

kaggle入門ハンズオン

~Titanic 編~

1/26

自己紹介

kaggleとは

- 世界中で30万人以上のデータサイエンティストが登録している世界最大の**データ解析コンペサイト**
- 企業から様々なお題が出され、その優勝者（精度がもっとも高かった人orチーム）に、優勝したプログラムコードを提供するかわりに、賞金が贈られる
- 今回は、kaggleのチュートリアルにあたる**Titanic tutorial**を進めていく

kaggleに登録しよう

We use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies.

kaggle

Search

Competitions Datasets Notebooks Discussion Courses ...

Sign in

Register

Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.



REGISTER WITH GOOGLE

[Register with Email](#)

Predict Malicious Websites: XGBoost
Python notebook using data from [Malicious and Benign Websites](#) · 4 views

Version 6
6 commits
forked from Predict Malicious Websites: XGBoost w/ GPU

Notebook
Data
Log
Comments

This kernel has an XGBoost model that predicts whether a website is malicious or not.

```
In [1]: import numpy as np
import pandas as pd
import xgboost as xgb

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.utils.multiclass import unique_labels

data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

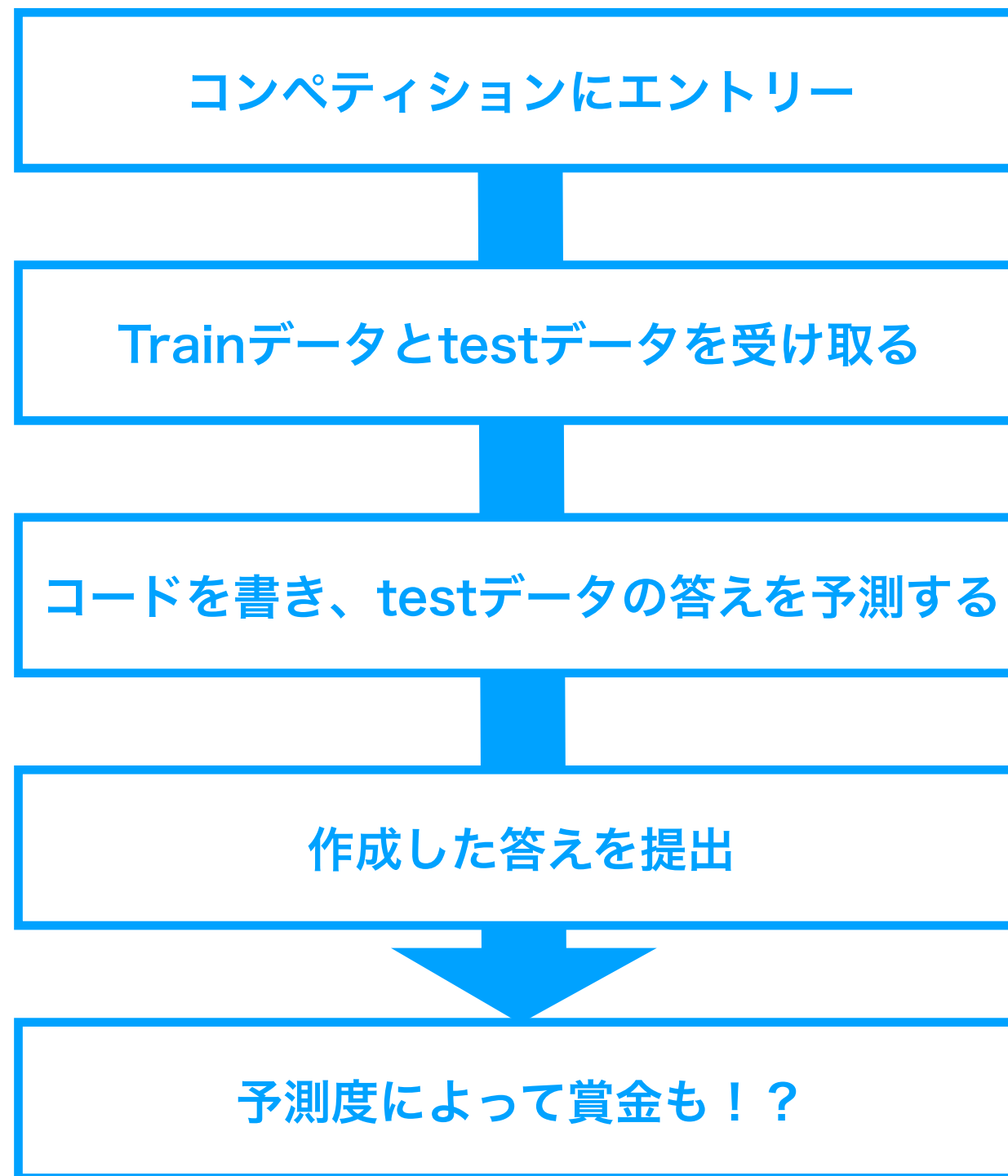
# peek @ dataframe
train.head()
```

Out[1]:

	url_length	number_special_characters	content_length	tcp_conversation_exchange	dist_remote_tcp_port	remote_ips	app_byt
344	37	9	162.0	1	0	1	66
77	26	6	NaN	0	0	0	0
923	51	10	231.0	7	1	2	769

