Pythonで行う機械学習プログラム

Lesson_0 Python の基礎: インストール、主要ライブラリ操作（pandas, numpy, matplotlib）

Lesson_1 データセット取り扱いと機械学習基礎

Lesson_2 Python: 機械学習（分類）評価と応用

Lesson_3 Python: 機械学習（回帰）評価と応用

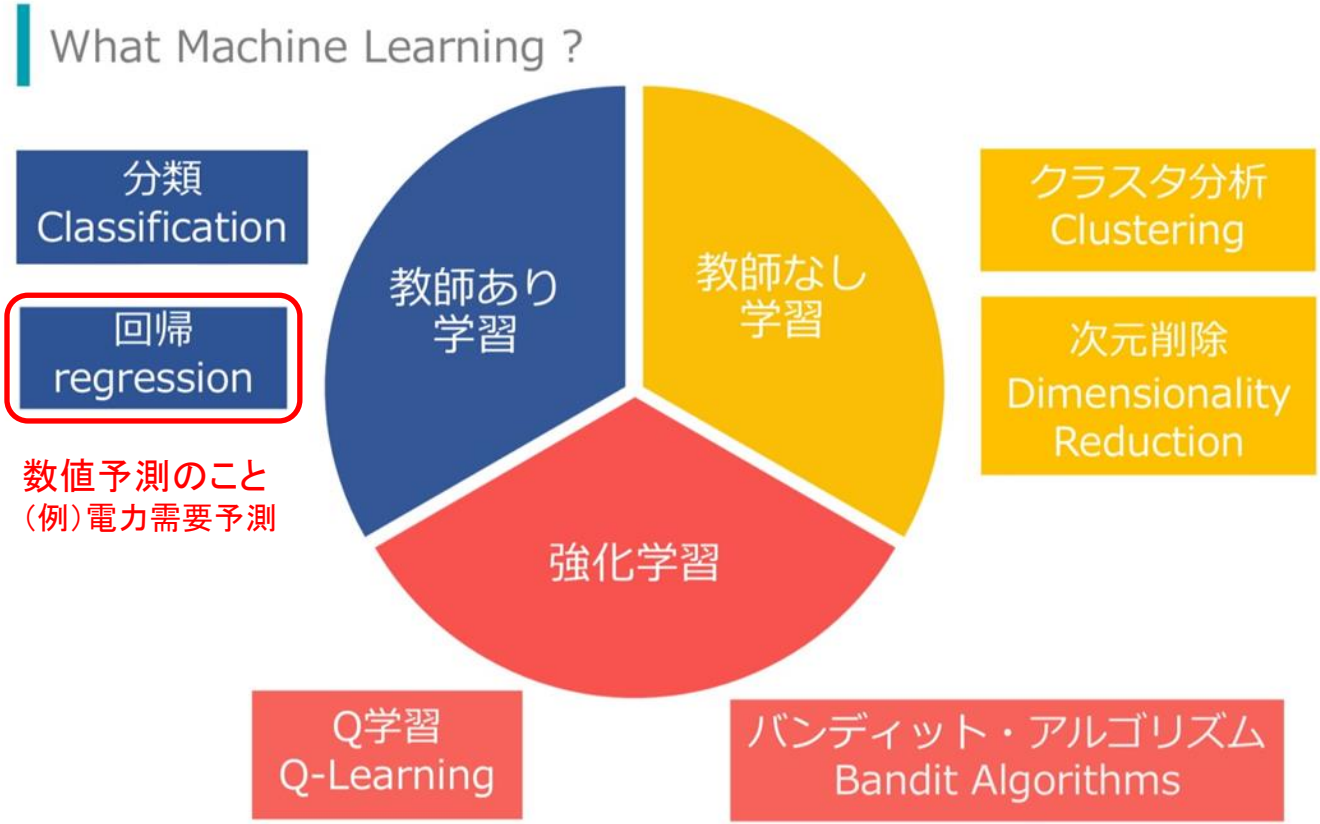
Lesson_4 Python: 機械学習 前処理、最適化

Lesson_5 Python: OpenCV による画像処理

Lesson_6 Python: DL 画像判別（１） MLP と CNN

Lesson_7 Python: DL 画像判別（２） 転移学習、学習データの再利用、結果の保存

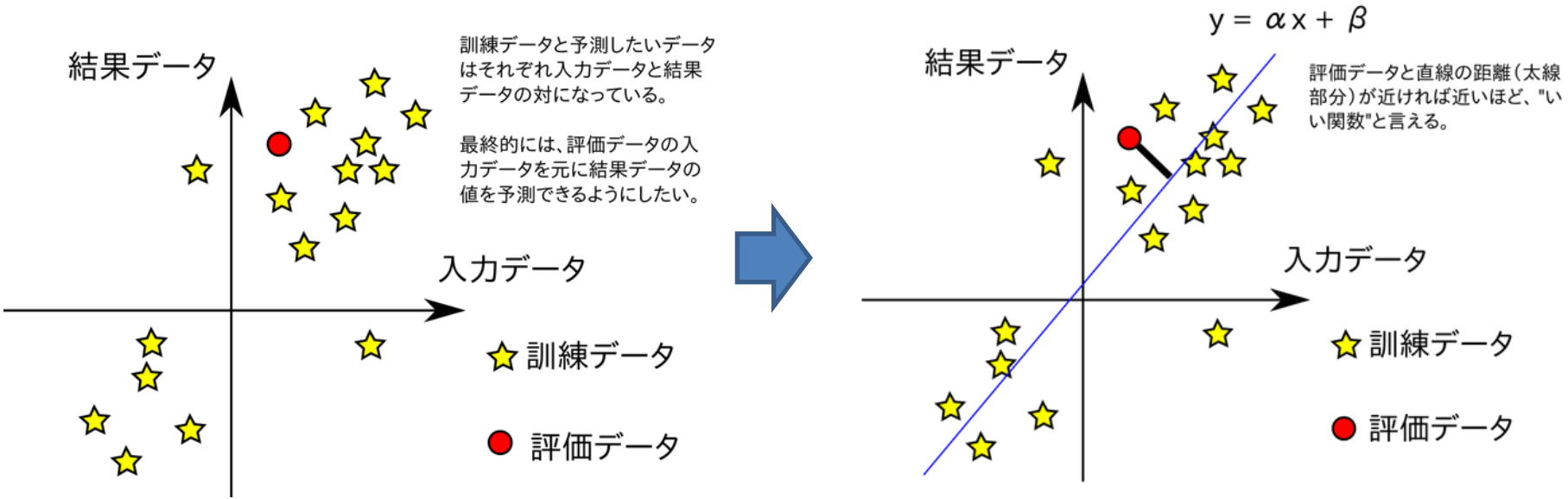
機械学習概要



今回は回帰を学びます

数値予測のこと
（例）電力需要予測

回帰とは？



教師（訓練）データをもとに学習した数式で、未知のデータの数値を予測する

回帰に用いられる主なアルゴリズム

・関数で回帰を行う
　　・線形回帰：１次関数で回帰を行う　→　（例）一般化線形回帰

・分類と同じアルゴリズムでも回帰にも使えるものが多くある
　　・ツリー系　→　（例）ランダムフォレスト
　　・SVM
　　・Neural Network系　→（例）Neural Networks, Deep Learning

KaggleとGitHub

Kaggle（カグル）は企業や研究者がデータを投稿し、世界中の統計家やデータ分析家がその最適モデルを競い合う、予測モデリング及び分析手法関連プラットフォーム及びその運営会社である。モデル作成にクラウドソーシング手法が採用される理由としては、いかなる予測モデリング課題には無数の戦略が適用可能であり、どの分析手法が最も効果的であるか事前に把握することは不可能であることに拠る。2017年3月8日、**Google**はKaggle社を買収すると発表した。（Wikipedia）

