rotation of hyperplane improves classification performance

ブール論理演算子とパーセプトロン

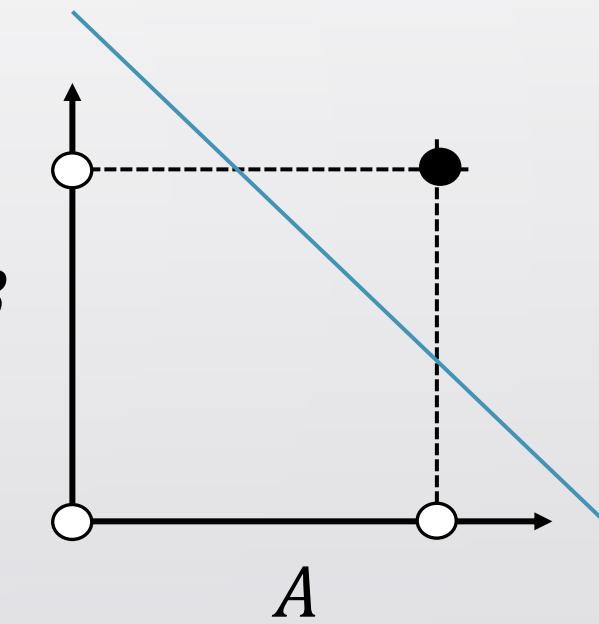
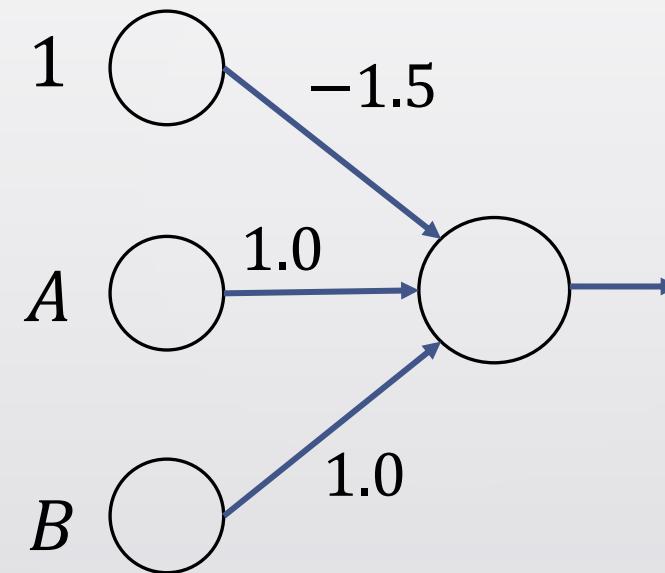
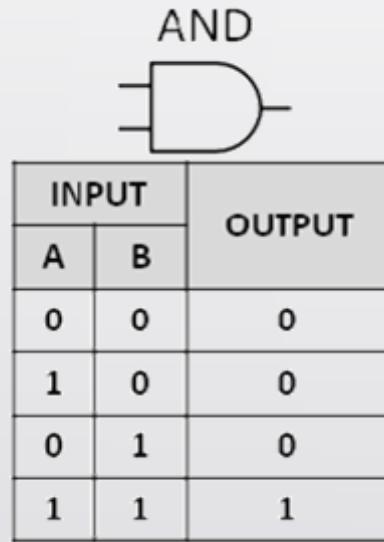
Boolean Logic Gate and Perceptron



Abels et al, 2015

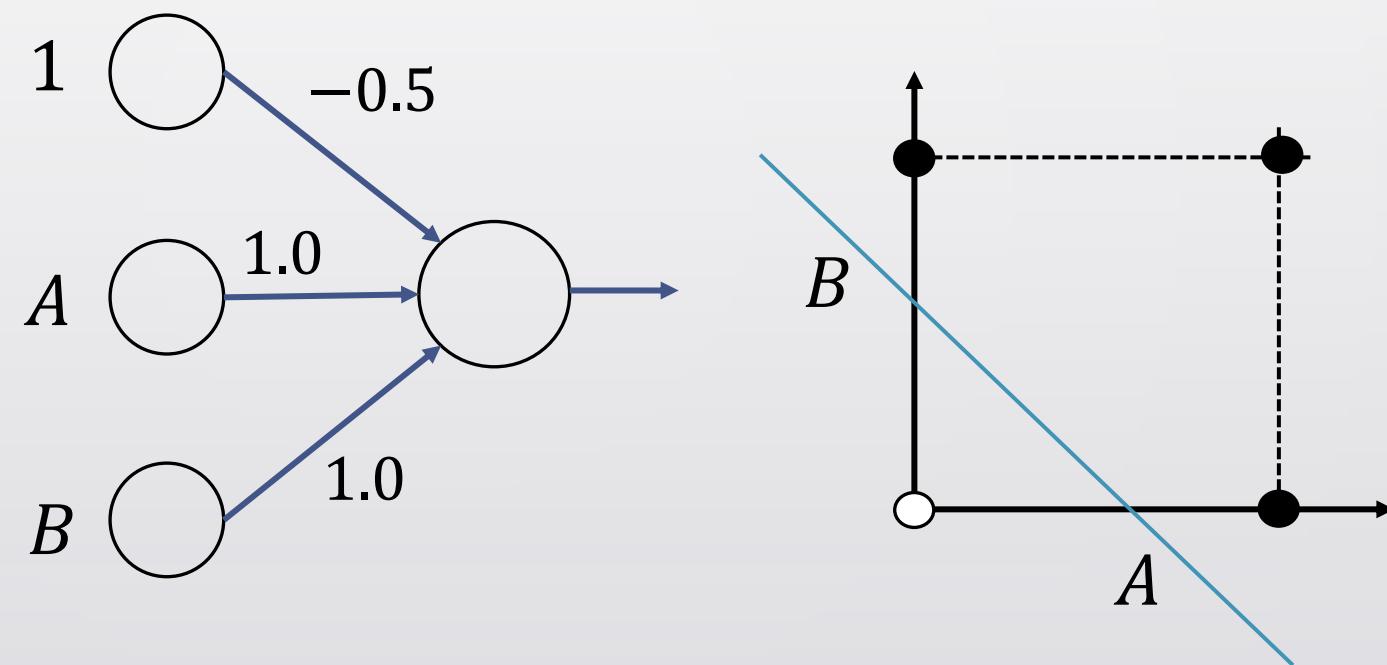
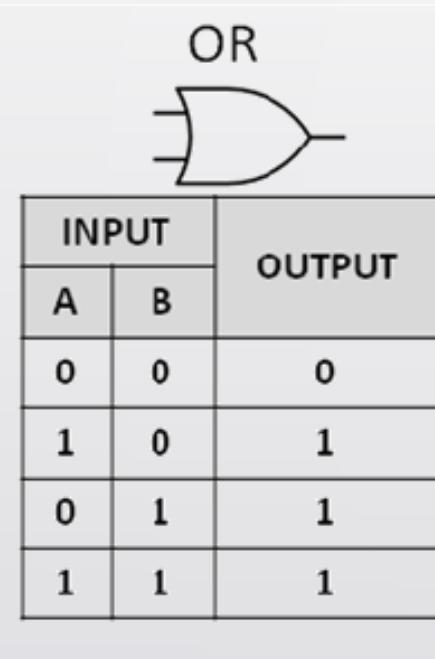
ブール論理演算子とパーセプトロン

Boolean Logic Gate and Perceptron



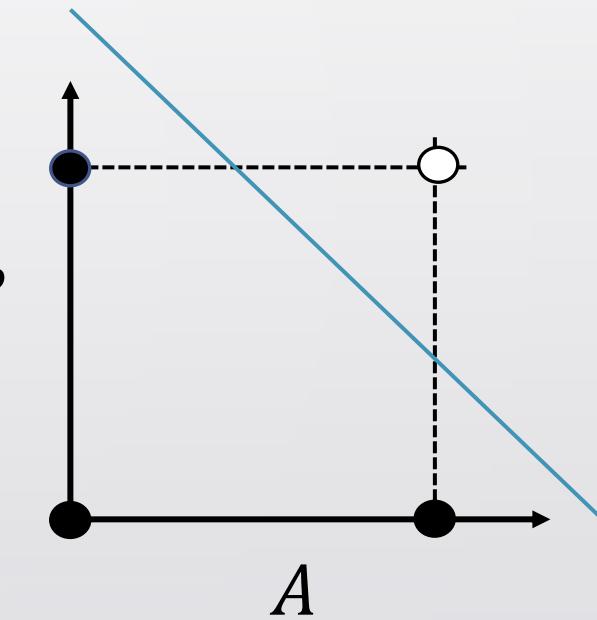
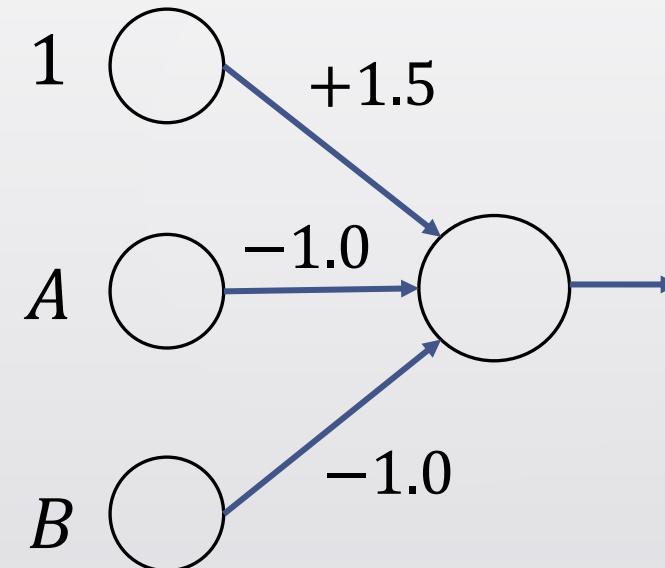
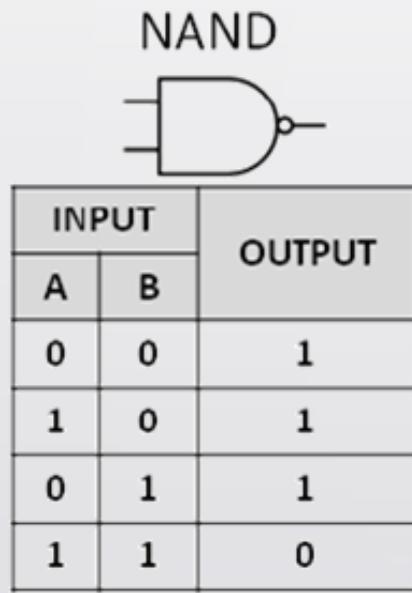
ブール論理演算子とパーセプトロン

Boolean Logic Gate and Perceptron



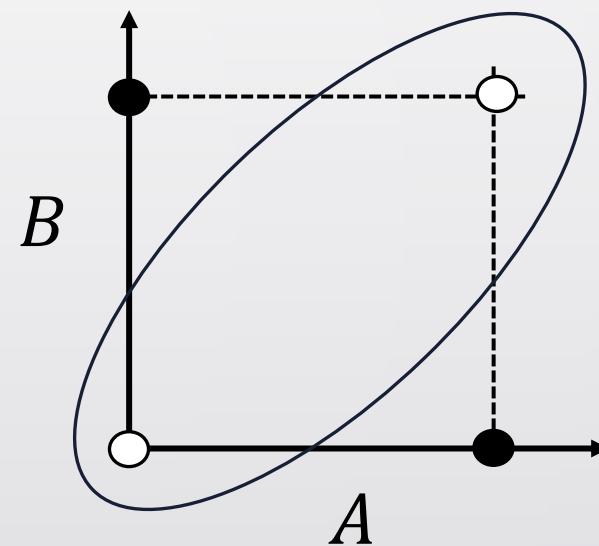
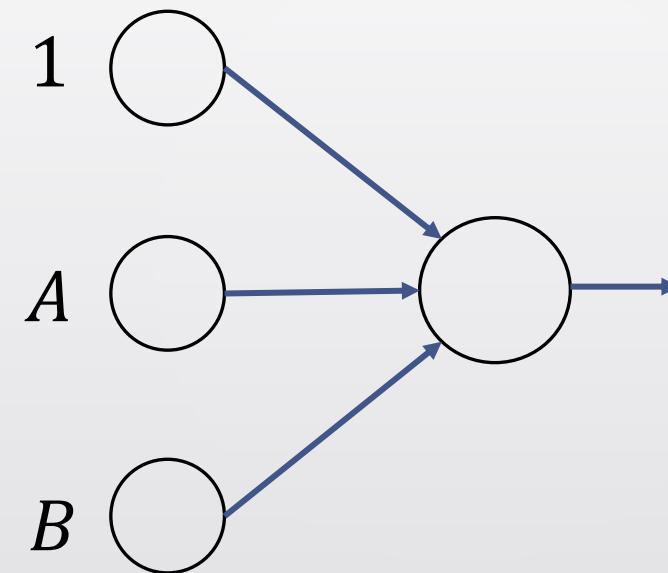
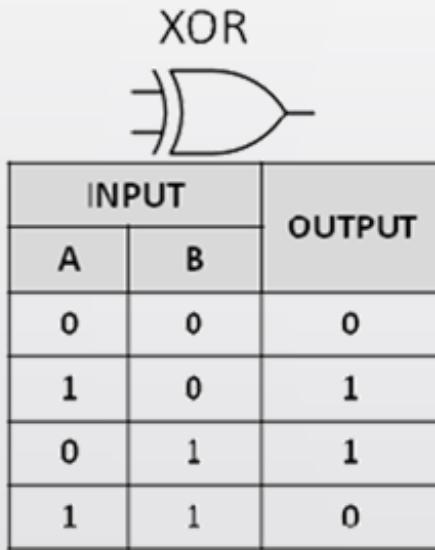
ブール論理演算子とパーセプトロン

Boolean Logic Gate and Perceptron



ブール論理演算子とパーセプトロン

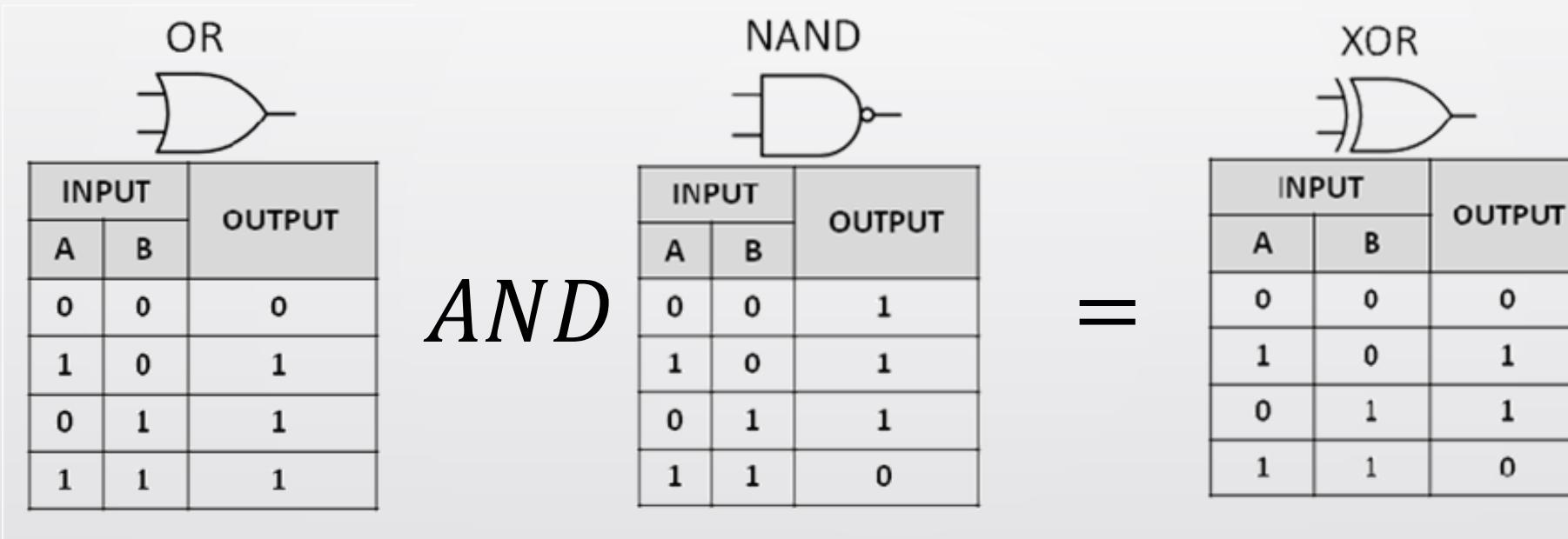
Boolean Logic Gate and Perceptron



単層パーセプトロンでは排他的論理和を構成できない
Single-layered perceptron cannot compose XOR

ブルー論理演算子とパーセプトロン

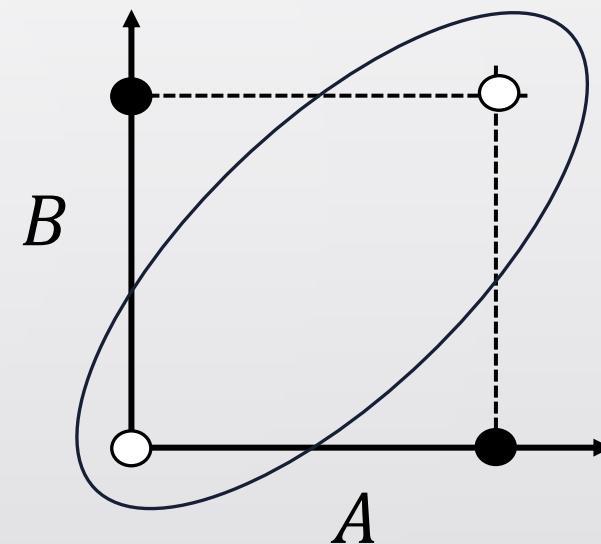
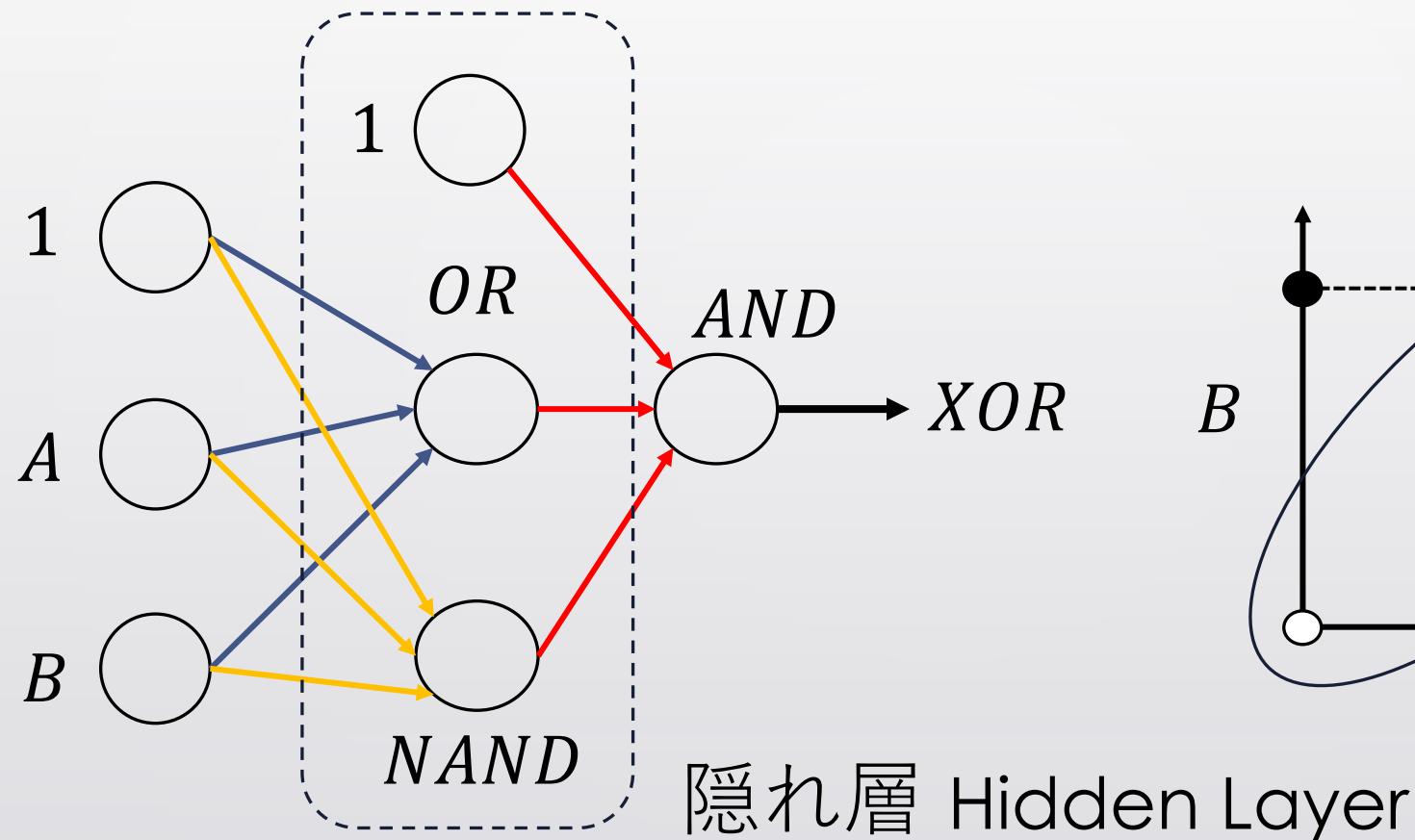
Boolean Logic Gate and Perceptron



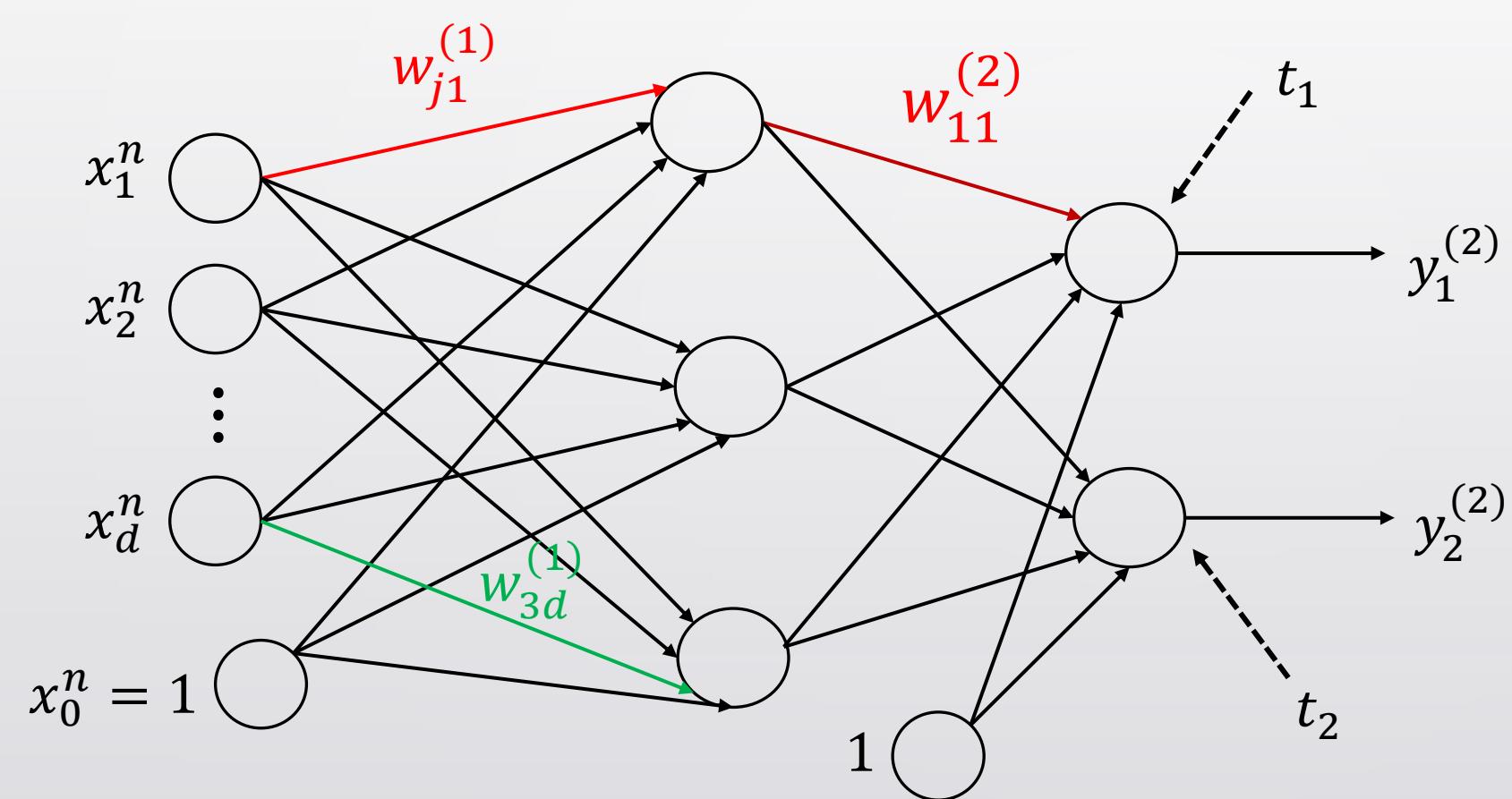
XORはORとNANDの論理積

XOR is logical conjunction of OR and NAND

XORと多層パーセプトロン XOR and Multi-layered Perceptron



多層パーセプトロン Multi-layered Perceptron



$w_{ji}^{(1)}$: 入力層から隠れ層への重み
Weights from input to hidden layer

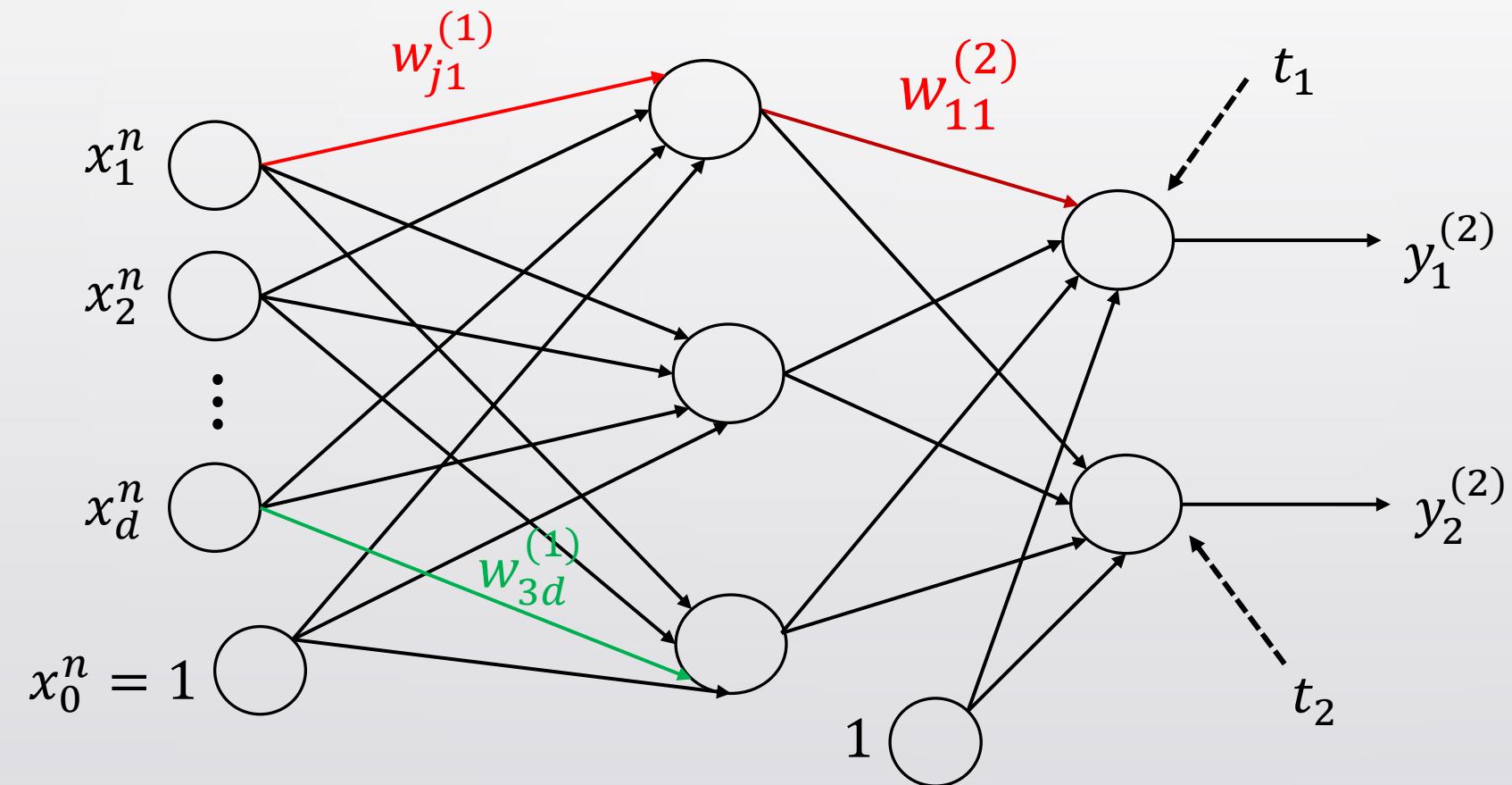
$w_{kj}^{(2)}$: 隠れ層から出力層への重み
Weights from hidden to output layer

$$i = 0, 1, 2 \dots d$$

$$j = 0, 1, 2 \dots M$$

$$k = 1, 2 \dots C$$

多層パーセプトロン Multi-layered Perceptron



$y_k^{(2)}$: 出力 Output

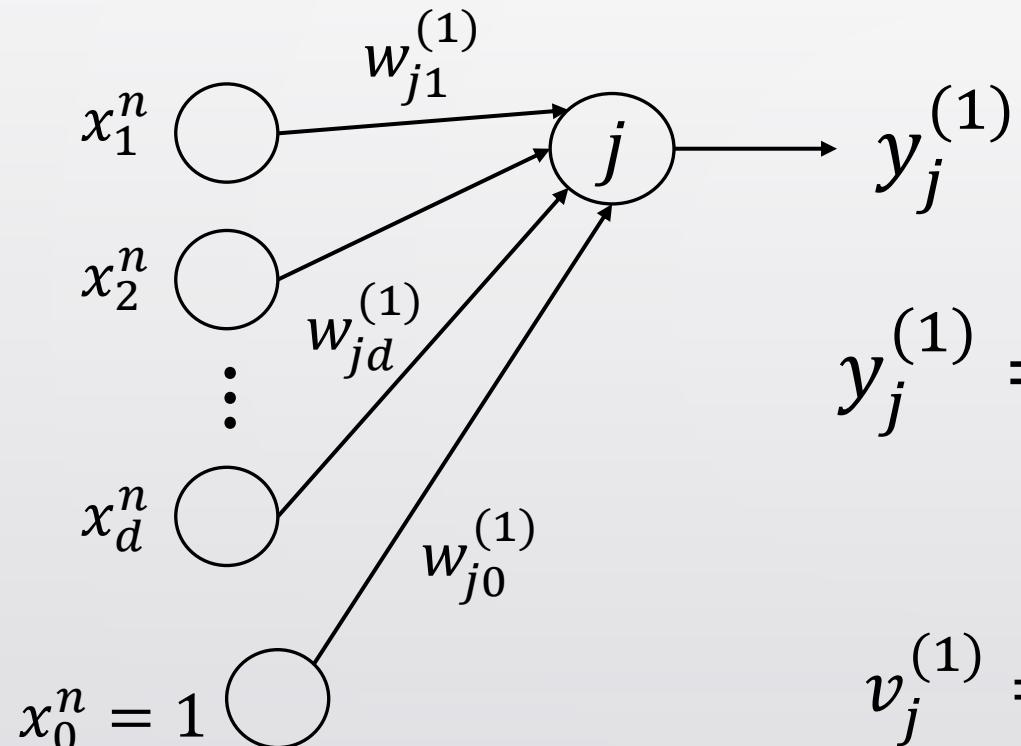
t_k : 教師信号

Teacher Signal

$$t_k \in \{0, 1\} \quad \sum_{k=1}^{k=C} t_k = 1$$

$$k = 1, 2 \dots C$$

隠れ層の出力 Output of Hidden Layer

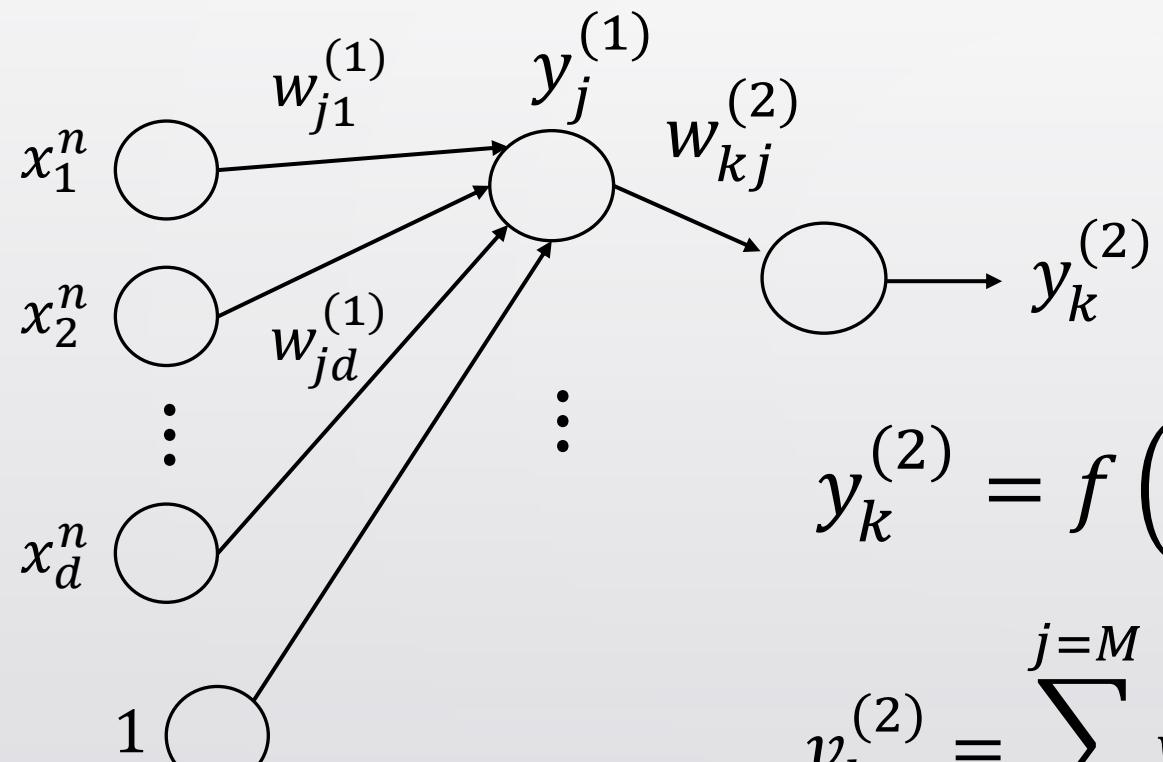


$$y_j^{(1)} = f \left(\sum_{i=0}^{i=d} w_{ji}^{(1)} x_i^n \right) = f \left(\mathbf{w}_j^{(1)} \mathbf{x}^{nT} \right)$$

$$v_j^{(1)} = \sum_{i=0}^{i=d} w_{ji}^{(1)} x_i^n$$



出力 Output



$$y_k^{(2)} = f \left(\sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)} \right) = f \left(\mathbf{w}_k^{(2)} \mathbf{y}^{(1)} \right)$$
$$v_k^{(2)} = \sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)}$$

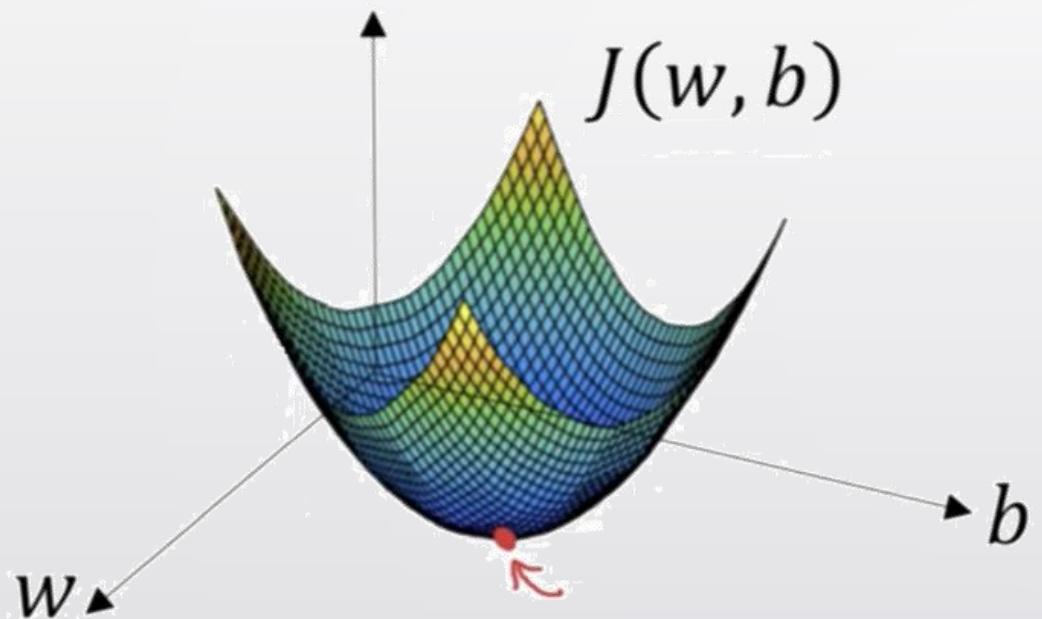
合成関数の微分 Derivative of Composite Function

$$\frac{df(g(x))}{dx} \quad u = g(x)$$

$$\frac{df(g(x))}{dx} = \frac{du}{dx} \frac{df(u)}{du}$$

$$= g'(x) f'(u) = g'(x) f'(g(x))$$

勾配降下 Gradient Descent



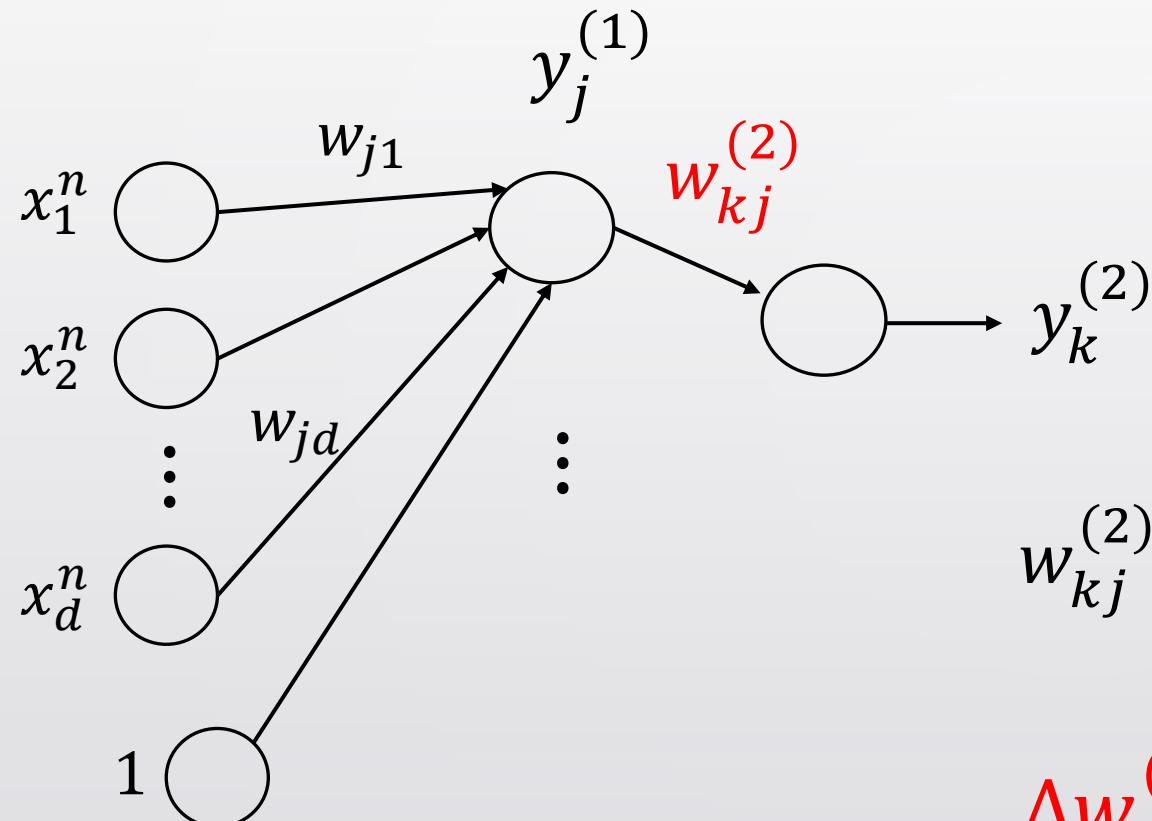
$$-\nabla J = \left(-\frac{\partial J}{\partial w}, -\frac{\partial J}{\partial b} \right)$$

$$w \leftarrow w - \eta \frac{\partial J}{\partial w}$$

$-\nabla J$ の方向に変数を変化させることで、関数 J の値を最も素早く減少させることが出来る

The output value of function J decreases most rapidly along the direction of $-\nabla J$

出力と教師信号の誤差 Error of network output



E を最小化するよう $w_{kj}^{(2)}$ を更新する
Update $w_{kj}^{(2)}$ so as to minimize E

$$E = \frac{1}{2} \sum_{k=1}^c (y_k^{(2)} - t_k)^2$$

$$w_{kj}^{(2)} = w_{kj}^{(2)} - \eta \frac{\partial E}{\partial w_{kj}^{(2)}} = w_{kj}^{(2)} + \Delta w_{kj}^{(2)}$$

$$\Delta w_{kj}^{(2)} = -\eta \frac{\partial E}{\partial w_{kj}^{(2)}}$$

重みの更新 Weight Updating

$$\Delta w_{kj}^{(2)} = -\eta \frac{\partial E}{\partial w_{kj}^{(2)}} = -\eta \left(y_k^{(2)} - t_k \right) \frac{\partial y_k^{(2)}}{\partial w_{kj}^{(2)}}$$

$$= -\eta \left(y_k^{(2)} - t_k \right) \frac{\partial v_k^{(2)}}{\partial w_{kj}^{(2)}} \frac{\partial f(v_k^{(2)})}{\partial v_k^{(2)}}$$

$$= -\eta \left(y_k^{(2)} - t_k \right) y_j^{(1)} f'(v_k^{(2)}) \quad v_k^{(2)} = \sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)}$$