Try Now

kaggleでの流れ

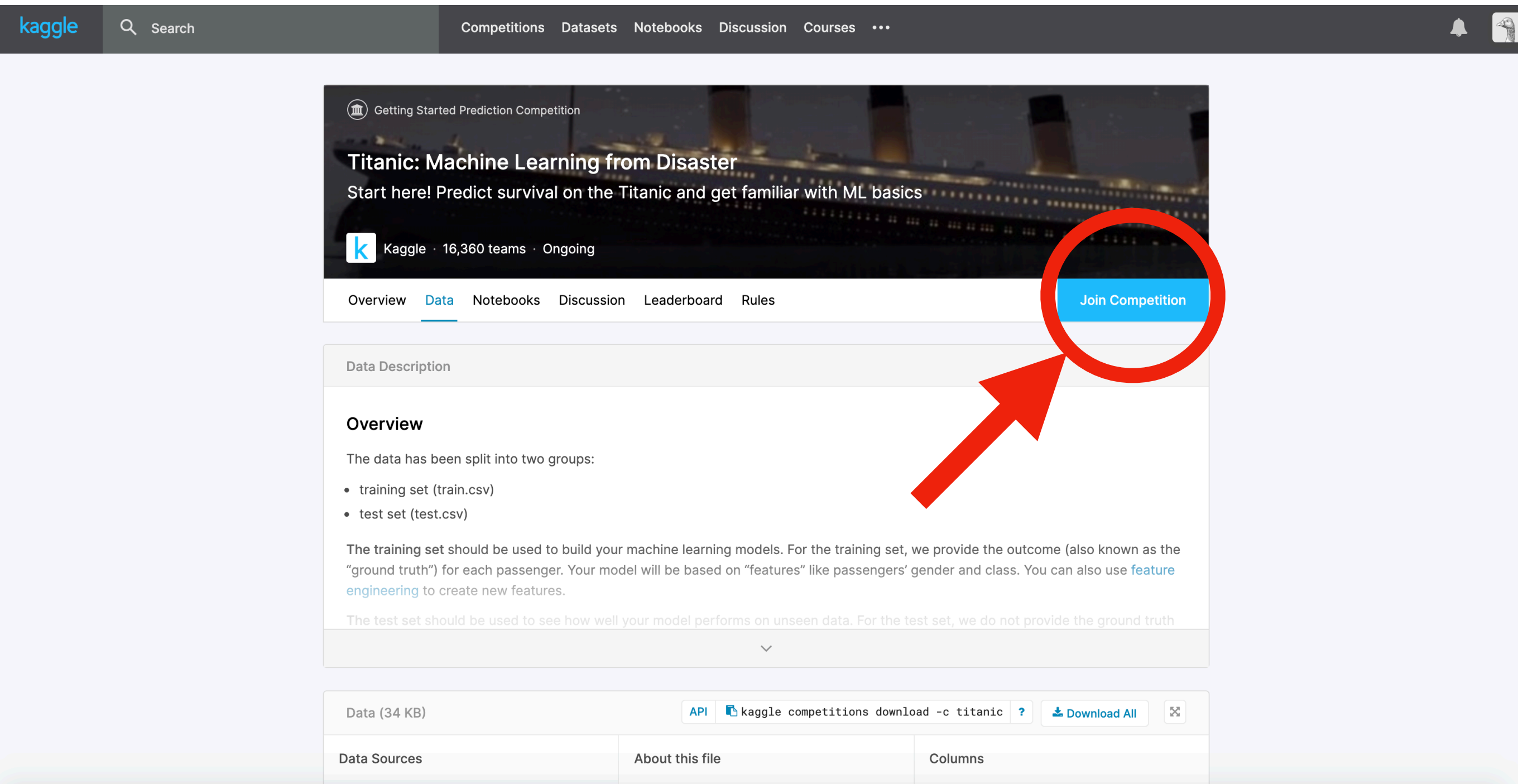


Titanic: Machine Learning from Disaster

1912年に起きた、かの有名な**タイタニック号沈没事件**を題材に、乗客の年齢、性別、社会階級ランク、などのデータから、**生死を予測する**、というもの

ちなみに、タイタニック号沈没事件は、**“若い乗客、女性の乗客から先に救命ボートに乗せた”**、などの史実があり、上記乗客のプロファイルデータからある程度生死を予測できることから、よくデータ分析の題材に用いられる

Titanic tutorial に エントリーしよう



The image shows the Kaggle website interface for the "Titanic: Machine Learning from Disaster" competition. The header includes the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Notebooks, Discussion, and Courses. The main content area features a large banner with the competition title and a "Join Competition" button, which is highlighted with a red circle and a red arrow. Below the banner, there is a "Data Description" section with an "Overview" tab selected. The overview text describes the data split into training and test sets and provides instructions on how to use the data for building machine learning models. At the bottom, there is a "Data (34 KB)" section with a table showing data sources, file information, and columns.

Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 16,360 teams · Ongoing

Overview **Data** Notebooks Discussion Leaderboard Rules

Join Competition

Data Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

The **training set** should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use [feature engineering](#) to create new features.

The **test set** should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth

Data (34 KB) [API](#) [kaggle competitions download -c titanic](#) [Download All](#)

Data Sources	About this file	Columns
	An example of what a submission file	

Data setとコードを入手

kaggleのdataタブから[train.csv](#)・[test.csv](#)をダウンロード

コードはGitHubのリンクから[TitanicHandsOn.ipynb](#)をダウンロード

[train.csv](#)・[test.csv](#)・[TitanicHandsOn.ipynb](#)を同じフォルダ（ディレクトリ）に入れる

データを見てみよう

与えられたtrain data

	Survived	Pclass		Name	Sex	Age	SibSp	Parch		Ticket	Fare	Cabin	Embarked
PassengerId													
1	0	3		Braund, Mr. Owen Harris	male	22.0	1	0		A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0		1	0		PC 17599	71.2833	C85	C
3	1	3		Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282		7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0		1	0		113803	53.1000	C123	S
5	0	3		Allen, Mr. William Henry	male	35.0	0	0		373450	8.0500	NaN	S

各変数の定義

Variable	Definition	Key
survival	生存	0 = No, 1 = Yes
pclass	チケットクラス	1 = 1st, 2 = 2nd, 3 = 3rd
sex	性別	
Age	年齢	
sibsp	Titanic号に乗船している兄弟の数	
parch	Titanic号に乗船している親・子の数	
ticket	チケット番号	
fare	乗船料金	
cabin	客室番号	
embarked	乗船港	C = Cherbourg, Q = Queenstown, S = Southampton

今回の目標は？

Train data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Train dataのsurvived変数と他の変数の関係を機械学習を用いて学習し、**test dataの各乗客のsurvived変数を予測する**

Test data

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

→survived?