GitHub（ギットハブ）は、ソフトウェア開発のプラットフォームであり、ソースコードをホスティングする。コードのバージョン管理システムにはGitを使用する。Ruby on RailsおよびErlangで記述されており、アメリカのカリフォルニア州サンフランシスコ市に拠点を置くGitHub社によって保守されている。2009年のユーザー調査によると、GitHubは最もポピュラーなGitホスティングサイトとなった。2018年に**マイクロソフト**による買収が発表されている。（Wikipedia)

今回用いるデータセット（Kaggle Datasetsより）

https://www.kaggle.com/wkirgsn/electric-motor-temperature

ambient	coolant	u_d	u_q	motor_speed	torque	i_d	i_q	pm	stator_yoke	stator_tooth	stator_winding
−0.75214297	−1.1184461	0.3279352	−1.2978575	−1.2224282	−0.2501821	1.0295724	−0.24586003	−2.522071	−1.8314217	−2.0661428	−2.0180326
−0.77126324	−1.1170206	0.3296648	−1.2976865	−1.2224293	−0.2491333	1.029509	−0.24583231	−2.5224178	−1.8309687	−2.0648587	−2.0176313
−0.78289163	−1.1166813	0.3327715	−1.3018217	−1.2224278	−0.24943107	1.0294477	−0.24581794	−2.5226731	−1.8304	−2.064073	−2.0173435
−0.78093535	−1.1167642	0.3336999	−1.301852	−1.2224301	−0.24863635	1.0328449	−0.2469548	−2.521639	−1.8303328	−2.0631368	−2.0176322
−0.7740426	−1.116775	0.3352061	−1.303118	−1.2224286	−0.24870083	1.0318071	−0.24660969	−2.5219002	−1.8304977	−2.0627947	−2.0181448
−0.7629362	−1.1169548	0.33490124	−1.3030168	−1.2224286	−0.248197	1.0310309	−0.24634062	−2.522203	−1.8319309	−2.0625494	−2.017884
−0.74922806	−1.1161705	0.33501354	−1.3020816	−1.2224296	−0.24791418	1.0304929	−0.24616154	−2.5225377	−1.8330117	−2.0621152	−2.0172427
−0.7384499	−1.1139864	0.3362563	−1.3051548	−1.2224321	−0.24832098	1.0301074	−0.24603489	−2.5228438	−1.8321822	−2.0619526	−2.0172133
−0.7309097	−1.1118276	0.3349053	−1.3037902	−1.2224315	−0.24778472	1.0298513	−0.24598087	−2.5228078	−1.8315759	−2.062443	−2.0177386
−0.72712964	−1.1094859	0.3359881	−1.3056333	−1.2224314	−0.24829444	1.0296358	−0.24588774	−2.5226767	−1.8314383	−2.062317	−2.0181801
−0.72371286	−1.1082836	0.33539978	−1.3045602	−1.2224283	−0.24791297	1.029509	−0.24583231	−2.5226302	−1.8314928	−2.0620575	−2.0176919
−0.71775365	−1.1085879	0.33443072	−1.3043374	−1.2224288	−0.24772124	1.0293863	−0.24580358	−2.5226388	−1.8318189	−2.0622487	−2.017435

目的変数：**ambient**
説明変数：**11項目**
今回用いたデータセットの行数：**33,426**

k-NN による回帰分析

✓目的変数の値を推定したいサンプル $\mathbf{x}_{new}$ について、すべての
モデル構築用サンプルとの間でユークリッド距離を計算する

✓最も距離の近い k 個のサンプルを選択する

✓k 個の目的変数の値の平均値を、$\mathbf{x}_{new}$ の推定された値とする

✓k 個の目的変数の値の標準偏差で推定値の信頼度を検討できる

- 標準偏差が小さい (k 個の値がばらついていない) 方が、
標準偏差が大きい (k 個の値がばらついている) 方より
目的変数の推定値を信頼できる