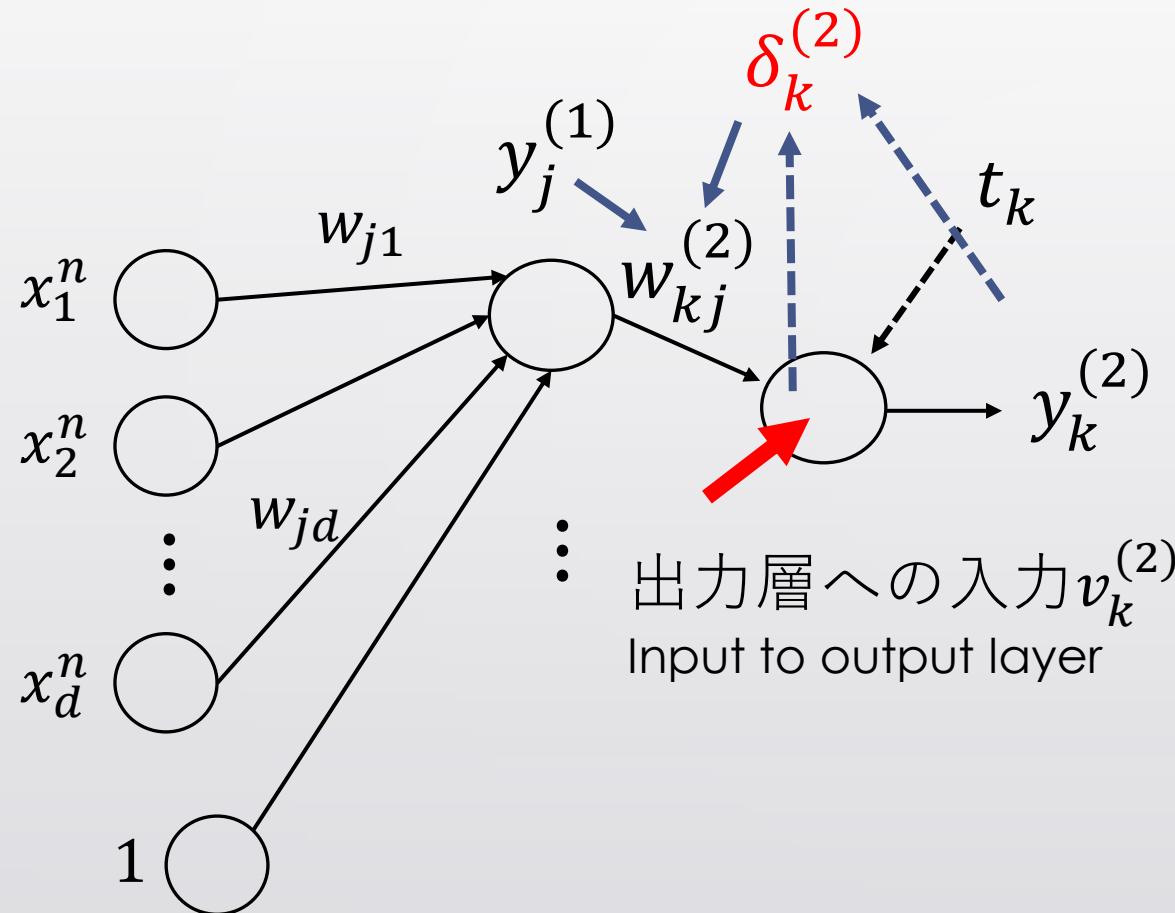
重みの更新 Weight Updating

$$\Delta w_{kj}^{(2)} = -\eta \left(\underbrace{y_k^{(2)} - t_k}_{\text{誤差 Error}} \right) y_j^{(1)} f' \left(\underbrace{v_k^{(2)}}_{\text{出力層への入力 Input to output layer}} \right)$$

↑

隠れ層の出力
Output of hidden layer

重みの更新 Weight Updating



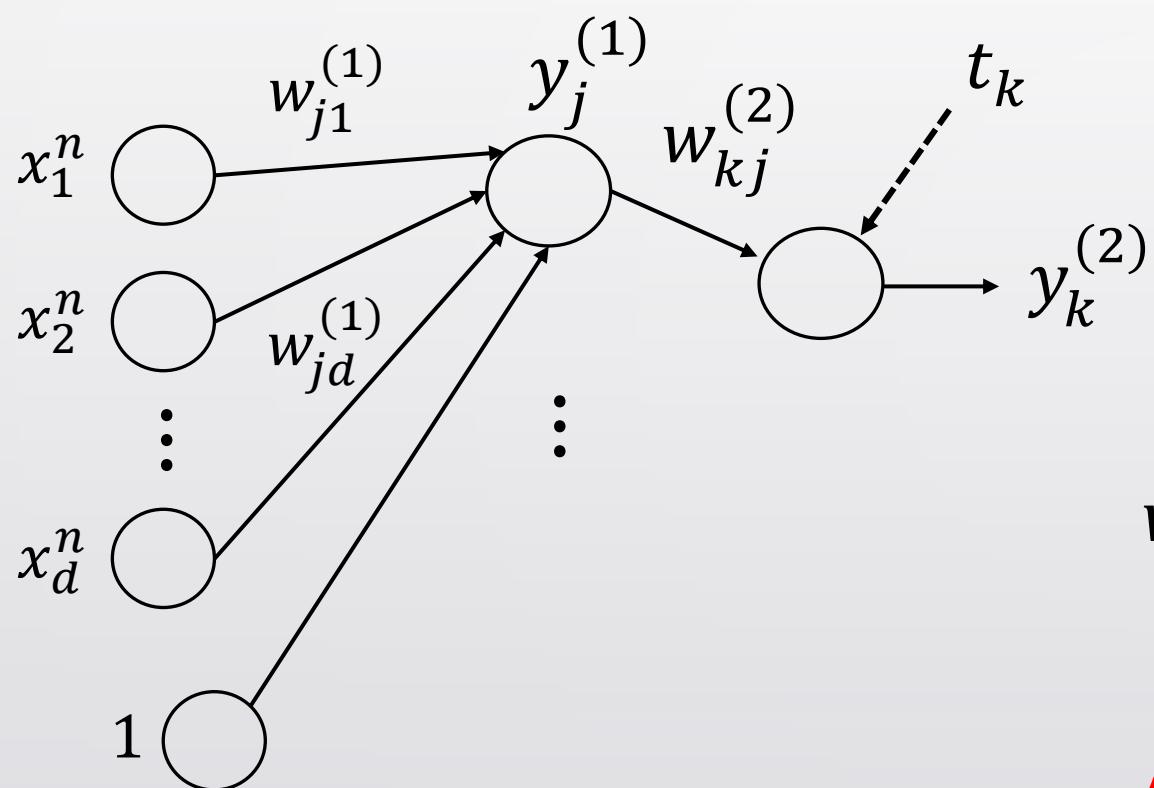
$$w_{kj}^{(2)} = w_{kj}^{(2)} + \Delta w_{kj}^{(2)}$$

$$\Delta w_{kj}^{(2)} = -\eta \delta_k^{(2)} y_j^{(1)}$$

$$\delta_k^{(2)} = \frac{(y_k^{(2)} - t_k)}{f'(v_k^{(2)})}$$

誤差 Error 出力層への入力
Input to output layer

重みの更新 Weight Updating



E を最小化するよう $w_{ji}^{(1)}$ を更新する
Update $w_{ji}^{(1)}$ so as to minimize E

$$E = \frac{1}{2} \sum_{k=1}^C (y_k^{(2)} - t_k)^2$$

$$w_{ji}^{(1)} = w_{ji}^{(1)} - \eta \frac{\partial E}{\partial w_{ji}^{(1)}} = w_{ji}^{(1)} + \Delta w_{ji}^{(1)}$$

$$\Delta w_{ji}^{(1)} = -\eta \frac{\partial E}{\partial w_{ji}^{(1)}}$$

重みの更新 Weight Updating

$$\Delta w_{ji}^{(1)} = -\eta \frac{\partial E}{\partial w_{ji}^{(1)}} = -\eta \frac{\partial}{\partial w_{ji}^{(1)}} \frac{1}{2} \sum_{k=1}^C \left(y_k^{(2)} - t_k \right)^2$$

$$= -\eta \sum_{k=1}^C \left(y_k^{(2)} - t_k \right) \frac{\partial y_k^{(2)}}{\partial w_{ji}^{(1)}}$$



重みの更新 Weight Updating

$$v_k^{(2)} = \sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)} \quad y_k^{(2)} = f(v_k^{(2)})$$

$$\Delta w_{ji}^{(1)} = -\eta \sum_{k=1}^C \left(y_k^{(2)} - t_k \right) \frac{\partial f(v_k^{(2)})}{\partial w_{ji}^{(1)}}$$

$$= -\eta \sum_{k=1}^C \left(y_k^{(2)} - t_k \right) \frac{\partial v_k^{(2)}}{\partial w_{ji}^{(1)}} \frac{\partial f(v_k^{(2)})}{\partial v_k^{(2)}}$$

重みの更新 Weight Updating

$$v_k^{(2)} = \sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)} \quad v_j^{(1)} = \sum_{i=0}^{i=d} w_{ji}^{(1)} x_i^n \quad y_j^{(1)} = f(v_j^{(1)})$$

$$\frac{\partial v_k^{(2)}}{\partial w_{ji}^{(1)}} = \frac{\partial v_j^{(1)}}{\partial w_{ji}^{(1)}} \frac{\partial v_k^{(2)}}{\partial v_j^{(1)}} = x_i^n \frac{\partial}{\partial v_j^{(1)}} \sum_{l=0}^{l=M} w_{kl}^{(2)} f(v_l^{(1)})$$

$$= x_i^n w_{kj}^{(2)} \frac{\partial f(v_j^{(1)})}{\partial v_j^{(1)}} = x_i^n w_{kj}^{(2)} f'(v_j^{(1)})$$

重みの更新 Weight Updating

$$\begin{aligned}\Delta w_{ji}^{(1)} &= -\eta \sum_{k=1}^C \left(y_k^{(2)} - t_k \right) \frac{\partial v_k^{(2)}}{\partial w_{ji}^{(1)}} \frac{\partial f(v_k^{(2)})}{\partial v_k^{(2)}} \\ &= -\eta \sum_{k=1}^C \left(y_k^{(2)} - t_k \right) x_i^n w_{kj}^{(2)} f'(v_j^{(1)}) \frac{\partial f(v_k^{(2)})}{\partial v_k^{(2)}} \\ \delta_k^{(2)} &= \left(y_k^{(2)} - t_k \right) f'(v_k^{(2)})\end{aligned}$$

重みの更新 Weight Updating

$$\Delta w_{ji}^{(1)} = -\eta \sum_{k=1}^C (y_k^{(2)} - t_k) x_i^n w_{kj}^{(2)} f'(v_j^{(1)}) \frac{\partial f(v_k^{(2)})}{\partial v_k^{(2)}}$$

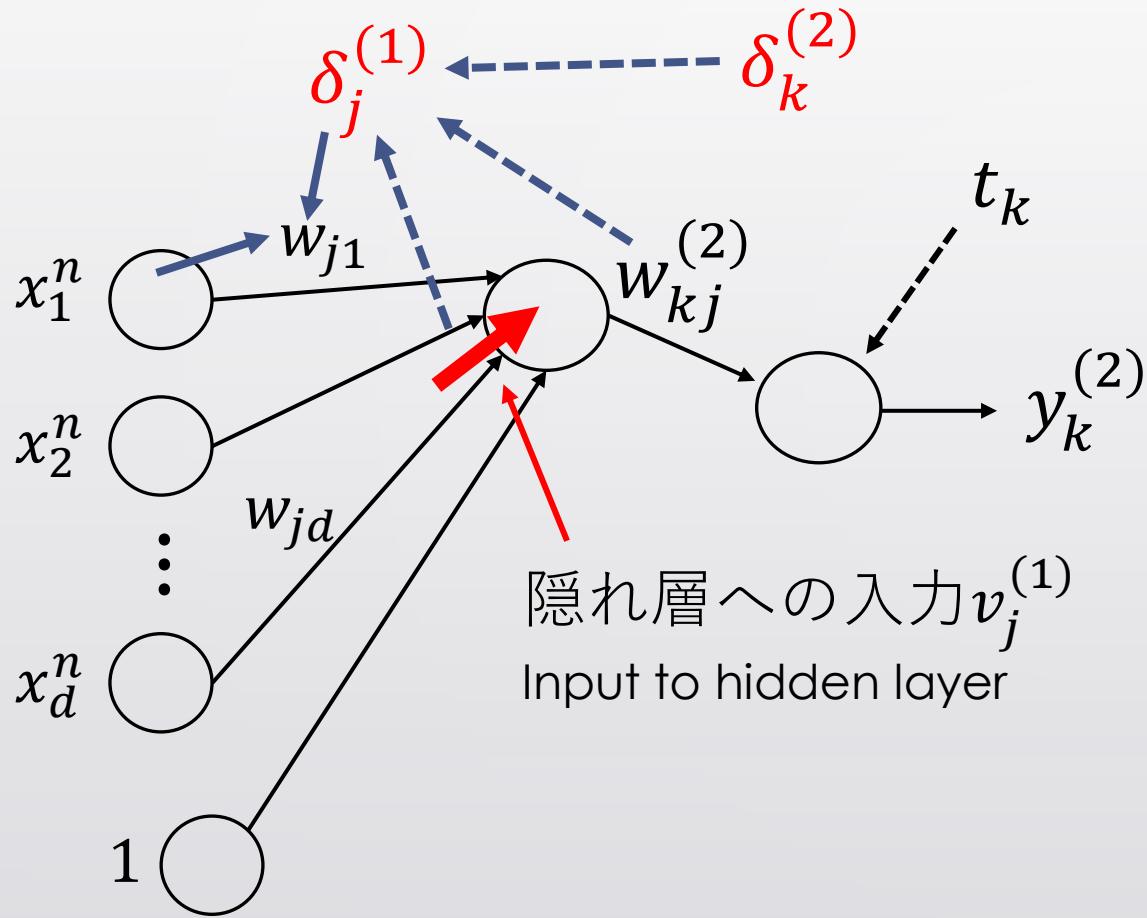
$$= -\eta \underbrace{x_i^n}_{\begin{array}{l} \text{↑} \\ \text{入力 Input} \end{array}} f'(v_j^{(1)}) \underbrace{\sum_{k=1}^C w_{kj}^{(2)} \delta_k^{(2)}}_{\begin{array}{l} \text{↓} \\ \text{隠れ層への入力} \\ \text{Input to hidden layer} \end{array}}$$

入力 Input

隠れ層への入力

Input to hidden layer

重みの更新 Weight Updating

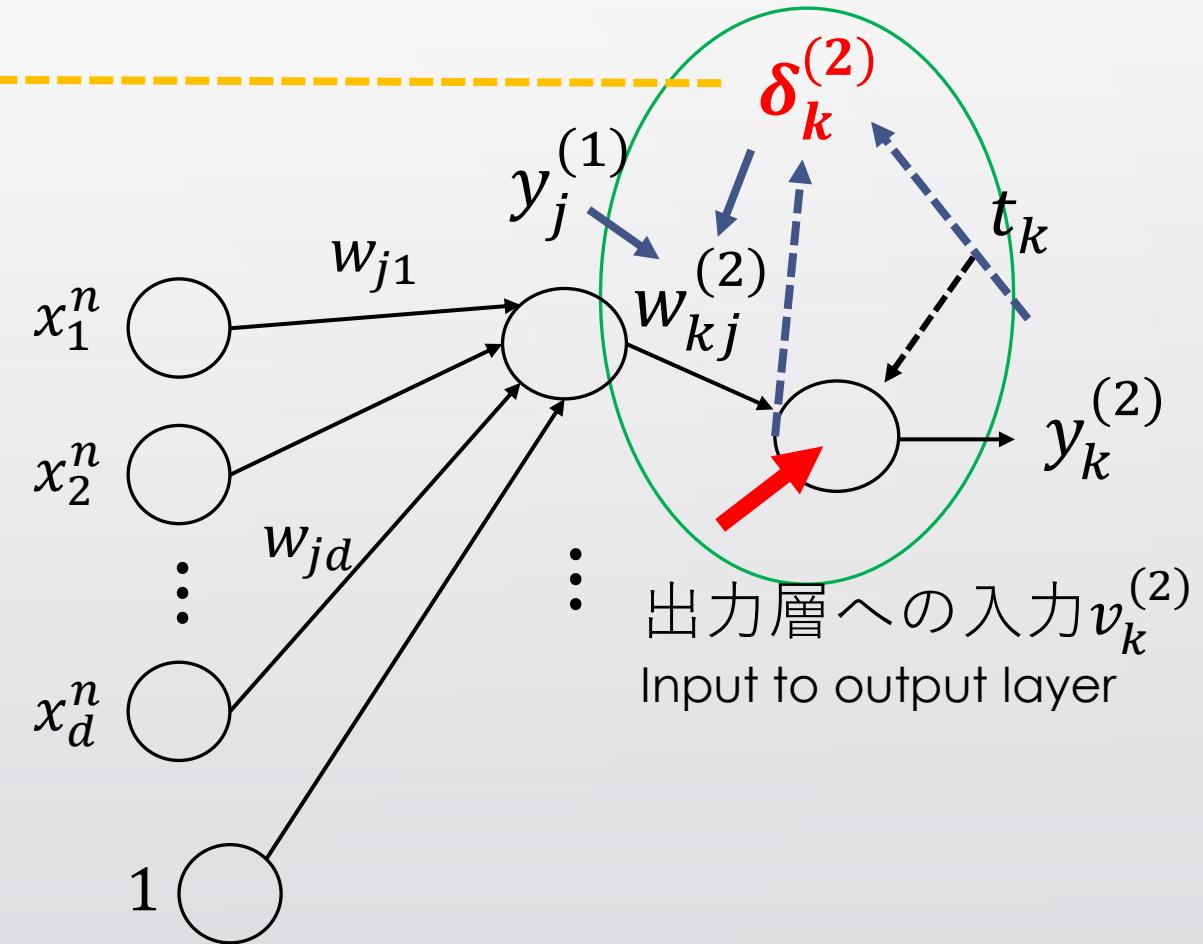
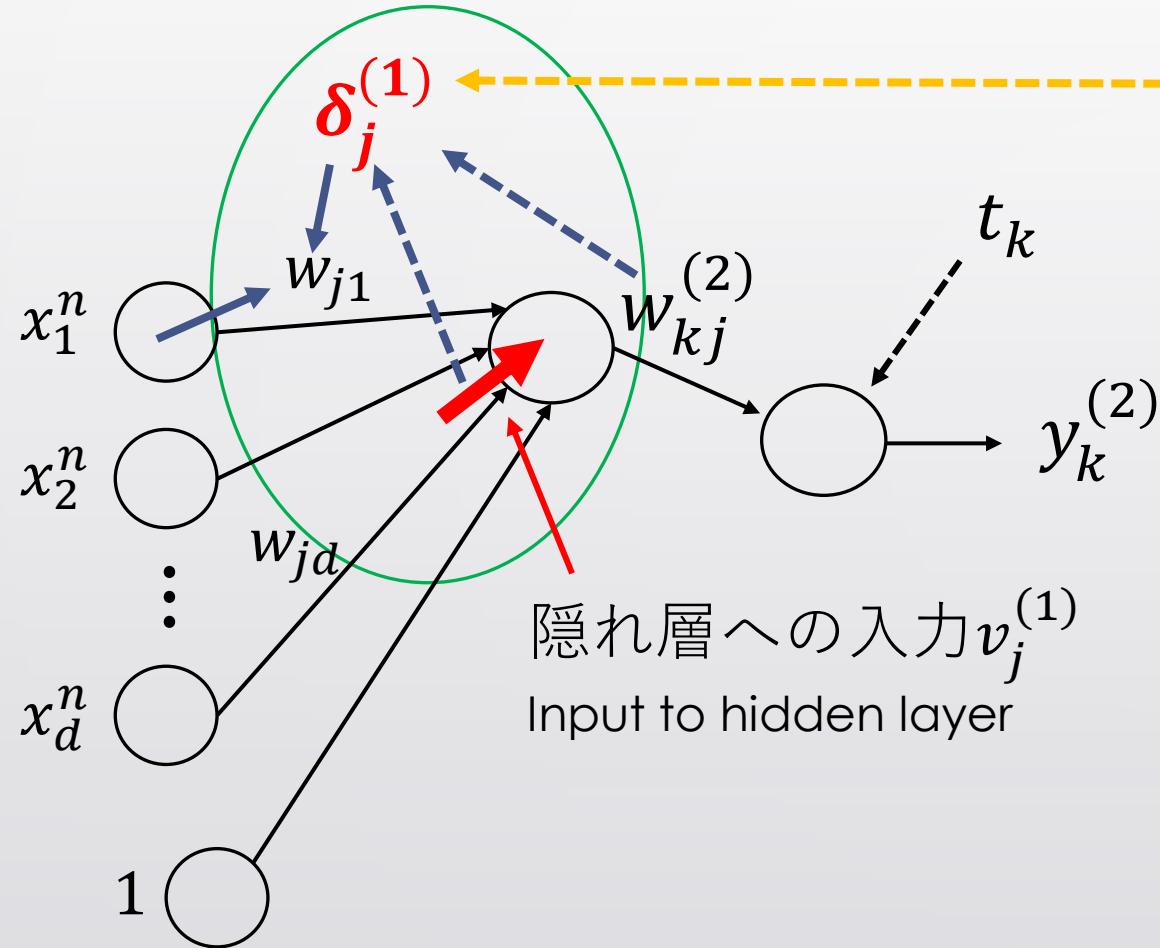


$$w_{ji}^{(1)} = w_{ji}^{(1)} + \Delta w_{ji}^{(1)}$$
$$\Delta w_{ji}^{(1)} = -\eta \delta_j^{(1)} x_i^n$$

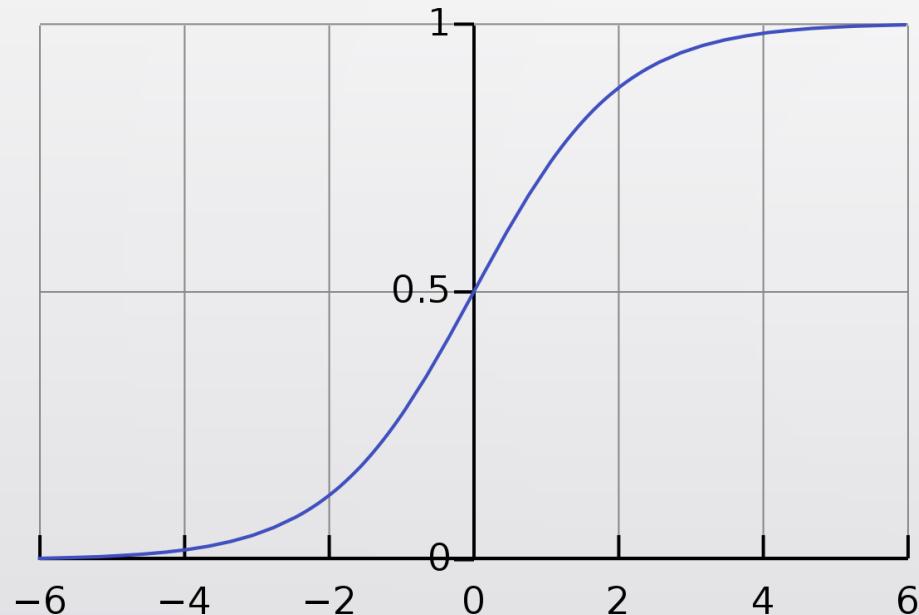
入力 Input

$$\delta_j^{(1)} = f' \left(v_j^{(1)} \right) \sum_{k=1}^{k=C} w_{kj}^{(2)} \cdot \delta_k^{(2)}$$

重みの更新 Weight Updating



シグモイド関数 Sigmoid Function

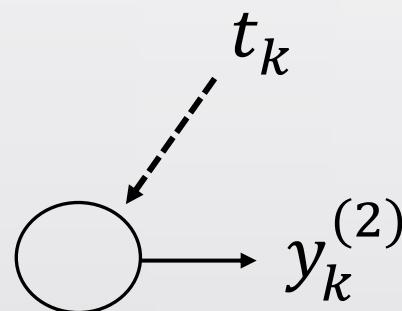


$$f(x) = \frac{1}{1 + e^{-x}}$$

$$f'(x) = \frac{1 + e^{-x}}{(1 + e^{-x})^2} = f(x)(1 - f(x))$$

https://en.wikipedia.org/wiki/Sigmoid_function

ソフトマックス関数 Softmax Function

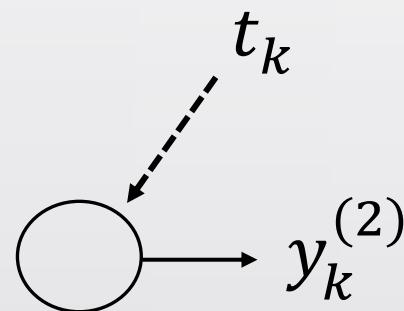


ソフトマックス関数で出力を各クラスに属する確率に変換する
Convert output of NN to the probability of belonging to each class
by softmax function

$$P_k = \frac{\exp(y_k^{(2)})}{\sum_{k=1}^C \exp(y_k^{(2)})}$$

C : クラスの総数
Total number of classes

損失関数 Loss Function



出力と教師データの誤差を反映する損失関数を最小化する

Minimize loss function that reflects divergence between
output of network and teacher signal

二乗誤差の和も損失関数の一つ

Sum of squared error is one type of loss function

$$E = \frac{1}{2} \sum_{k=1}^{k=C} \left(y_k^{(2)} - t_k \right)^2$$

交差エントロピー – Cross Entropy

KLダイバージェンスを拡張した損失関数

Loss function based on the concept of KL divergence

$$CE = - \sum_{k=1}^{k=C} t_k \log(P_k) \quad P_k = \frac{\exp(y_k^{(2)})}{\sum_{k=1}^{k=C} \exp(y_k^{(2)})}$$



交差エントロピー – Cross Entropy

$$KL(p(x) \| q(x)) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

$$p = (t_1, t_2 \dots t_C) \quad t_k \in \{0, 1\} \quad \sum_{k=1}^{k=C} t_k = 1$$

$$q = (P_1, P_2 \dots P_C)$$

交差エントロピー – Cross Entropy

$$KL(p|q) = \sum_{k=1}^{k=C} t_k \log \left(\frac{t_k}{P_k} \right)$$

$$= \underbrace{\sum_{k=1}^{k=C} t_k \log(t_k)}_{\text{教師信号にのみ関係}} - \underbrace{\sum_{k=1}^{k=C} t_k \log(P_k)}_{\text{交差エントロピー}}$$

教師信号にのみ関係
Relevant only to
teacher signal

交差エントロピー
Cross Entropy



データマイニング

Data Mining

14: ニューラルネットワーク② Neural Network

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

2次元の畳み込み 2D-Convolution

0	2	0	1	2	0
1	0	3	0	0	2
3	1	0	2	2	1
2	2	2	1	0	1
1	0	1	3	2	2
0	3	1	0	2	0

1	0	0
0	1	0
0	0	1

0		

$$0*1 + 2*0 + 0*0 + \\ 1*0 + 0*1 + 3*0 + \\ 3*0 + 1*0 + 0*1 = 0$$

→

0	2	0	1	2	0
1	0	3	0	0	2
3	1	0	2	2	1
2	2	2	1	0	1
1	0	1	3	2	2
0	3	1	0	2	0

1	0	0
0	1	0
0	0	1

0	7	

$$2*1 + 0*0 + 1*0 + \\ 0*0 + 3*1 + 0*0 + \\ 1*0 + 0*0 + 2*1 = 7$$

<https://axa.biopapyrus.jp/deep-learning/cnn/convolution.html>

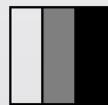


画像フィルター Image Filter

$$\begin{array}{|c|c|c|c|c|c|} \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline 10 & 10 & 10 & 0 & 0 & 0 \\ \hline \end{array} \quad * \quad \begin{array}{|c|c|c|} \hline 1 & 0 & -1 \\ \hline 1 & 0 & -1 \\ \hline 1 & 0 & -1 \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline -0 & 30 & 30 & 0 \\ \hline 0 & 30 & 30 & 0 \\ \hline 0 & 30 & 30 & 0 \\ \hline 0 & 30 & 30 & 0 \\ \hline \end{array}$$

3×3 4×4

6×6



<https://datahacker.rs/edge-detection/>

中央のフィルターで畳み込むと、縦方向のエッジが強調される

When an image is convoluted with a filter in the center, vertical edges in the image is extracted



画像フィルター – Image Filter



1	0	-1
1	0	-1
1	0	-1



Vertical edges

-1	-1	-1
0	0	0
1	1	1



Horizontal edges

<https://datahacker.rs/edge-detection/>



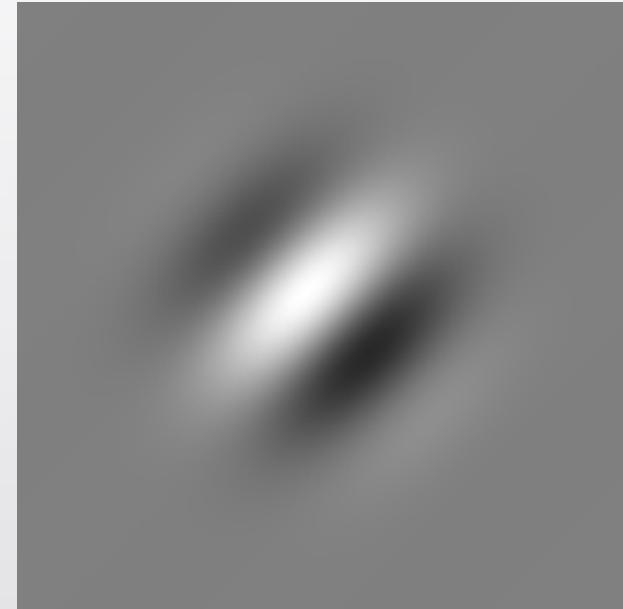
ガボールフィルター – Gabor Filter

$$g(x, y) = K \exp\left(-\frac{x'^2 + y'^2}{2}\right) \exp\{i(2\pi f x' + P)\}$$

$$(x', y') = (x, y) \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_x} & 0 \\ 0 & \frac{1}{\sigma_y} \end{pmatrix}$$

座標回転した後、正規化

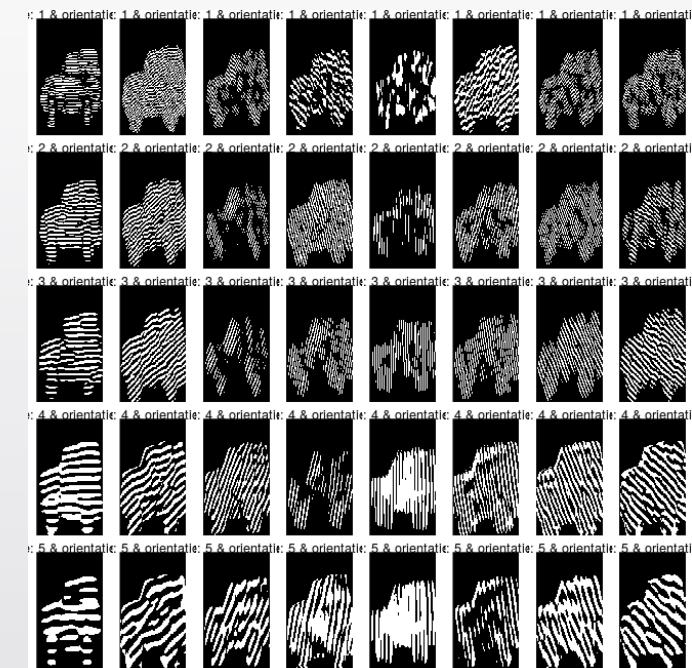
Coordinate rotation and
then scaling





ガボールフィルター – Gabor Filter

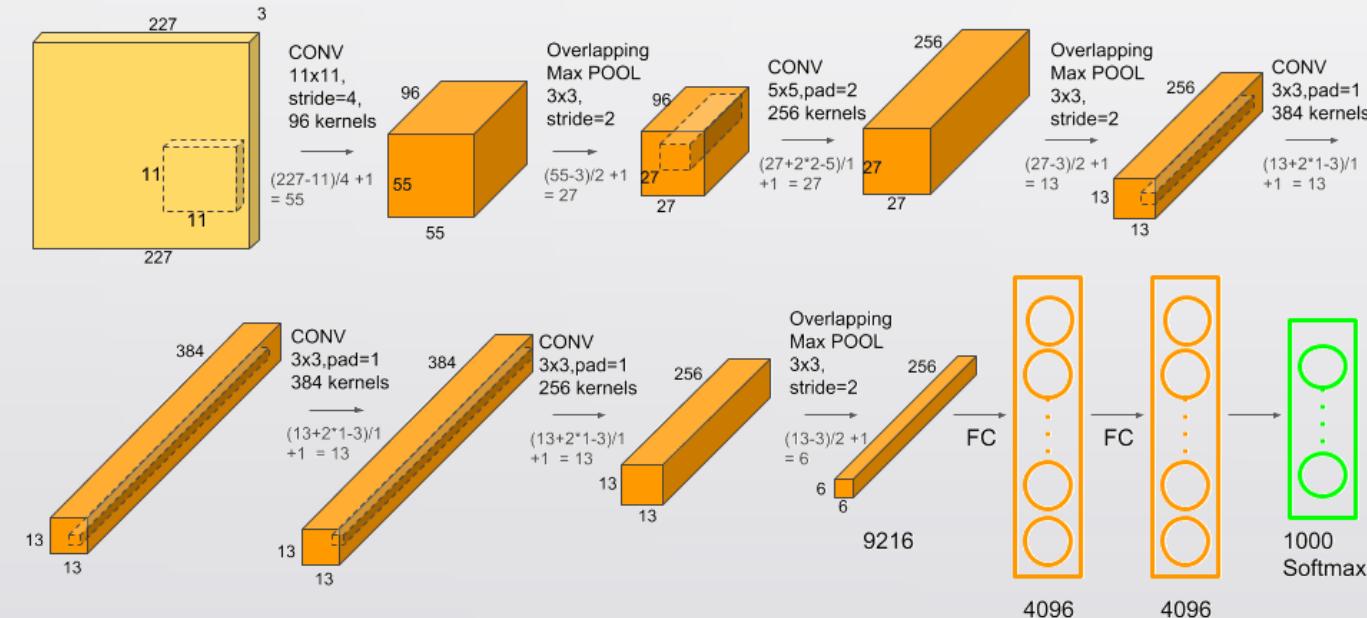
Soons et al., 2016



https://cran.r-project.org/web/packages/OpenImageR/vignettes/Gabor_Feature_Extraction.html

畳み込みニューラルネットワーク CNN Convolutional Neural Network

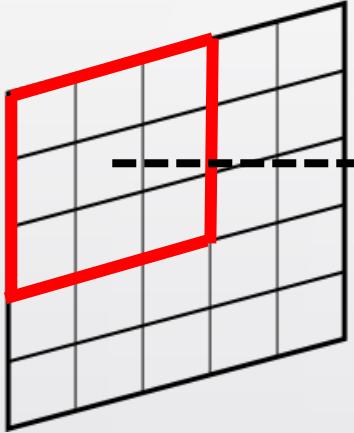
Alex Net



Layer		Feature Map	Size	Kernel Size	Stride	Activation
Input	Image	1	227x227x3	-	-	-
1	Convolution	96	55 x 55 x 96	11x11	4	relu
	Max Pooling	96	27 x 27 x 96	3x3	2	relu
2	Convolution	256	27 x 27 x 256	5x5	1	relu
	Max Pooling	256	13 x 13 x 256	3x3	2	relu
3	Convolution	384	13 x 13 x 384	3x3	1	relu
4	Convolution	384	13 x 13 x 384	3x3	1	relu
5	Convolution	256	13 x 13 x 256	3x3	1	relu
	Max Pooling	256	6 x 6 x 256	3x3	2	relu
6	FC	-	9216	-	-	relu
7	FC	-	4096	-	-	relu
8	FC	-	4096	-	-	relu
Output	FC	-	1000	-	-	Softmax

<https://medium.com/@siddheshb008/alexnet-architecture-explained-b6240c528bd5>

畳み込み層 Convolution Layer



i 番目の赤枠内の画素値のベクトル x_i
Vector of pixels within i -th red square

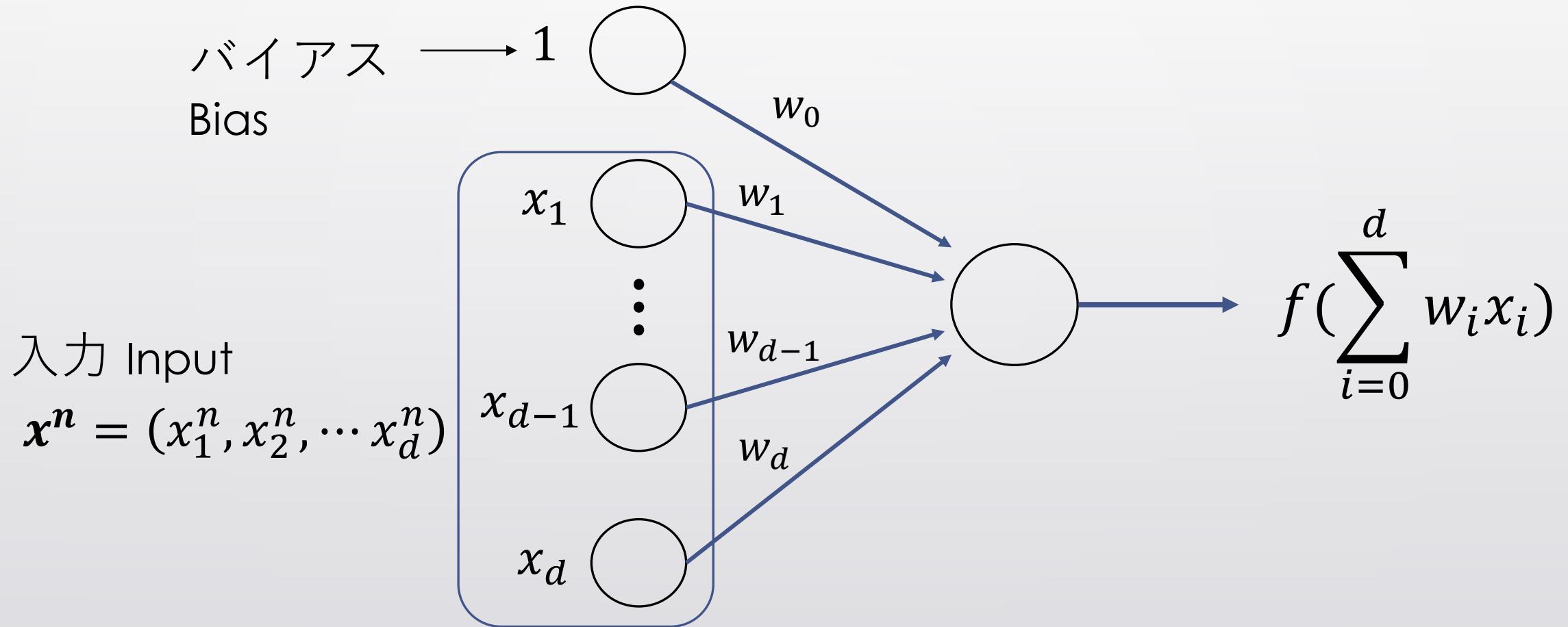


j 番目のフィルターの重みベクトル w_j
Weight vector of j -th filter

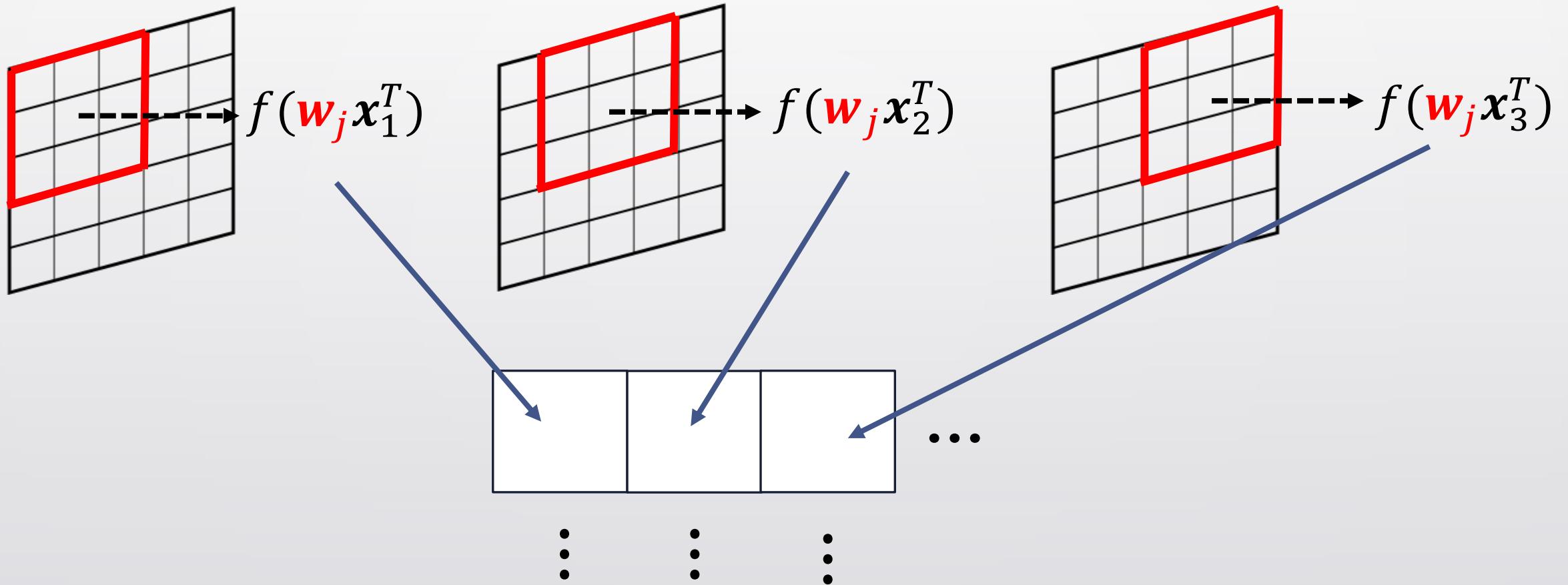


$$\text{Activation} = f(w_j x_i^T)$$

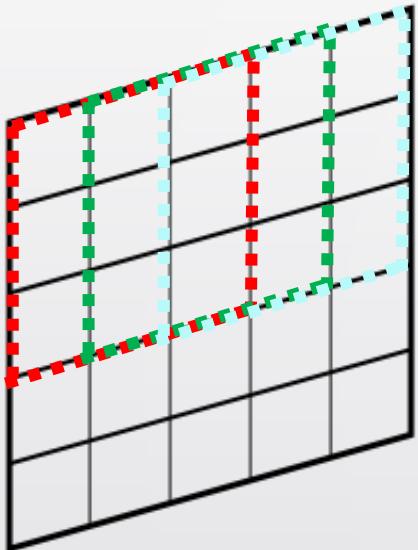
パーセプトロン Perception



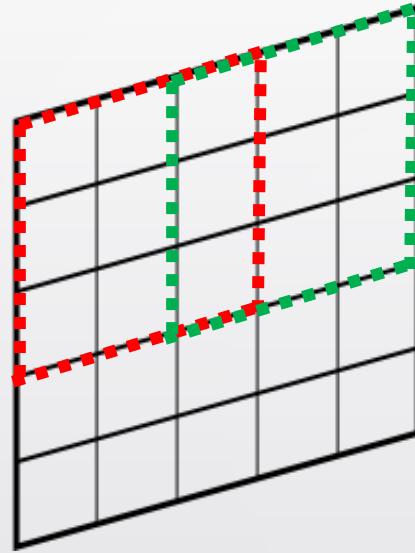
畳み込み層 Convolution Layer



パディングとストライド Padding and Stride



Stride = 1



Stride = 2

ストライドが大きいと畳み込みの結果出力される画像が小さくなる

Larger stride makes size of output image smaller

0	0	0	0	0	0
0					0
0					0
0					0
0	0	0	0	0	0

出力画像の周りを数値で埋めることで、画像サイズを復元する

Restore image size by adding numerical values around output image

パディングとストライド Padding and Stride

H, W : 入力画像のサイズ
Size of input image

O_h, O_w : 出力画像のサイズ
Size of output image

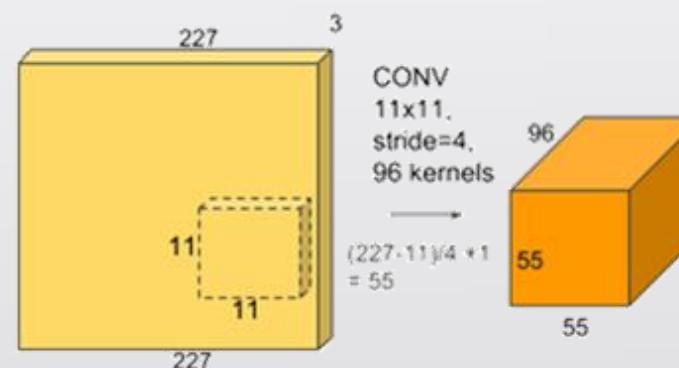
F_h, F_w : フィルターのサイズ
Size of output image

