機械学習

「機械学習」は人工知能と呼ばれているジャンのひとつです。
機械学習が得意な分野は、データを分析してその傾向を掴むことです。したがって、過去のデータを学習して未来を予測したりもします。

■人工知能の変遷

1959 年代：推論、検索ベースでパズルを解く
1980 年代：知識ベースのエキスパートシステム
　　専門家の知識・技術をコンピュータに教え込む。ここで機械学習の必要性がでてきた。
2000 年代：深層学習（ディープラーニング）を取り込む

■人工知能の注目されるのは

・インターネットの普及により、データの収集が容易になった
・CPU の高速化で、PC でも解析が可能なレベルになってきた
・GPU 機能の上昇・・・人工知能処理には GPU も使われている

■機械学習の準備と処理

問題解決のためには、次のような準備または処理が必要
・データの収集
・データの前処理
・精度評価
・システムへの取り込み

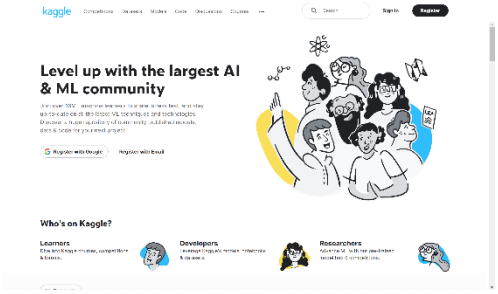
■データの収集

機械学習を満足に果たすためには大量のデータが必要になります。このことを「**データ収集**」といいます。

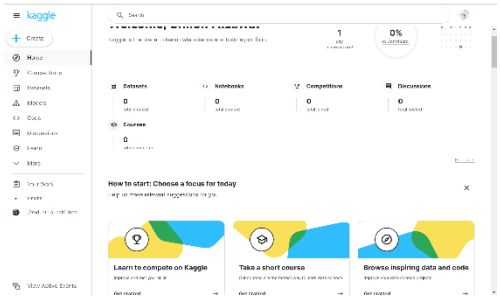
データは、「データセット」として公開されている場合があります。「オープンデータ」と呼ばれている場合おあります。そのほか、WEB API が公開されているサイトでは、API 経由でデータを取得します。ない場合には「**Web スクレイビング**」手法で集める。

■公開されているデータセット

Kaggle は世界中のデータサイエンティストが集まるコミュニティです。
企業などとのマッチングを Competition でおこない、成果報酬が賞金という形で支払われます。



Kaggle のトップページ



ログイン後のページ

また、初心者向けにも、データの前処理の方法や解析モデルを使った予測や精度の出し方などが解説されています。

ここに登録されている世界中 18 万人がコンペティションで競い合っている。中でも優秀な人材は「Kaggle Competition Master」の称号が与えられ、日本国内にも 200 人程度いる。また、さらに上のクラス「Kaggle Competition Grand Master」になると世界中に 263 人しか存在しない。

Kaggle は、ビジネスのマッチングの場であると同時に、データサイエンスの学習の場であり、メンバー同士のコミュニケーションの場であり、遊び場でもある。

◆よくあるコンペ出題問題

・タイタニック号の生存者の予測
　どんな人が生き残れたのか
・住宅の販売価格予測
　アイオワ州の住宅販売価格を予測する
・GAN（敵対的生成ネットワーク）による芸術作品の創造
　モネの描いたような絵を創作する

■前処理

収集したデータは、機械学習で使える形に変換する必要がある。

・**欠損値**：欠けたデータ

・**外れ値**：大きく値が外れているデータ

・**画像の数値化**：

・**ラベル付け**：画像などの前処理で名前をつける

◆代表的なデータ操作

こうした前処理ではデータベースの知識や技術も使われます。

・データ結合

・ソート処理

・グループ化

・データ形式の変換

・行や列の抽出

・行や列の追加

・欠損値の対処

・外れ値の対応

◆文字列の前処理

英語などと違い、日本語は単語単位で切れていない。そこで、文章を単語単位に分解して、その品詞と原形を示す必要がある。

「日本語/の/学習/は/むずかしい」

こうした前処理を「**形態素解析**」といいます。

◆画像の前処理

画像データは事前に、「ピクセル数」や「色の階調」、「データ形式」を統一する必要がある。また、画像内の「向き」も揃えておく必要がある。その上で、画像を数値へ変換する。

◆ラベル付け

機械学習を予測のために使う場合は、正解をもった**学習データ**が必要です。

たとえば、犬と猫を区別する場合には、正解データに「犬」「猫」とラベルをつける必要があります。基本的に、この作業は手作業になる。

■機械学習の手法

機械学習には 3 つの手法があり、それぞれ用途がある

・教師あり学習：データの分類、数値予測
・教師なし学習：データのクラスタリング（似たものを集める）
・強化学習：ルールを得て、自ら学んで強くなる

◆教師あり学習

正解データを元に機械学習をおこなう。
前処理でラベル付けが必要。

教師あり学習の種類

・分析　　：画像の分類、文字認識など
・回帰　　：売上予測、気温の予測など（連続したデータ）

◆教師なし学習と強化学習

正解がない状態で似たデータを探す「クラスタリング」で使われる。

「強化学習」はある環境の中で取った行動によって、得られる報酬が最大になるように学習する手法。囲碁や将棋のプログラムなどとして近年盛んに研究が成されている分野。ほかにも、ロボットや自動車の自動運転などにも応用されている。

■機械学習のアルゴリズム

これまで学習してきたように、アルゴリズムはコンピュータや数学で、問題を解くための決まった手順。「並べ替え」や「探索」のような古典的なアルゴリズムもあるが、機械学習にもそれに特化したいくつかのアルゴリズムがある。機械学習では、そうしたアルゴリズムを組み合わせて使用する。

◆教師あり学習のアルゴリズム

これらのアルゴリズムはすべて機械学習ライブラリ「scikit-learn」（サイキットラーン）で提供されている。

＜回帰＞
・線形回帰
・サポートベクターマシン（SVM）－SVR
＜分類＞

・ロジスティック回帰ーいくつかの要因（説明変数）から「2値の結果（目的変数）」が起こる確率を説明・予測することができる統計手法

・サポートベクターマシン（SVM）－SVC

・決定木

・ランダムフォレスト