実践！機械学習のプロセス
徹底解説

Part3-2:データの基本情報を把握する

16T4017N
高久雅史

1. データセットの入手とアップロード

Kaggleから、乗客名簿のデータセットを入手し、
Google Colab上へアップロード

```
#Google Colabへファイルをアップロード
from google.colab import files
uploaded = files.upload()
```

<ファイル選択>から’train.csv’と’test.csv’を選択して、
アップロード

2. ライブラリのインポート

```
#データ処理ライブラリ
import pandas as pd
import numpy as np

#データ可視化ライブラリ
import matplotlib.pyplot as plt

#機械学習ライブラリ
from sklearn.preprocessing  import LabelEncoder
from sklearn.metrics import accuracy
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
```

3. CSV形式ファイルから
データフレーム形式に変換

Pands:数表および時系列データを操作するためのデータ構造を
提供するライブラリ

```
#訓練データとテストデータをread_csv()で変換
train_set = pd.read_csv('train.csv')
test_set = pd.read_csv('test.csv')
```

3. CSV形式ファイルから データフレーム形式に変換

訓練データとテストデータの最初2行だけ見てみると

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

図1. train_set.head(2)

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S

・Survivedの項目はない
・モデルを利用して
予測を行うのが目的

図2. test_set.head(2)

特徴量に関する補足

●Pclass:旅客等級
　値1,2,3の数値で表現し、1から順に乗客の経済的地位が高い。

●Age:年齢
　乗客の年齢を表す。小数値で表されるデータは推定年齢。

●Embarked:出発港
　・C：シェルプール
　・Q：クイーンズタウン
　・S：サウサンプトン