P : パディングの幅 Width of Padding

S : ストライド Stride

$$O_w = \frac{W+2P-F_w}{S} + 1$$

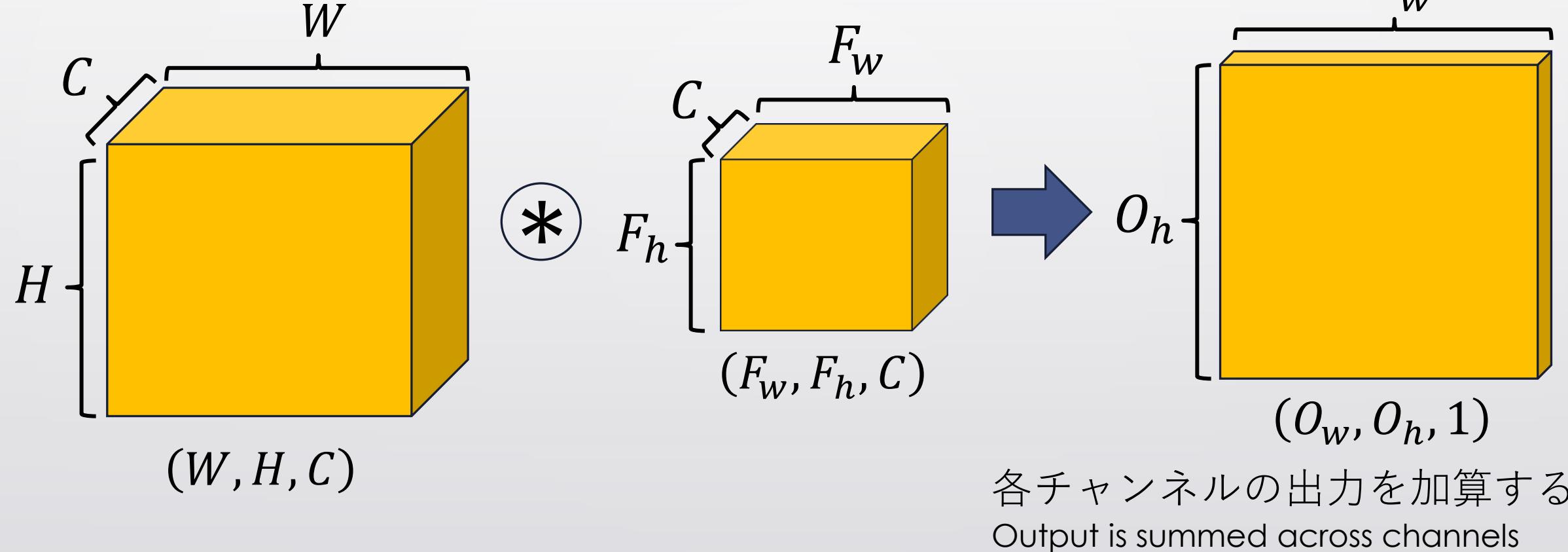
$$O_h = \frac{H+2P-F_h}{S} + 1$$



AlexNetでは96枚のフィルターを使用
96 filters are used in the first convolution layer in Alex Net

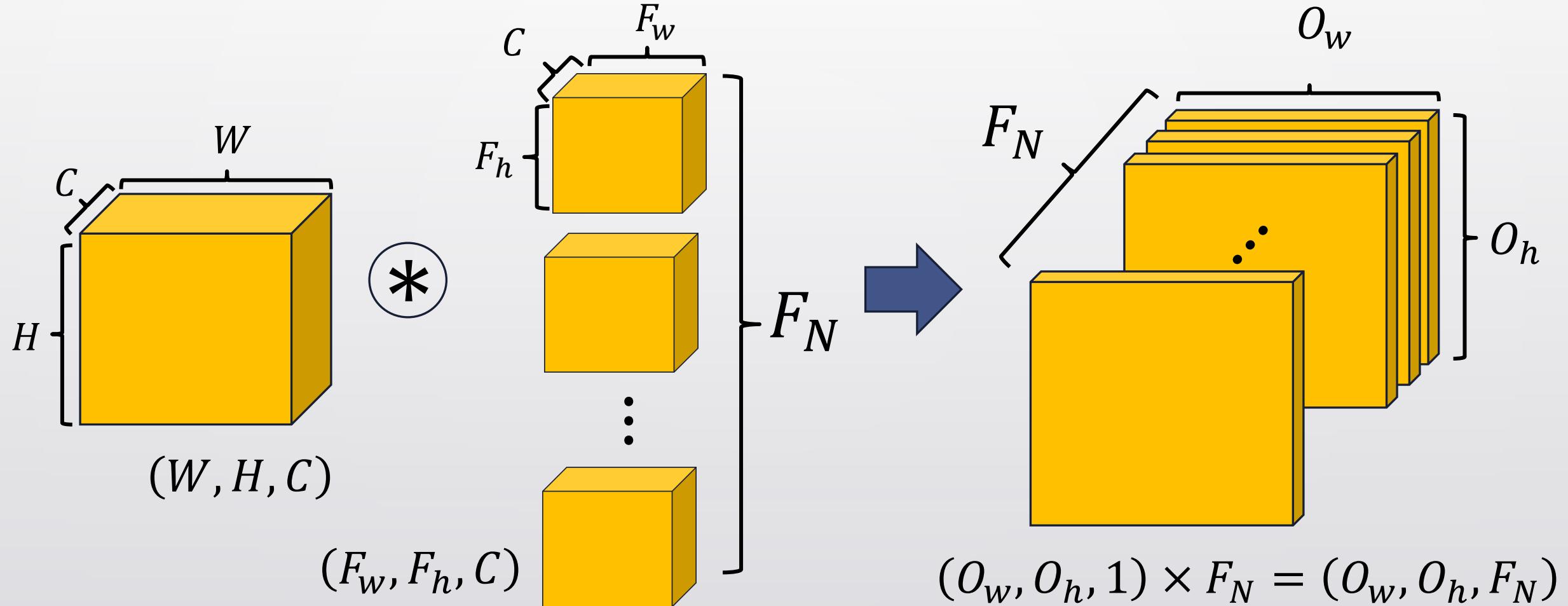


複数チャンネルの畳み込み Convolution of Multiple Channels

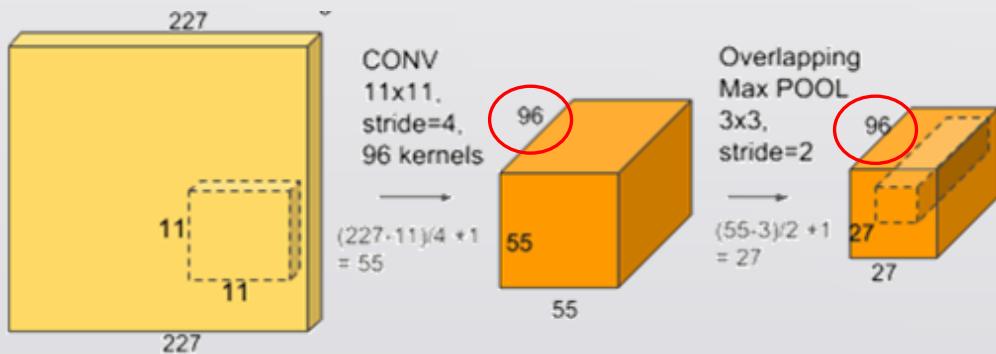
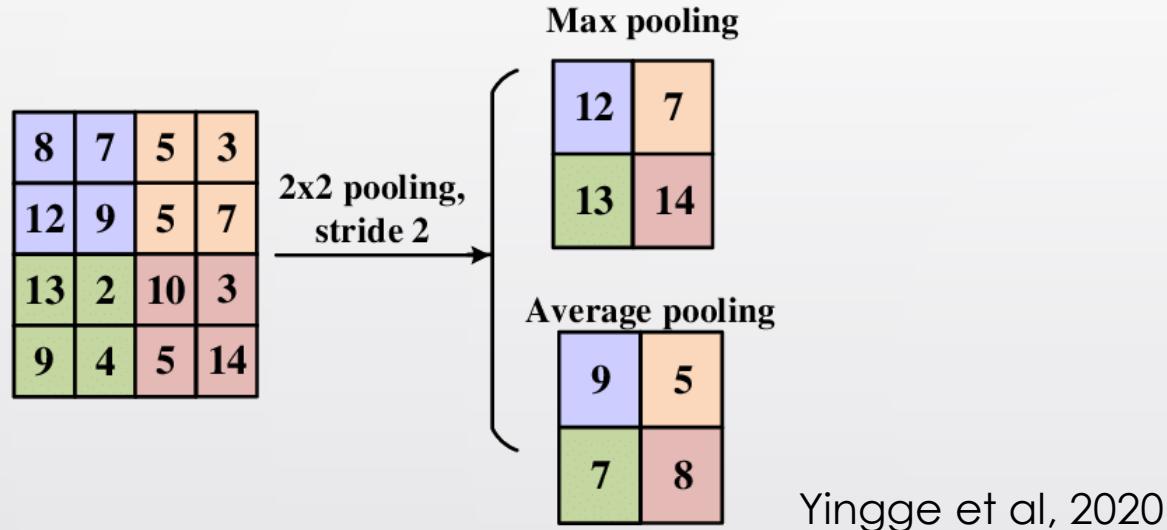




複数チャンネルの畳み込み Convolution of Multiple Channels



プーリング層 Pooling Layer



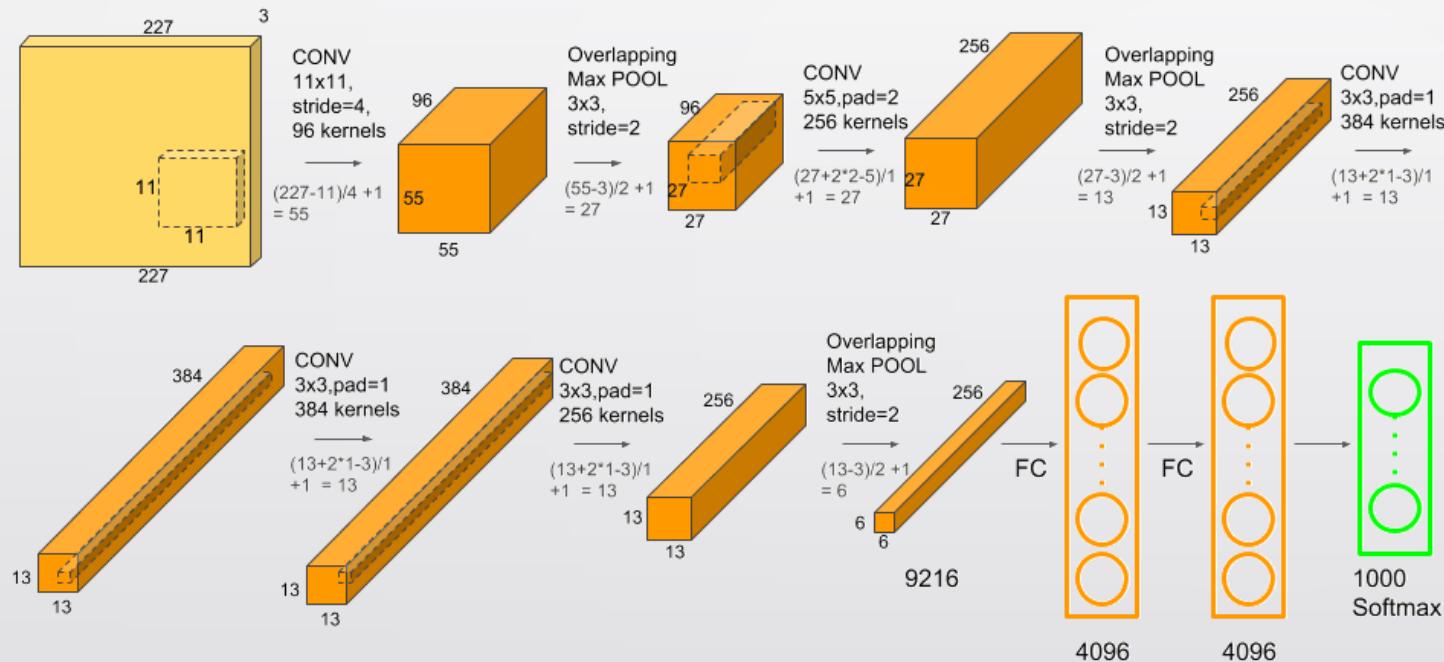
プーリングで位置への依存性を低減させる

Dependency on location within an image is mitigated by pooling

プーリングでも出力画像サイズは小さくなるので、パディングを行うことがある

Padding is often used to avoid shrinkage of image by pooling

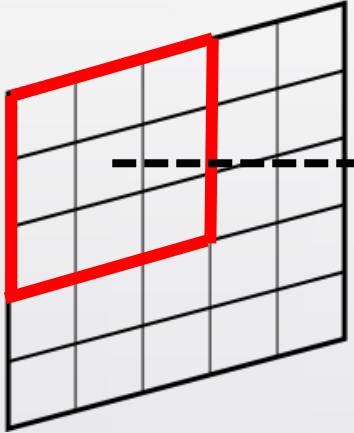
Alex Net



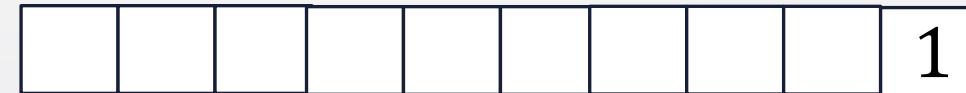
畳み込み層とプーリング層の後に、全結合層を配置する

Fully connected layers are located after repetition of convolution and pooling layers

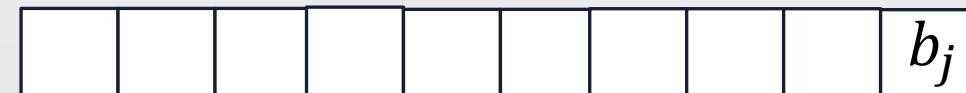
畳み込み層 Convolution Layer



i 番目の赤枠内の画素値のベクトル x_i
Vector of pixels within i -th red square

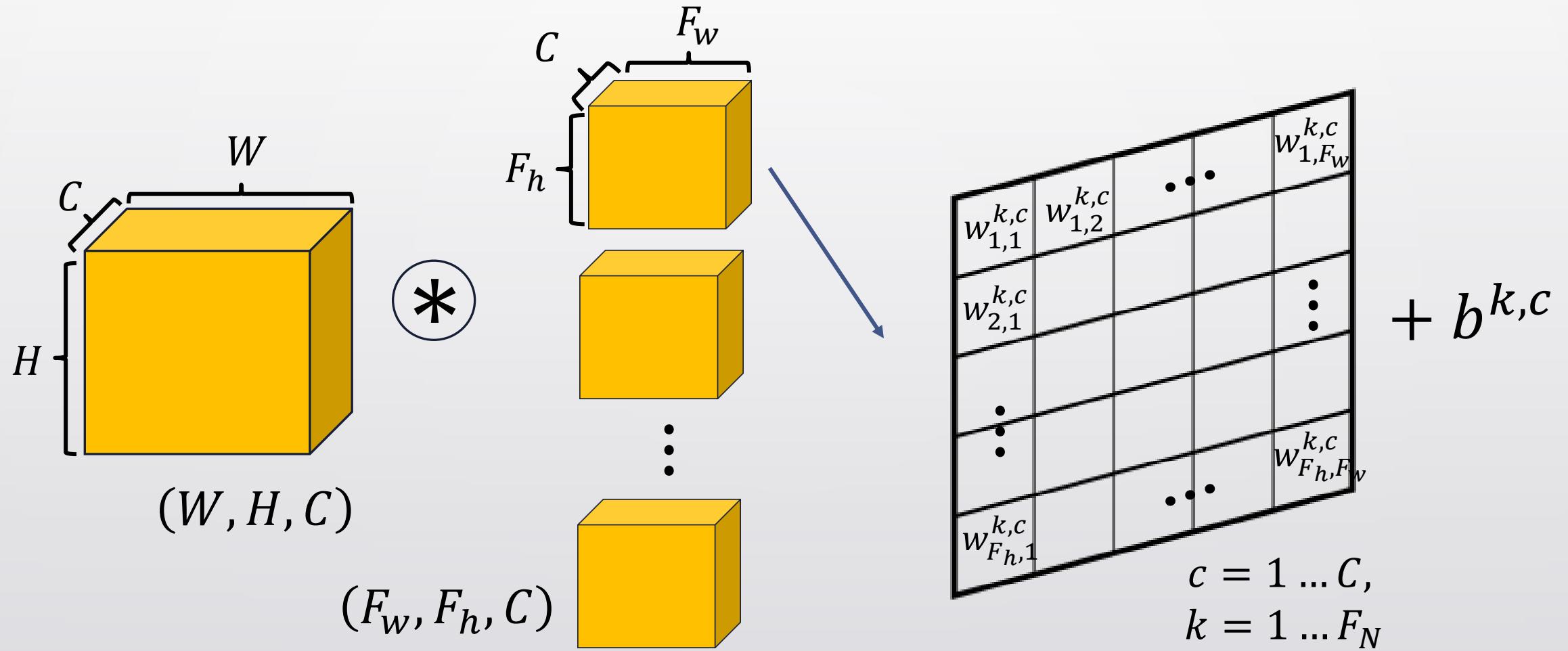


j 番目のフィルターの重みベクトル w_j
Weight vector of j -th filter



$$\text{Activation} = f(w_j x_i^T)$$

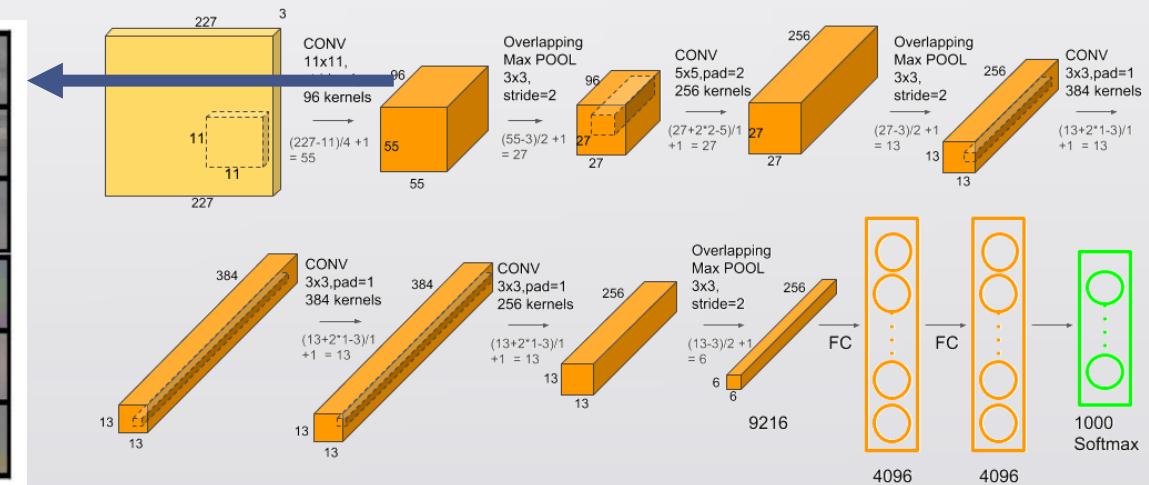
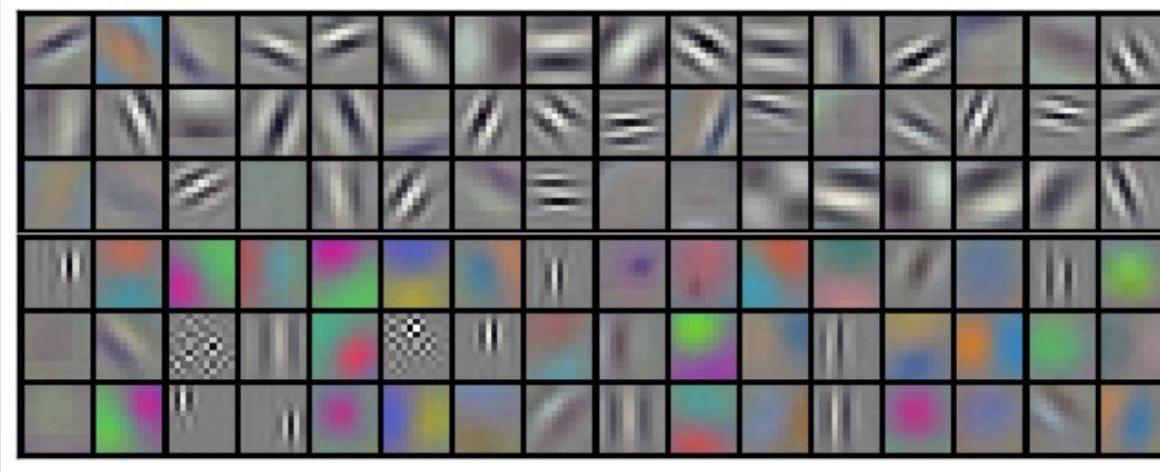
フィルターの学習 Acquisition of Filters



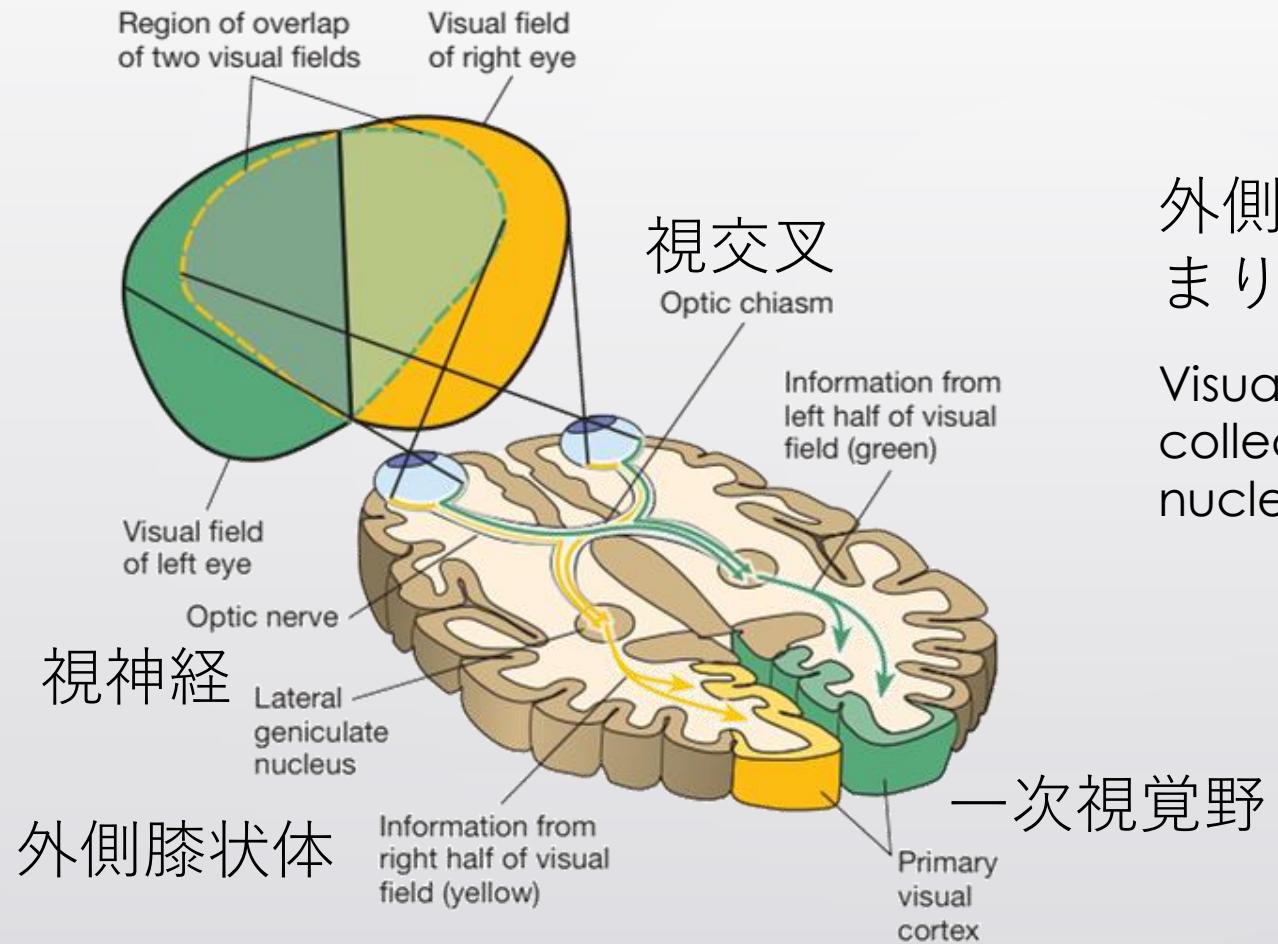
フィルターの学習 Acquisition of Filters

誤差逆伝搬法による学習の結果、Alex Netの第一の畠み込み層でガボールフィルターが獲得された

Weight updating by backpropagation led to acquisition of Gabor filter at the first convolution layer of Alex Net



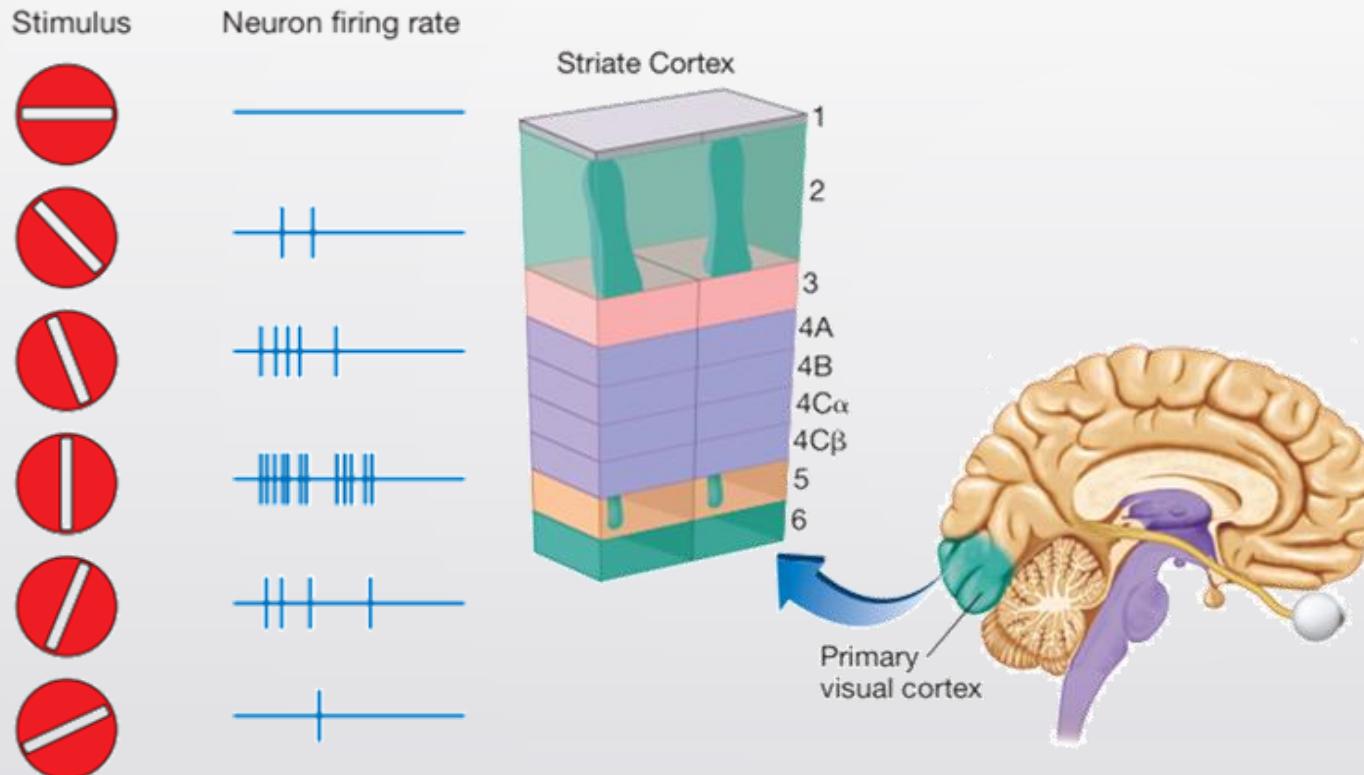
視覚伝導路 Visual Pathway



外側膝状体では、視覚像は画素の集まりとして表象される

Visual image is represented as a collection of pixels at lateral geniculate nucleus

單純・複雜細胞 Simple and Complex Cells at V1

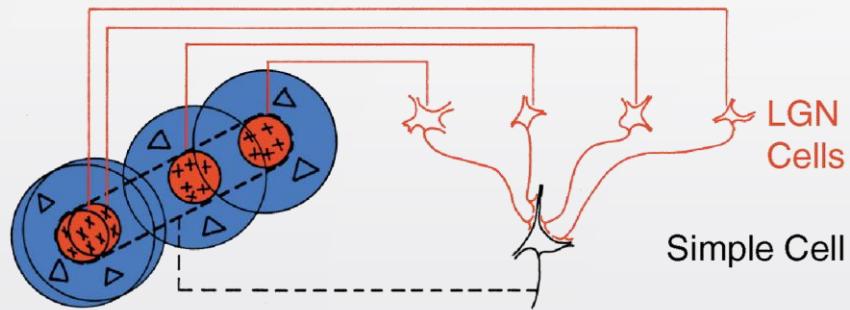


一次視覚野には、受容野内の色・傾きなどの単純な知覚特徴量に反応する細胞がある

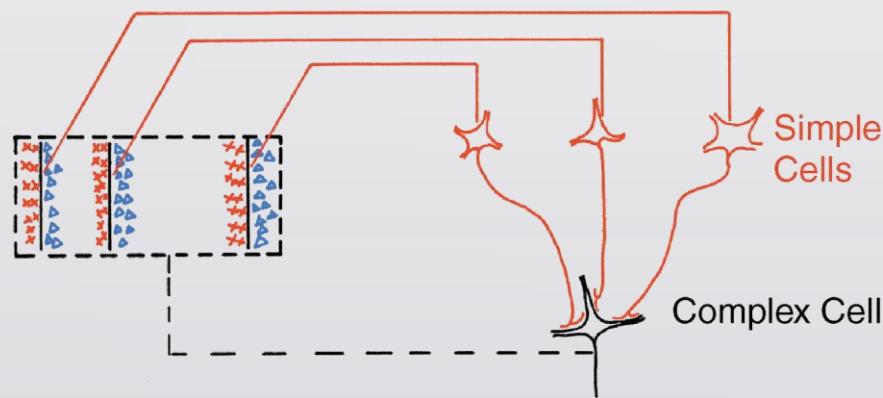
Primary visual cortex contains neurons responsive to low-order perceptual features such as color and orientation within corresponding receptive field

單純・複雜細胞 Simple and Complex Cells at V1

Circuit Building a Simple Cell from LGN Cells



Building a Complex Cell from Simple Cells

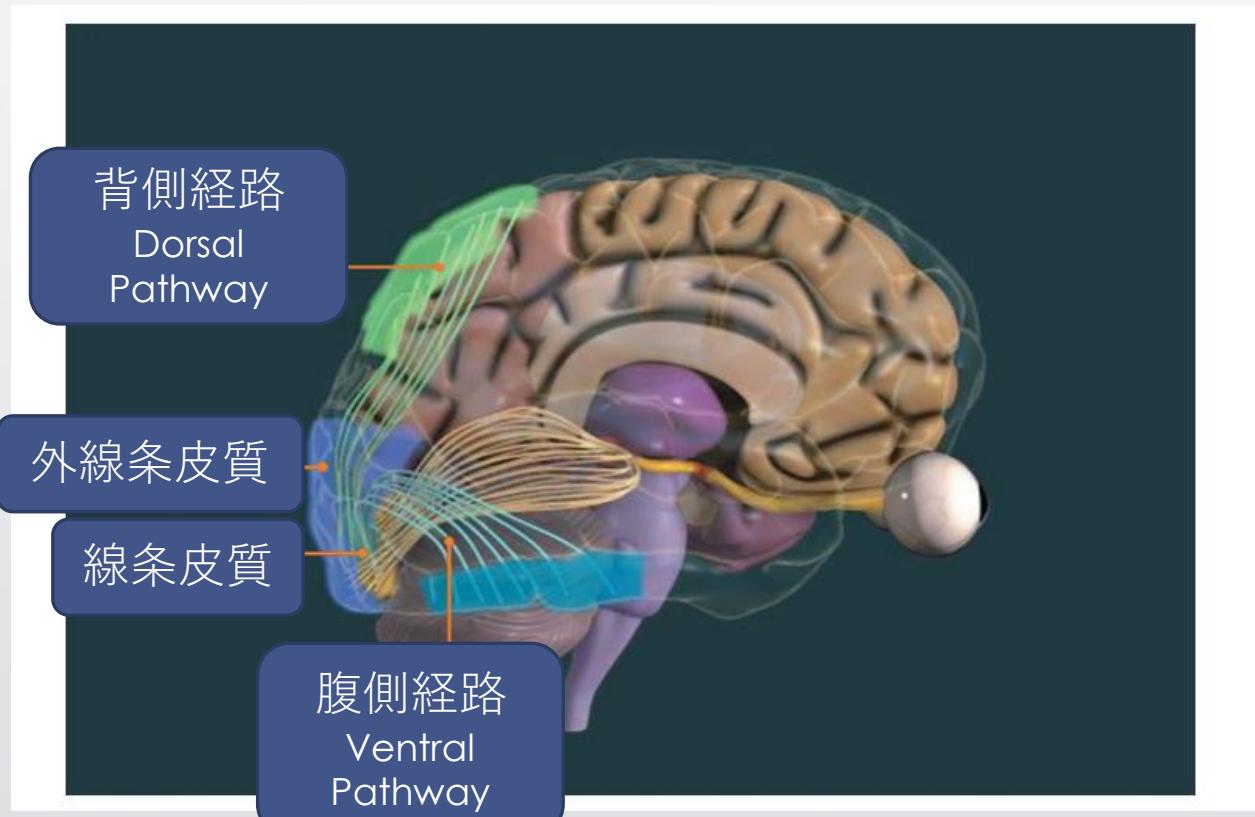


広域の情報を統合することで、
複雑な画像情報が計算される

Complex visual information is
computed by integrating features
from broad receptive field

Callaway, 2001

高次視覚野 Higher-Order Visual Regions

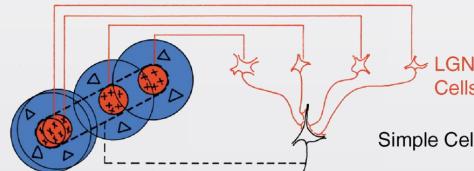


視覚情報処理の下流で、物体
カテゴリーが識別される
Object category is classified at
the downstream of visual
processing

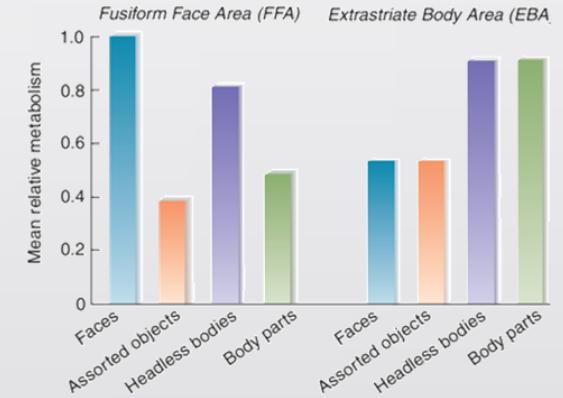
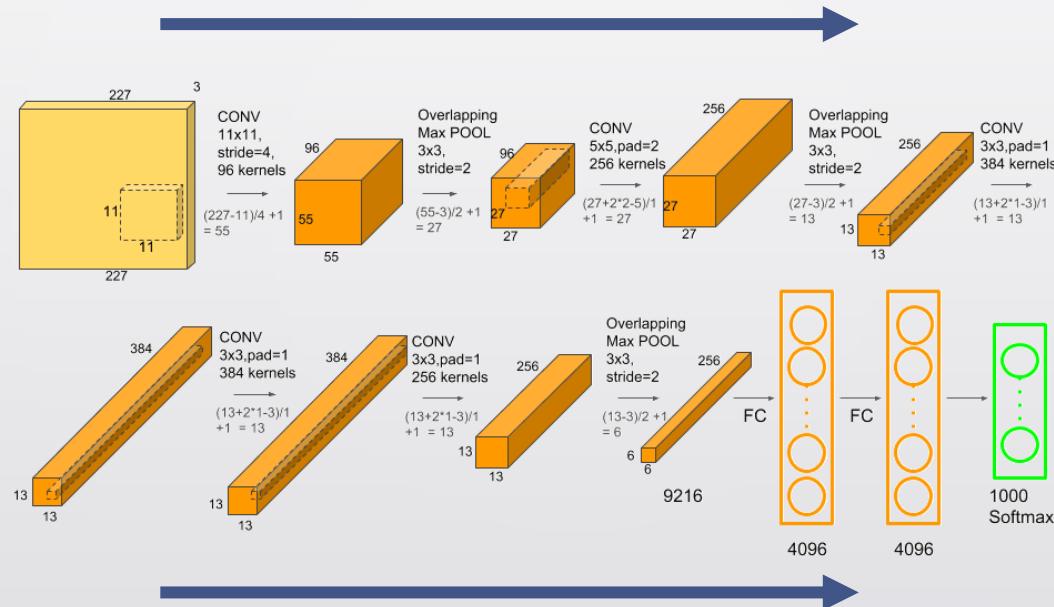
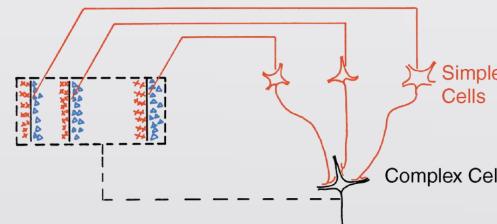
CNNと脳の類似性 Similarity between CNN and Brain

プーリングの効果で受容野が広くなる
Receptive field gets broader as a result of pooling

Circuit Building a Simple Cell from LGN Cells



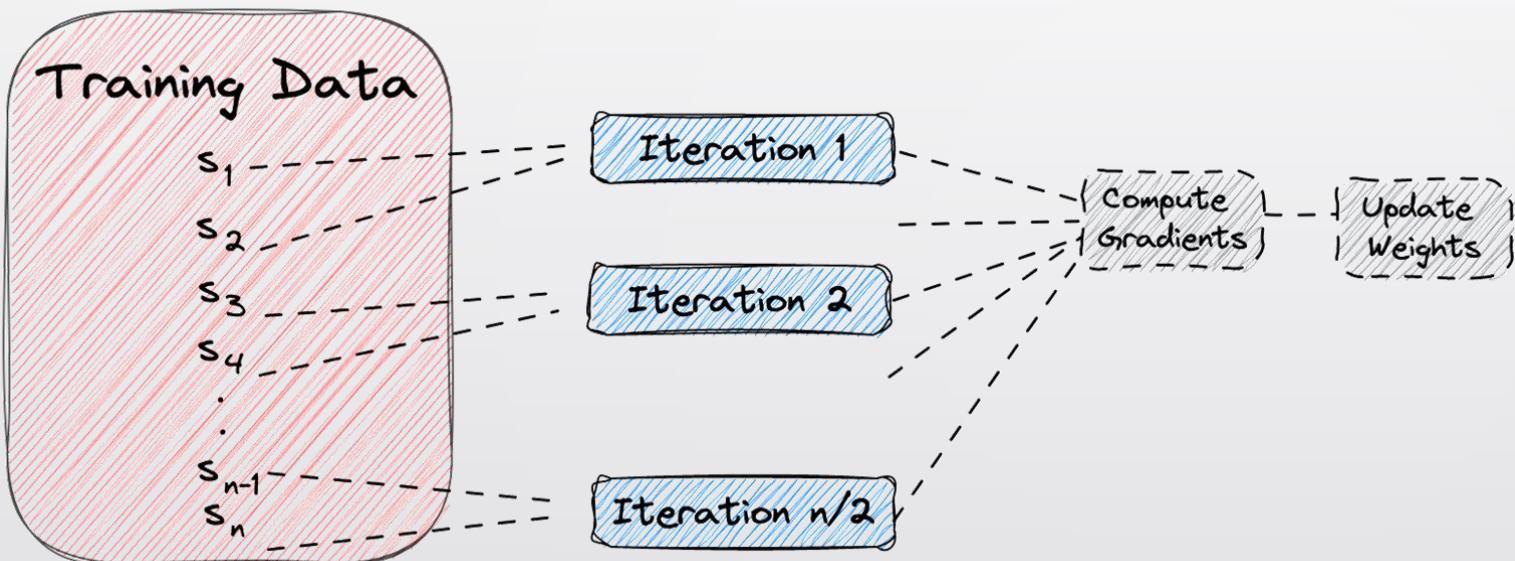
Building a Complex Cell from Simple Cells



畳み込み層で複雑な情報を表現する

More complex information is represented at convolution layers

バッチ学習 Batch Learning



<https://www.baeldung.com/cs/epoch-vs-batch-vs-mini-batch>

ランダムに選択した n 個のデータをまとめて $batch\ size = n$ のミニバッチを作る

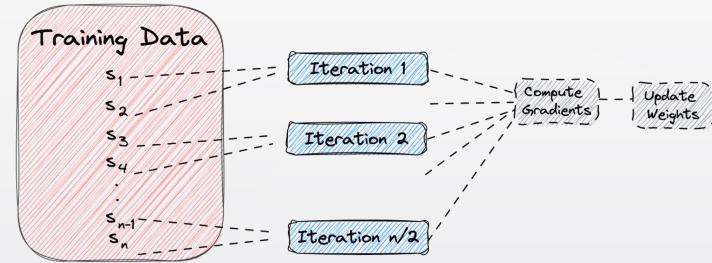
Create mini-batches with $batch\ size = n$ from randomly-selected n data

各ミニバッチのデータで重みを更新する

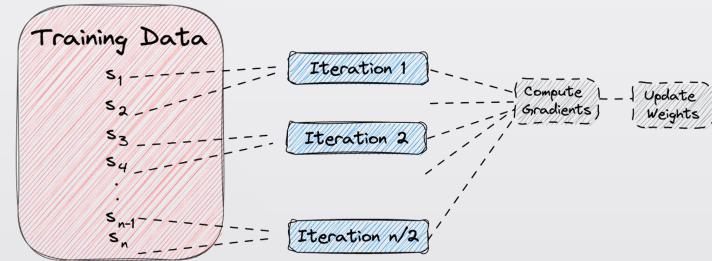
Update weights based on data from single minibatch in each iteration