→survived?

→survived?

→survived?

→survived?

データラングリング (data wrangling)

- 生データはそのままでは使いづらいので、データを整えていく
- 今回は、乗船クラス・性別・年齢・兄弟の数・親子の数・料金を採用する（本来は、どの変数を使用するか決めるために、可視化したり統計解析したりするが、割愛）
- 性別の名義尺度を数値にする
- 後に操作しやすいように配列データの形式を変える
- 欠損値を補う
- 正規化する
- etc...

結果4つのデータを得る

x

Pclass	Sex	Age	SibSp	Parch	Fare
0.827377	0.737695	-0.592481	0.432793	-0.473674	-0.502445
-1.566107	-1.355574	0.638789	0.432793	-0.473674	0.786845
0.827377	-1.355574	-0.284663	-0.474545	-0.473674	-0.488854
-1.566107	-1.355574	0.407926	0.432793	-0.473674	0.420730
0.827377	0.737695	0.407926	-0.474545	-0.473674	-0.486337

t

PassengerId	Survived
1	0
2	1
3	1
4	1
5	0

学習

test

Pclass	Sex	Age	SibSp	Parch	Fare
0.873482	0.755929	0.344284	-0.499470	-0.400248	-0.498258
0.873482	-1.322876	1.334655	0.616992	-0.400248	-0.513125
-0.315819	0.755929	2.523099	-0.499470	-0.400248	-0.464940
0.873482	0.755929	-0.249938	-0.499470	-0.400248	-0.483317
0.873482	-1.322876	-0.646086	0.616992	0.619896	-0.418323

