第3章 世界で支持されている補完関係 TensorFlow & scikit-learn をおさえておく

使えそうなAIライブラリを知る

ご購入はこちら

佐藤 聖

今回紹介する機械学習ライブラリ scikit-learnの特徴

● かゆいところに手が届くツールが揃っている

scikit-learnは機械学習だけのライブラリではなくデータ・アナリストが利用するツールとしての役割もあるため，一般的な機械学習ライブラリよりも汎用的に利用できます．

基本的な機械学習アルゴリズムが利用できるので教師あり学習による未来予測，クラス・ラベルを予測するための分類，連続値を予測するための回帰ができます．教師なし学習にはクラスタリングによるグループの発見，データ圧縮のための次元削減があります．アルゴリズムの側面から見ると多層パーセプトロン（ニューラル・ネットワーク），ロジスティック回帰，SVM（サポート・ベクタ・マシン），決定木，ランダム・フォレスト，k近傍法，K-meansなどがあるので目的に応じて自由に選択ができ，時には組み合わせて複雑なシステムを開発するのにも役立ちます．

単に機械学習を活用したシステムを作れるだけでなく，開発後に評価できないとモデルの良しあしが単に識別率だけで見られがちです．scikit-learnにはハイパ・パラメータのチューニング，予測を定量評価するためのツールも含まれています．これらのツールを駆使すればモデル構築が適切であったかの評価基準にできます．システムが単純なときには識別率が良ければ問題なかったことでも，複雑なシステムになると極端に汎用性が低い部分が存在し，連結して動かしたときに思ったほど識別率が上がらないなどの問題が出るかもしれません．バランスを見ながらチューニングしたり，モデルで行われる予測を評価したりを手作業で行っていてはかなり大変です．汎用的に利用する機械学習ライブラリはモデルを構築する部分だけでなく，その前後のツールが整備されていることも重要です．

scikit-learnは歴史あるツールなので必要なライブラリがそろっており機械学習を初めて学ぶときにとてもよいと思います．将来的にTensorFlowも学んでみたいと思う方もいるかと思いますが，scikit-learnを知っているとデータ・セットの自動生成などにも役立つと思います．

● 得意とする分野が広い

scikit-learnは，人工知能のフレームワークとして有名なグーグルのTensorFlow（後述）と比べて劣っているわけではなく，得意とする領域が異なります．

機械学習ライブラリとして汎用的で簡単にコードを記述できます．サポート範囲は広く，分類，回帰，クラスタリング，次元圧縮，モデル選択，事前処理が主な機能です．機械学習に必須のクロス・バリエーション，ハイパ・パラメータのチューニング，モデル評価，分析結果のビジュアル化などのツールはほとんどそろっています．

scikit-learnでも巨大な行列を扱うことができます．行列をそのままメモリに格納しようとすると大抵の場合処理速度の低下やメモリ不足になる可能性があります．scikit-learnの多くのアルゴリズムはスパース行列に対応しているのでより少ないメモリで処理できます．scikit-learnに限ったことではないのですがより小回りの利く機械学習ライブラリの方がさまざまなコンピュータで実行可能となるため実行環境を選ばないという利点があります．

例えば100万×100万の行列があり，ほとんどの数値が0で，わずかに0以外の値が格納されているような場合，全体としてはほぼ0の行列になっています．そうした行列からは特徴量がほぼ得られないため巨大な行列を扱うには処理上無駄が多くなります．効率的にデータを取り扱うためには0以外の値とその位置をデータとして保持すればずっと小さな情報量になり，100万×100万の行列を保持するよりもずっと効率的です．

● 小型コンピュータで動く

ラズベリー・パイのような組み込み系小型コンピュータだとCPU処理性能やメモリ・サイズに制約があります．scikit-learnはコンパクトなライブラリなのでIoT端末で機械学習するのに最適なツールだと