エポック Epoch

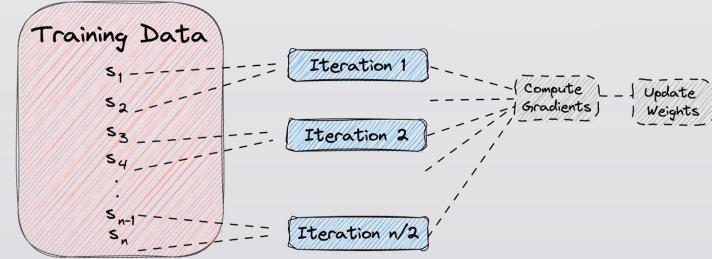
エポック1
Epoch 1



エポック2
Epoch 2



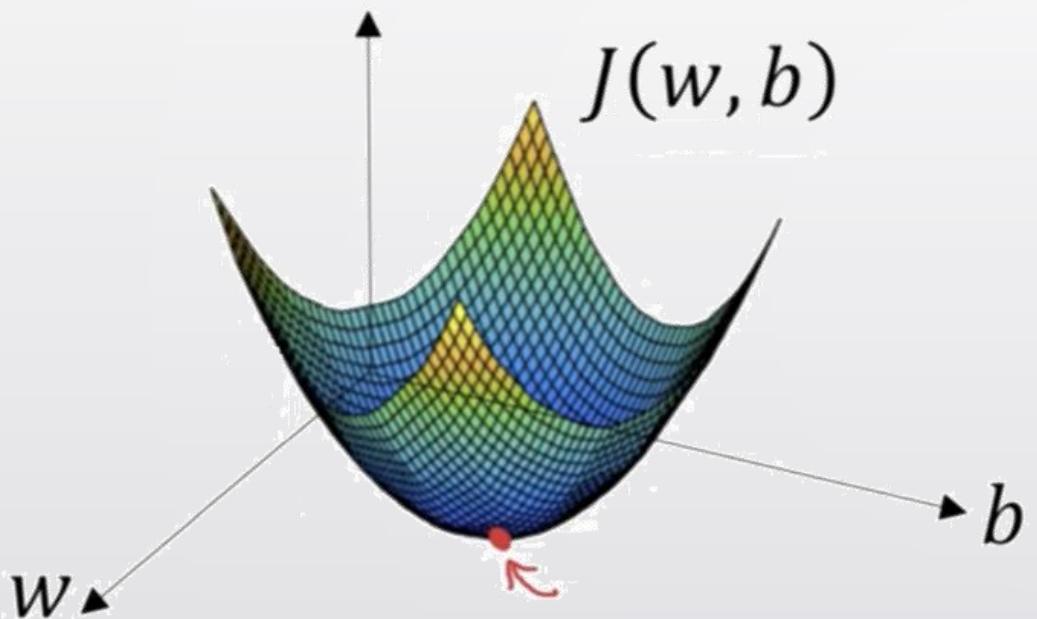
エポック3
Epoch 3



各エポックでは、ミニバッチすべてを用いた重みの更新を一巡する

A cycle of interactions using all the mini-batches is completed in single epoch

勾配降下 Gradient Descent



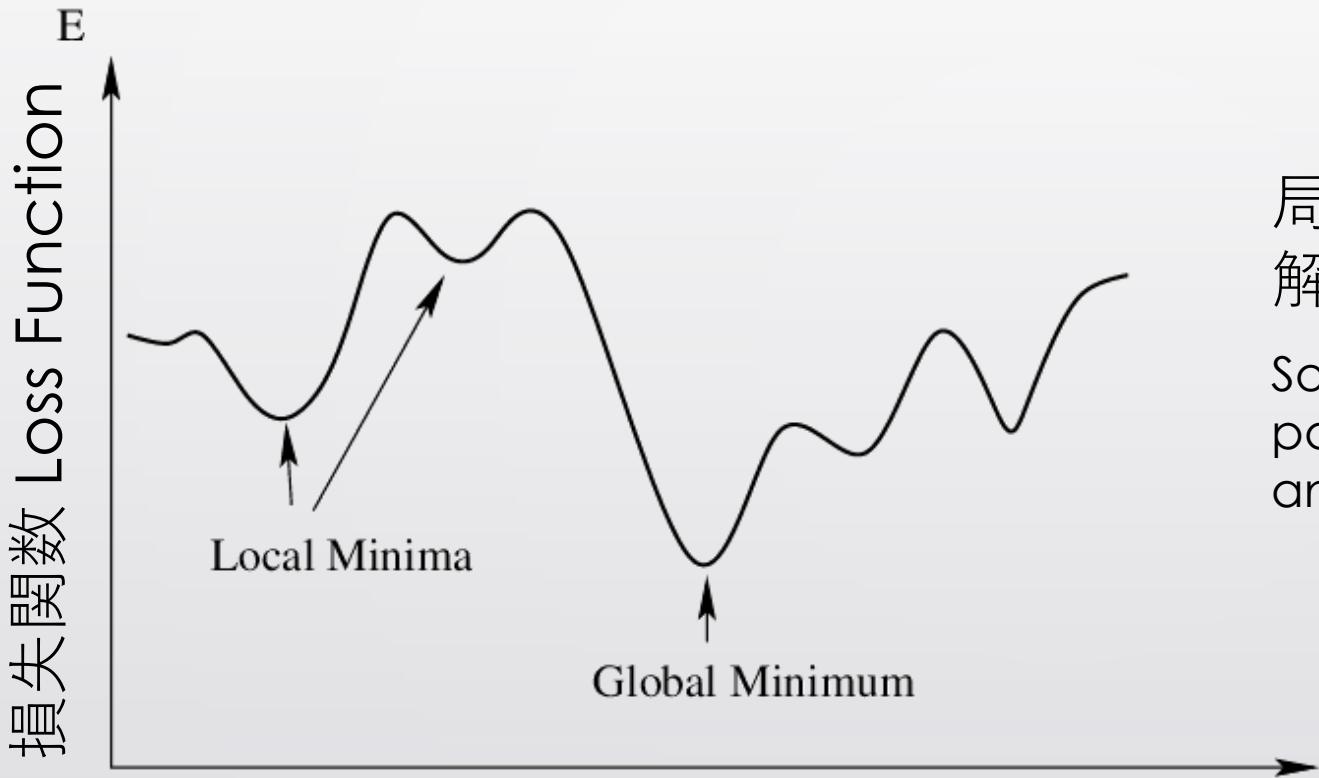
$$-\nabla J = \left(-\frac{\partial J}{\partial w}, -\frac{\partial J}{\partial b} \right)$$

$$w = w - \eta \frac{\partial J}{\partial w}$$

$-\nabla J$ の方向に変数を変化させることで、関数 J の値を最も素早く減少させることが出来る

The output value of function E decreases most rapidly along the direction of $-\nabla J$

局所最適と全体最適 Local and Global Minimum



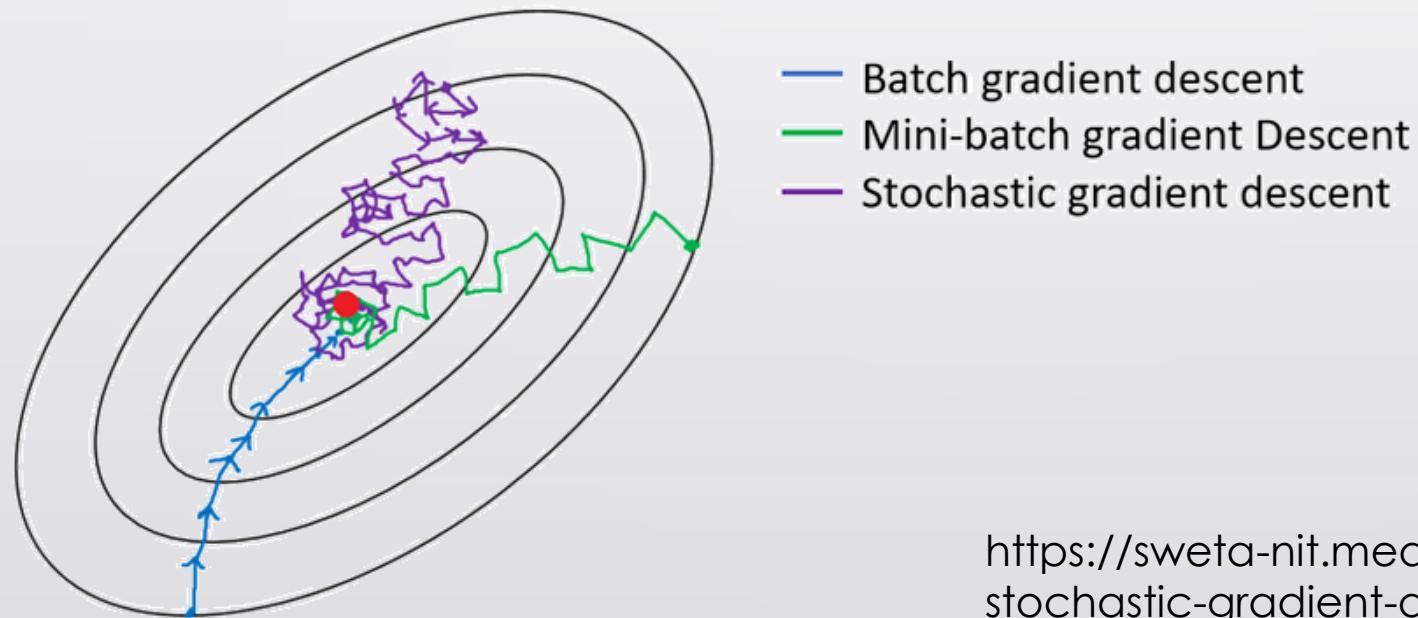
局所最適解にはまって、全体最適解にたどり着かない場合がある

Some times the search for the optimal parameters is caught by local minimum and fails to reach the global minimum

確率的最急降下法 Stochastic Gradient Descent (SGD)

ランダムに選択したデータセットによる重み更新の繰り返しで、局所最適解に陥るのを防ぐ

Avoid local minimum by weight updating using a set of randomly-selected data



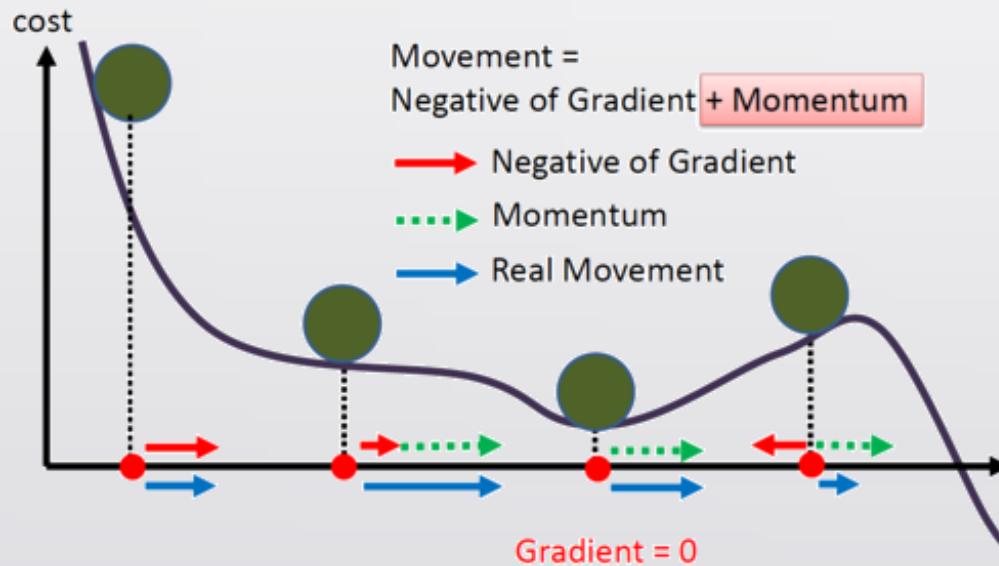
<https://sweta-nit.medium.com/batch-mini-batch-and-stochastic-gradient-descent-e9bc4cacd461>

モメンタム Momentum

$$w \leftarrow w - \eta \nabla E$$

$$w \leftarrow w + m$$

$$m \leftarrow \alpha m - \eta \nabla E$$



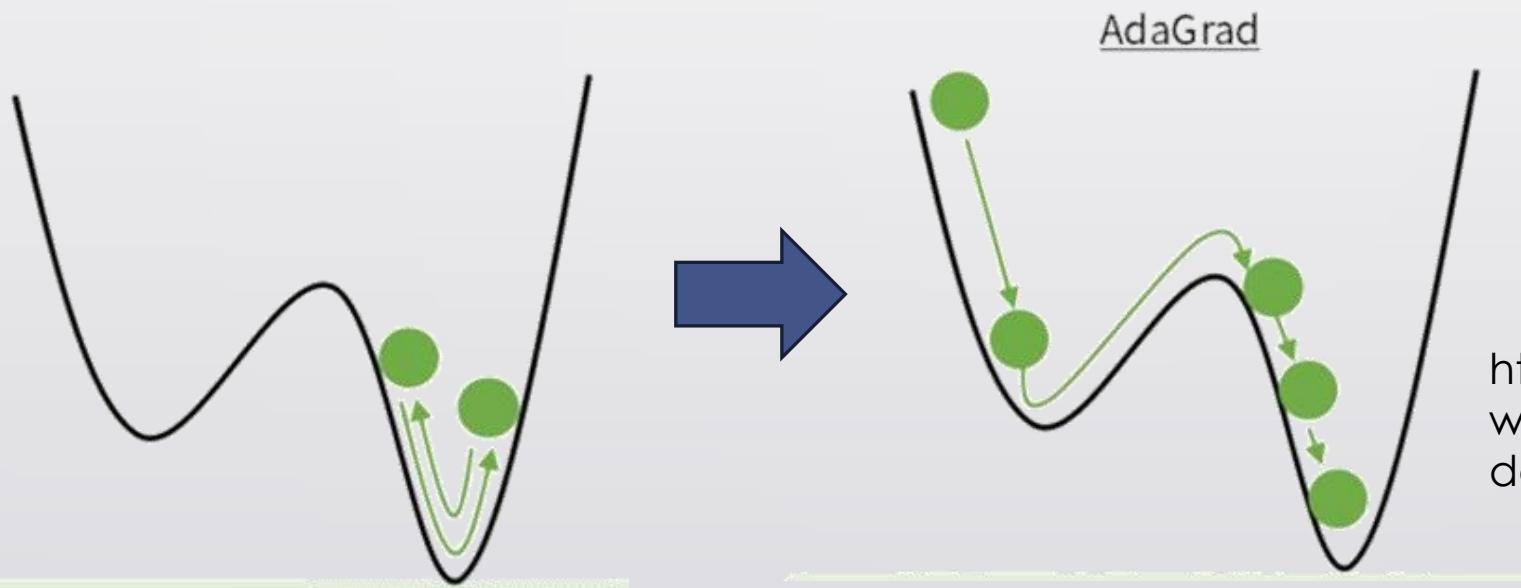
前の更新と同じ方向に勾配降下を行わせる慣性項

Momentum to induce gradient descent in the same direction as the preceding step

<https://medium.com/analytics-vidhya/momentum-rmsprop-and-adam-optimizer-5769721b4b19>

AdaGrad

$$h \leftarrow h + (\nabla E)^2 \quad w \leftarrow w - \frac{\eta}{\sqrt{h + \varepsilon}} \nabla E$$



ステップ毎に h が大きくなるので、学習率が減衰する

Learning rate decays as h increases at each step

<https://zero2one.jp/ai-word/problems-in-gradient-descent-methods/>

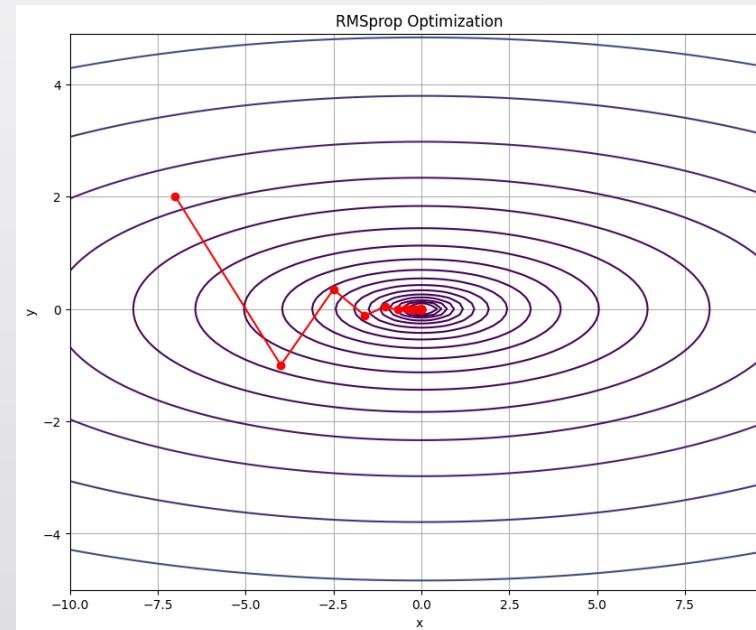
RMSProp

$$h \leftarrow \beta h + (1 - \beta)(\nabla E)^2$$

$$w \leftarrow w - \frac{\eta}{\sqrt{h + \varepsilon}} \nabla E$$

勾配の二乗の移動平均を取り、学習率の急激な変化を抑制

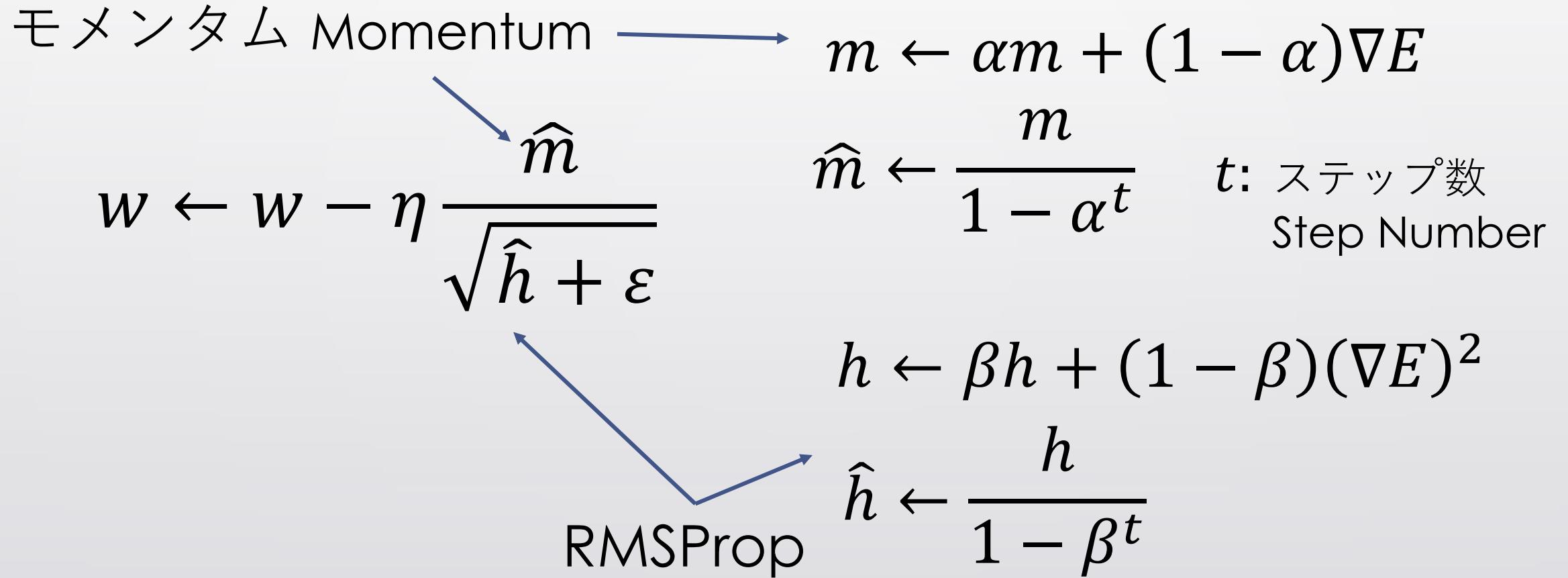
Avoid abrupt change of learning rate by computing moving average of squared gradient



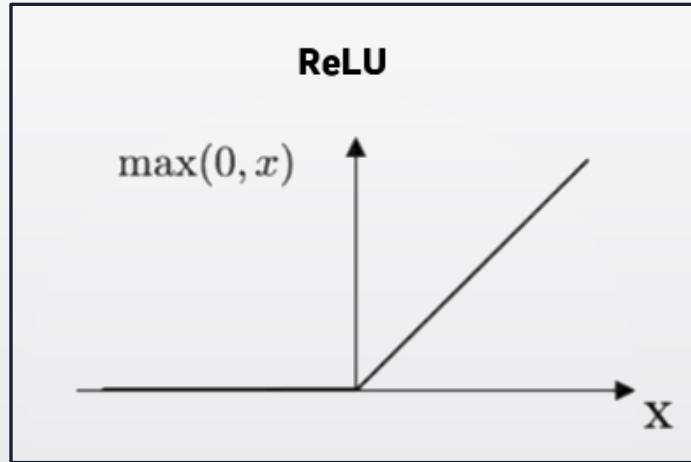
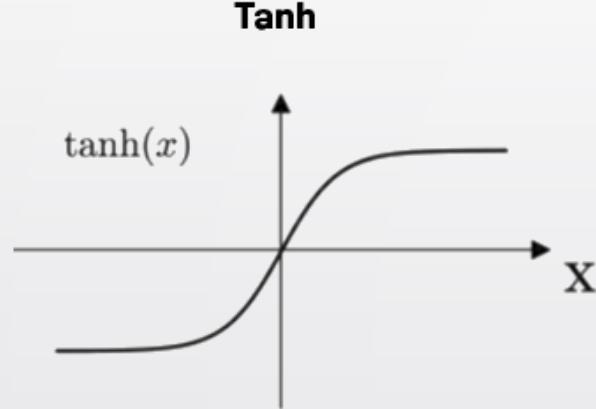
<https://tsukumochi.com/archives/8866>



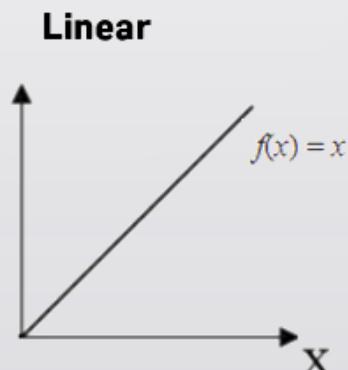
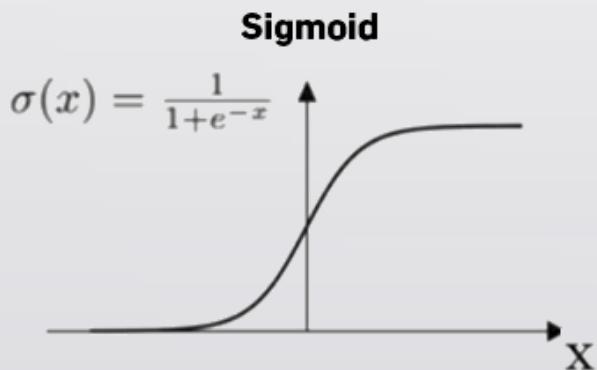
Adam (Adaptive Moment)



活性化関数 Activation Function



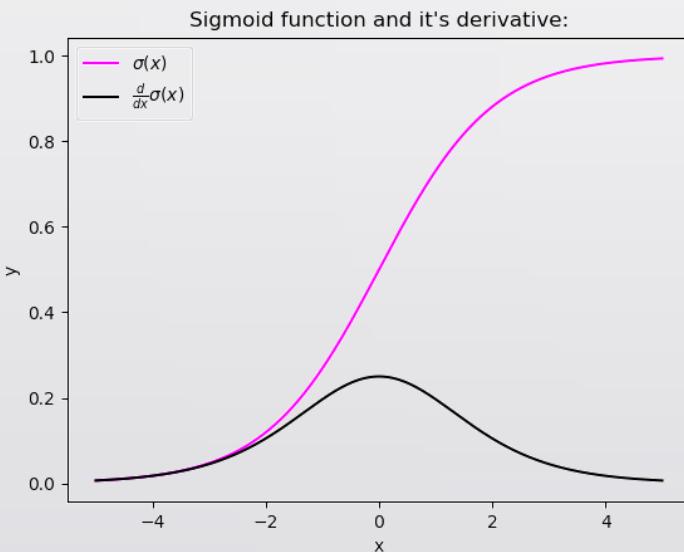
$$f(x) = \begin{cases} x & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$



<https://machine-learning.paperspace.com/wiki/activation-function>

勾配消失問題 Vanishing Gradient Problem

$$\frac{\partial v_k^{(2)}}{\partial w_{ji}^{(1)}} = x_i^n w_{kj}^{(2)} f'(v_j^{(1)})$$



各層への入力が 0 から遠いと、シグモイド関数の微分が 0 に漸近する

If input to each layer deviates from zero, derivative of sigmoid function gets close to zero



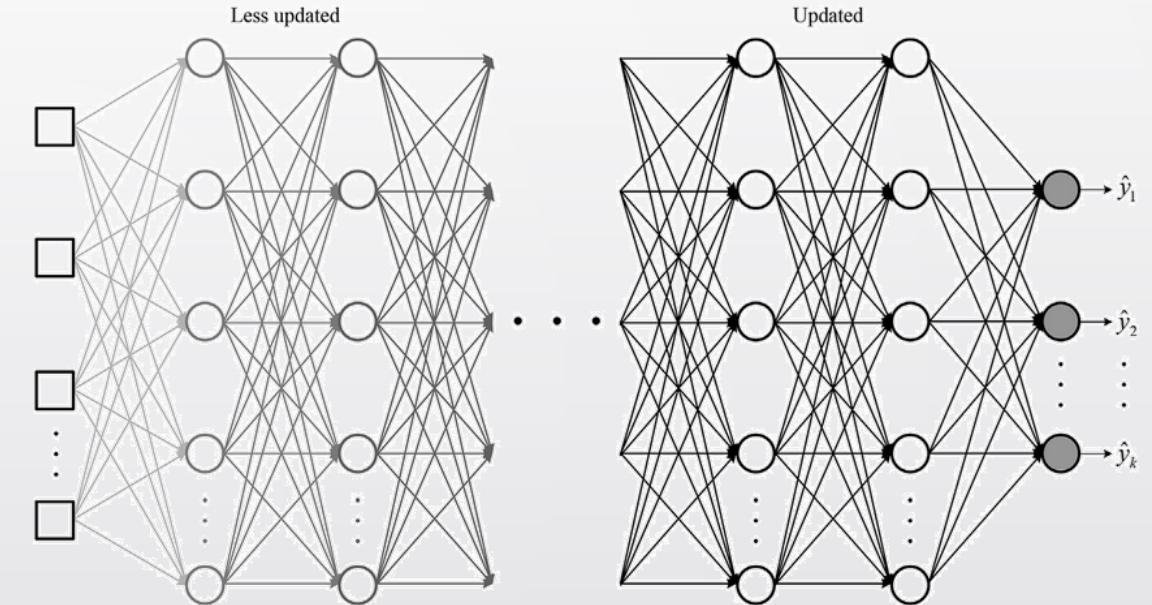
重みが更新されなくなる

Wight updating gets almost halted

勾配消失問題 Vanishing Gradient Problem

$$\delta_k^{(2)} = \left(y_k^{(2)} - t_k \right) f' \left(\sum_{j=0}^{j=M} w_{kj}^{(2)} y_j^{(1)} \right)$$

$$\delta_j^{(1)} = f' \left(v_j^{(1)} \right) \sum_{k=1}^{k=C} w_{kj}^{(2)} \cdot \delta_k^{(2)}$$



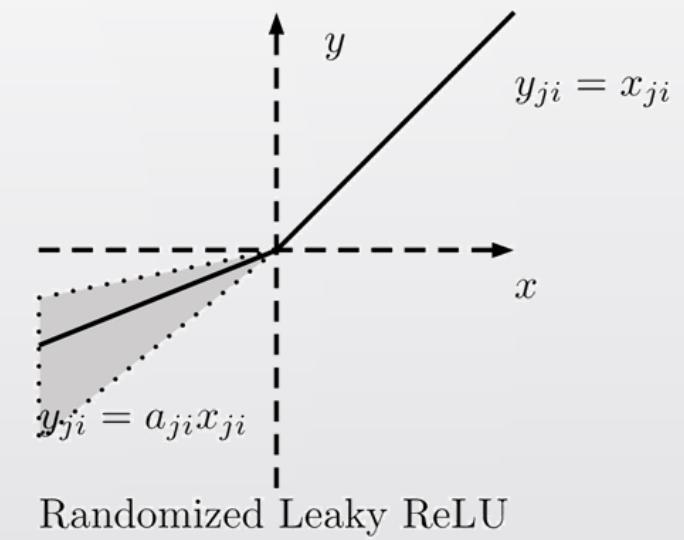
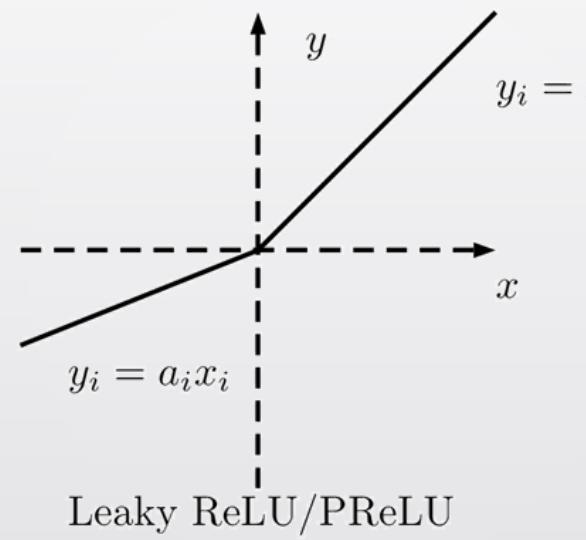
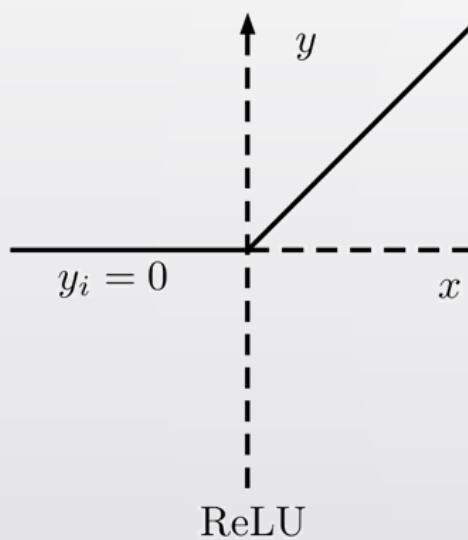
Koo et al, 2018

勾配消失問題は入力層に近い層でより深刻

Vanishing gradient problem is severer in layers closer to the input layer

ReLUとLeaky ReLU

Rectified Linear Unit (ReLU) and Leaky ReLU

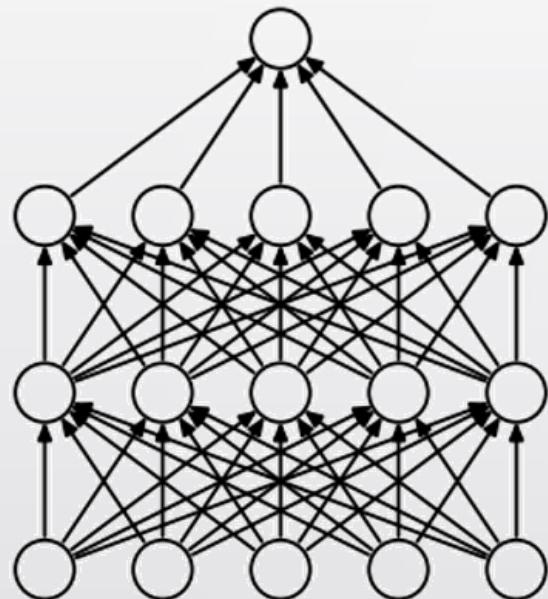


<https://paperswithcode.com/method/rrelu>

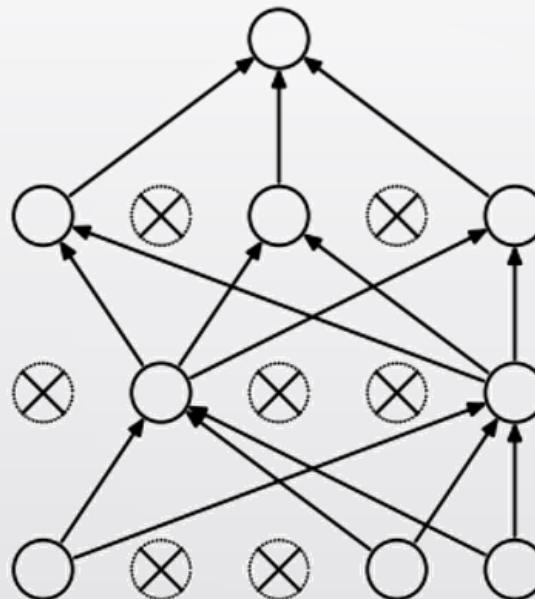
Leaky ReLUでは入力が 0 以下でも勾配が発生する

In contrast to ReLU, Leaky ReLU retains non-zero gradient for input below zero

ドロップアウト Dropout



(a) Standard Neural Net



(b) After applying dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting", JMLR 2014

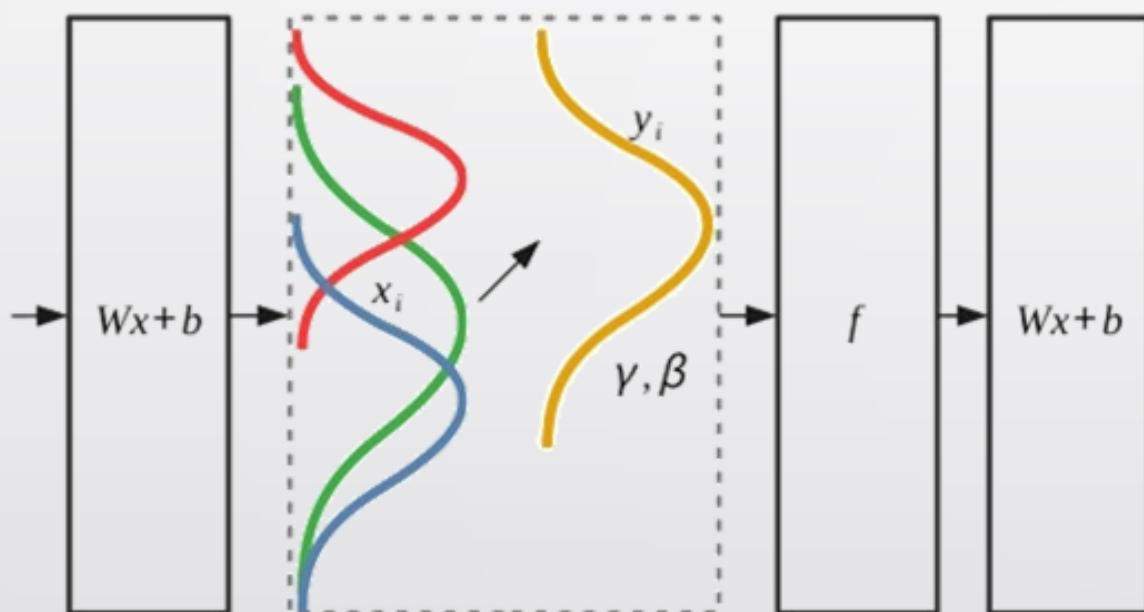
ミニバッチごとにネットワークのノードをランダムに取り除く

Remove randomly-selected nodes from network for each mini-batch

過学習を抑制する効果がある
Effective in suppressing overfitting

バッチ正規化 Batch Normalization

Ensure the output statistics of a layer are fixed.



バッチ毎に各層への入力を正規化する
Normalize inputs to layers within each single batch

$$y = \gamma \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

学習の安定化、過学習の抑制、勾配消失問題の緩和に効果がある

Effective in stabilization of learning process,
suppression of overfitting and vanishing gradient
problem

<https://www.srose.biz/wp-content/uploads/2020/08/Deep-Learning-Performance-Part-3-Batch-Normalization-Dropout-Noise.html>



データマイニング

Data Mining

15:ニューラルネットワーク③ Neural Network

講義のまとめ Conclusion

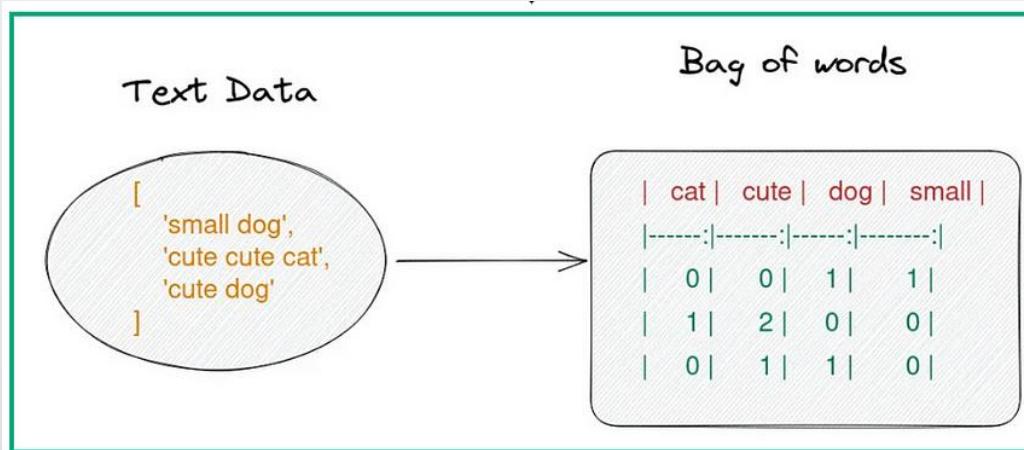
土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

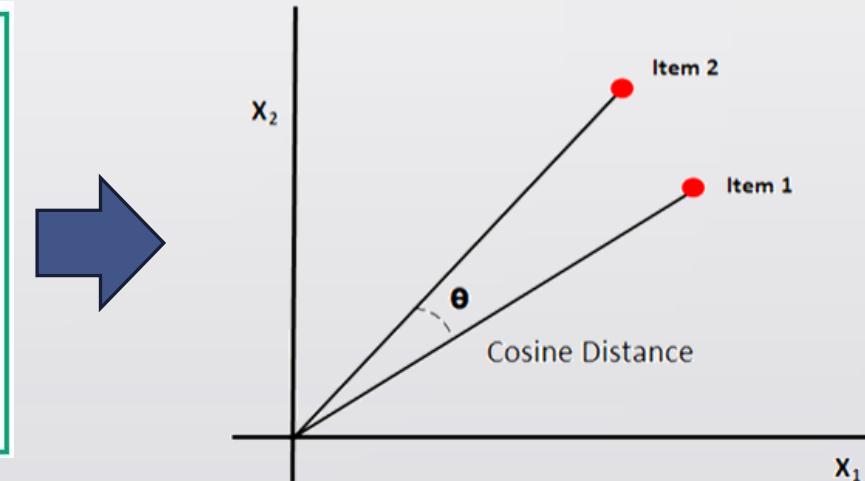
Bag of Words

形態素解析をした後、各単語の出現頻度によりテキストの特徴量ベクトルを生成する

Generate feature vector of a text by counting frequency of each morpheme after morphological analysis



<https://ayselaydin.medium.com/4-bag-of-words-model-in-nlp-434cb38cdd1b>



Term Frequency-Inverse Document Frequency (TF-IDF)

ある文書における特定の単語の重要度を評価する
Measure of importance of a certain word in document

$$TF - IDF(t, d) = TF(t, d) * IDF(t, d)$$

$$TF(t, d) = \frac{\text{number of } t \text{ in document } d}{\text{total number of words in document } d}$$

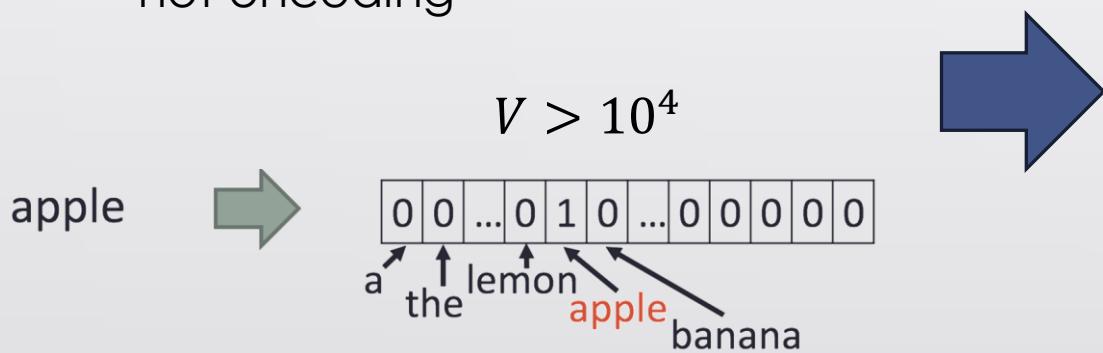
$$IDF(t, d) = \log \left(\frac{N}{1 + df} \right) \quad \begin{aligned} N &: \text{Total number of documents} \\ df &: \text{Number of documents containing word } t \end{aligned}$$

分布仮説と分散表現

Distributional Hypothesis and Distributed Representation

One-Hot Encodingによる
単語表現

Representation of words by one-hot encoding



<https://qiita.com/kouhara/items/e895f6350aa1ebe77133>

分布仮説に基づく低次元ベクトルでの単語の意味表現

Representation of word meaning by low dimensional vector based on "Distributional Hypothesis"

$$\text{apple} = [\text{○} \text{●} \dots \text{○}]$$

※分布仮説 Distributional Hypothesis

類似の文脈に登場する単語は似た意味を持つ

Words occurring in similar contexts have similar meanings

分散表現の獲得 Acquisition of Distributed Representation



潜在意味解析 Latent Semantic Analysis

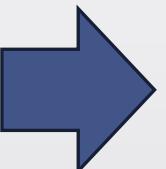
共起行列 Co-occurrence Matrix

2-sized window for **cat**

... I saw a **cute** **grey** **cat** **playing** **in** the garden ...

contexts for **cat**

	bird	sitting	wall	cat	fence
bird	1	1	1	0	0
sitting	1	1	1	1	0
wall	1	1	1	0	0
cat	0	1	0	1	1
fence	0	1	0	1	1