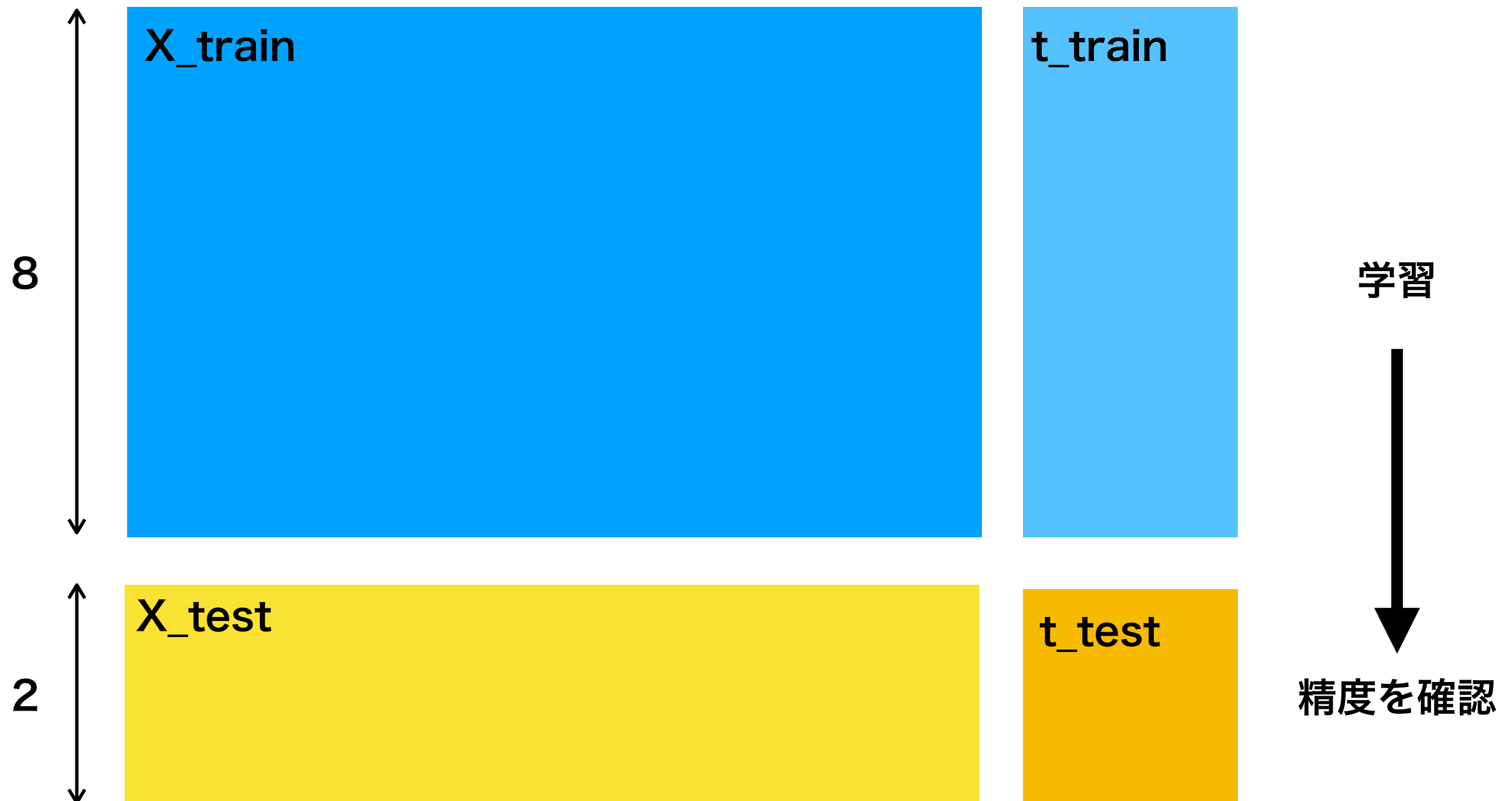
予測

passenger_id

PassengerId
892
893
894
895
896

学習の前に

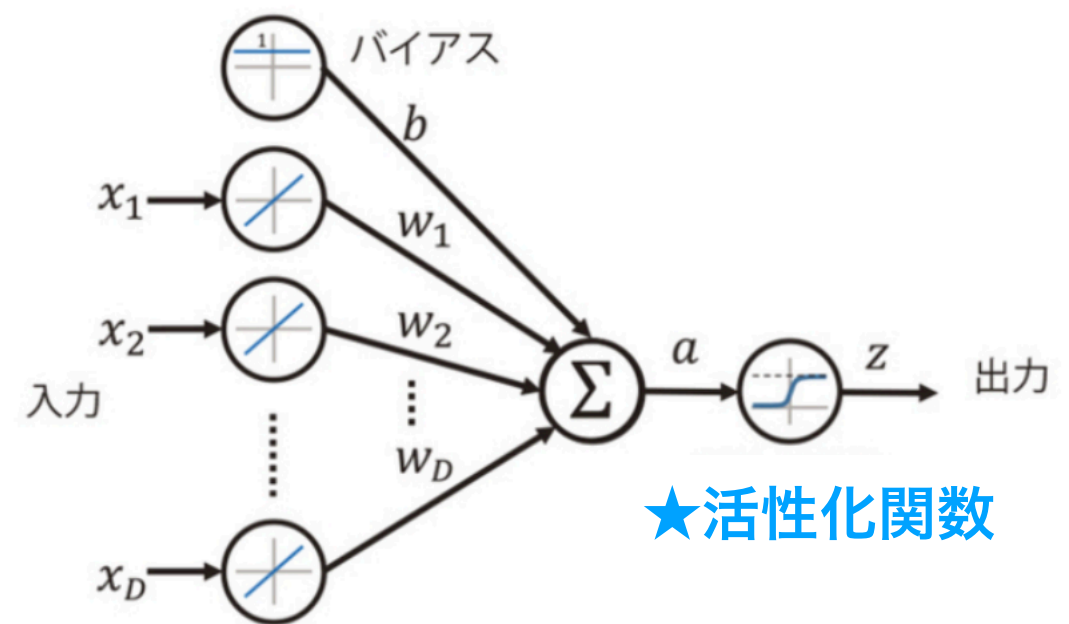
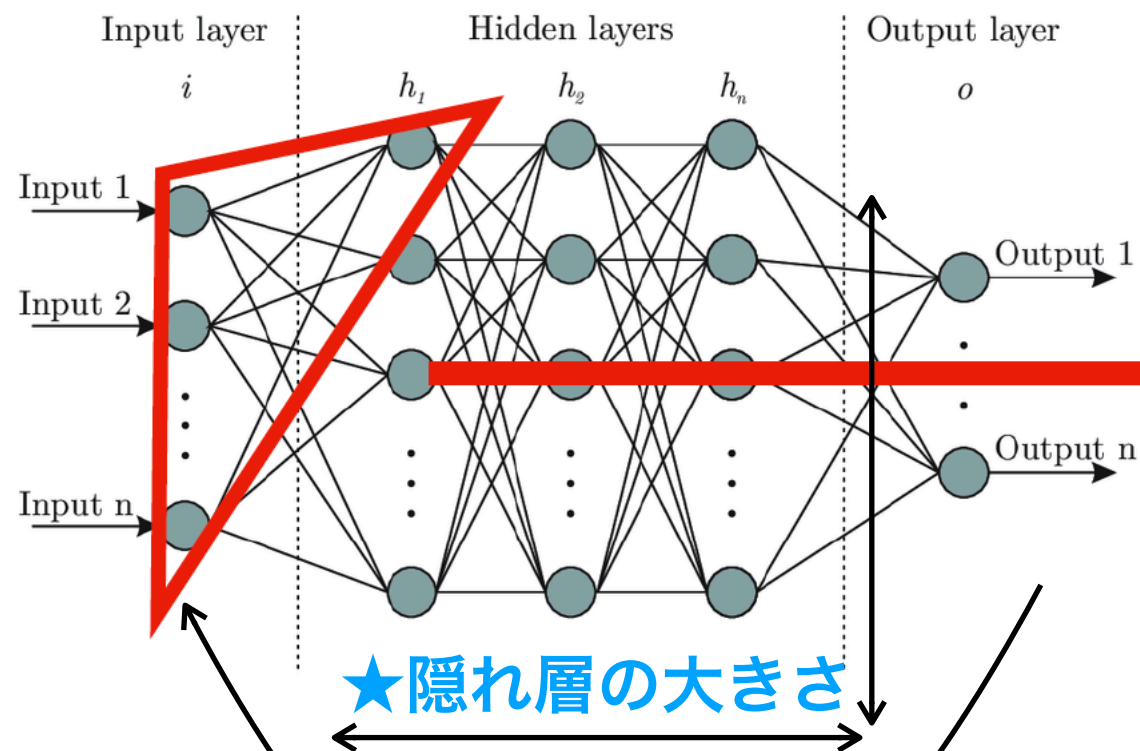
与えられたtrainデータを8:2の比率で、trainデータとtestデータに分ける。



ハイパーパラメータとは

- 学習プロセスを決定するパラメータのこと。学習中に更新されないので、初めに決定しないといけない。例えるなら

ex) Neural Network



★反復回数

★ペナルティ etc.

