16 バッチ学習とオンライン学習

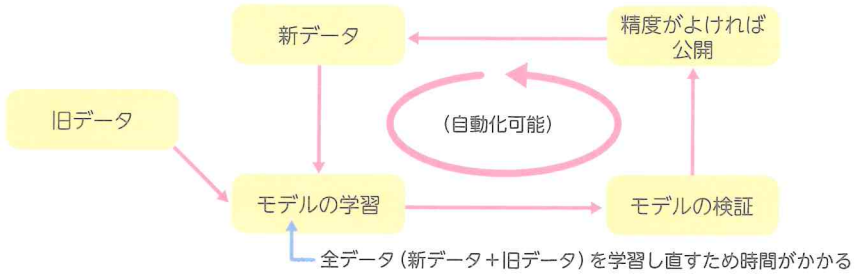
バッチ学習は全データを一括で処理する手法であるため、モデルの更新に時間がかかります。一方のオンライン学習は、データを少しずつ処理しつつモデルを高速で更新していくため、モデルの更新をひんぱんに行う必要のある状況で役立ちます。

バッチ学習

バッチ学習では、すべてのデータを使ってモデルの学習を行う必要があります。そのため計算時間は非常に長くなり、モデルの学習とモデルによる予測は切り離して行われます。このように、予測を切り離して学習する方法を**オフライン学習**と呼びます。

またバッチ学習では、新しいデータをモデルに適用したい場合、新旧データすべてを入力として学習をやり直す必要があります。そのようなやり直しを経て新旧データ両方を学習させた新モデルができたら、それまで稼働させていた予測モデルを停止させて置き換えます。データの学習には時間がかかってしまうため、リアルタイムでモデルを更新することは不可能です。そのため、たとえば状況が刻々と変化する株式市場で、この手法を用いた機械学習トレードシステムは不利になる可能性があります。また、全データをひんぱんに学習し直すと計算資源を多く消費するため、コストがかさむのも難点です。

■ バッチ学習



オンライン学習

オンライン学習は、モデルに少数のデータ(**ミニバッチ**と呼ばれる小さな単位か、1つのデータ)を投入し続けて次々に学習させる方法です。これは学習サイクルが速く、新しいデータが手に入るとすぐにそのデータが学習されたモデルが手に入ります。そのため、先ほどのようなトレードシステムにも適しているといえます。また、計算資源が限られている場合にも有効です。モデルにそのデータが学習されてさえいれば、過去のデータを保存する必要がないためです。

オンライン学習の欠点は、異常なデータが入力されるとモデルの予測能力が低くなることです。これは、新しく与えられたデータは例外なく正しい分類としてパラメータを更新するためです。これを防ぐには、異常検出アルゴリズムを使うなどして異常なデータの入力を監視する必要があります。

また、オンライン学習では、新しいデータにモデルを適応させる割合を意味する**学習率**が重要になってきます。学習率が高いと新しいデータに適応しやすくなりますが、古いデータの情報が失われやすくなります。学習率が低いと古いデータの情報は保たれやすくなる一方、新しいデータへ適応しにくくなります。

なお、データが大きすぎてバッチ学習を行えない場合に、データを小さな単位に分割した上で、オンライン学習のアルゴリズムを使って学習を行うことがあります。この学習方法を**アウトオブコア学習**といいます。

■ オンライン学習

まとめ

- **バッチ学習は一括学習、オンライン学習は逐次学習**