相互情報量行列

Point-wise Mutual Information(PMI) Matrix

$$PMI(w_1, w_2) = \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

$P(w_i)$: 文書に単語 w_i が登場する確率

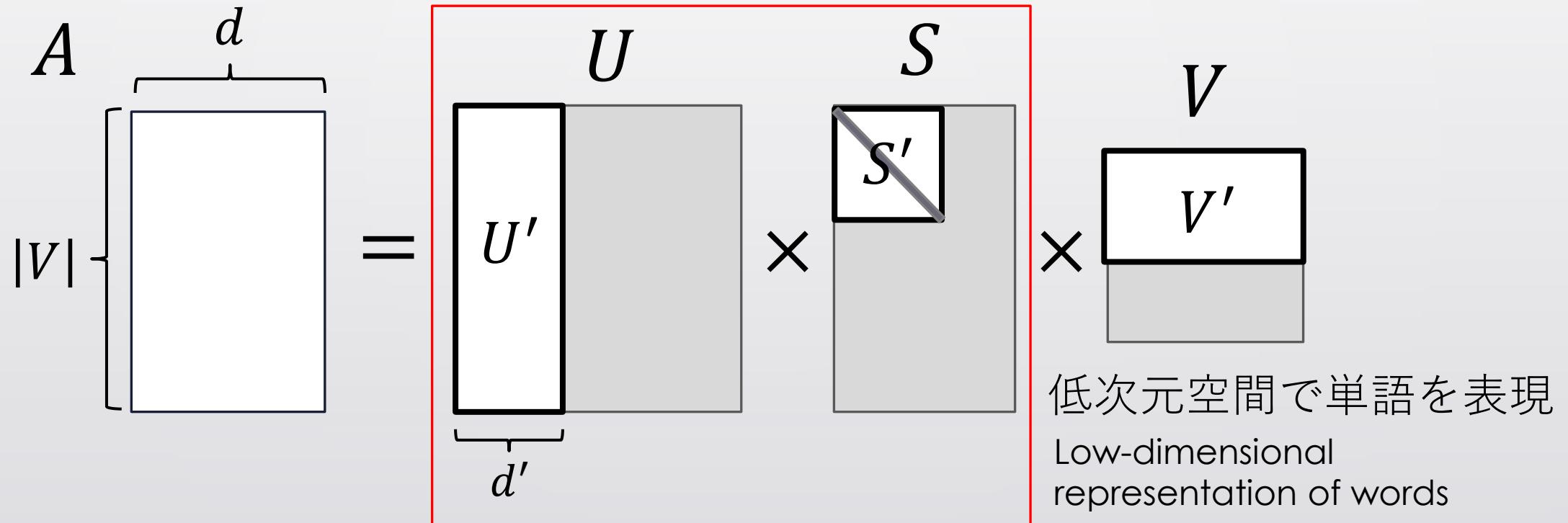
$P(w_i, w_j)$: 文書に単語 w_i と w_j ($i \neq j$)が同時に登場する確率

<https://medium.com/@imamitsehgal/nlp-series-distributional-semantics-co-occurrence-matrix-31283629951e>

潜在意味解析 Latent Semantic Analysis

PMI行列を特異値分解で次元削減

Dimension reduction by singular value decomposition of PMI matrix



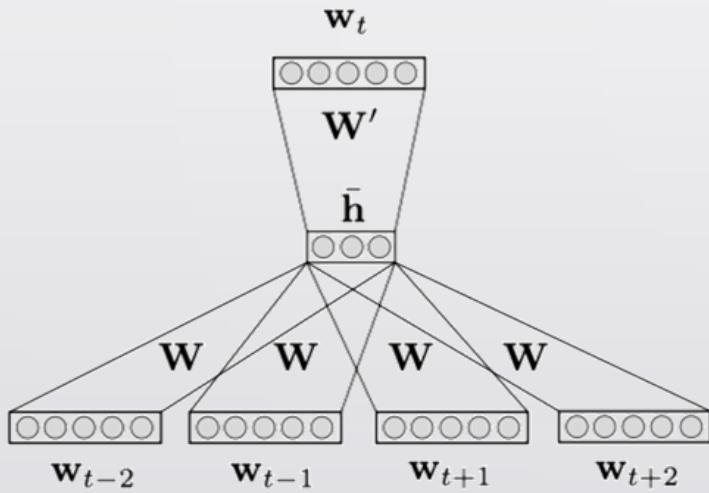


Word2vec

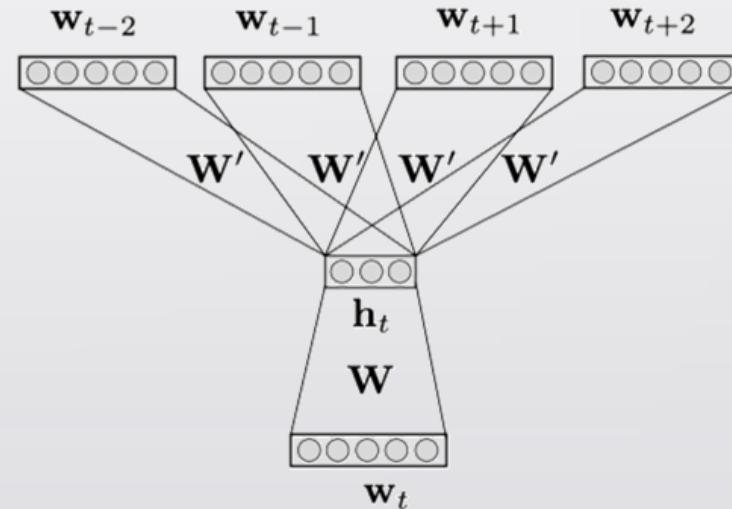
分布仮説に基づき単語埋め込みを行うニューラルネットワークモデル

Neural network models that conduct word embedding based on distributional hypothesis

CBOW



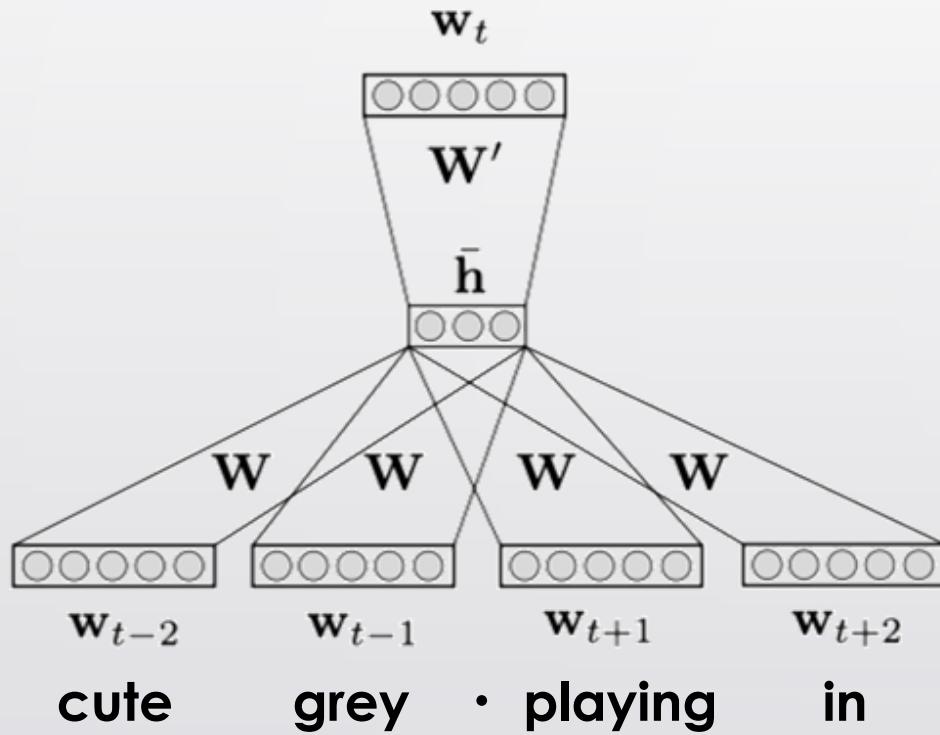
Skip-gram



<https://cvml-expertguide.net/terms/nlp/word2vec/>

Continuous Bag of Words (CBOW)

“**cat**”を予測 Predict the word “cat”



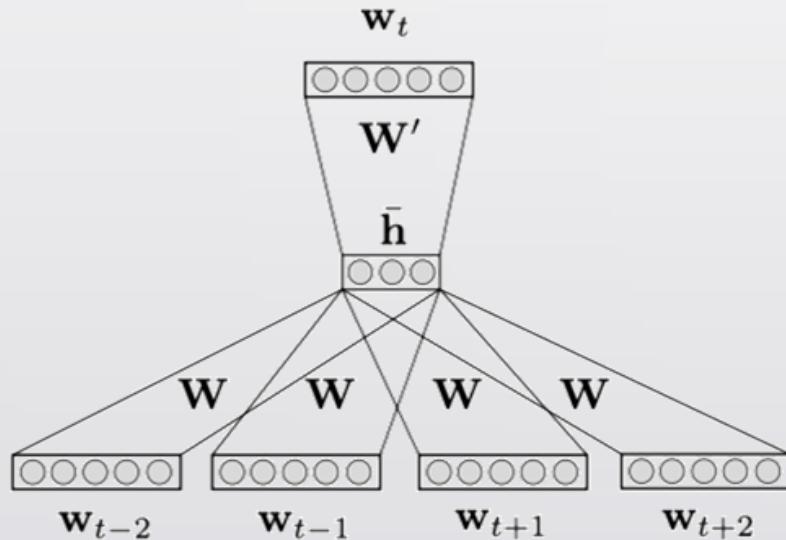
One-hot Encodingされた文脈単語の入力

前後の単語から中心にくる単語を推測する
Train the network so that it can predict a target word flanked by context words preceding or following the target word

Continuous Bag of Words (CBOW)

重み行列 W に単語の意味が低次元で表現される

Low-dimensional representation of word meaning is stored in weight matrix W



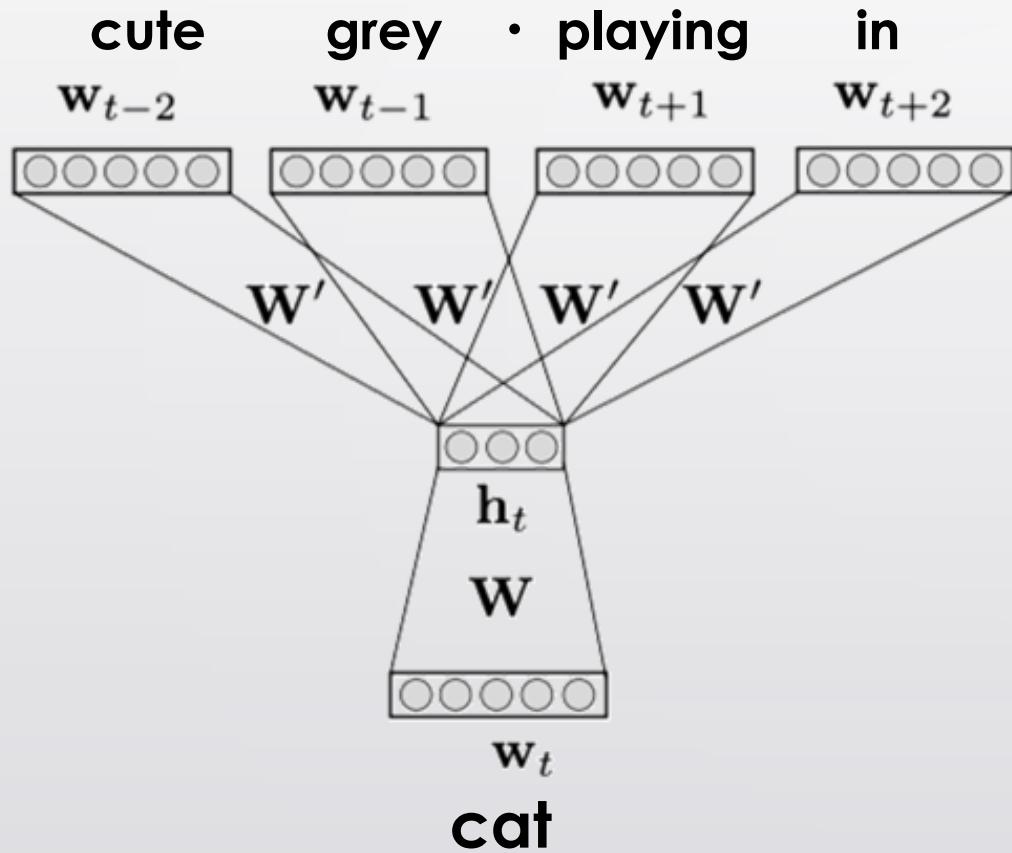
$$\begin{array}{ll} w_{t-2} & \boxed{0, 0, 0, 1, 0, \dots, 0, 0, 0, 0} \\ w_{t-1} & \boxed{0, 0, 0, 0, 0, \dots, 1, 0, 0, 0} \\ w_{t+1} & \boxed{1, 0, 0, 0, 0, \dots, 0, 0, 0, 0} \\ w_{t+2} & \boxed{0, 0, 1, 0, 0, \dots, 0, 0, 0, 0} \end{array}$$

$(1, |V|)$

$$\begin{matrix} W \\ \bullet \\ \begin{matrix} h \\ = \\ \boxed{} \\ (1, 3) \end{matrix} \end{matrix}$$

$(|V|, 3)$

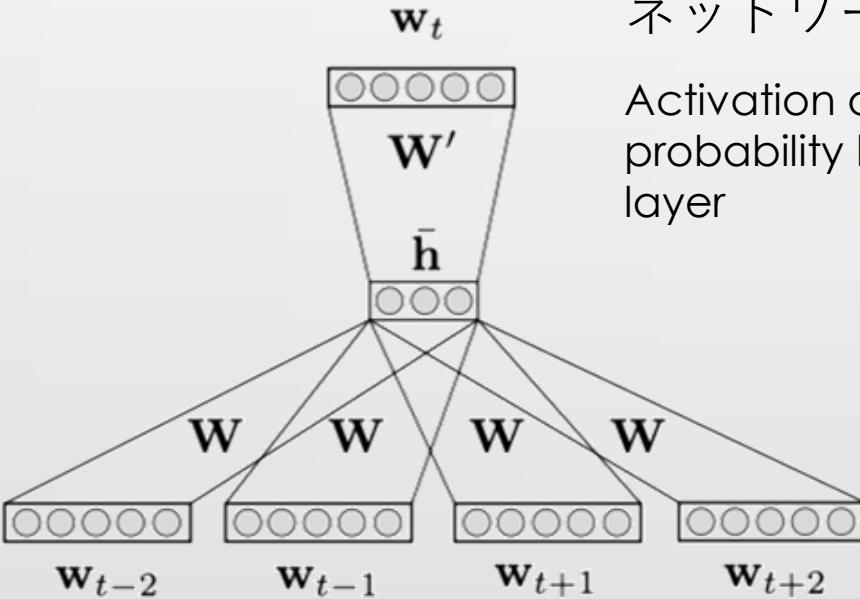
Skip-gram



CBOWとは逆に、入力後の前後の文脈単語を予測する

In contrast to CBOW, a network is trained so that it predicts context words flanking the input word

Word2Vecの損失関数 Loss Function of Word2Vec



出力層ではソフトマックス関数により、
ネットワークの活性が確率に変換される

Activation of network is converted into
probability by softmax function in the output
layer

$$P(w_k) = \frac{\exp(y(w_k))}{\sum_{|V|} \exp(y(w_i))}$$

損失関数は交差エントロピー
Loss function is cross entropy

$$L = - \sum_{i=1}^{|V|} t_i \log [P(w_i)] \quad \begin{aligned} t &= (t_1, t_2 \dots t_{|V|}) \\ t_k &\in \{0, 1\} \end{aligned}$$

負例サンプリング Negative Sampling

コーパスデータを用いた学習では $|V|$ が巨大な数値になる

The value of $|V|$ is huge in training based on corpus data

$$P(w_k) = \frac{\exp(y(w_k))}{\sum_{|V|} \exp(y(w_i))} \quad L = - \sum_{i=1}^{|V|} t_i \log[P(w_i)]$$

二値分類問題に置き換えることで計算を高速化

Accelerate computation by replacing the multiclass classification with binary classification



データマイニングの流れ Steps in Data Mining

1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis・モデリング Modeling
6. 性能評価 Performance Evaluation
7. (ディプロイメント Deployment)

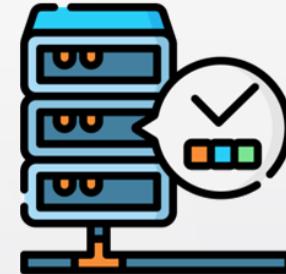
データ収集 Data collection



Confidentiality



Integrity



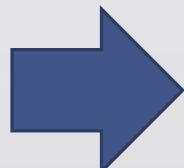
Availability

機密性

完全性

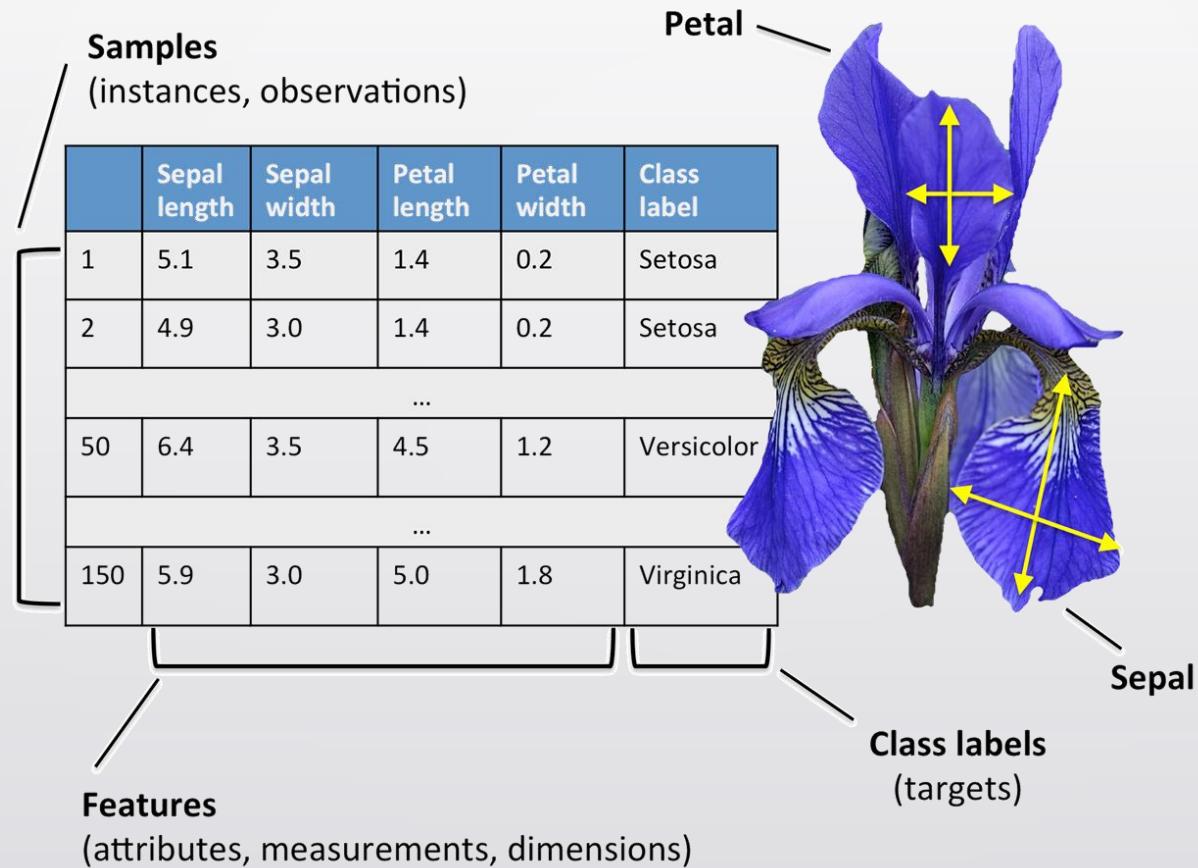
可用性

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円



性別	年齢	年収	
男	[20-29]	[300-499]	万円
女	[30-39]	[500-699]	万円
女	[30-39]	[500-699]	万円
男	[20-29]	[300-499]	万円

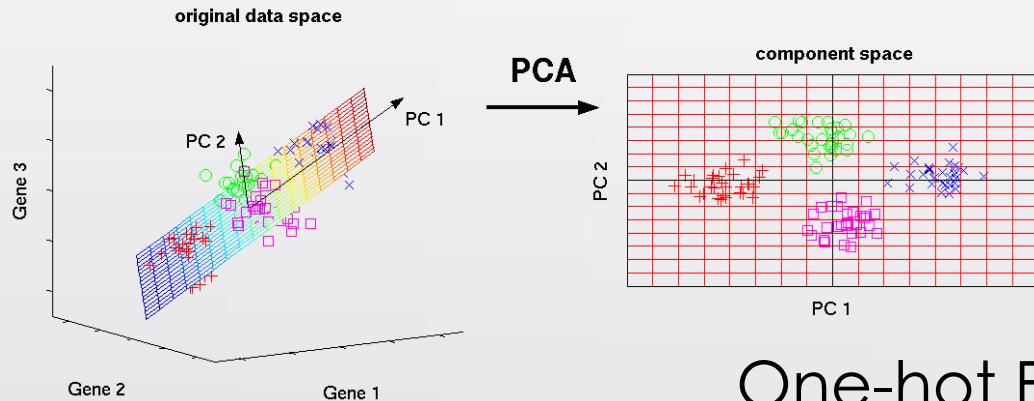
特徴量選択 Feature Selection





前処理 Pre-processing

次元削減 Dimension Reduction



スケーリング Scaling

$$x' = \frac{x - \mu}{\sigma}$$

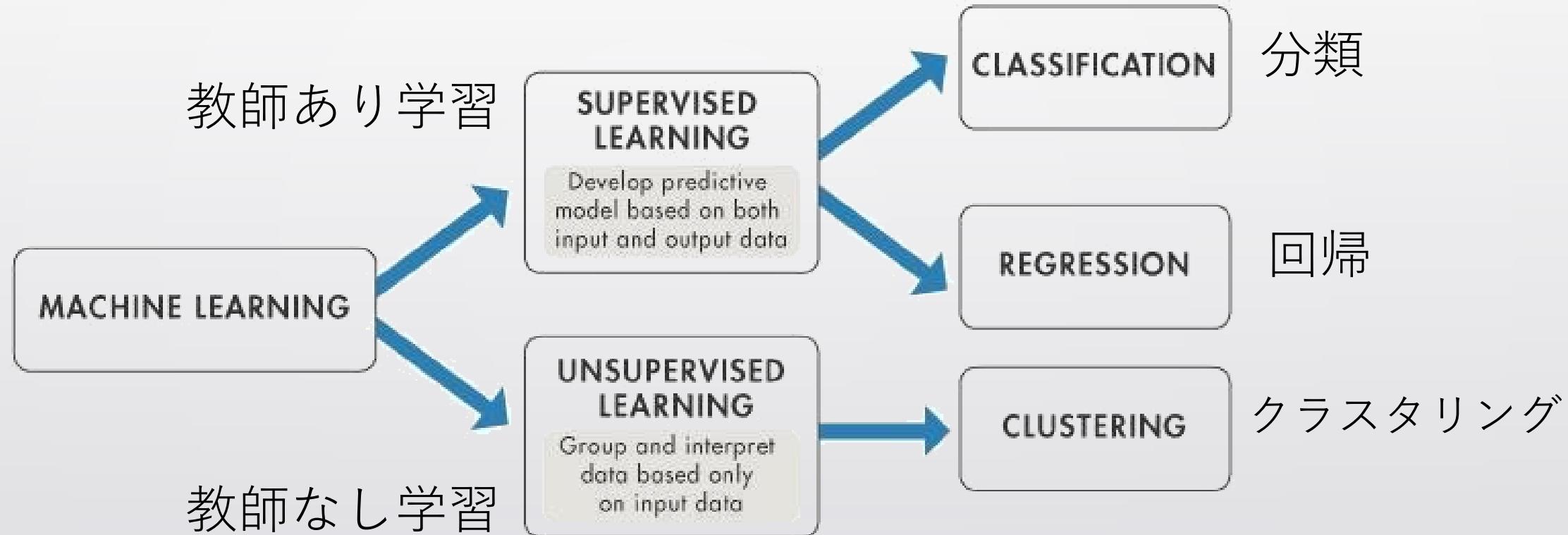
$$x' = \frac{x - median}{NIQR}$$

One-hot Encoding

The diagram illustrates the process of One-hot Encoding. On the left, a table shows a list of colors: Red, Red, Yellow, Green, and Yellow. A yellow arrow points from this table to the right, where another table shows the corresponding one-hot encoding. The second table has columns for 'Red', 'Yellow', and 'Green'. The rows show binary values (0 or 1) indicating the presence of each color: Red has a 1 in the first column and 0s in the others; Yellow has a 1 in the second column and 0s in the others; Green has a 1 in the third column and 0s in the others; and the two Yellow entries also have 0s in the third column.

Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

データ分析・モデリング



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

多項式回帰 Polynomial Regression

Simple
Linear
Regression

$$y = b_0 + b_1 x_1$$

Multiple
Linear
Regression

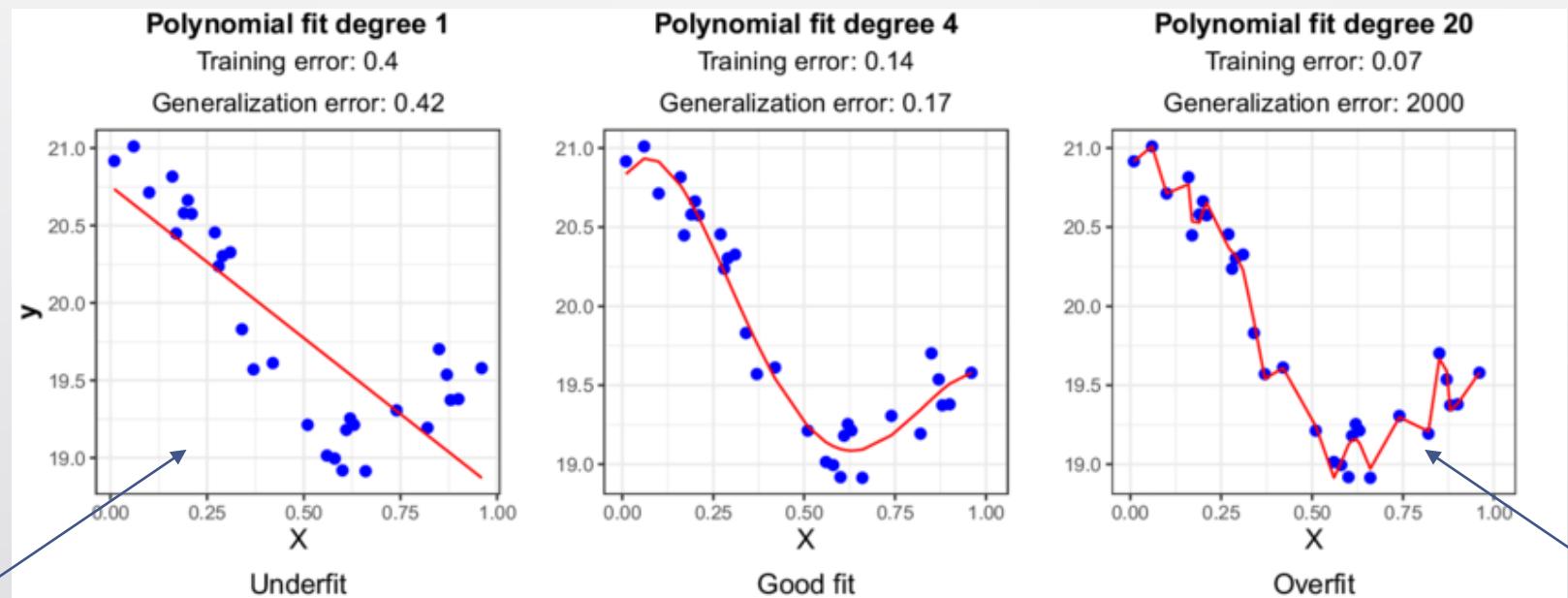
$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

Polynomial
Linear
Regression

$$y = b_0 + b_1 x_1 + \underline{b_2 x_1^2} + \dots + b_n x_1^n$$

性能評価 Performance Evaluation

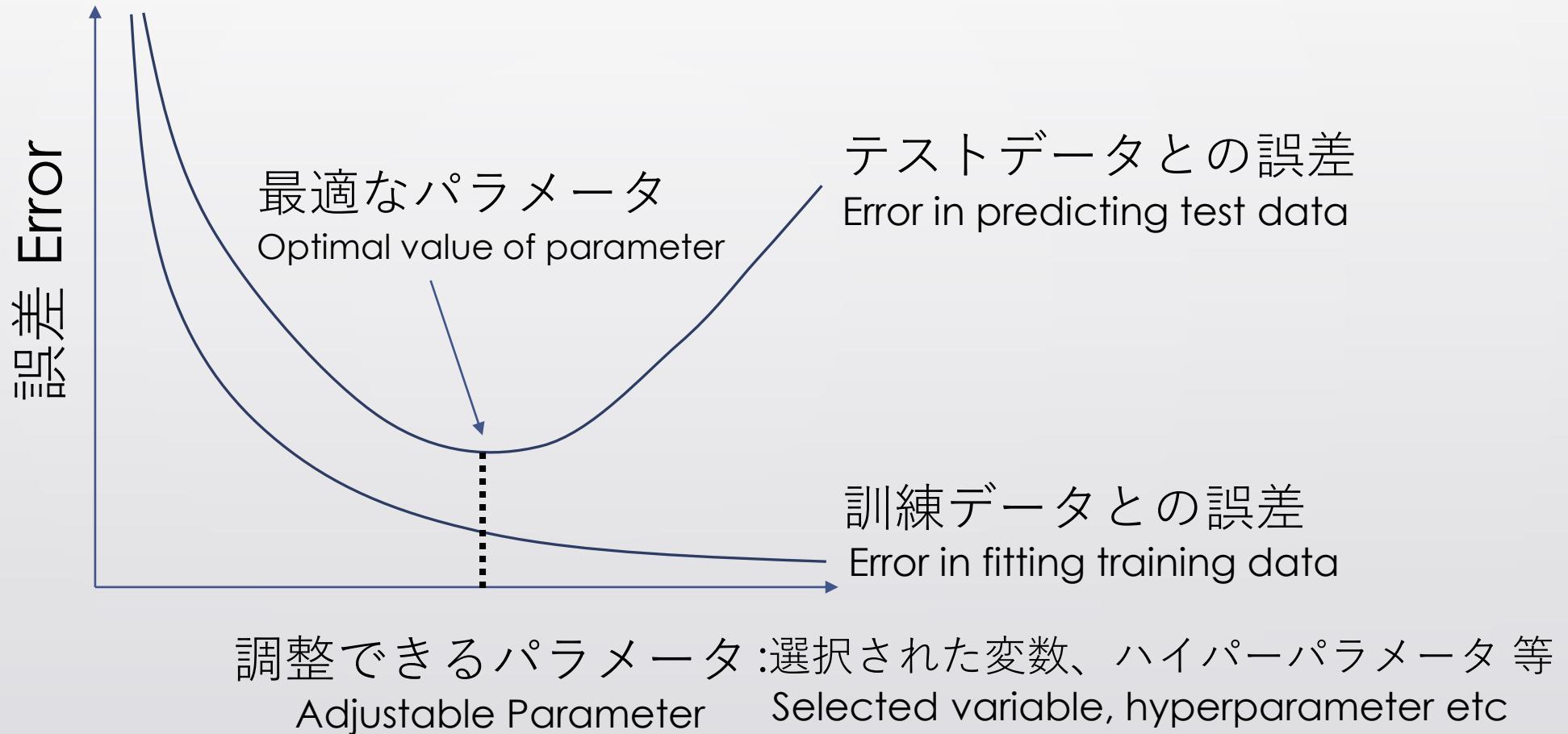
観測されたデータはノイズを含む Observed data contains random noise



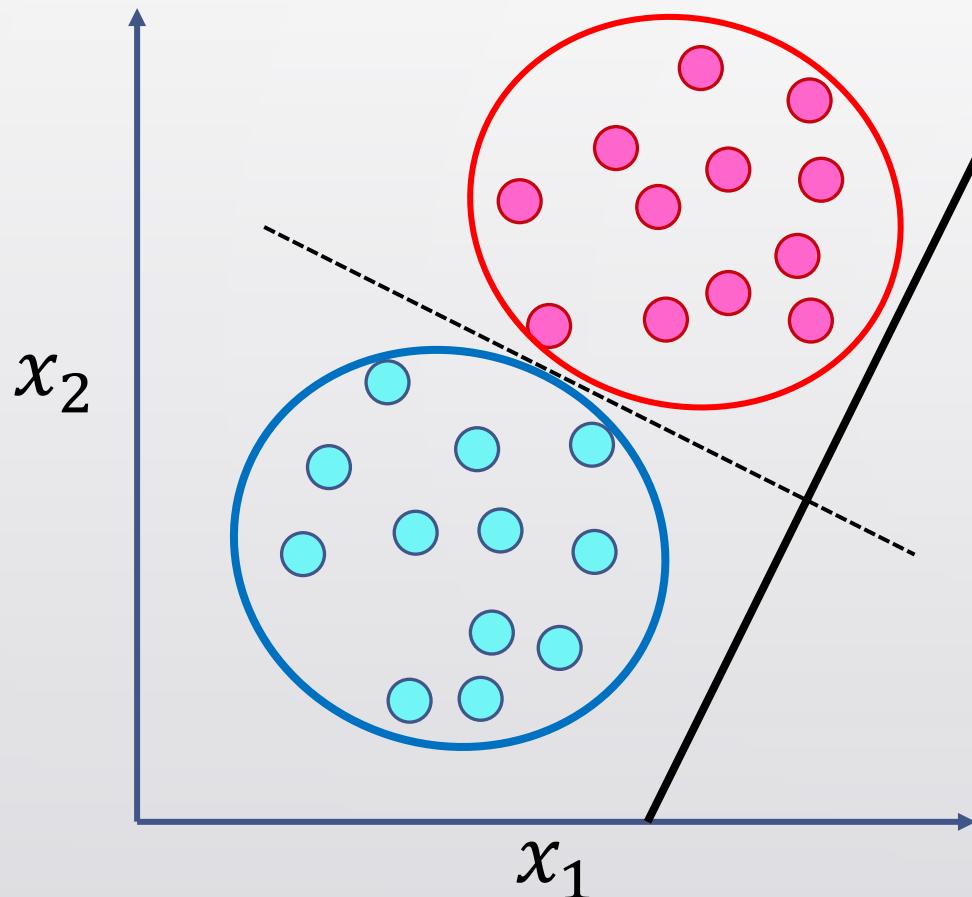
単純すぎてはダメ
Should not be too simplistic

複雑すぎてはダメ
Should not be too complex

性能評価 Performance Evaluation



線型判別分析 Linear Discriminant Analysis (LDA)



よい決定境界は、下の二つの条件を満たす

A good decision boundary meets the two conditions below

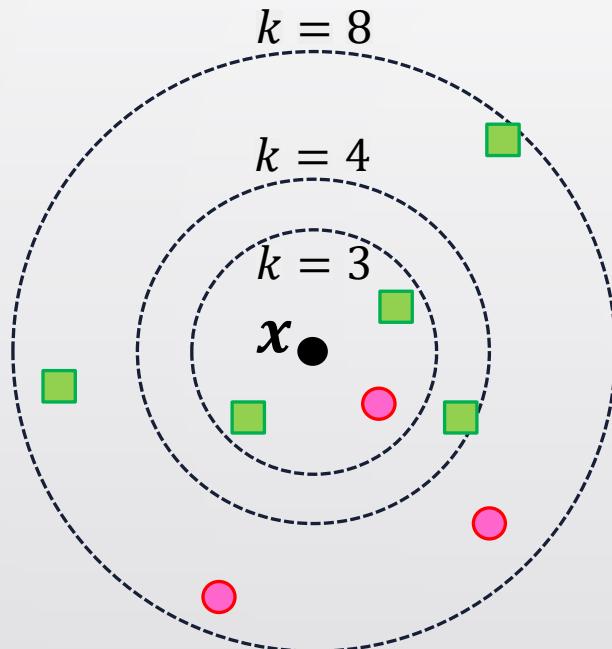
1. 2 クラスの中心が離れている

Centers of the two classes are distant from each other

2. 各クラスのクラス内分散が小さい

Within-class variance of each class is small

k 最近傍法 k Nearest Neighbor Method



データ x のクラスを最近傍にある k 個のデータの多数決投票により決定する

Class of data x is determined by majority voting of k data points closest to x

ナイーブベイズ Naïve Bayes

メールに“秘密”“技術”“大当たり”という3つの単語が含まれていた。

An e-mail contains three words, “Secret”, “Technology” and “Jackpot”

$$\frac{P_3(H_s|W_3)}{P_3(H_a|W_3)} = \frac{P(W_3|H_s)P(W_2|H_s)P(W_1|H_s)P_1(H_s)}{P(W_3|H_a)P(W_2|H_a)P(W_1|H_a)P_1(H_a)} = \frac{628 \times 10^{-4}}{128 \times 10^{-4}}$$

$$Probability\ of\ being\ a\ spam = \frac{628}{628 + 128} = 0.83$$

スパムメールと判定するかどうかは閾値による

It depends on the threshold whether the e-mail is judged to be a spam or not

閾値と偽陽性 Threshold and False Positives

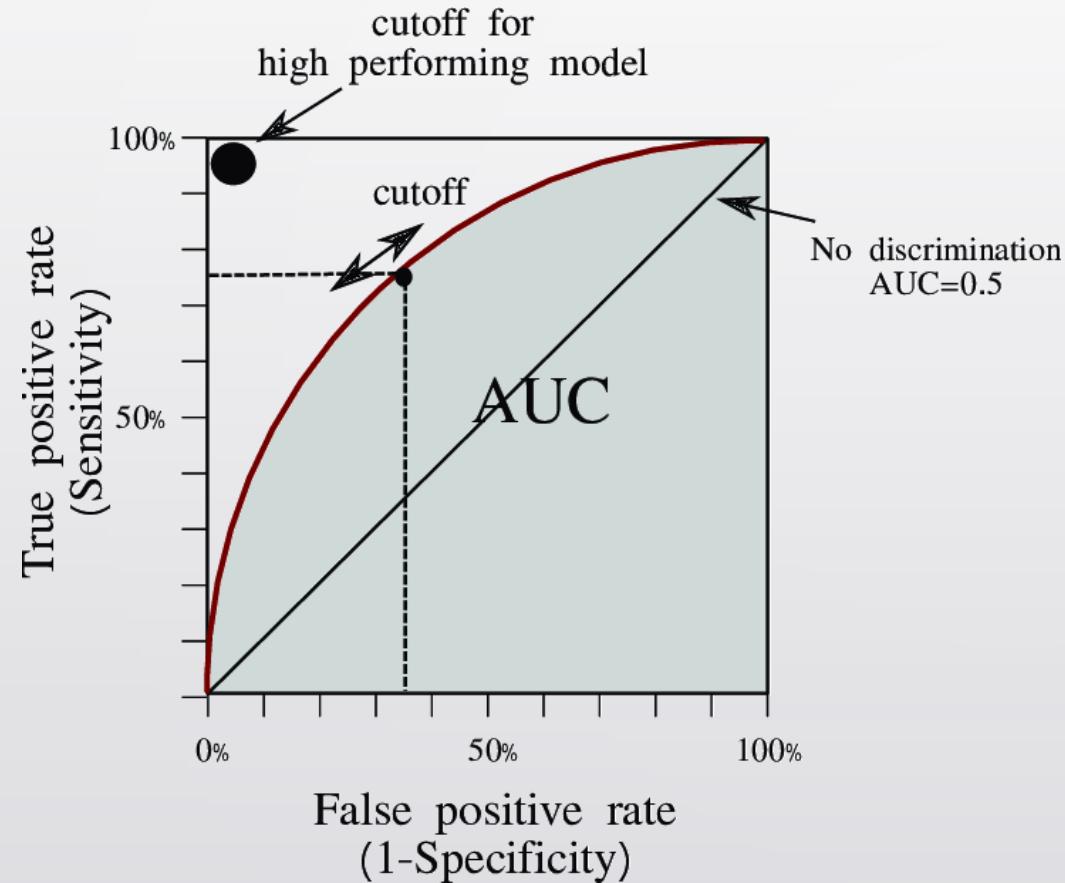
正解 Answer

判定
Judgment

	スパムメール Spam Mail	普通のメール Authentic Mail
スパムメール Spam Mail	真陽性 True Positive	偽陽性 False Positive
普通のメール Authentic Mail	偽陰性 False Negative	真陰性 True Negative

スパムと判定する閾値を下げる → 偽陽性率が上がる
Lowering threshold for judging to be a spam Higher false positive rate

ROC曲線 ROC(Receiver-Operator Characteristics) Curve



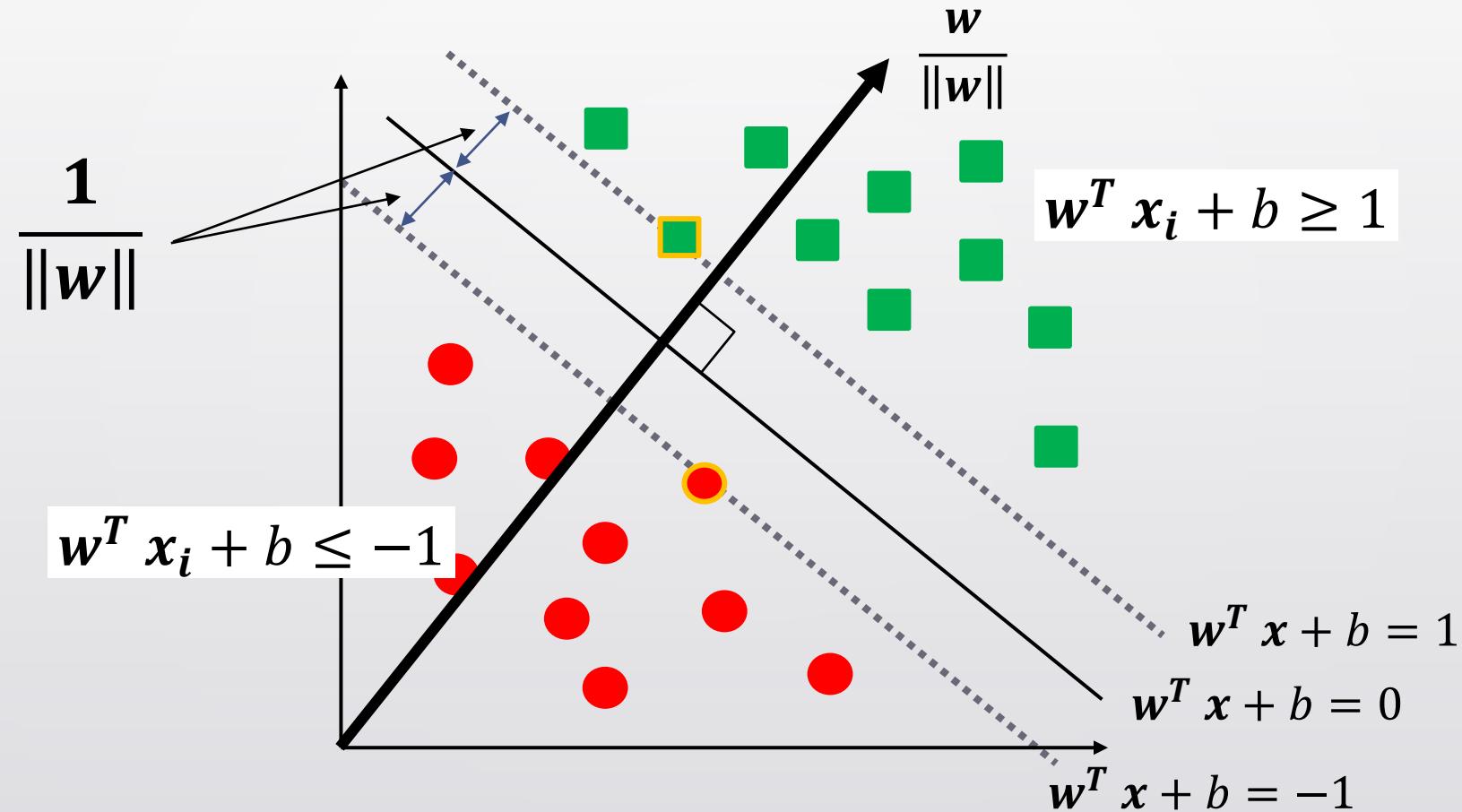
AUC: Area Under Curve

AUCが大きいほど、分類器の性能が良い

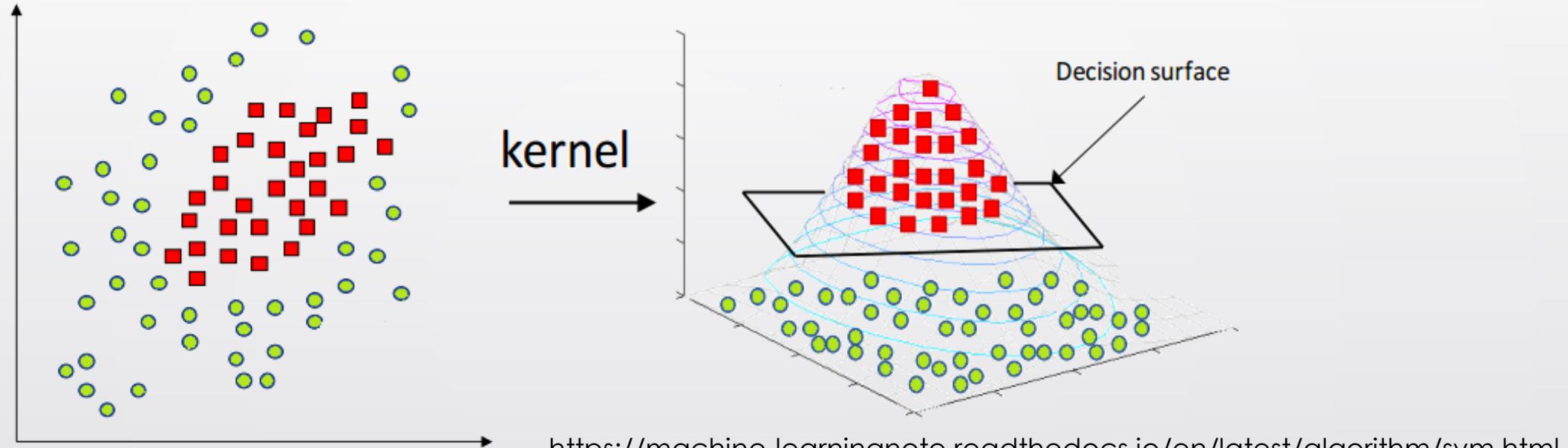
Larger AUC indicates better performance of classifier