Validation (検定)

先ほど、trainデータと精度測定用のtestデータに分けたが、最も優れたハイパーパラメータを決定するために、trainデータにvalidation(検定)領域を設けハイパーパラメータ更新用testデータとする必要がある



K Fold Cross Validation

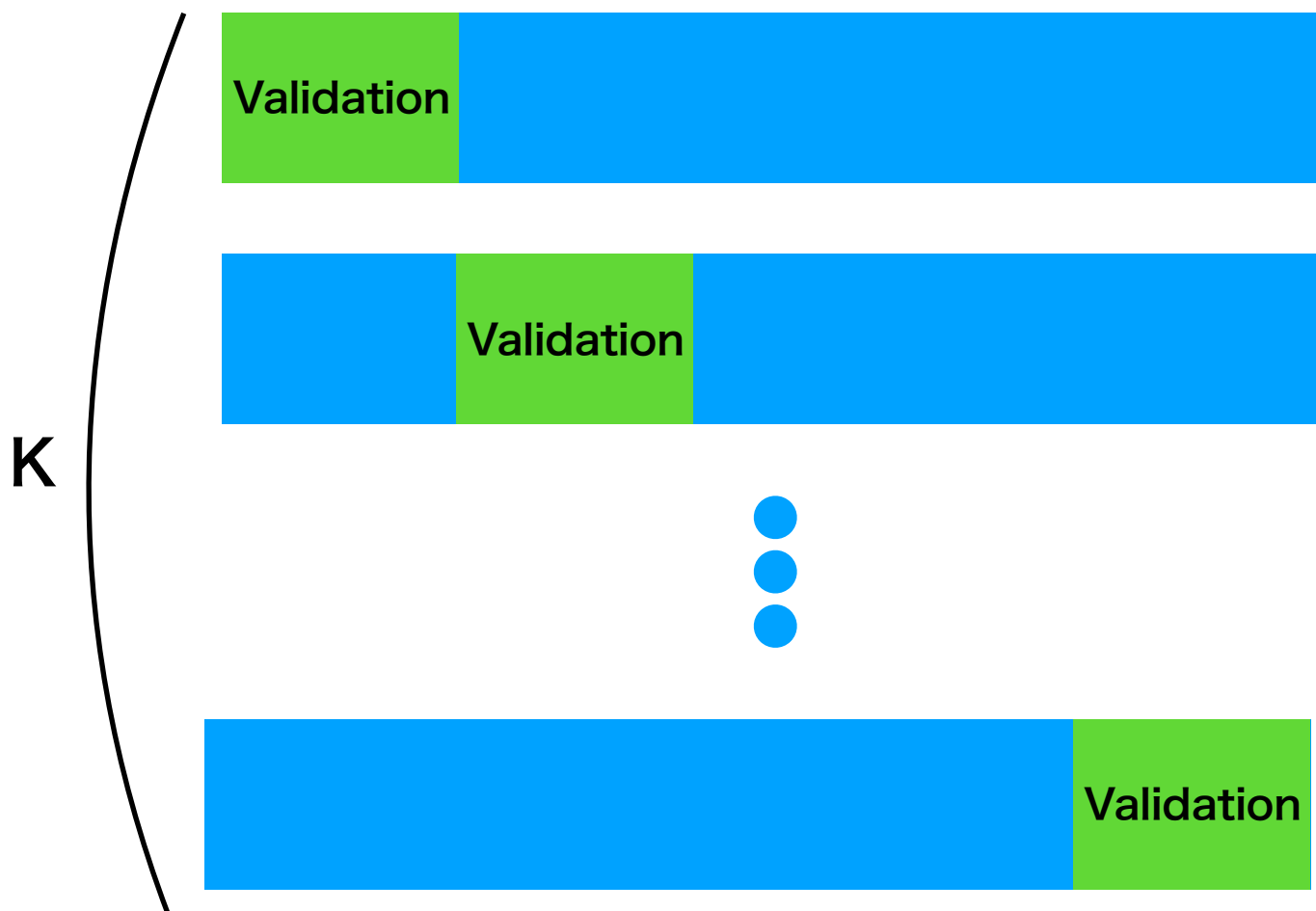
データをk個のブロックに分ける。これを分割 (fold) という。

最初の分割1 を test set、残りの分割2~k を training set とし、モデルの学習と評価を行う。

