機械学習プロセス
徹底理解

PART-1, 2

16t4032n　田中裕隆

PART-1　環境設定

- 使用するデータセット　タイタニック号乗客乗員名簿データ
  - Kaggleに無料会員登録して使える
- 実行環境　Webサービス　Google Colab
  - Jupyter Notebook が使える
- 言語　Python3

PART-1　ライブラリ

- numpy　数値計算を効率的に行うライブラリ
- pandas　データ解析を支援するライブラリ
- matplotlib　numpyのためのグラフ描写ライブラリ
- scikit-learn　多数の手法／アルゴリズムを備えた機械学習ライブラリ

PART-2　機械学習の基礎知識

- 機械学習の一般的な定義の一つ

- 機械学習とは人工知能アプリケーションの一つであり，
  明示的にプログラミングされることなく，
  データ(経験)から学習し改善を行うことが可能なシステム

PART-2　機械学習の基礎知識

- 機械学習の分類
  - 教師なし学習
  - 教師あり学習

- 今回は教師あり学習
  - 訓練データに正解ラベル(教師)を付与して学習を行う

PART-2　特徴量とターゲット

- 特徴量
  - 予測する値を説明する変数

- ターゲット
  - 特徴量で説明される変数

PART-2　特徴量とターゲット

- 例　マンションの家賃を予測

- 特徴量
  - 築年数
  - 専有面積

- ターゲット
  - 家賃

PART-2　特徴量とターゲット

- タイタニック号データ

- 特徴量
  - 性別
  - 年齢
  - etc...

- ターゲット
  - 生還　した or しなかった

PART-2　訓練データ　テストデータ

- データセットを2種類に分ける

- 訓練データ
  - 機械学習モデルの学習に使用するデータ

- テストデータ
  - 機械学習モデルの予測に使用するデータ
  - 機械学習モデルの精度を評価

PART-2　Kaggle

- データサイエンティスト／機械学習エンジニアのための
  コミュニティWebサイト

- Kaggleの特徴
  - データ分析競技
  - データセット
  - 初心者のための豊富なリソース