AUC	
0.9 - 1.0	High accuracy
0.9 - 0.7	Moderate accuracy
0.5 - 0.7	Low accuracy

サポートベクターマシン Support Vector Machine



カーネル法 Kernel Methods

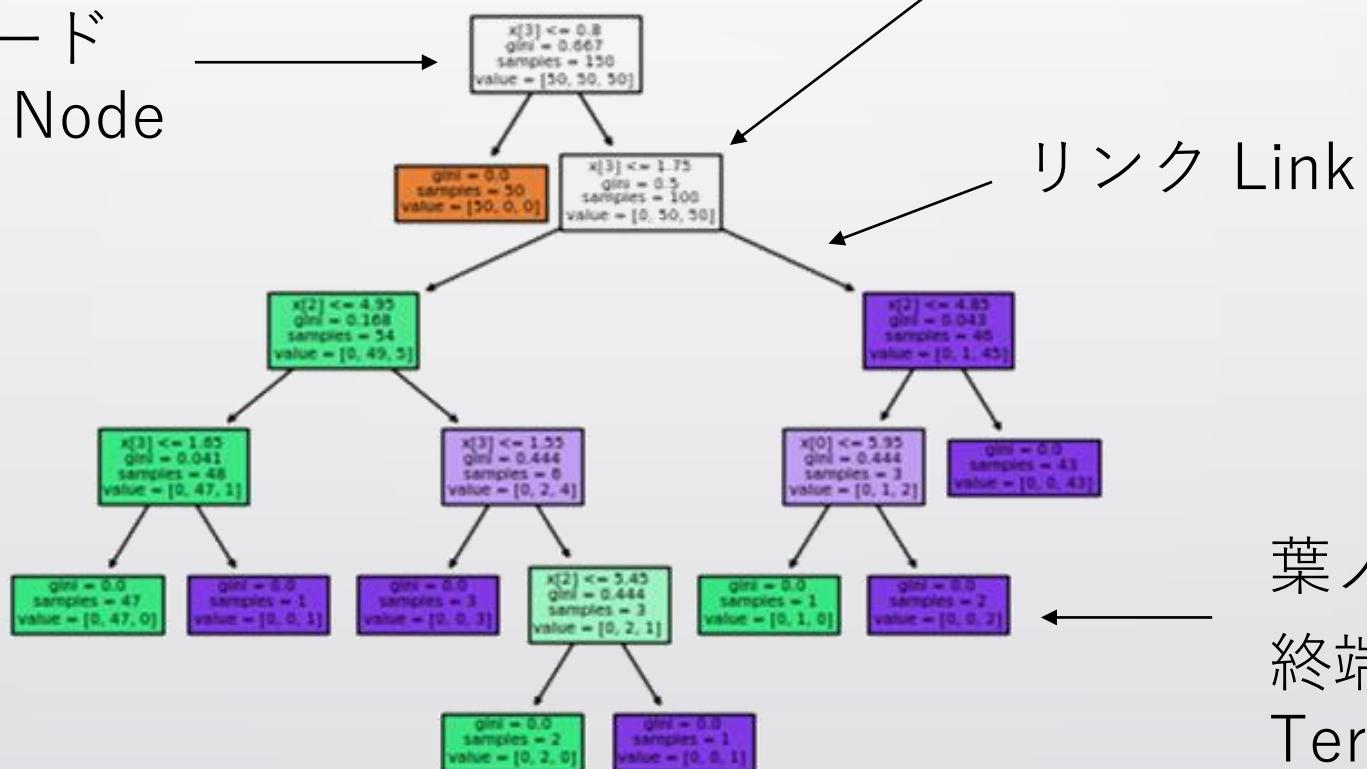


データを高次元空間に写像することで、線型分離不可能な問題を線型分離可能にする

Transforming linearly inseparable problem to linearly separable one by mapping data to higher-dimensional space

決定木 Decision Tree

根ノード
Root Node



ノード Node

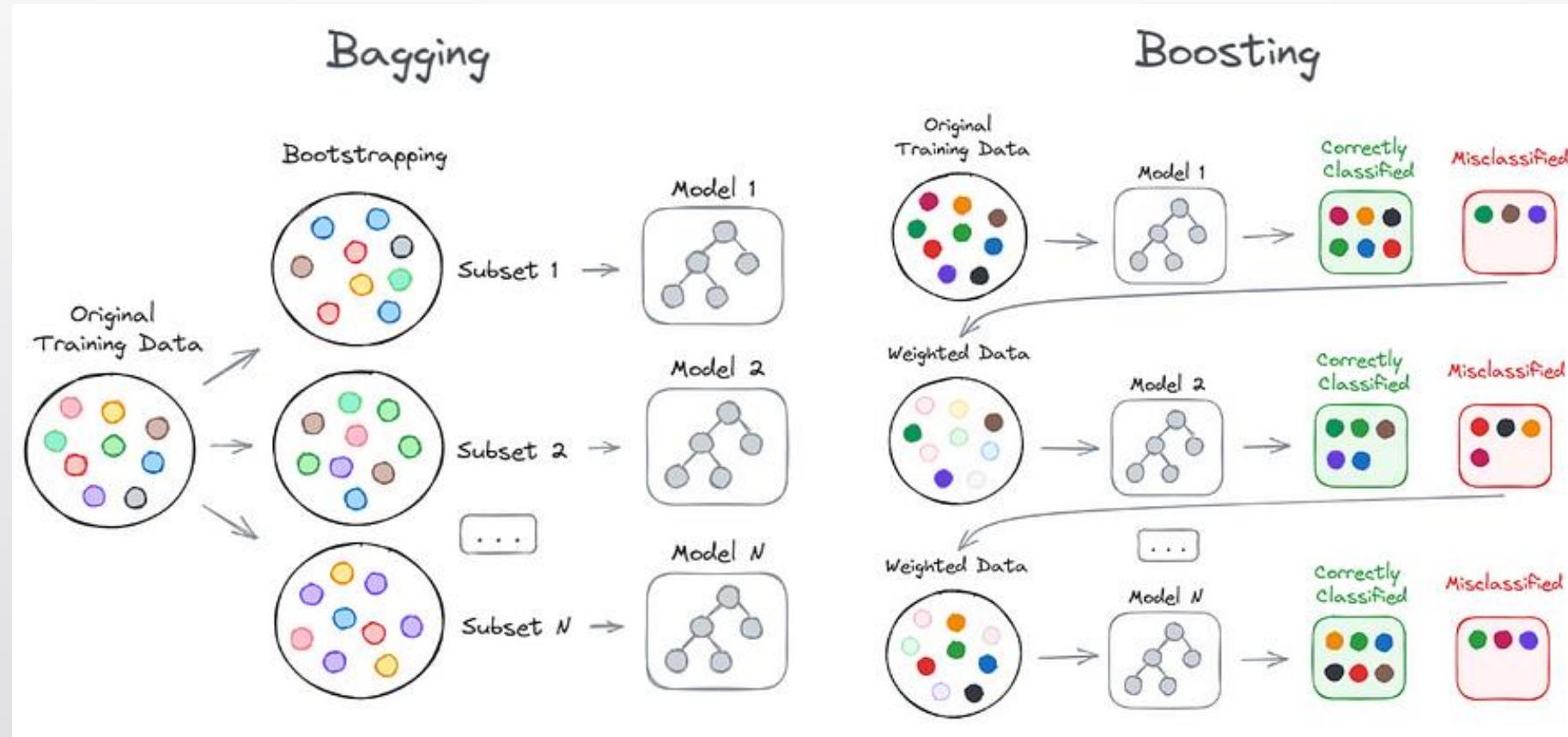
リンク Link

葉ノード Leaf Node

終端ノード

Terminal Node

バギングとブースティング Bagging and Boosting



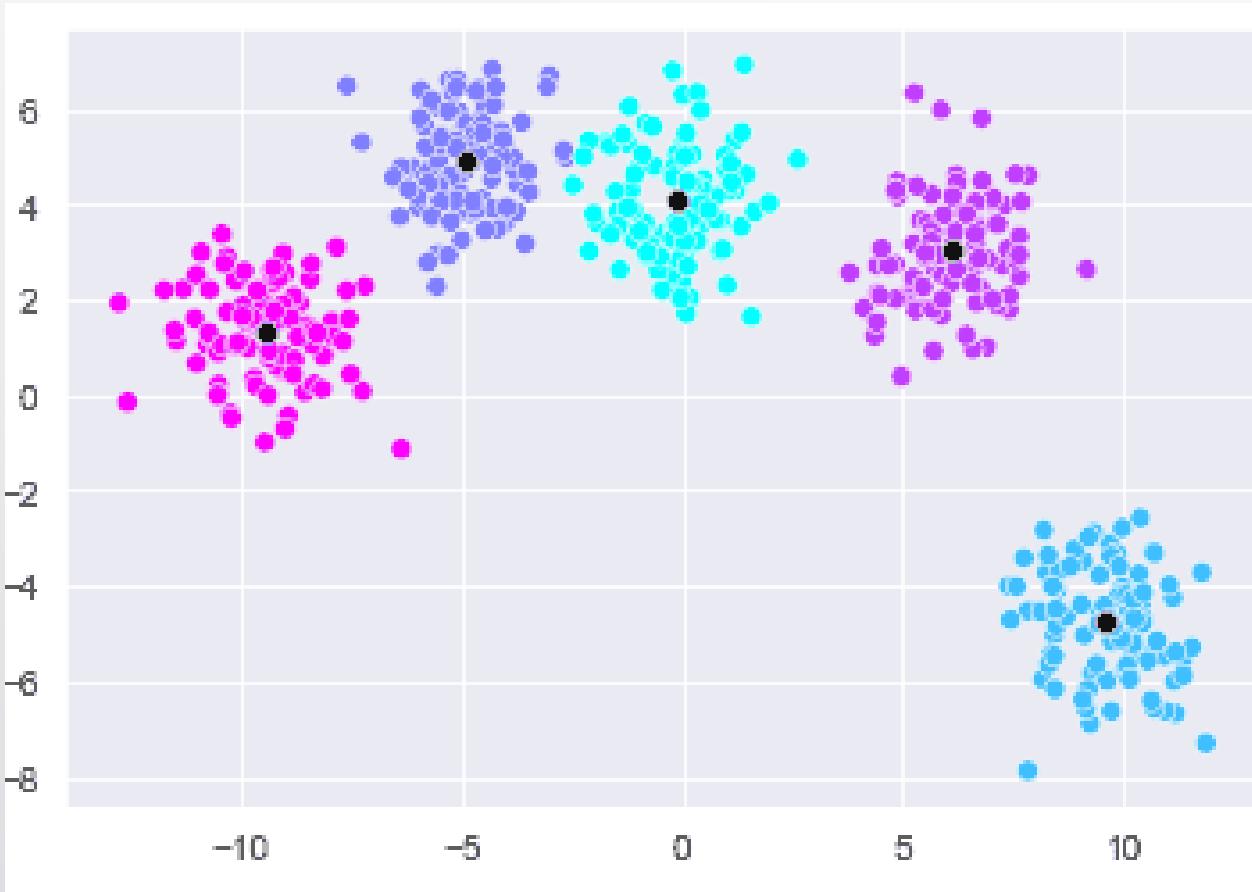
<https://pub.towardsai.net/bagging-vs-boosting-the-power-of-ensemble-methods-in-machine-learning-6404e33524e6>

クラスタリングの種類 Types of Clustering

- 非階層的クラスタリング
Non-Hierarchical Clustering
- 階層的クラスタリング
Hierarchical Clustering
- モデル・ベース・クラスタリング
Model-Based Clustering

データの統計的分布についての仮定をおく
Make presumptions about statistical distribution of data

K平均クラスタリング k-means clustering



クラスターの数を指定しなくては
いけない

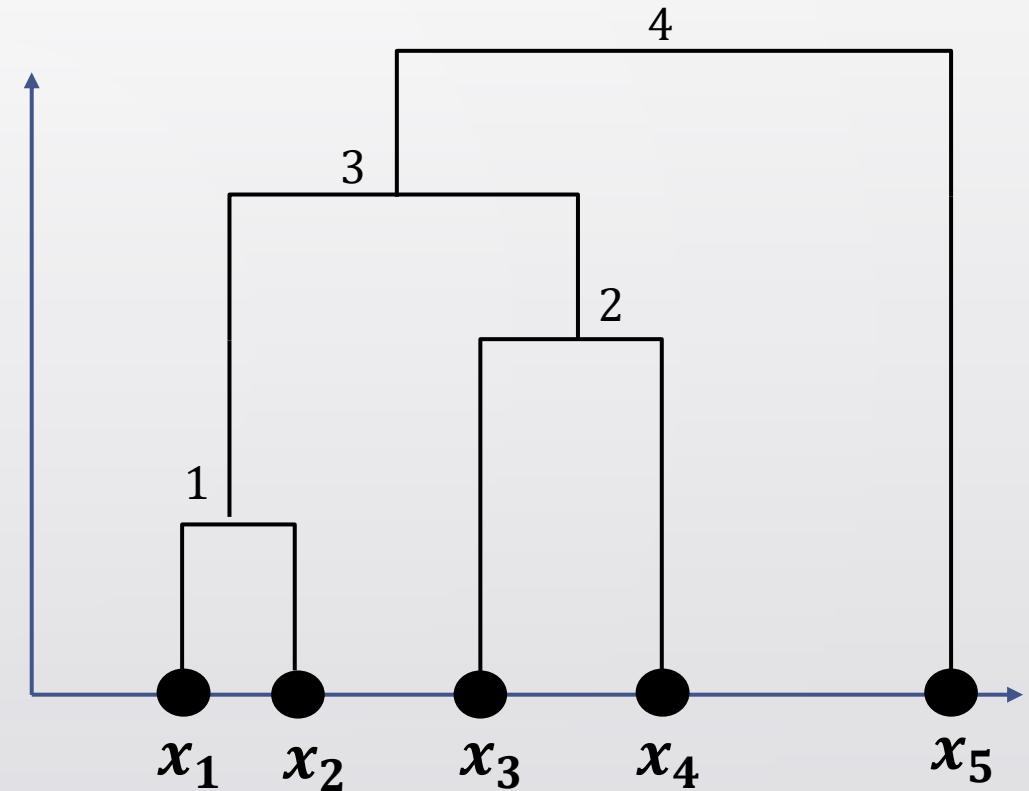
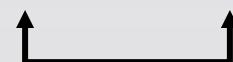
You have to specify the number of
clusters, k .



凝聚性階層的クラスタリング Agglomerative Hierarchical Clustering

	$\{x_1, x_2, x_3, x_4\}$	x_5
$\{x_1, x_2, x_3, x_4\}$	0	
x_5	4	0

x_1	x_2	x_3	x_4	x_5
1	2	5	7	11

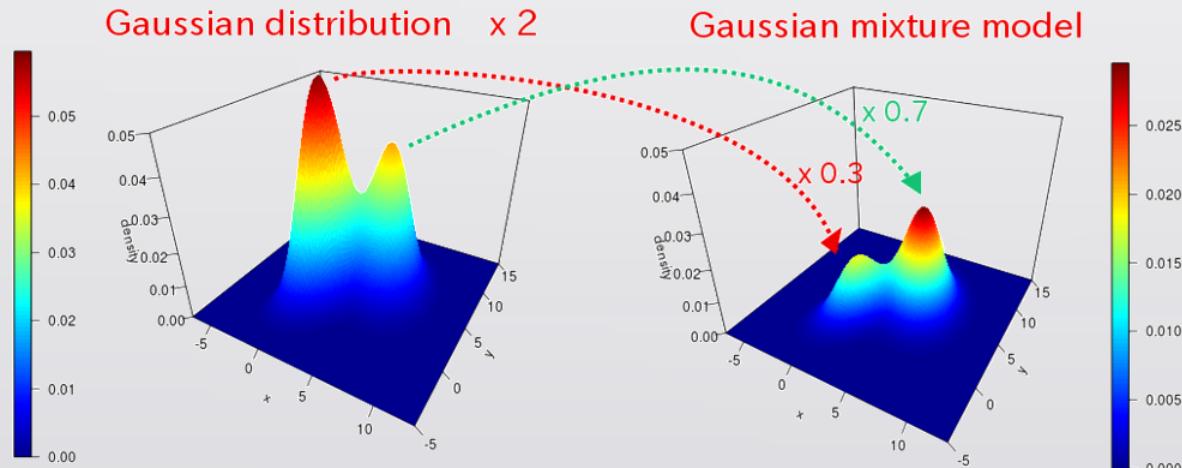


混合ガウス分布 Gaussian Mixture Distribution

M 個の正規分布の重ね合わせにより確率分布を表現する

Represent probability distribution as weighted mixture of M normal distributions

$$p(x) = \sum_{m=1}^M \pi_m N(x|\mu_m, \sigma_m) \quad 0 \leq \pi_m \leq 1 \quad \sum_{m=1}^M \pi_m = 1 \quad \pi_m : \text{混合比 Mixing Ratio}$$

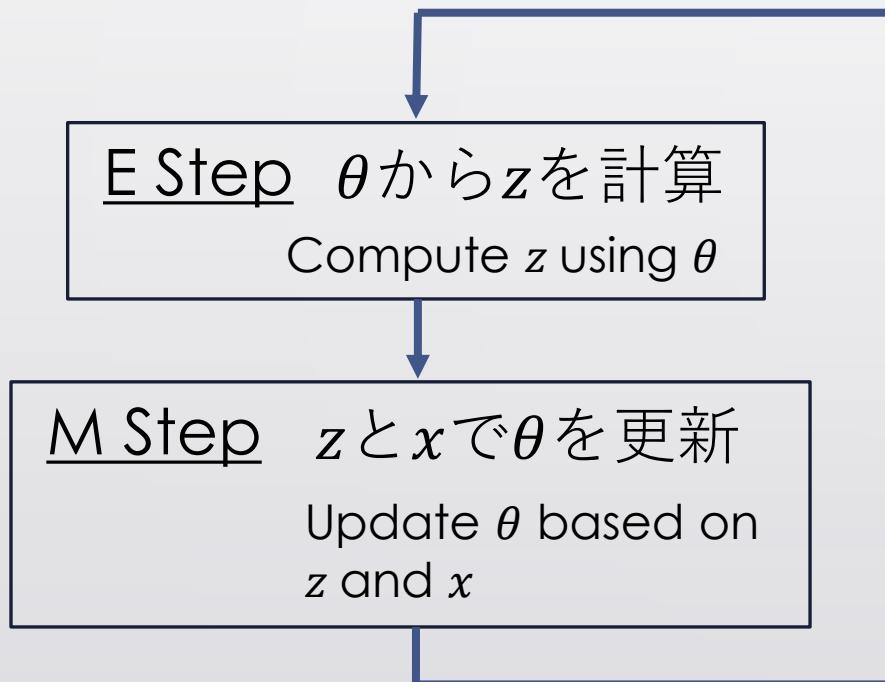


<https://work-in-progress.hatenablog.com/entry/2018/11/08/224826>

EM アルゴリズム Expectation-Maximizing Algorithm

潜在変数を含むモデルの代表的なパラメータ推定法

Algorithm for parameter estimation of models including latent variables



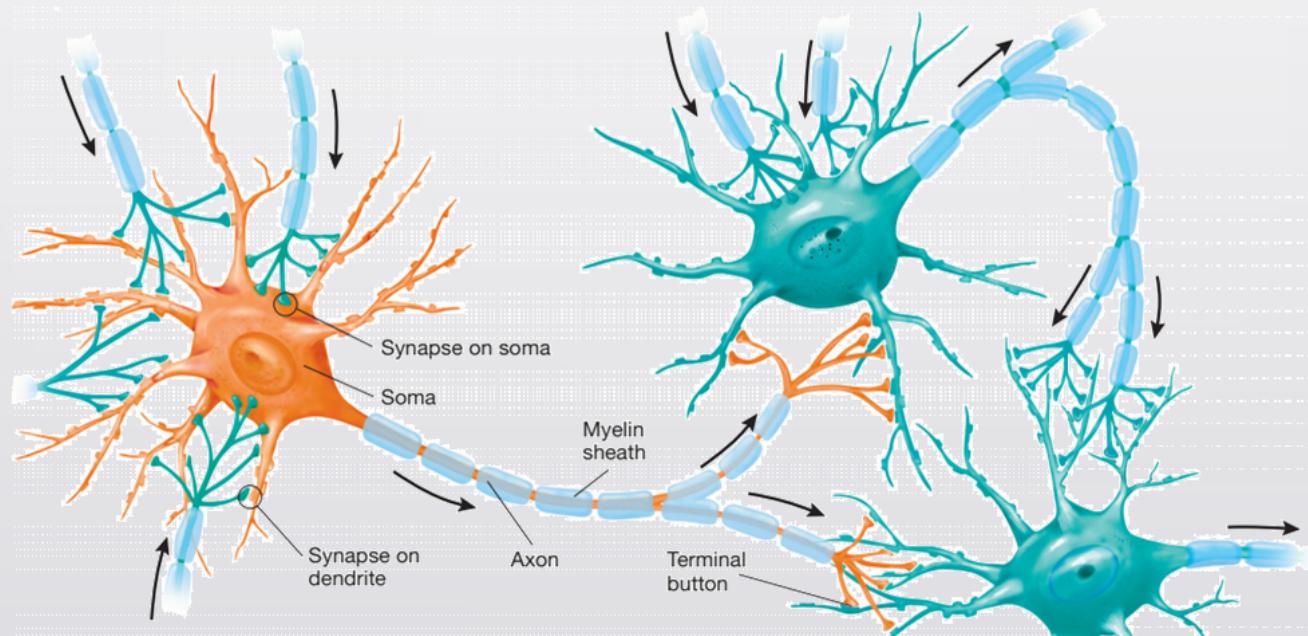
x : 観測 Observations

θ : 確率密度関数のパラメータセット
Parameter set of probability distribution functions

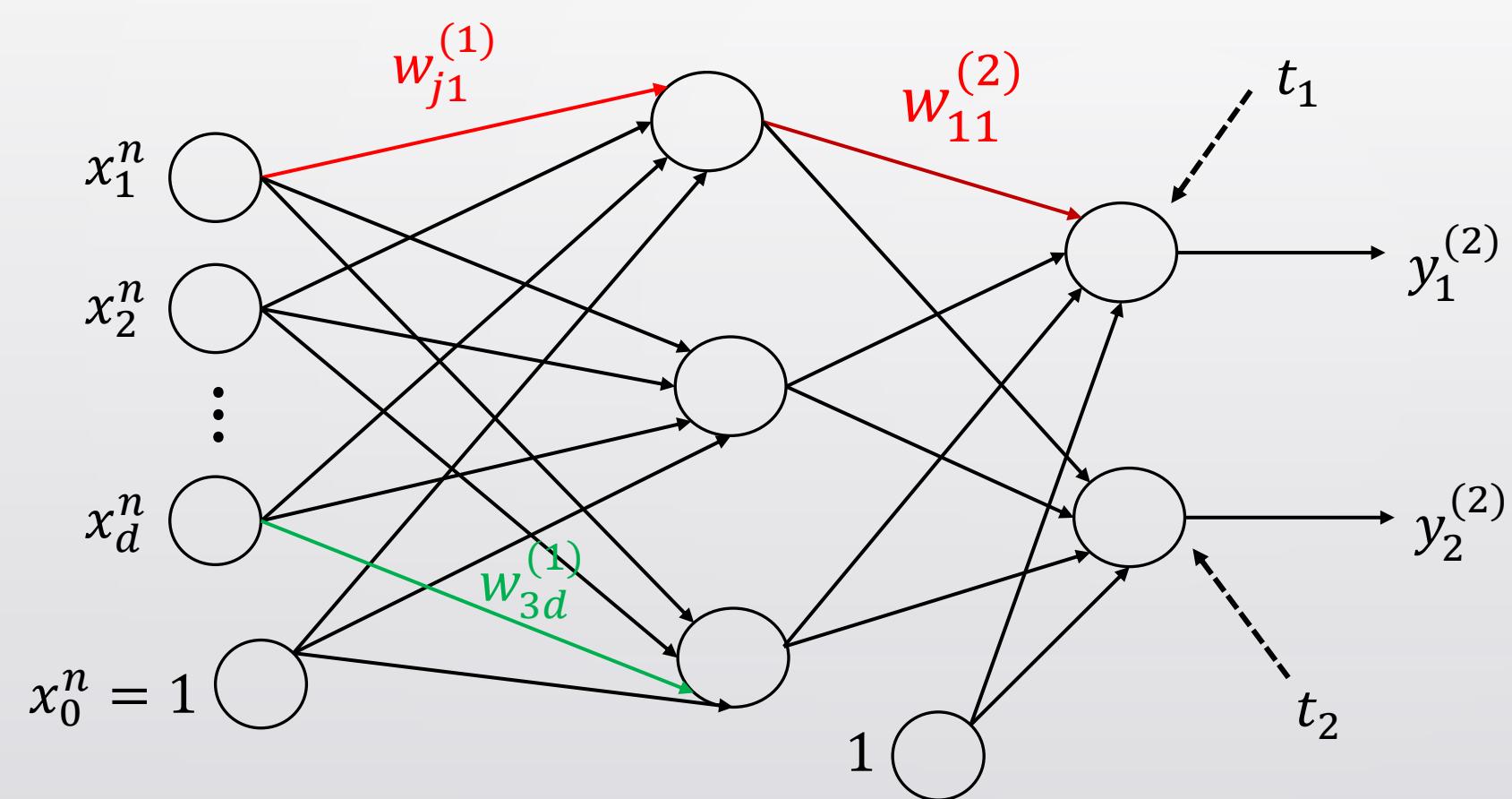
z : 潜在変数 Latent Variables

神経細胞の興奮 Neuronal Excitation

神経系の活動 = 神経細胞が電気活動を発生させ、神経細胞間で
伝えていくこと
Inter-neuronal transmission of electrical activity



多層パーセプトロン Multi-layered Perceptron



$w_{ji}^{(1)}$: 入力層から隠れ層への重み
Weights from input to hidden layer

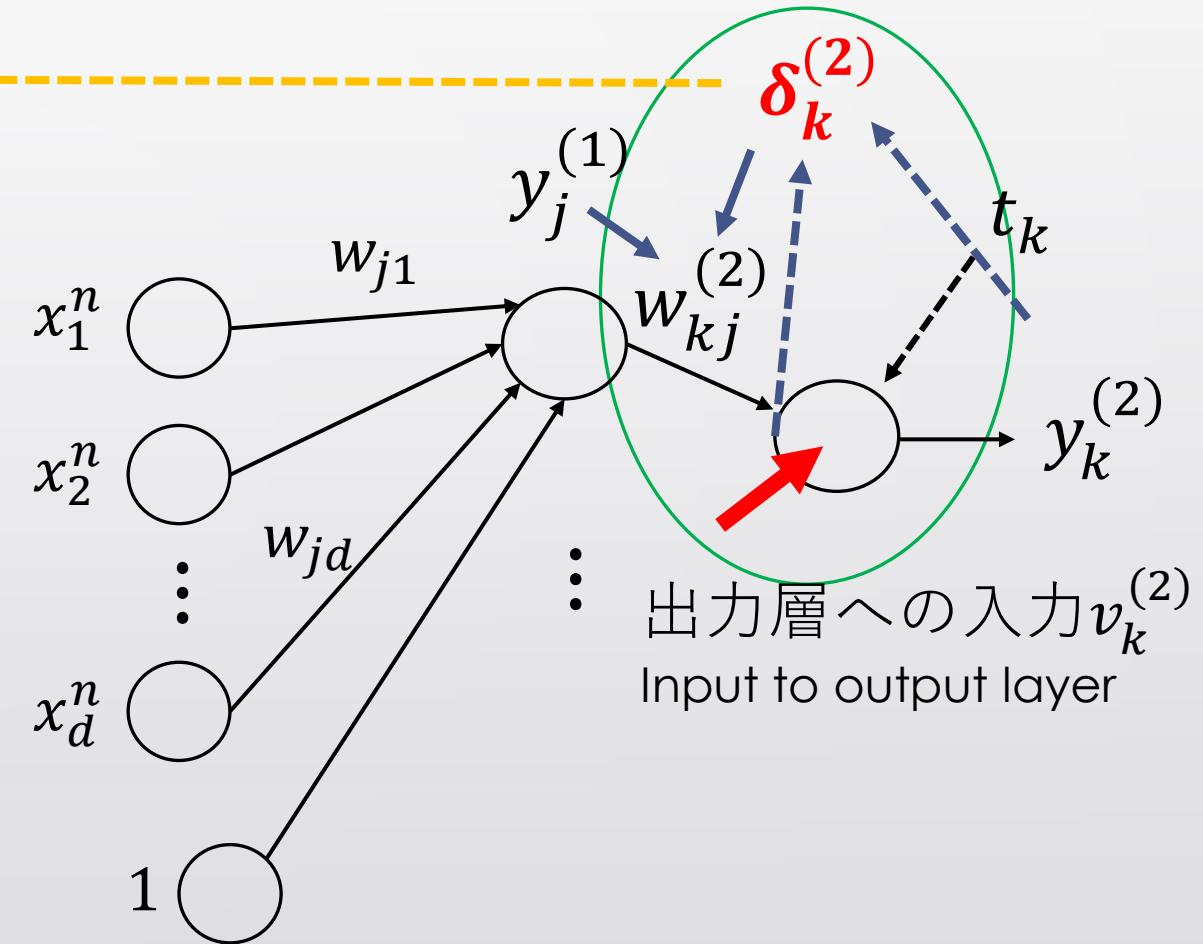
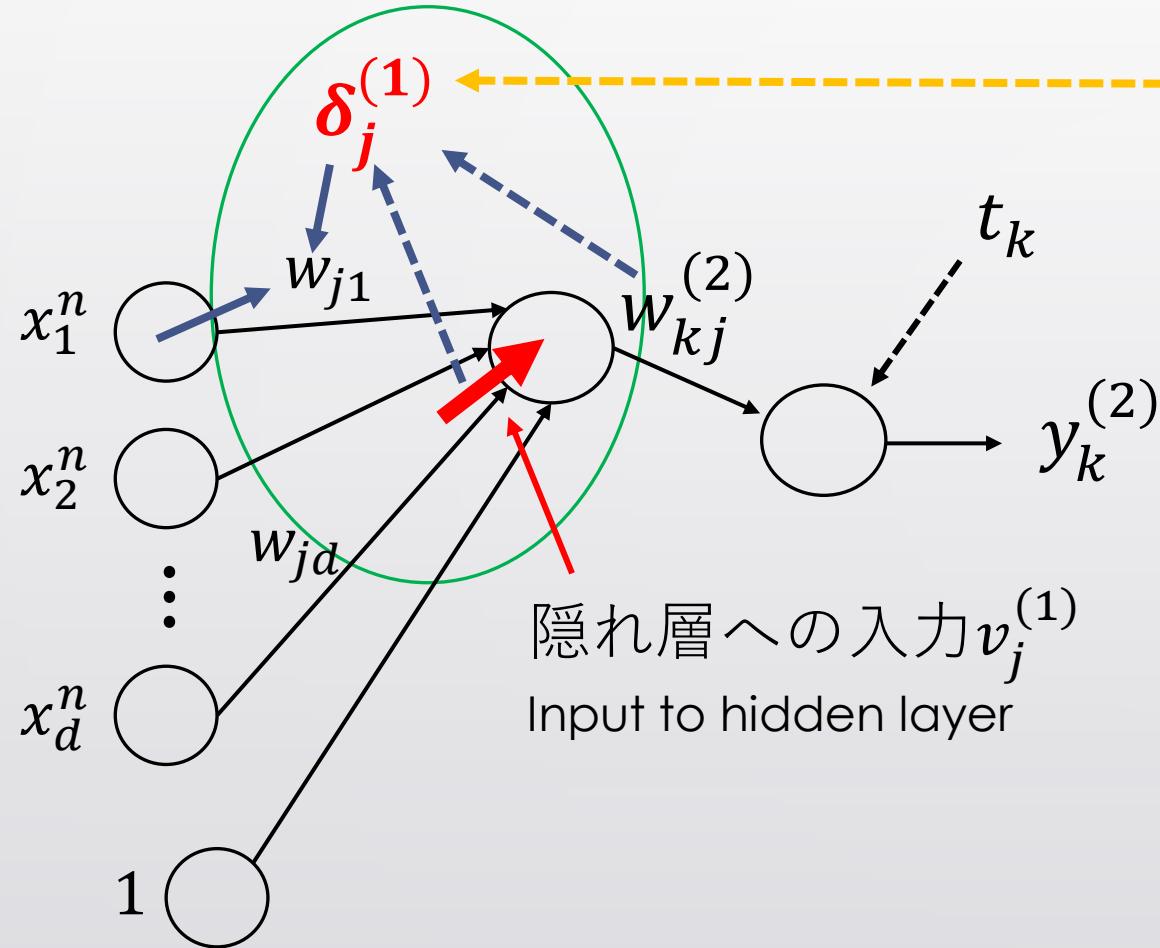
$w_{kj}^{(2)}$: 隠れ層から出力層への重み
Weights from hidden to output layer

$$i = 0, 1, 2 \dots d$$

$$j = 0, 1, 2 \dots M$$

$$k = 1, 2 \dots C$$

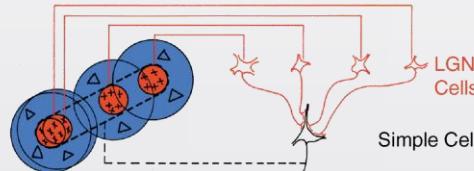
重みの更新 Weight Updating



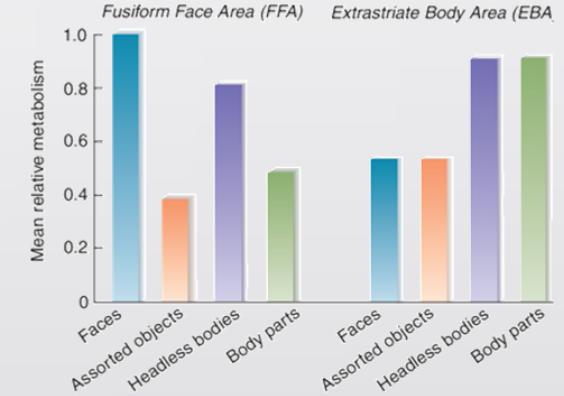
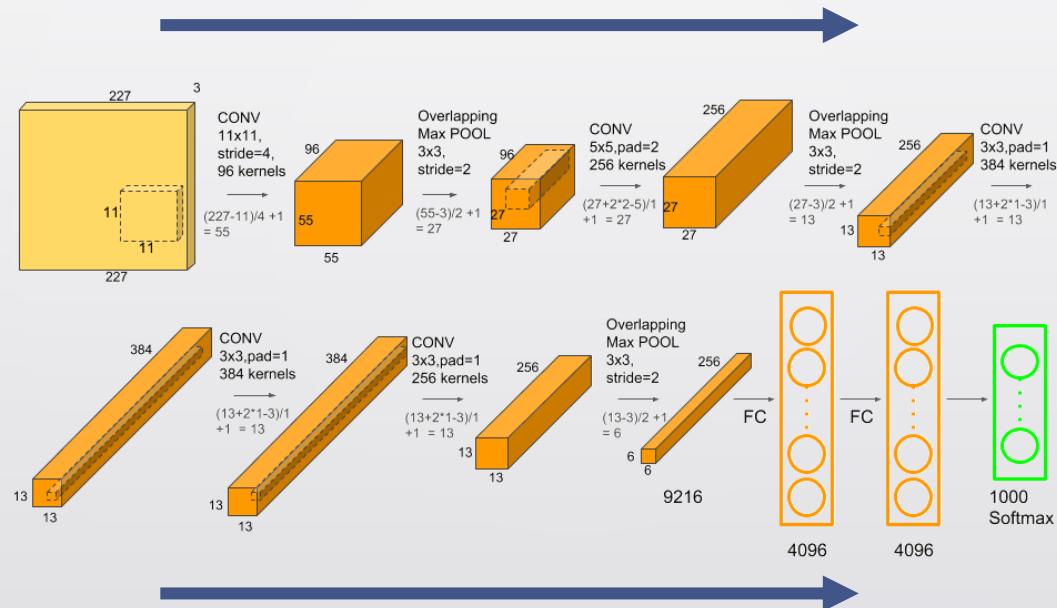
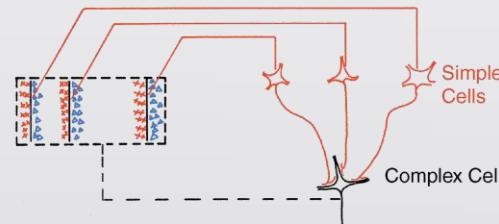
CNNと脳の類似性 Similarity between CNN and Brain

プーリングの効果で受容野が広くなる
Receptive field gets broader as a result of pooling

Circuit Building a Simple Cell from LGN Cells



Building a Complex Cell from Simple Cells

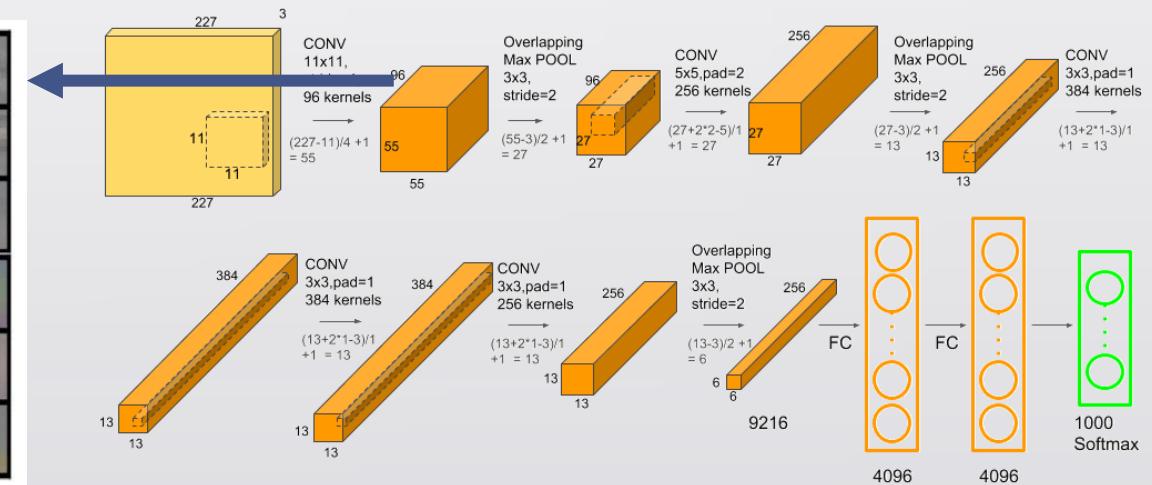
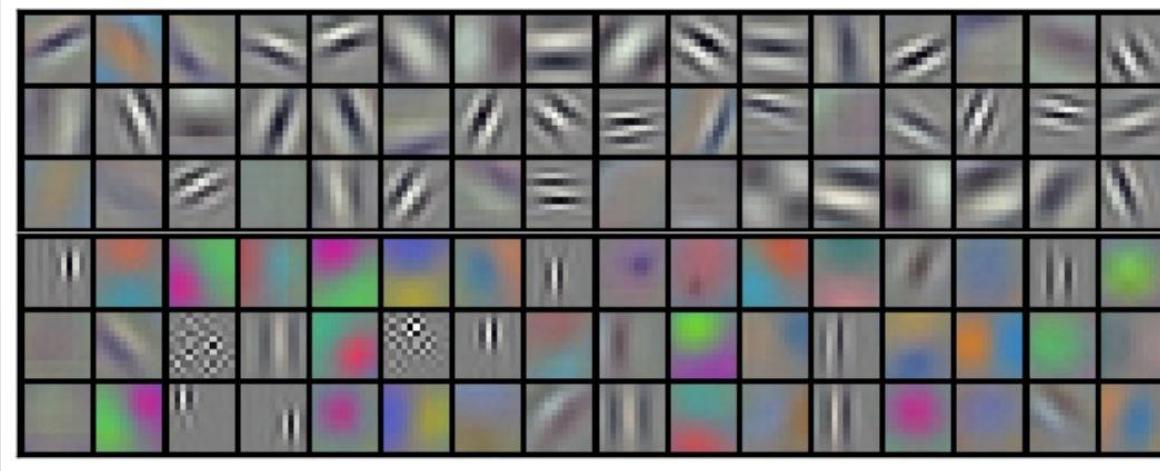