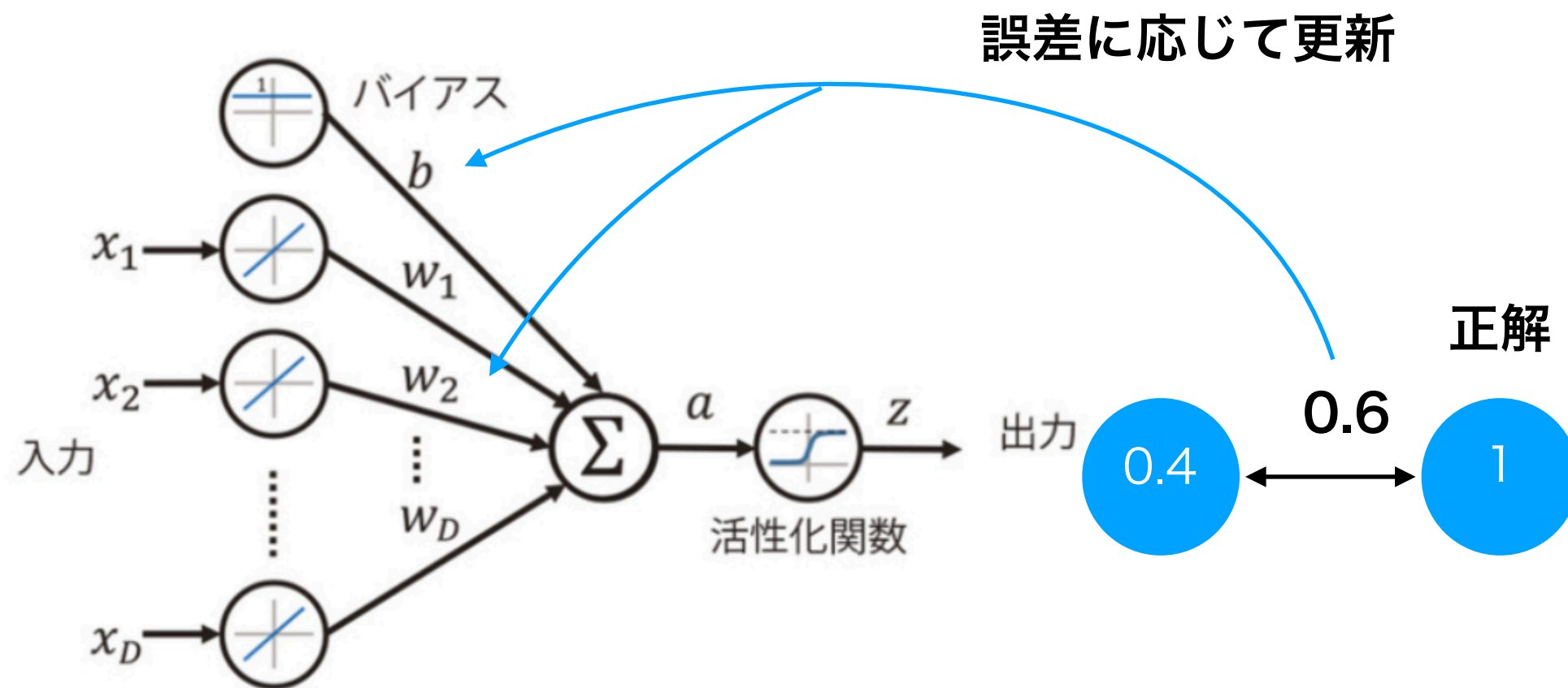
分割2 を test set、残りの分割1、3~k を training set として、モデルの学習と評価を行う。

この過程を、分割3, 4, ..., k を test set として繰り返す。

得られたk個の精度の平均値をモデルの評価値とする。

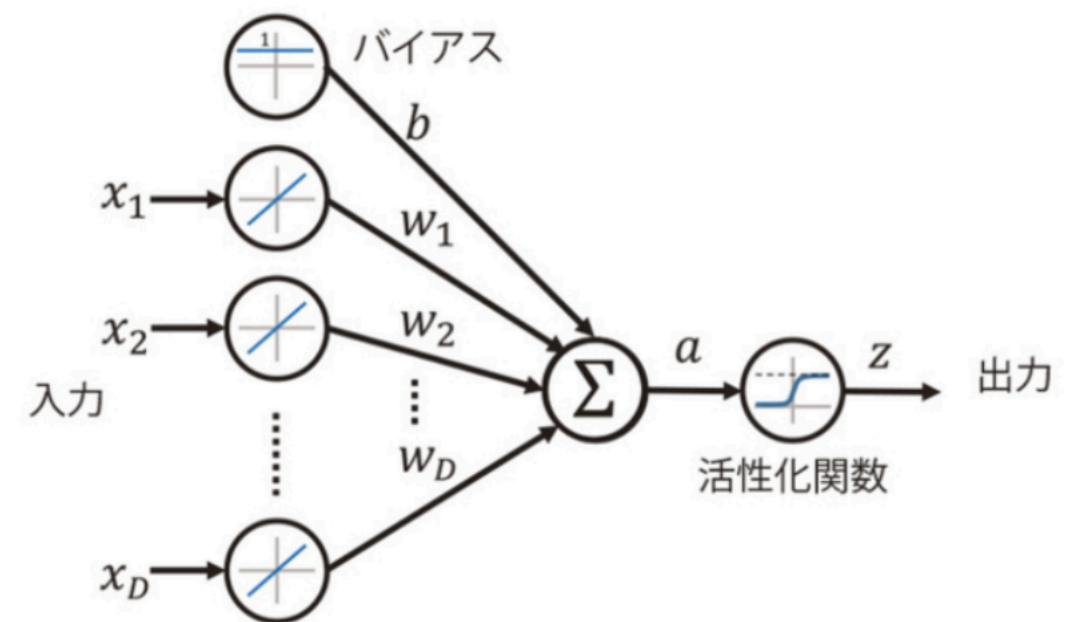
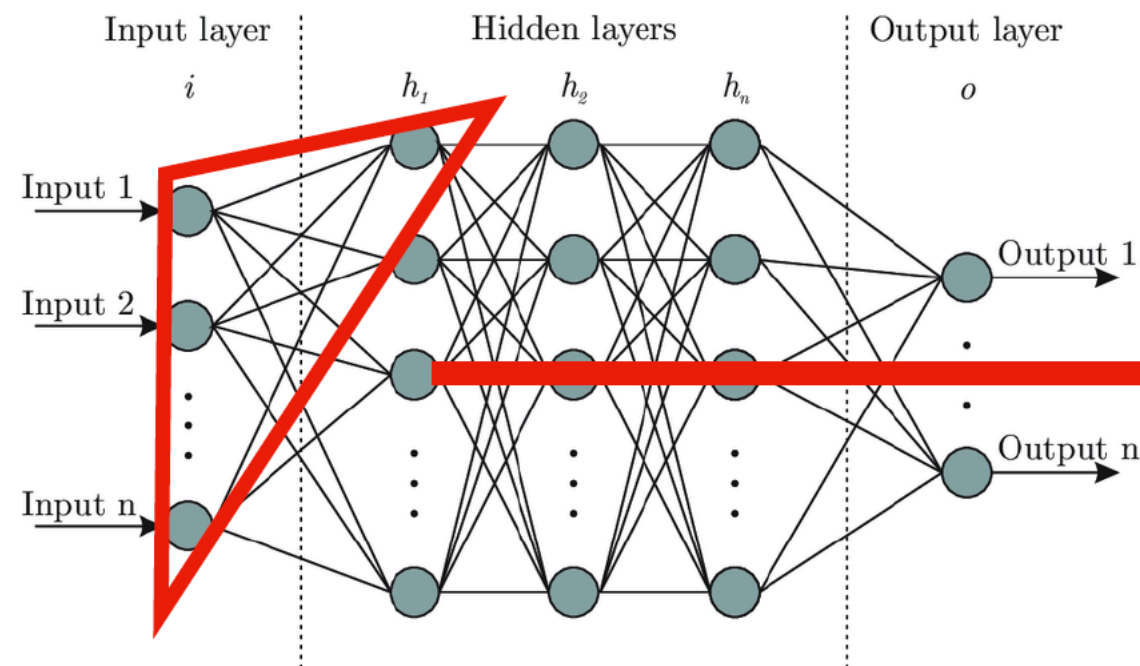