畳み込み層で複雑な情報を表現する
More complex information is represented at convolution layers

フィルターの学習 Acquisition of Filters

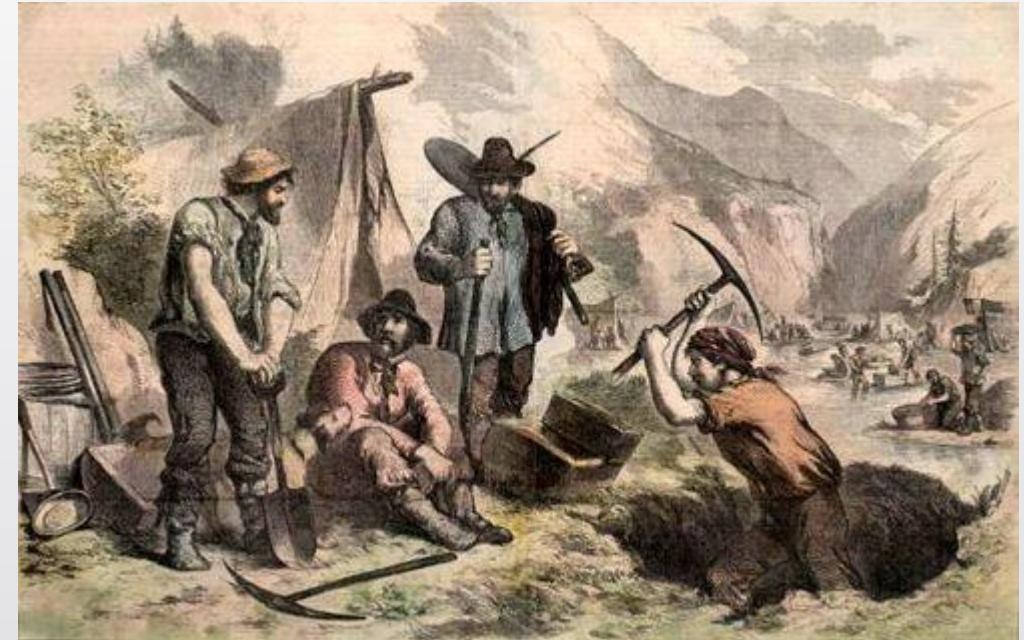
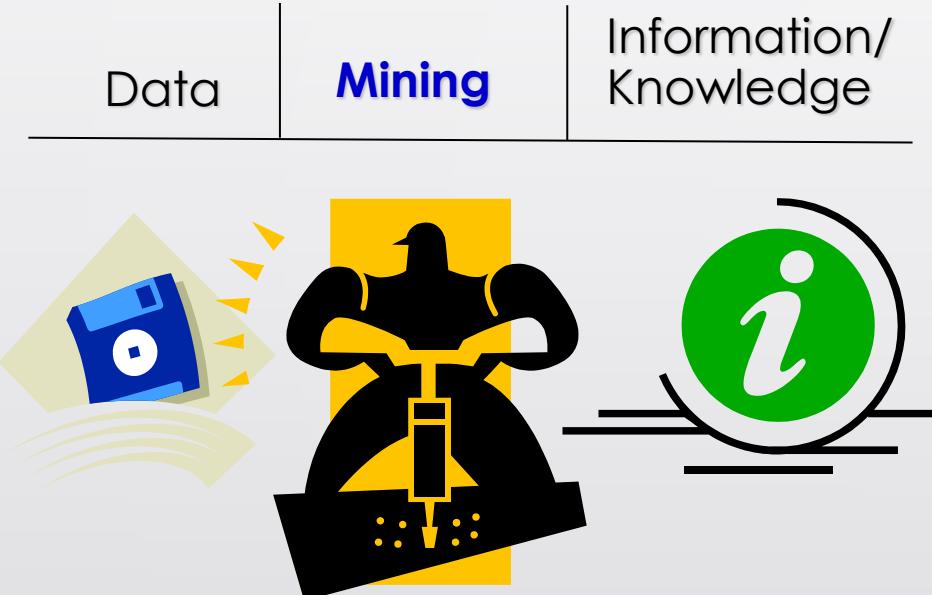
誤差逆伝搬法による学習の結果、Alex Netの第一の畠み込み層でガボールフィルターが獲得された

Weight updating by backpropagation led to acquisition of Gabor filter at the first convolution layer of Alex Net





データマイニング ≒ 金鉱の採掘
Data Mining ≒ Gold Mining



<https://www.legendsofamerica.com/mining/>



データマイニング

1: データマイニングの全体像

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

自己紹介 Self Introduction



准教授 土居 裕和 (Associate Professor Hirokazu DOI)

連絡先 e-mail doidoih@vos.nagaokaut.ac.jp

略歴 Professional Career

Sep. 2022 - Present 長岡技術科学大学 技学研究院 情報・経営システム系 准教授 Associate Prof. Department of Information and Management Engineerings, Nagaoka University of Technology

May. 2022 - Present 滋賀大学 データサイエンス・AIイノベーション研究推進センター 特任准教授 Specially Appointed Associate Prof. Data Science and AI Innovation Research Promotion Center, Shiga University

Apr. 2019 - Aug. 2022 国士館大学 理工学部 人間情報学系 准教授 Associate Prof. School of Science and Engineering, Kokushikan University

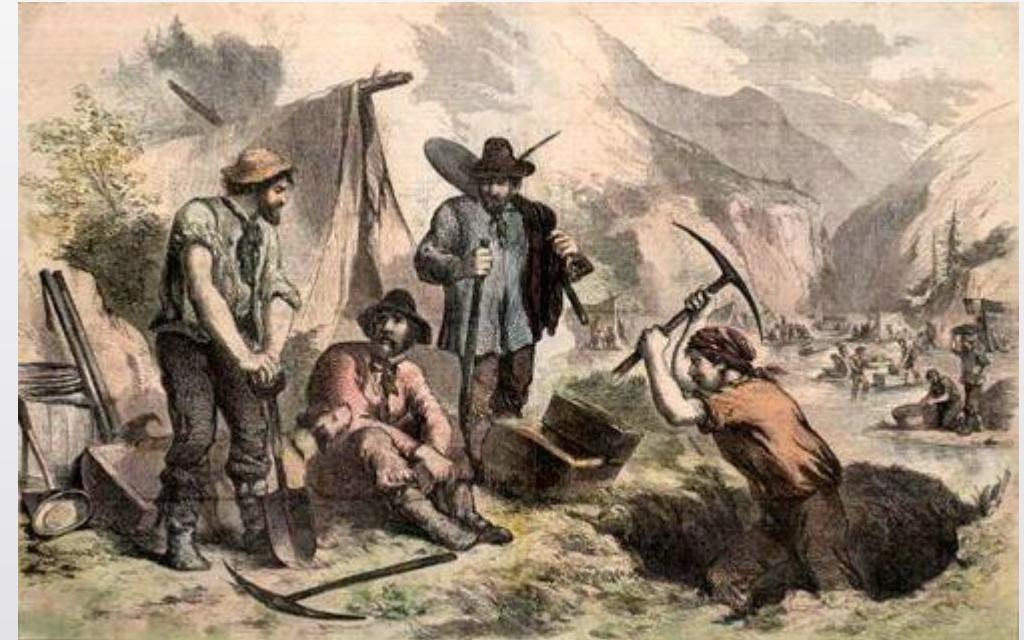
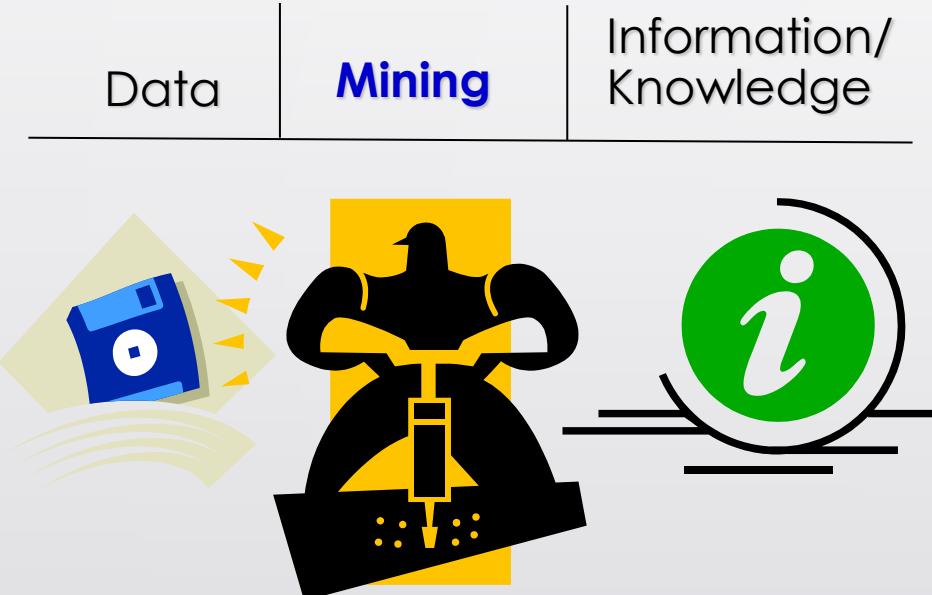
Arr. 2013 - Mar. 2019 長崎大学大学院 医歯薬学総合研究科 講師 Lecturer. Graduate School of Biomedical Sciences, Nagasaki University

Apr. 2006 - Mar. 2011 長崎大学大学院 医歯薬学総合研究科 助教 Assistant Prof. Graduate School of Biomedical Sciences, Nagasaki University

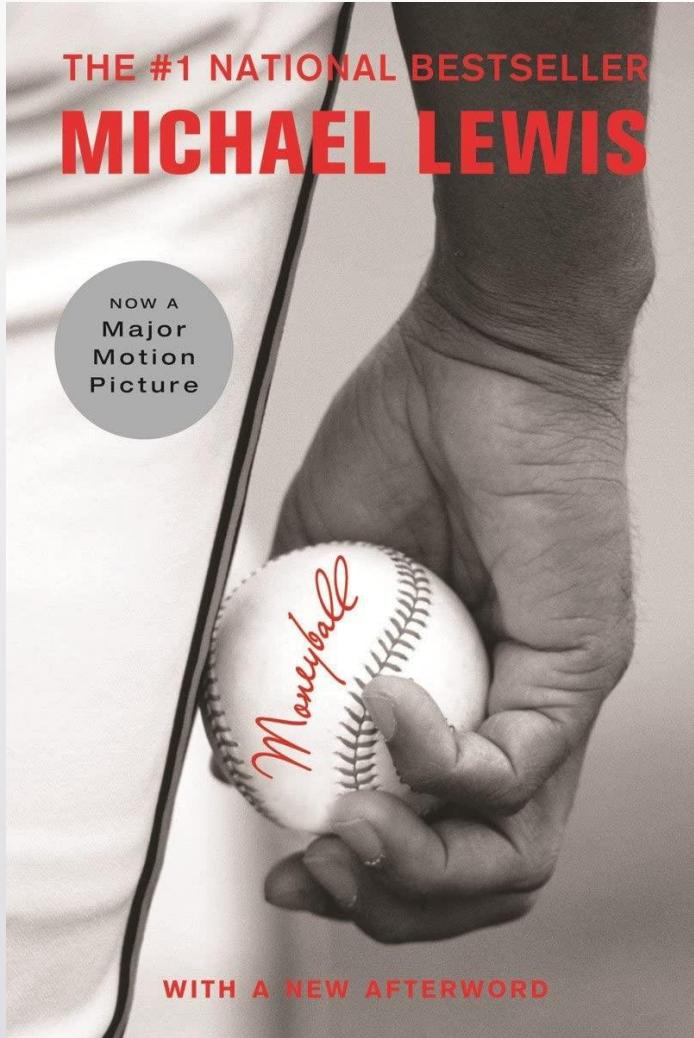




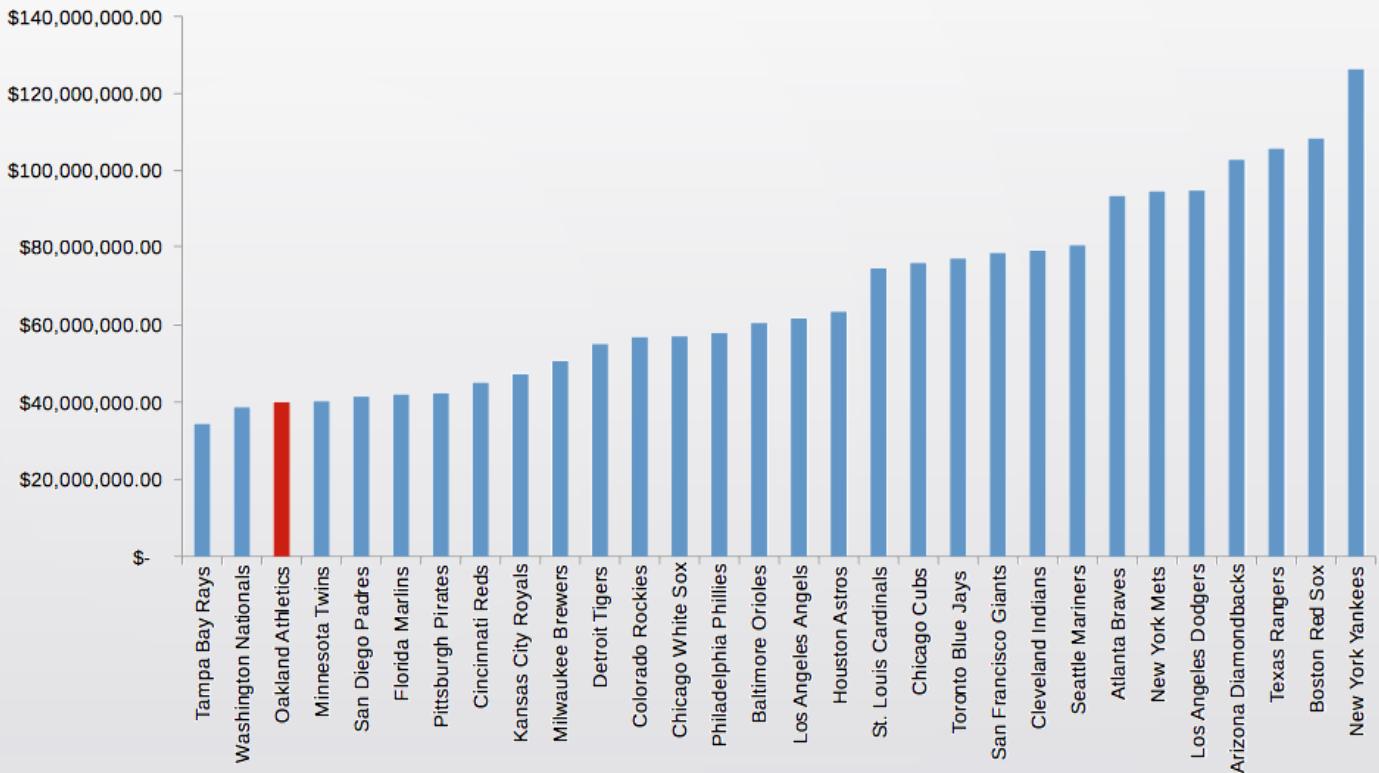
データマイニング ≒ 金鉱の採掘
Data Mining ≒ Gold Mining



<https://www.legendsofamerica.com/mining/>



Moneyball Year (2002) MLB Team Salaries



Darryl Leewood, CC BY-SA 3.0



講義の予定

【授業目的】

データマイニングとは、統計学・パターン認識・機械学習等を利用して大規模データから有用な知見を抽出する技術を指す。本講義では、データマイニングに使用される各手法について、その応用例も含め理解を深めることを目的とする。

【達成目標】

- (1) 大規模データの収集法と、データセキュリティ・安全性を高めるための基本的手法を説明できる。
- (2) データマイニングを行うための情報の前処理・変換法を説明できる。
- (3) データマイニングの代表的な手法を理解し、実データ解析に応用できる。



講義の予定

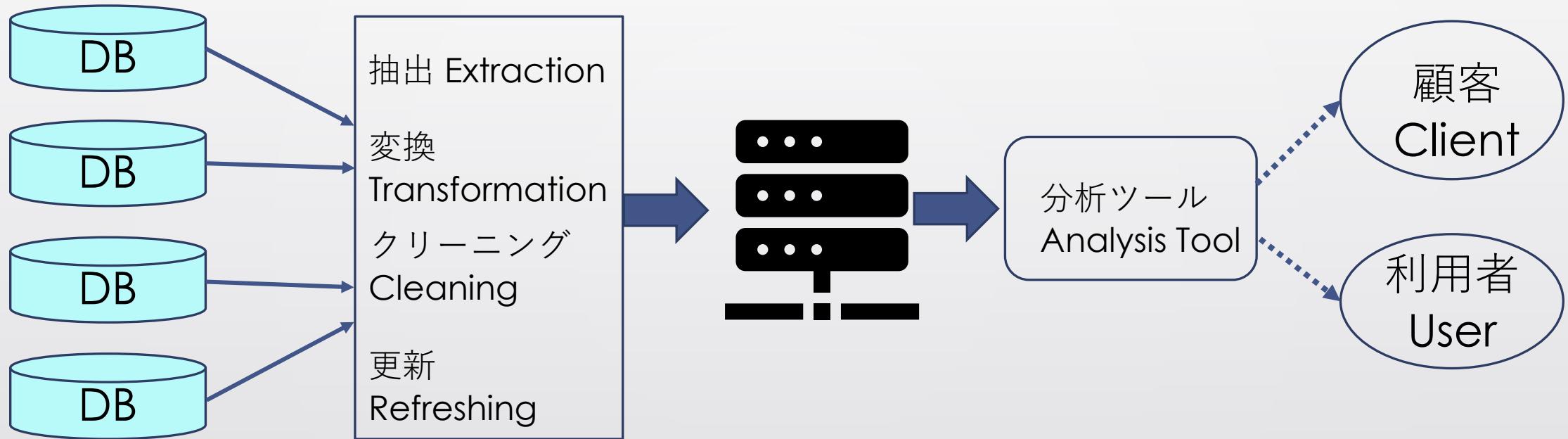
1. データマイニングの概要
2. データセキュリティとデータサイエンスの倫理
3. 前処理と次元削減
4. 記述統計
5. 回帰
6. 予測と性能評価
7. 分類
8. クラスタリング
9. ニューラルネットワーク



注意 2024年度

- ・第二回以降は、完全オンデマンドです。
- ・オンデマンド教材は、講義当日から2週間に限り視聴できます。
- ・試験・レポートについては、別途メールで案内を流します。

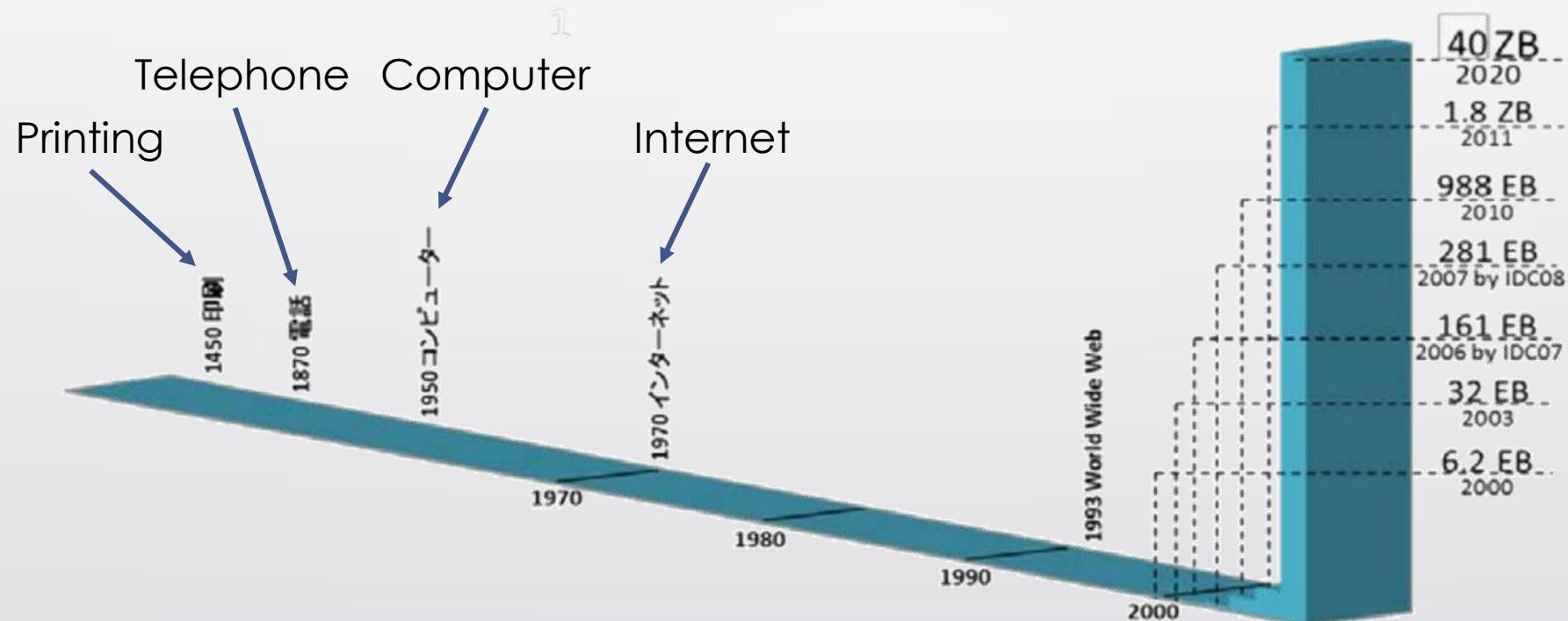
データウェアハウス Data Warehouse



複数のソースから収集したデータを統一的なフォーマットで保管

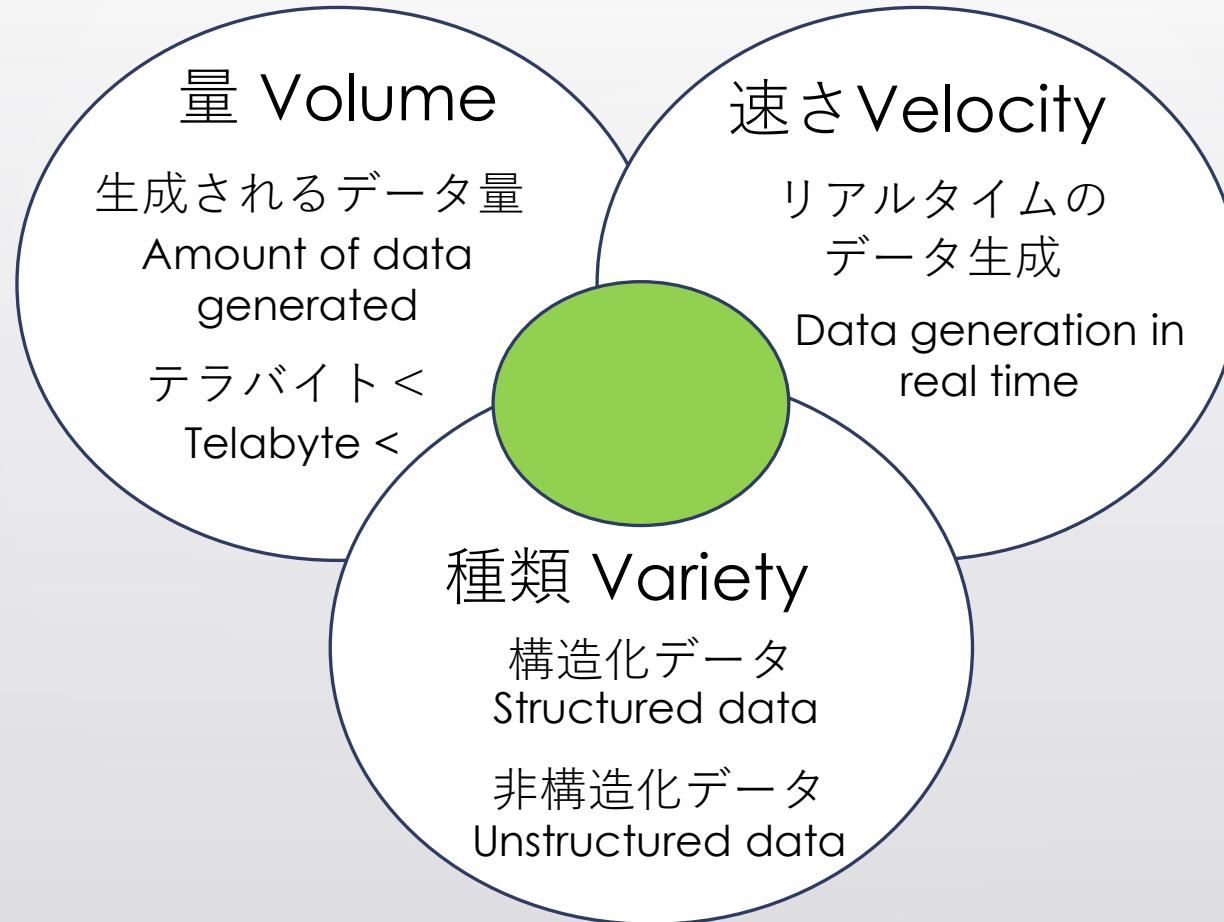
ビッグデータの蓄積

Accumulation of Big-data

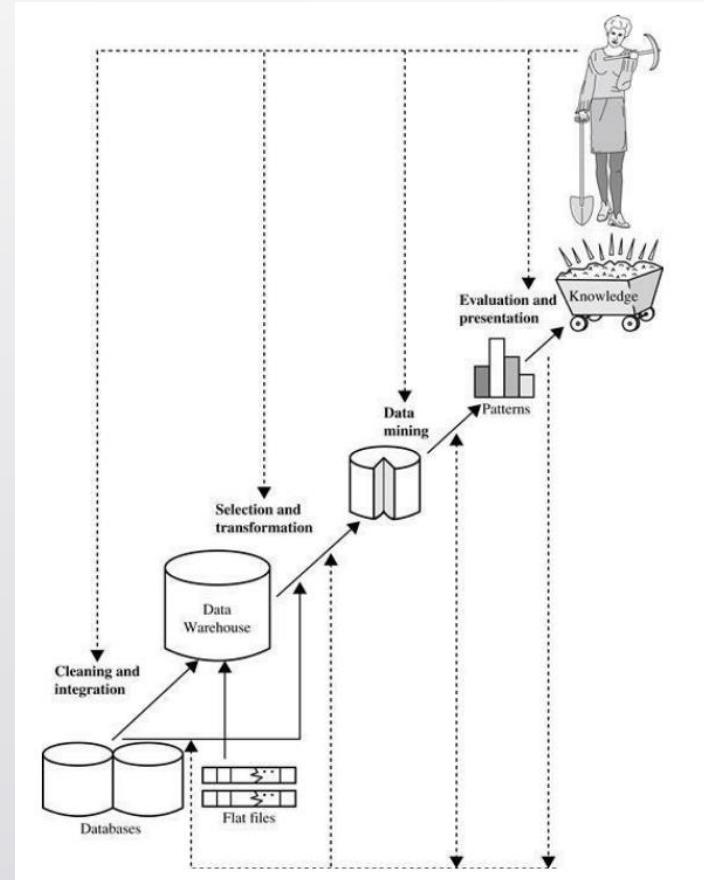


出典：IDC 「The Digital Universe Decade in 2020」（平成 24 年 12 月）等

3つの"V" Three Vs



データマイニングの目的 Goal of Data Mining



データマイニングとは、大量のデータから、興味深いパターンや知識を発見するプロセスのことである。

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data."

Han, Jiawei; Pei, Jian; Kamber, Micheline. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (p.8). Elsevier Science. Kindle 版.

データの種類

構造化データ Structured Data

ある目的のために、特定の様式で収集されたデータ

Data collected for certain purpose in a pre-defined format

データウェアハウスは、一般に、構造化データを扱う

Data warehouse generally deals with structured data.

<i>trans_ID</i>	<i>list_of_item_IDs</i>
T100	I1, I3, I8, I16
T200	I2, I8
...	...

非構造化データ Unstructured Data

画像データ、音声データ等 Image data, voice recordings, etc

非構造化データの分析には、メタデータが必要

In many cases, analysis of unstructured data requires meta-data.

メタデータ Metadata

データを取得した状況についての情報

Information about the situation under which the data was collected.

画像ファイルの場合・・・

In case of image files, metadata includes

撮影場所 Location,
撮影日時 Date,
撮影者 Photographer
画像サイズ Image Size
カメラの機種 Type of Camera etc

Basic Image Information

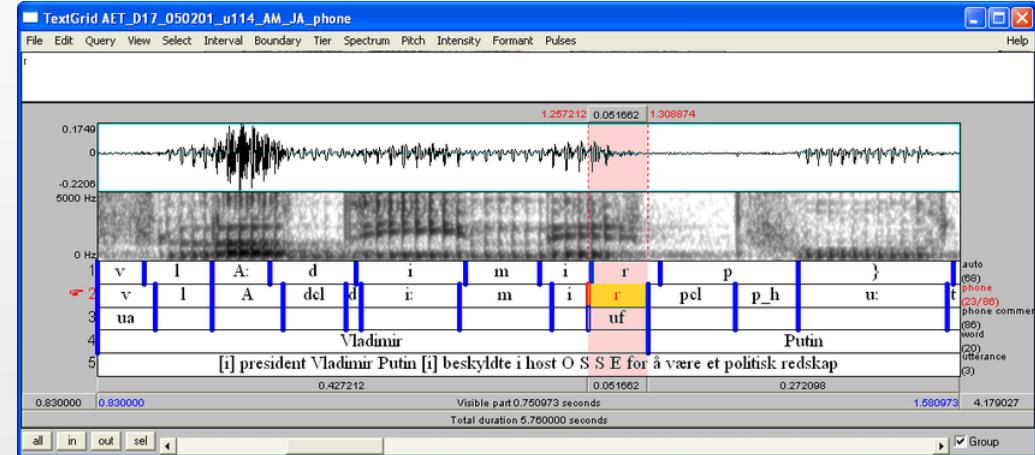
Target file: IMG_20161101_112511.jpg

Location:	Latitude/longitude: 39° 5' 7" North, 94° 35' 1" West (39.085278, -94.583611)
	Location guessed from coordinates: 2301 Main St, Kansas City, MO 64108, USA
	Map via embedded coordinates at: Google , Yahoo , Wikimapia , OpenStreetMap , Bing (also see the Google Maps pane below)
	Timezone guess from earthtools.org: 6 hours behind GMT
File:	5,312 × 2,988 JPEG (15.9 megapixels) 5,272,745 bytes (5.0 megabytes)
Color Encoding:	WARNING: No color-space metadata and no embedded color profile: Windows and Mac web browsers treat colors randomly. Images for the web are most widely viewable when in the sRGB color space and with an embedded color profile. See my Introduction to Digital-Image Color Spaces for more information.



Stewart & Dawson, 2018

アノテーション Annotation



Amdal et al, 2008

<https://annotationlabs.com/annotation-tools/semantic-panoptic-segmentation/>

タグ付け/ラベル付けとも呼ばれる

It is also called “tagging” or “labelling”

データ量が多い場合、作業をクラウドソーシングすることがある

When the data amount is huge, annotation is crowdsourced.

一次データと二次データ

Primary Data and Secondary Data

一次データ Primary Data

自分自身でデータを収集する必要がある You have to collect the data by yourselves

自分の目的に適したデータが手に入る You can get the data suitable for your purpose

二次データ Secondary Data

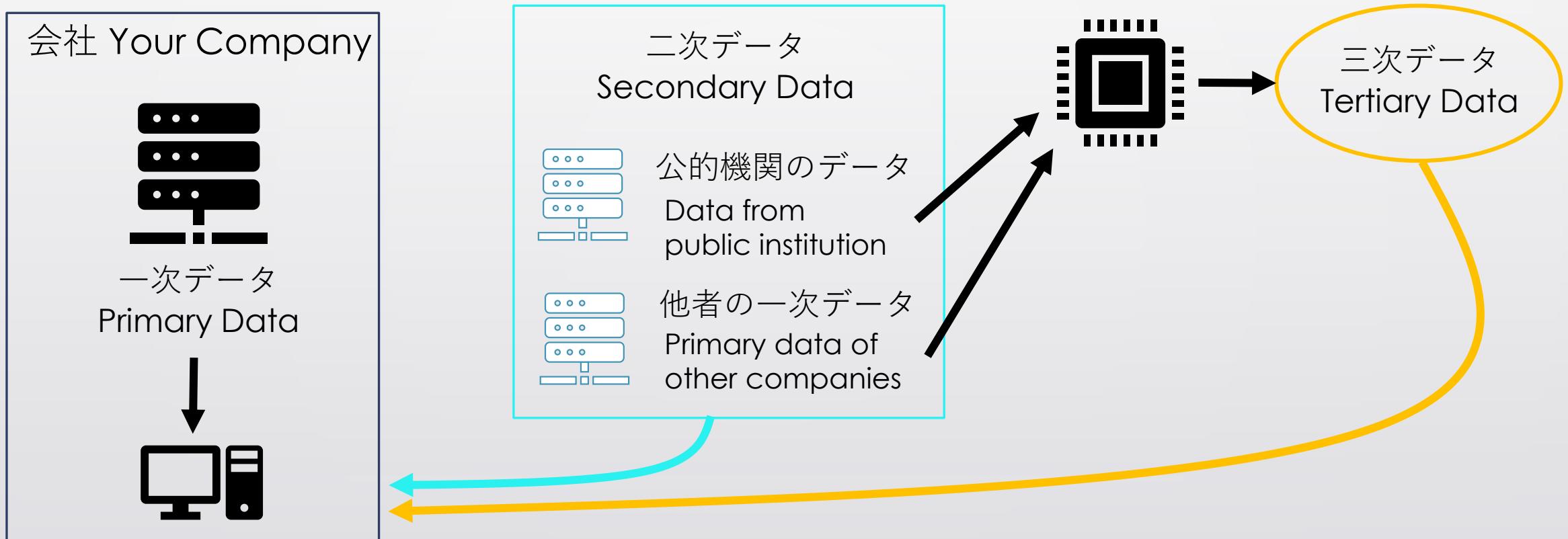
データを購入する、もしくは、オープンデータを入手する

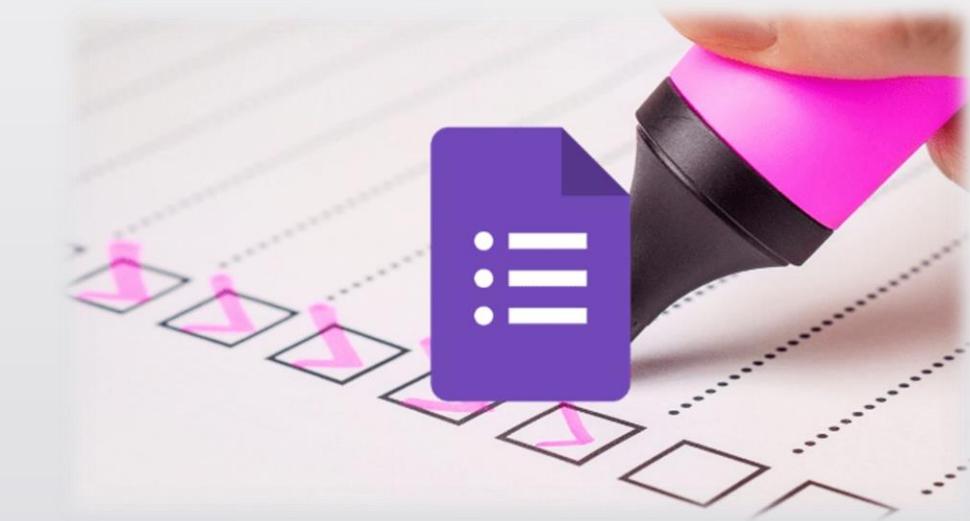
Purchasing data or searching for open data source

自分の目的に沿ったデータが手に入るとは限らない

May not match the purpose of the survey

3次データ Tertiary Data





データの種類

質問紙データ Questionnaire data

質問作成・分析のための統計的手法が確立されている

Statistical methods are established for creation and analysis of questionnaires.

質問 回答

満足度アンケートのお願い

DEKIRUサービスご利用の方にアンケートのご協力をお願いしています。

メールアドレス *

有効なメールアドレス

このフォームでは回答者のメールアドレスを収集しています。 [設定を変更](#)

年齢 *

10代

20代

30代

データの種類

行動ログデータ Behavioral Log Data



Top 10 E-Commerce Websites



ユーザが意識しているかどうかに關係なく、計算機上に自動的に蓄積されるデータ
Data automatically registered in computers irrespective of users' awareness

オープンデータ Open Data



国や研究機関が集めたデータを、個人情報に

配慮して公開したもの

Dataset made publicly available by national and academic institutes while keeping anonymity.

オープンデータを公開するための学術誌もある

There are academic journals for curating open research dataset.

<https://www.data.jma.go.jp/gmd/risk/obstd/>

データの信頼性 Data Reliability



<https://journalistsresource.org/health/trust-science-pseudoscience-misinformation/>

科学的に信頼できるデータか？ Is the data reliable from scientific standpoint?

自治体など、信頼できる機関から提供されたデータか？
Is the data retrieved from reliable source such as local municipality?

データ収集の方法は適切か？
Is the methodology of data collection appropriate enough?

データマイニングで得られる情報
Information obtained by Data-Mining

説明的データ分析 Descriptive Data Analysis

データの分布・性質を簡潔に記述する

Describe distribution and characteristics of data pattern in a succinct manner

データを分析することで、出来事の原因解明に役立てる

Clarify the cause of an event/incident by analyzing the data

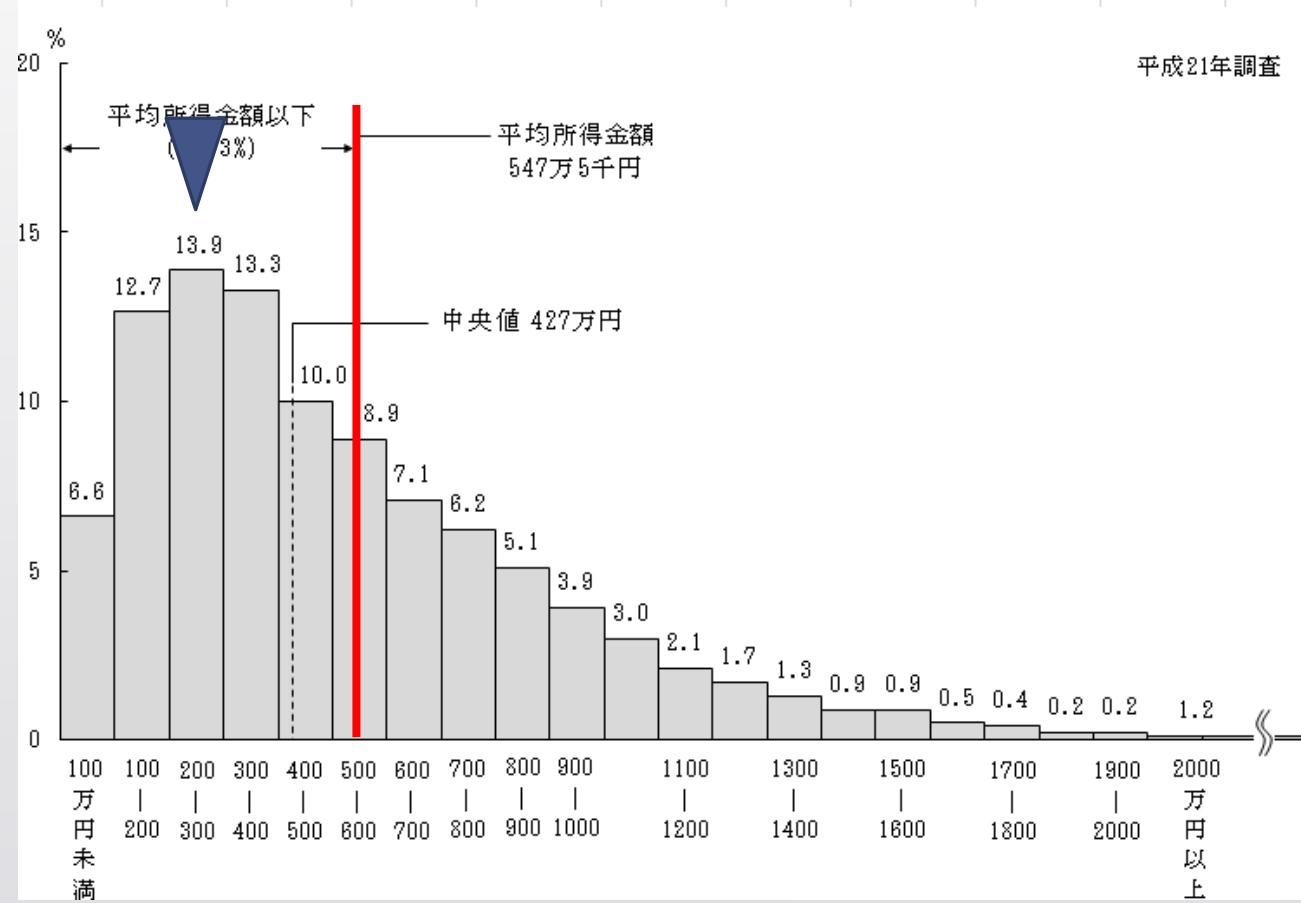
用いられる手法 Methods for descriptive data analysis

記述統計 Descriptive Statistics, 推測統計 Inferential Statistics

クラスタリング Clustering, 可視化 Visualization etc

※クラスタリングは予測にも使えます Clustering can be used in predictive data analysis as well

記述統計 Descriptive Statistics



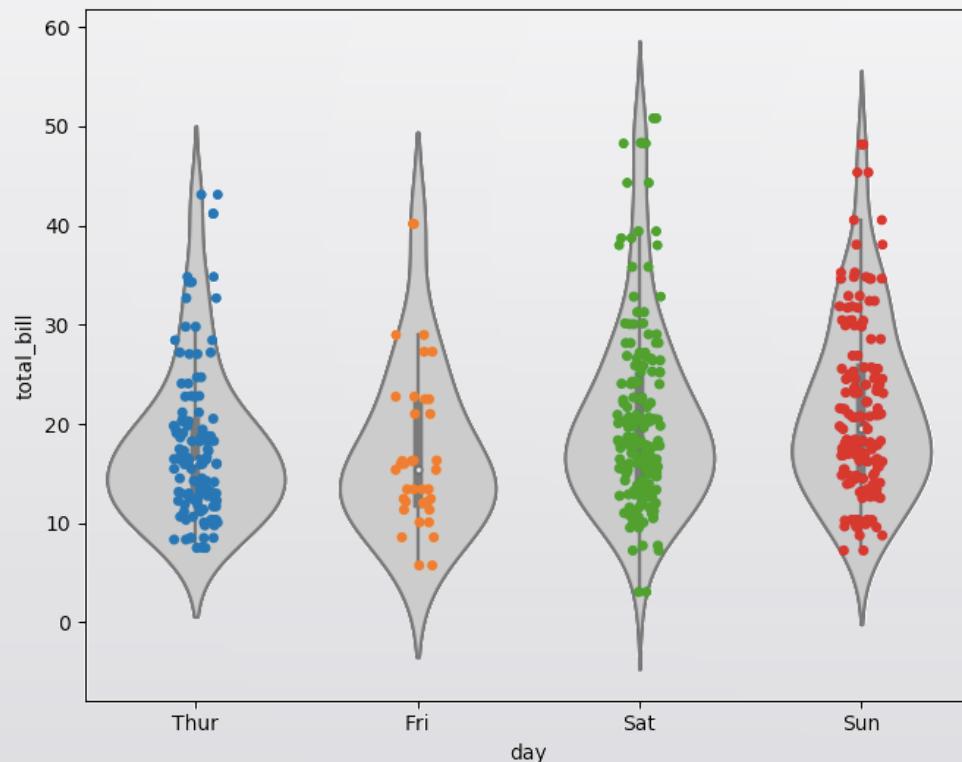
日本における収入の分布
Household income distribution
in Japan

可視化 Data visualization



可視化 Data Visualization

バイオリンプロット Violin Plot



多次元データの可視化
Visualization of Multidimensional Data

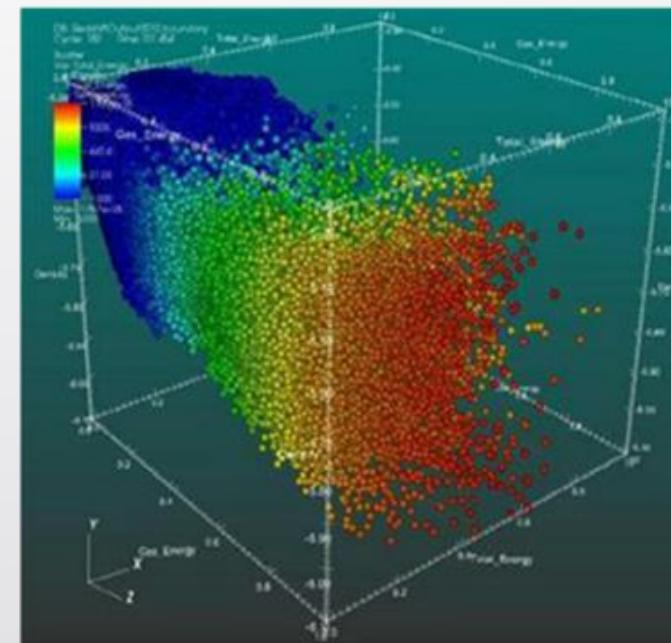


FIGURE 2.14 Visualization of a 3-D data set using a scatter plot. Source: http://upload.wikimedia.org/wikipedia/commons/c/c4/Scatter_plot.jpg.

データマイニングで得られる情報
Information obtained by Data-Mining

予測的データ分析 Predictive Data Analysis

データ分析結果に基づいて、未知の性質・出来事を予測する
Make predictions about unknown attributes/events based on the results of data analysis

用いられる手法 Methods for predictive data analysis

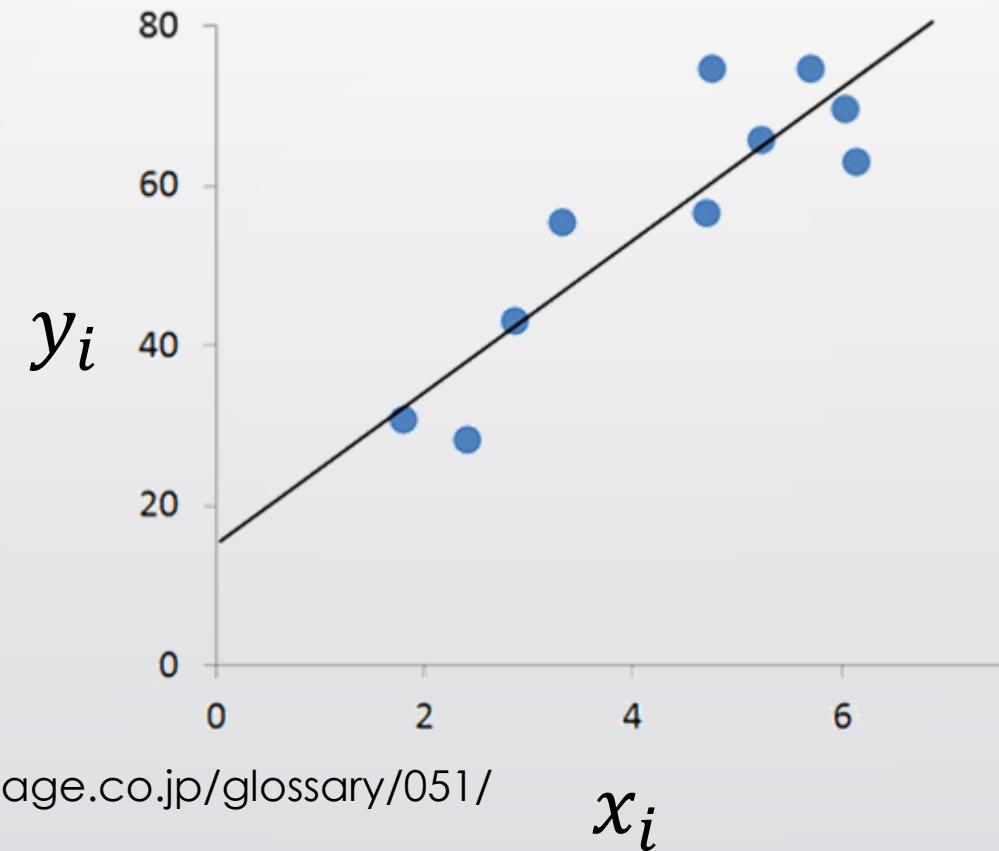
回帰 Regression, クラスタリング Clustering, 分類 classification,
統計的機械学習 Statistical Machine Learning、時系列解析 time series analysis

予測とは何か？ What is “Prediction”?

(x_i, y_i)

x_i : 家族の人数
Number of Family Member

y_i : 購入数
Number of purchased Items



未知データ
Unknown Data

$(x_i,)$

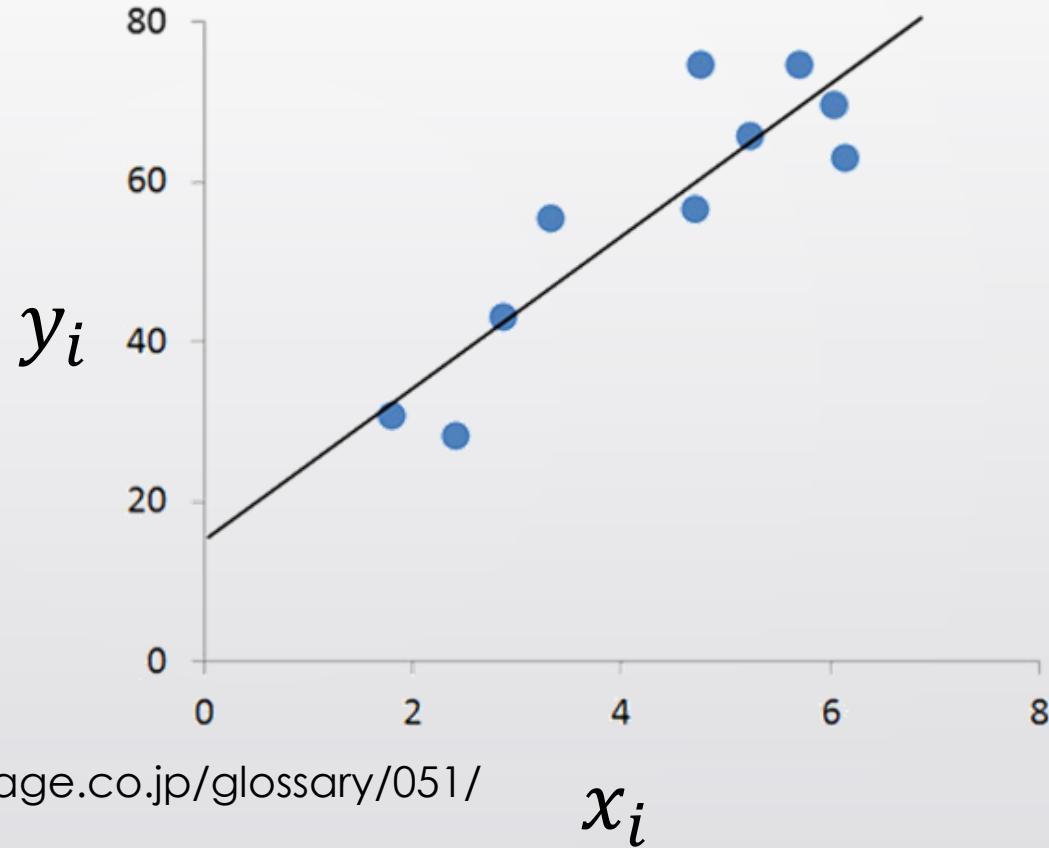
モデルに基づいて、未知の変数を予測する
Predict unknown variable based on the model

回帰 Regression

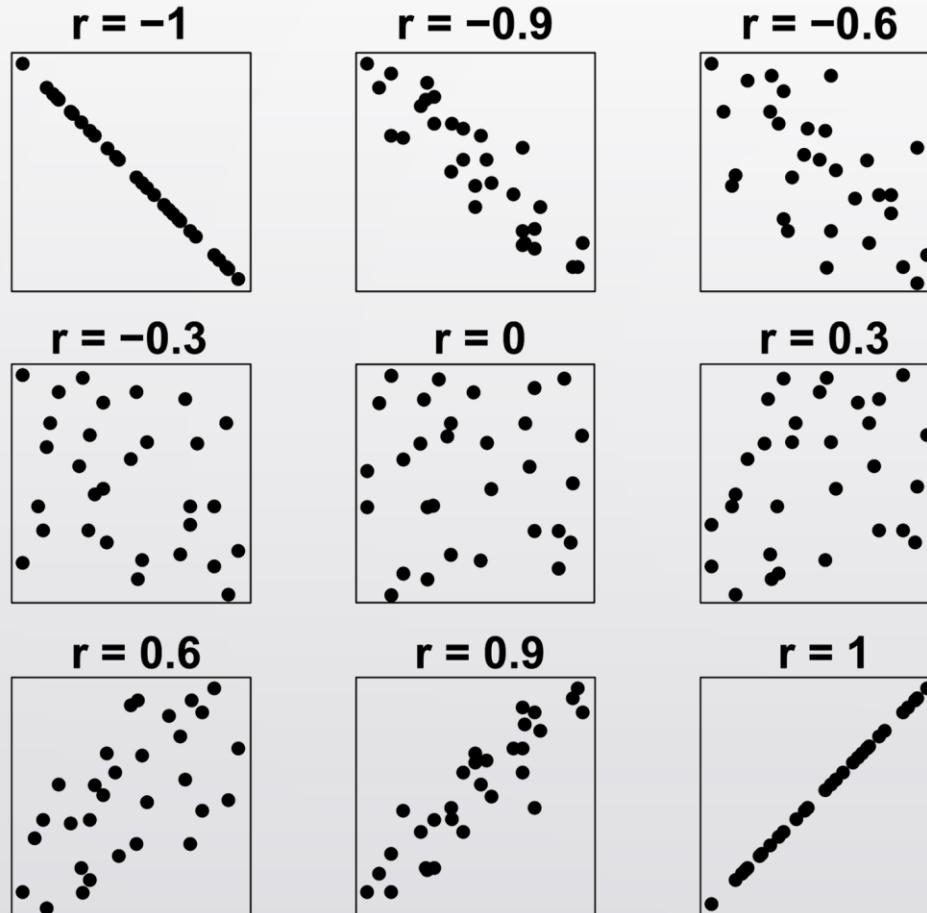
(x_i, y_i)

x_i : 家族の人数
Number of Family Member

y_i : 購入数
Number of purchased Items



相関係数 Correlational Coefficients

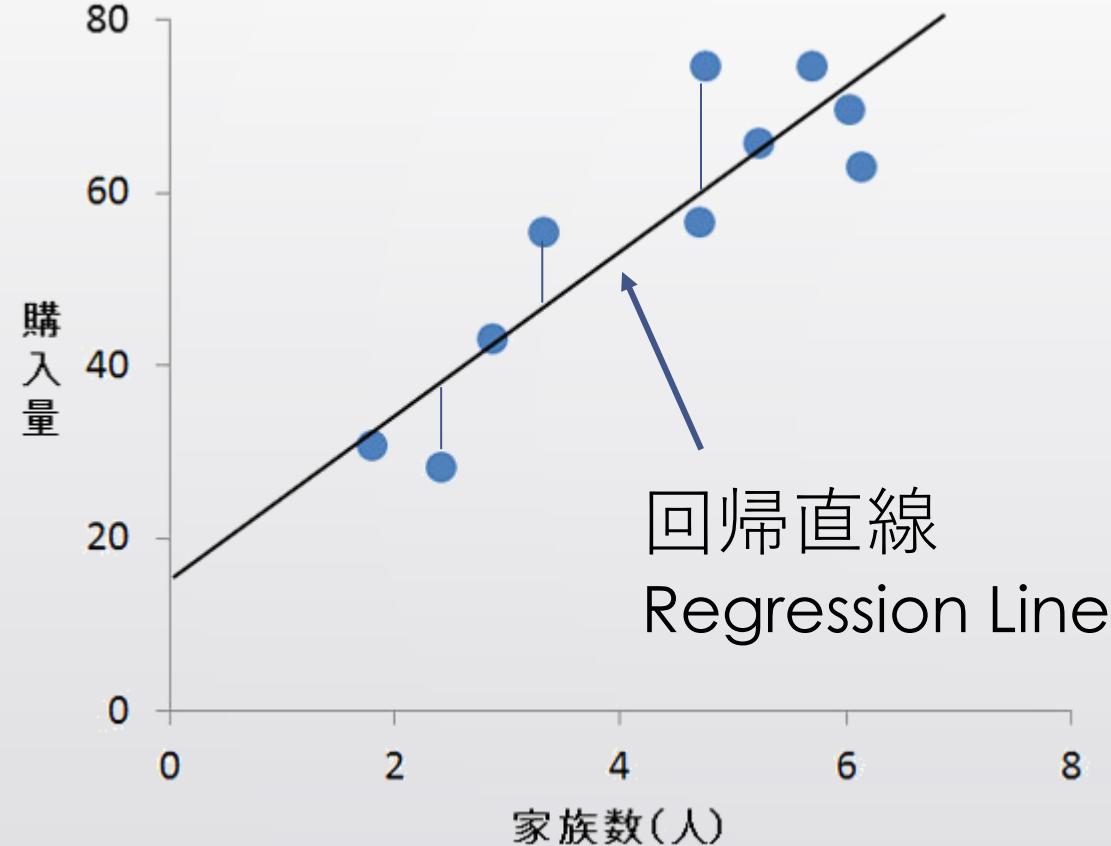


2つの変数の間の関連性の強さを表す

Quantifies the strength of association between two variables

$[-1, 1]$ の間で変動するよう標準化されている
standardized between -1 to 1

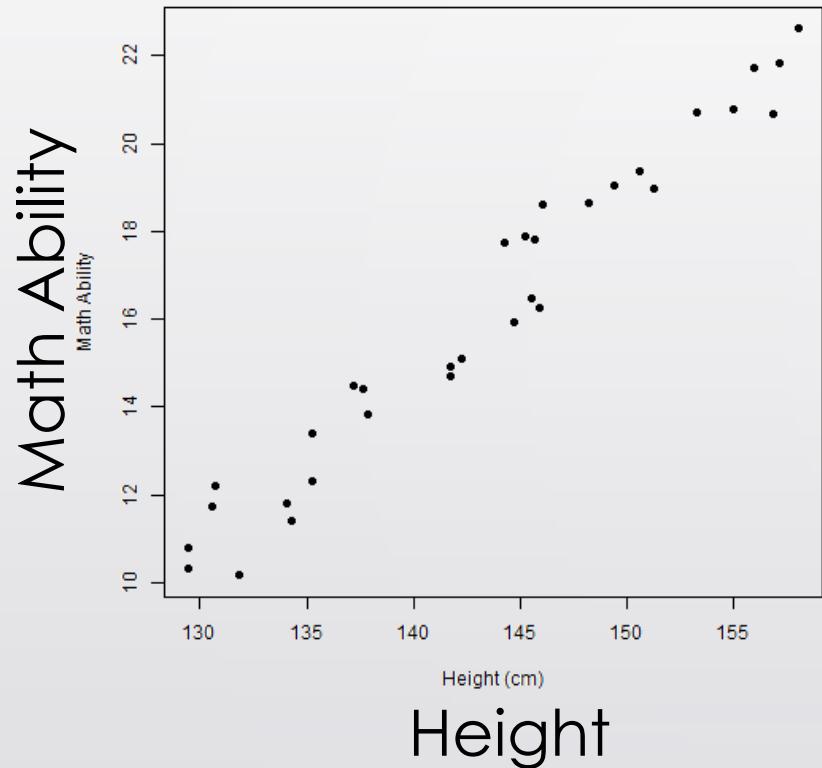
回帰直線 Linear Regression Line



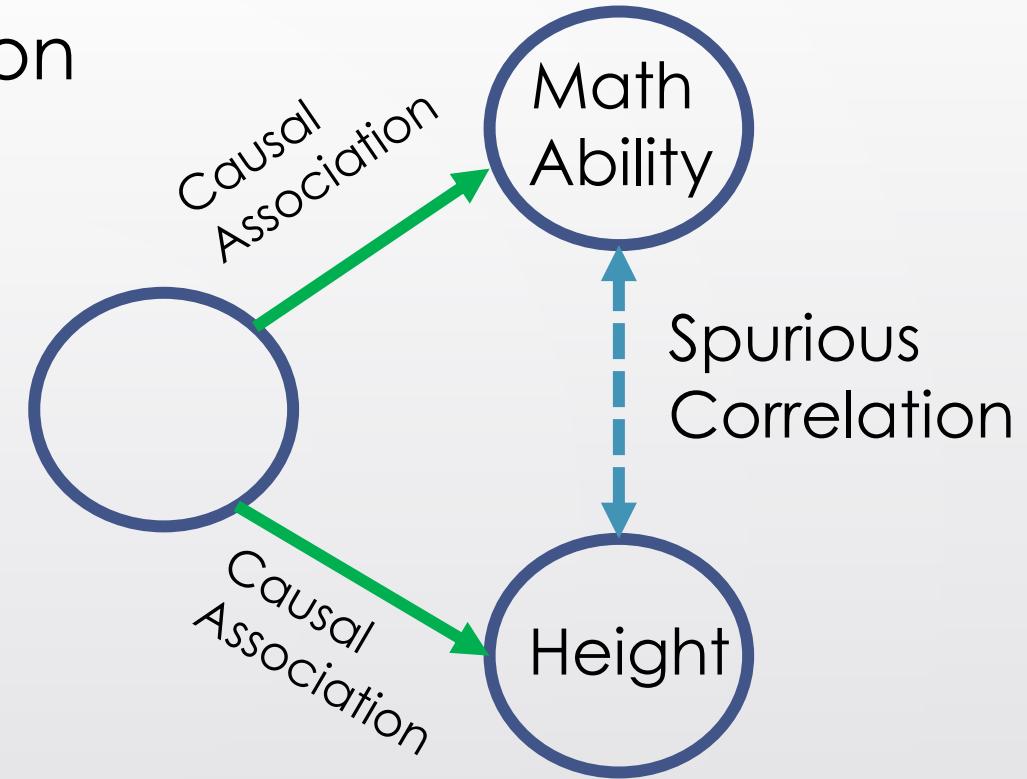
相関係数 r は、回帰直線の傾きを標準化した値

Correlational coefficient R
represents standardized slope
of linear regression line

疑似相関 Spurious Correlation



<https://hoxo-m.hatenablog.com/entry/20130711/p1>

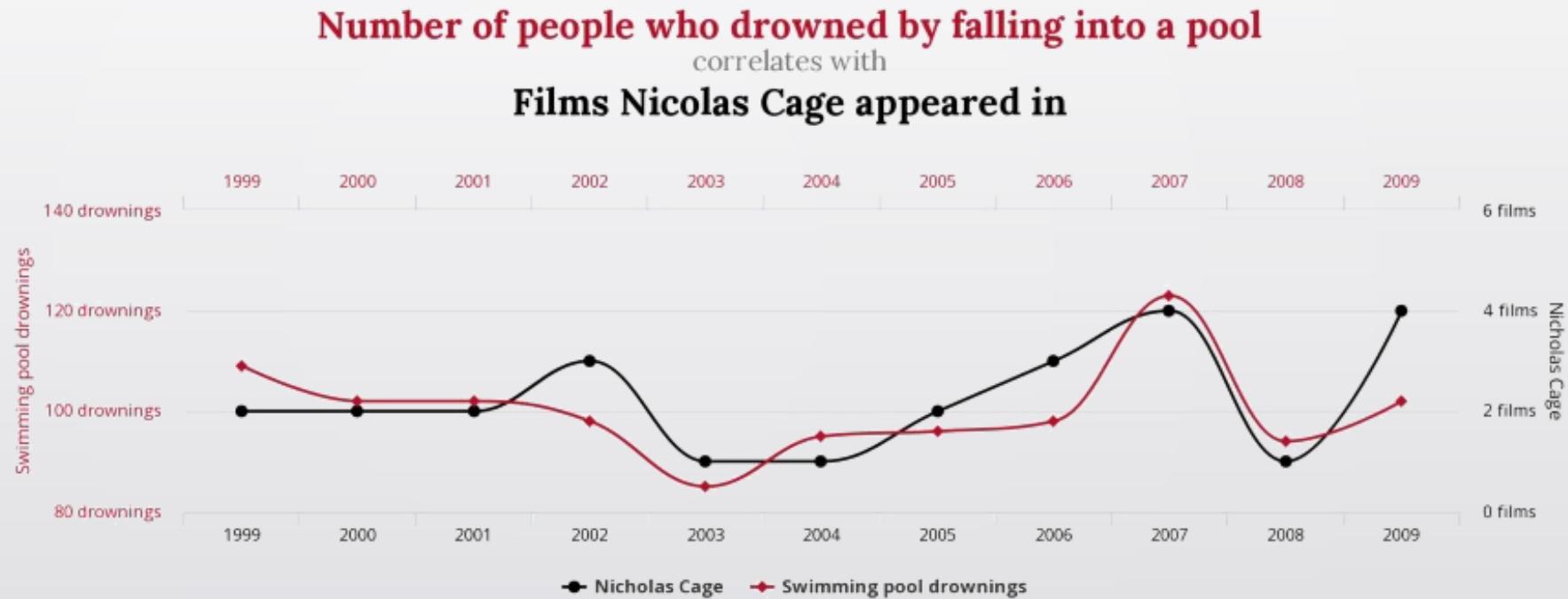
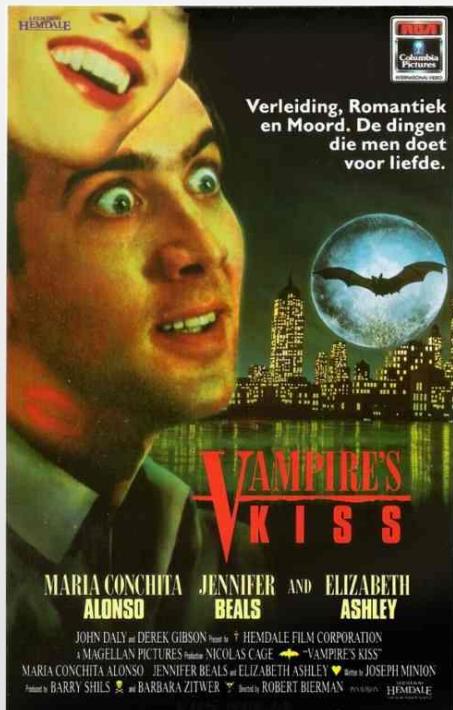


データ分析では、見かけの関係性に注意

Be aware of spurious association in any kinds of data analysis

ニコラス・ケイジと水難事故

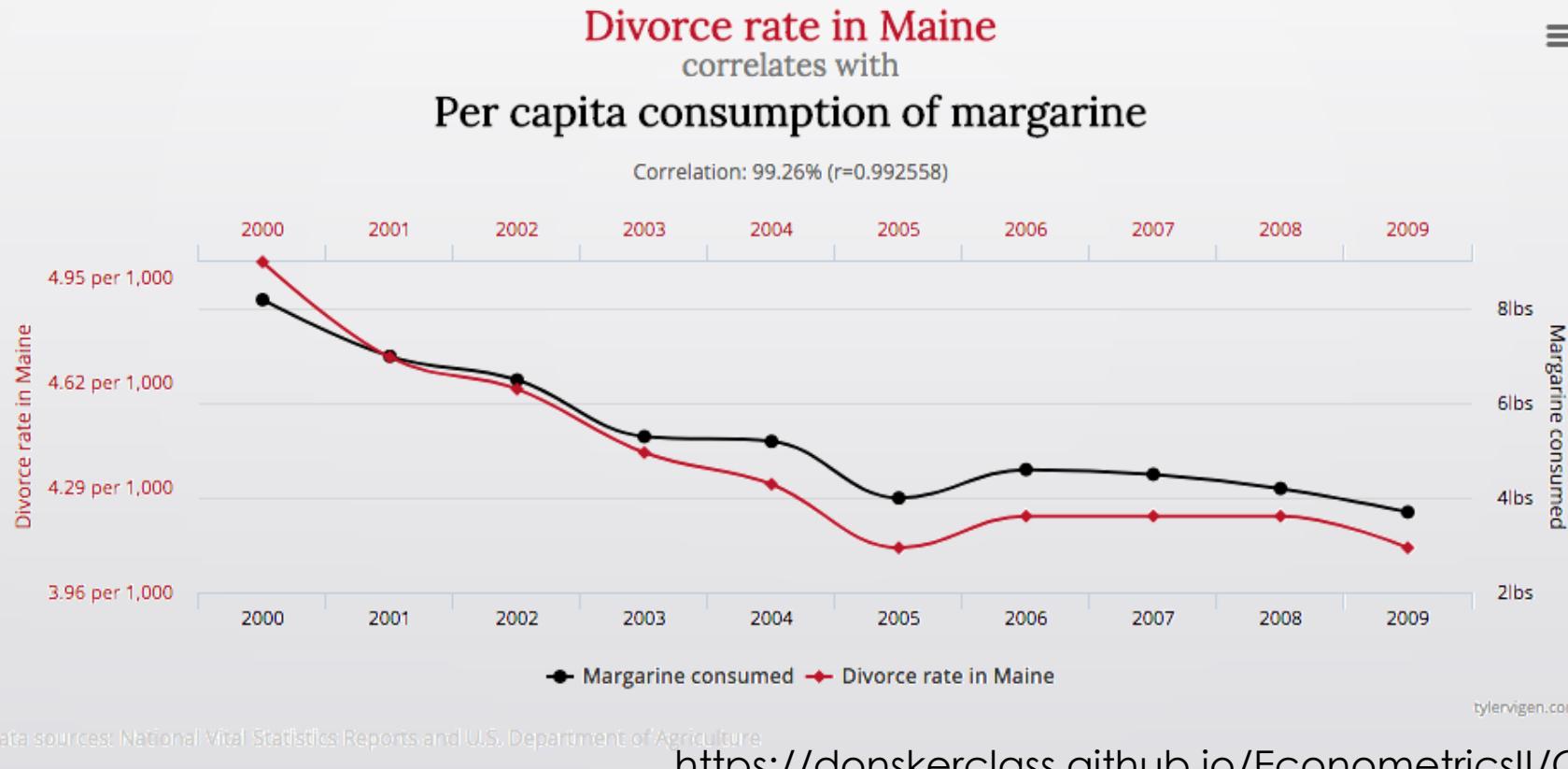
Nicholas Cage and Drowning Accident



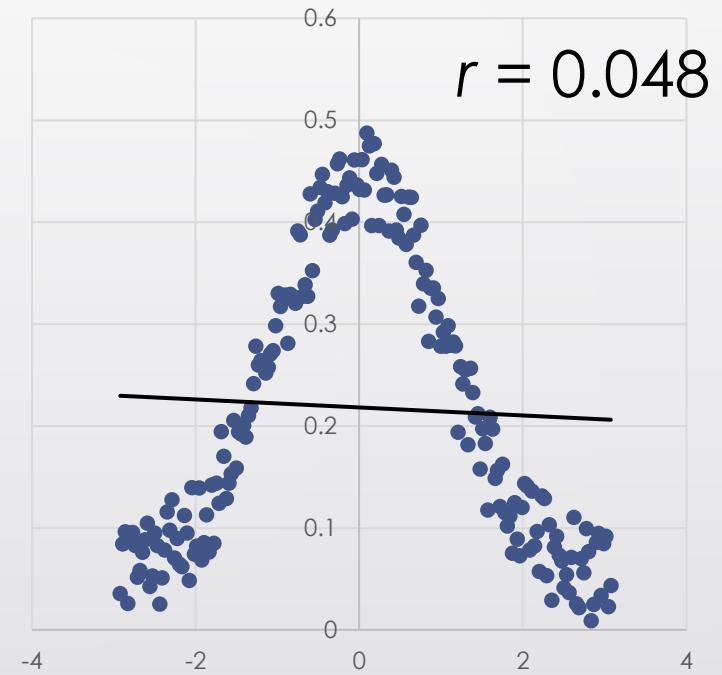
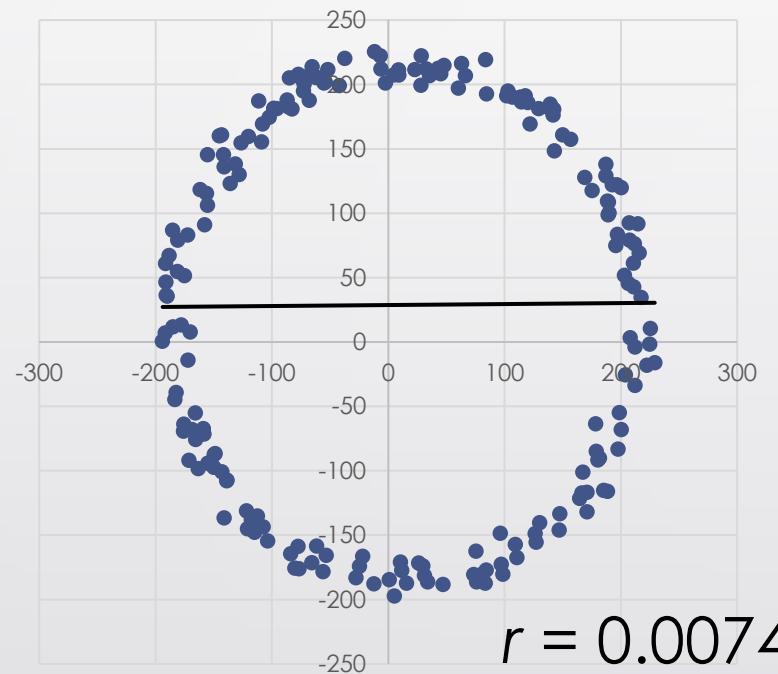
<http://tylervigen.com>

時系列データにおける疑似相関

Spurious Correlation in Time-Series data



非線形的な関係性 Nonlinear Association

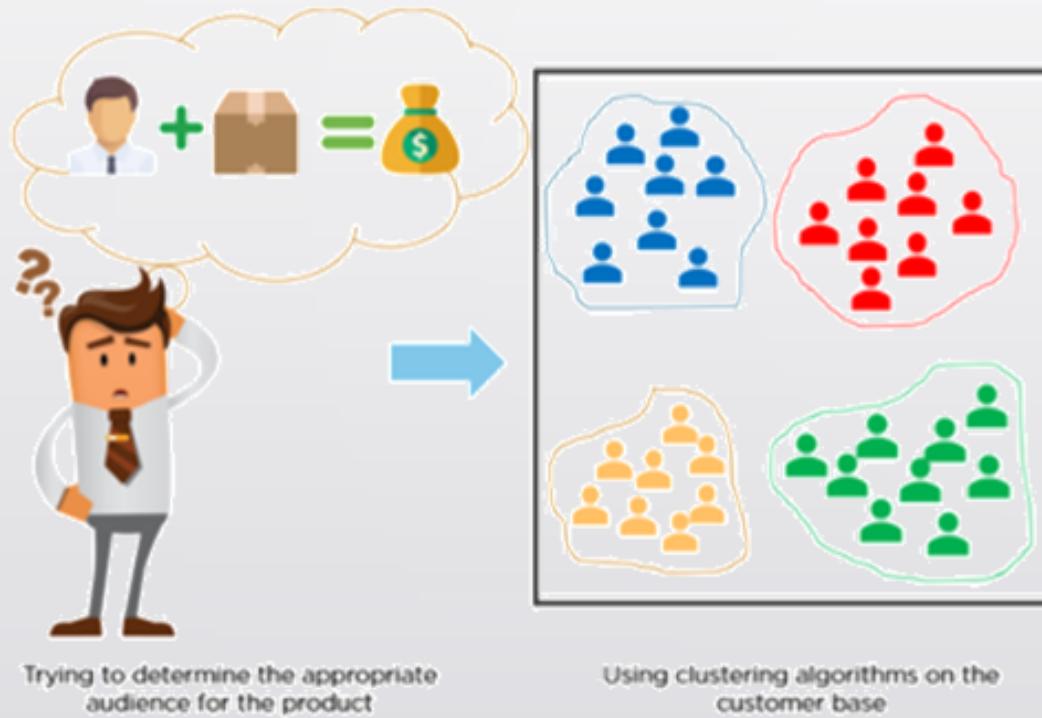


変数間の関係は、適切なモデルで評価する必要がある
Association between variables should be estimated by appropriate model

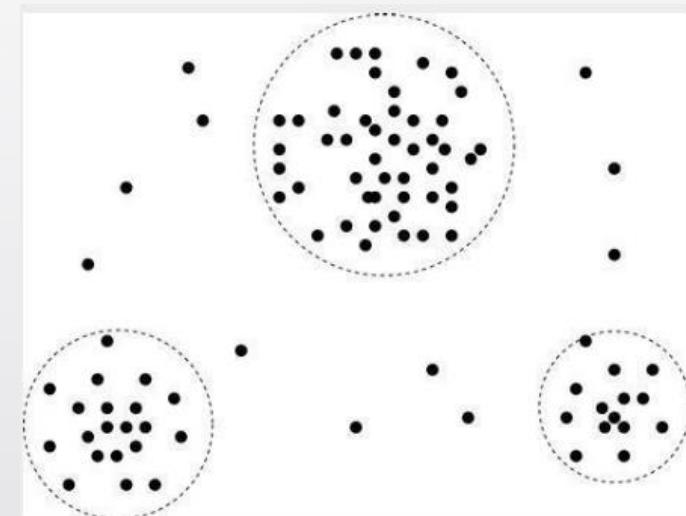


クラスタリングの一例

An example of clustering

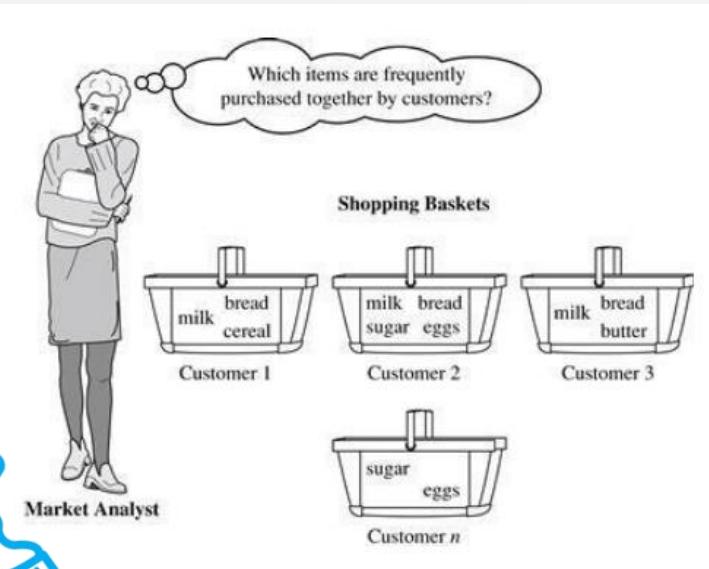


<https://www.quora.com/What-is-clustering>



性質が似たデータポイントの集団をみつける
Finding groups of data-points with homogenous characteristics

頻出パターン発見 Frequent Pattern Discovery



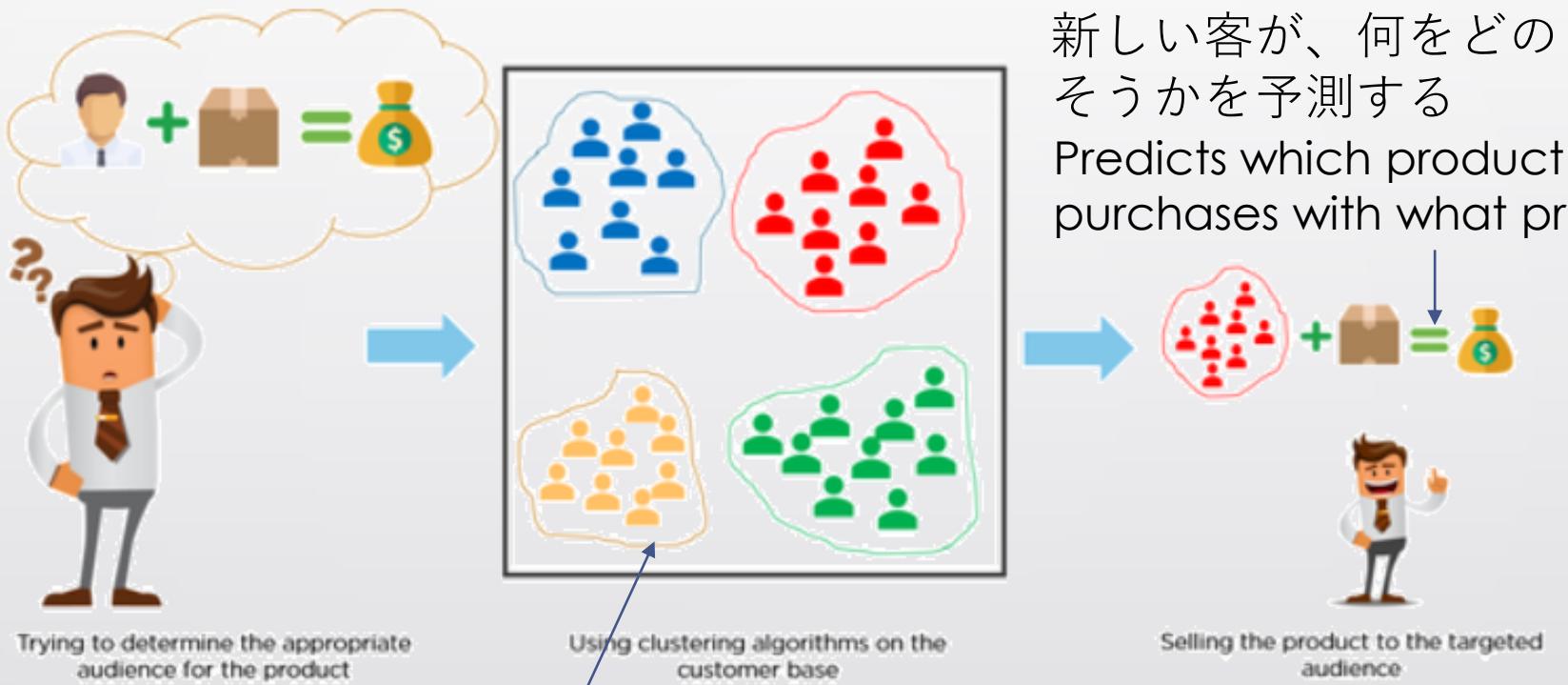
$age(X, "20..29") \wedge income(X, "40K..49K") \Rightarrow buys(X, "laptop")$
[support = 2%, confidence = 60%].

客の2%は20-29歳、かつ、年収が40K-49kである
この人達は、60%の確率でノートPCを購入する

2% of customers aged 20-29 yrs old and earns 40K to 49K dollars a year. They have purchased a laptop with the probability of 60%

カスタマーセグメンテーションによる予測

Prediction by Customer Segmentation

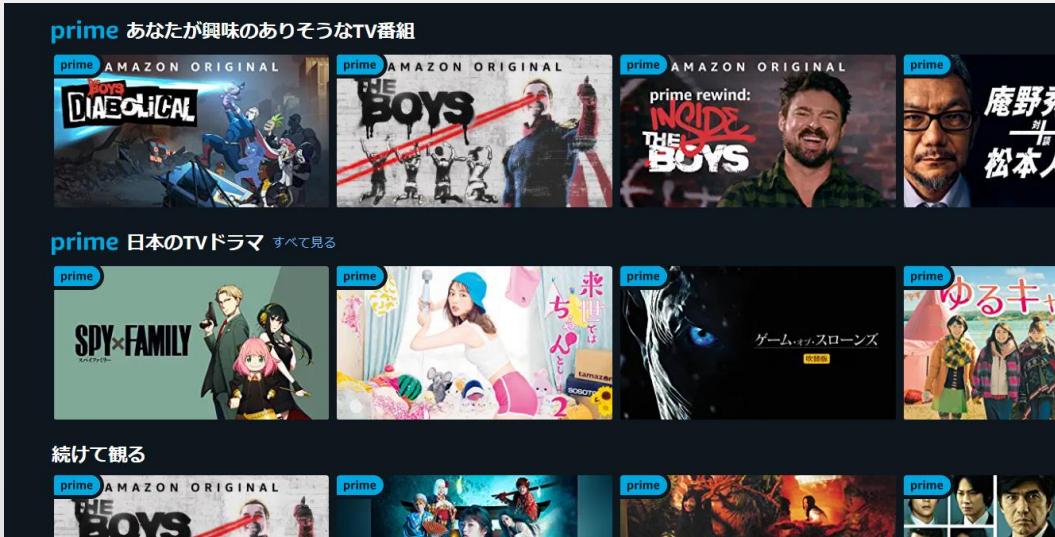


新しい客が、何をどのくらいの確率で買いかを予測する
Predicts which product a new customer purchases with what probability.

パーソナライゼーション Personalization

レコメンデーション

Recommendation



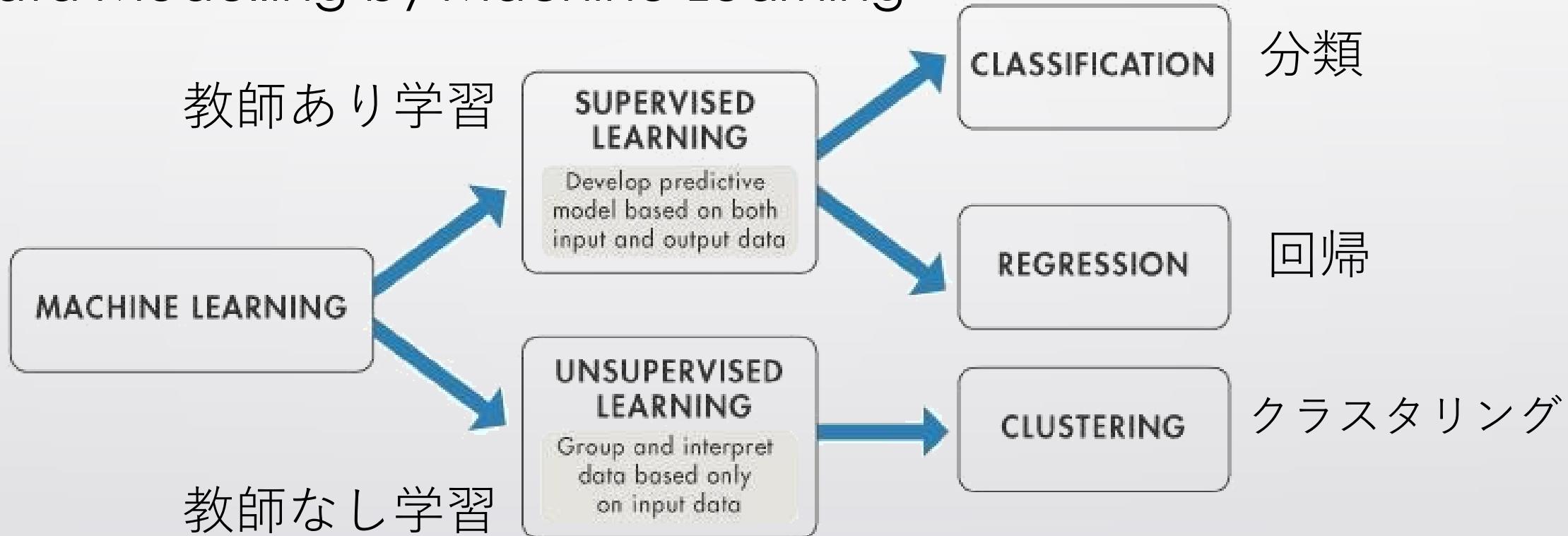
ターゲティング広告

Targeted Advertising



機械学習によるモデル化

Data Modelling by Machine Learning



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)