ロジスティック回帰 (Logistic regression)



多層パーセプトロン

(Multilayer perception classifier: MLPC)



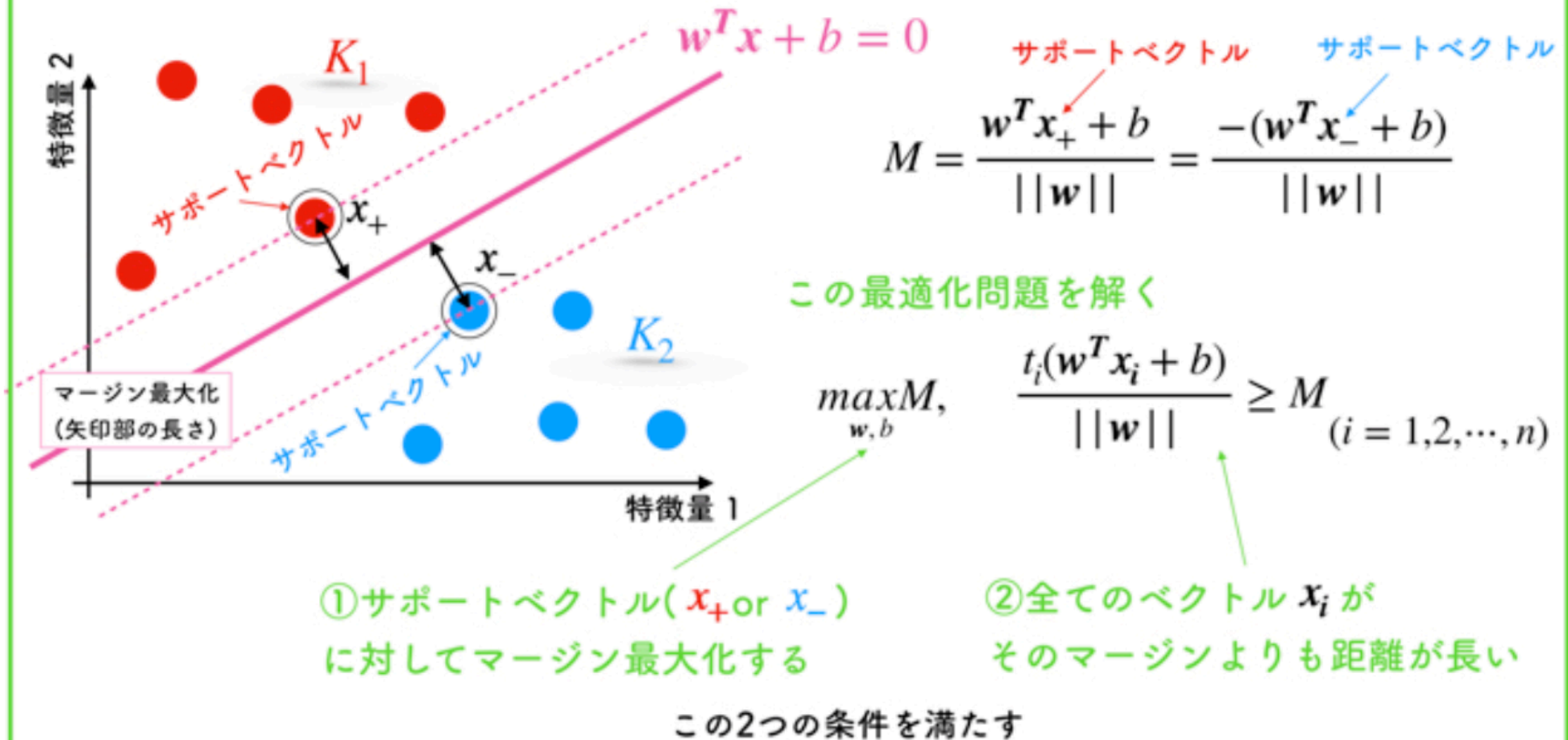
Logistic regressionと同じ構造のものを大量に並べて、全体としての出力と答えとの誤差の平均から、全パラメータを更新していく。

隠れ層を多くしていく（深くしていくと）Deep learningとも言える

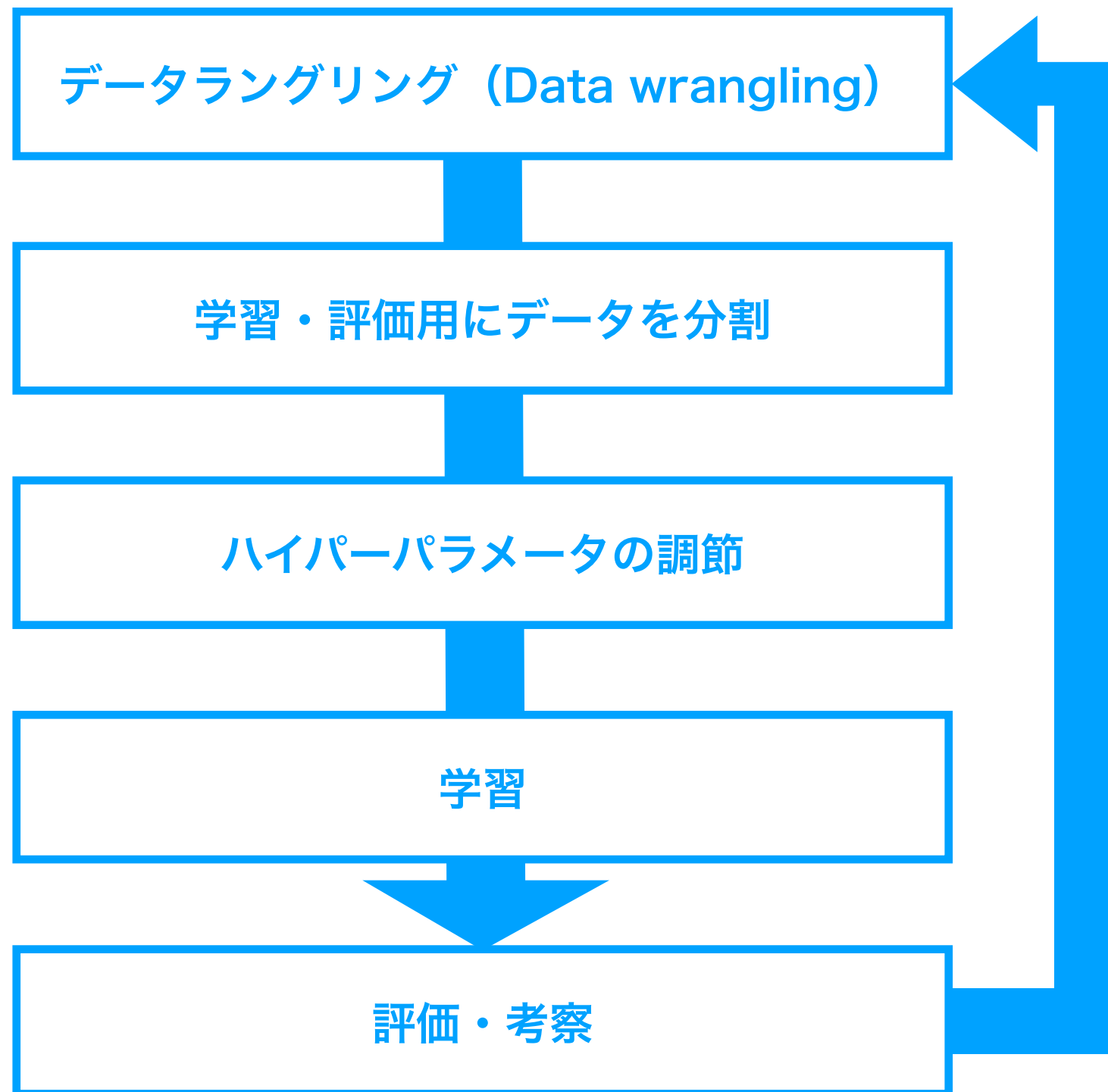
Support vector machine: SVM

ネットの解説を拝借

2つのクラスを分ける超平面に最も近いデータ(サポートベクトル)への距離を考える



機械学習の流れ



Hands-on

色々といじってみてください

Titanicコンペに提出しよう

- コードやデータセットを入れたフォルダの中に各学習結果のcsvファイルがあるはず
- Submit predictionタブから提出
- 自分の順位を確認しよう

The screenshot shows the submission interface for the Titanic competition. At the top, a navigation bar includes links for Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and a circled **Submit Predictions** link. Below the navigation bar, a status message states: "You have 10 submissions remaining today. This resets a day from now (00: 00 UTC)." The main content area is divided into two steps. **Step 1: Upload submission file** features a large dashed box with a circled upload icon (an arrow pointing up from a document). Below this box, two columns of text provide instructions: "File Format" (submission should be in CSV format, or zip/gz/rar/7z archive) and "Number of Predictions" (expect 418 prediction rows with a header row, with a link to a sample submission file). **Step 2: Describe submission** includes a rich text editor with a toolbar (undo, redo, bold, italic, link, quote, code, list, table, image, emoji) and a text input field with the placeholder "Briefly describe your submission". At the bottom of the form is a circled blue button labeled **Make Submission**.

アンケートにご協力ください

https://docs.google.com/forms/d/1bwf_sY9cS0nZR0m0qEo8IpEM-y_dYyGdoepX314OP0k/edit

