13 データの収集

機械学習をするためには、アルゴリズムの学習や予測を行うためのデータを取得する必要があります。このセクションではさまざまなデータの取得方法について解説していきます。

自分でデータを記録する

データを取得する方法として一番に考えられるのは、**自分でデータを記録する**ことです。特に企業などが社内の問題を解決するために機械学習を利用する場合には、必要なデータを記録するようなしくみを作ることで、より目的に沿った機械学習モデルを作成できます。しかし、自分で記録するからこそ注意しなければならない点もあります。特に以下の点に注意しましょう。

・十分なデータ量が確保できるか

外部からデータを取得する場合と異なり自分でデータを記録する場合には、「データの蓄積にかかる期間」も考慮に入れる必要があります。例として、ある顧客がサービスを解約する可能性を機械学習で予測することを考えてみましょう。年間数件しか解約が発生しない場合には、5年間収集したとしてもせいぜい数十件程度しか集まりません。

・途中で条件が変わったデータではないか

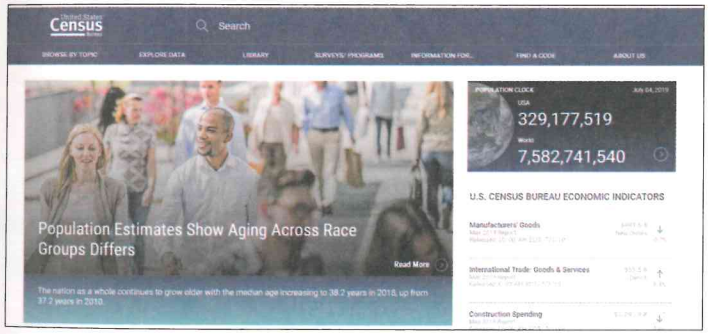
データ量が十分であったとしても、実は途中で取得環境が変化してしまっていた、というケースがあります。たとえば、顧客アンケートは実施時期により対象の顧客層やアンケート項目等が変化していることがあります。またセンサのデータは、センサの位置や数などが不変であったことを確認する必要があります。

官公庁や企業が公開しているデータを利用する

官公庁や企業は、保有しているデータベースをインターネットなどで公開している場合があります。こういったデータベースは多くの場合、行政や企業活動、学術研究などで利用することができるよう、データ項目が充実しています。またインターネット等で取得できる一般的なデータは、年月日や対象によってデータのまとめの形式が異なっている場合が多いものの、データベースとして公開されているものはまとめの形式が統一されているものが多く、扱いやすいデータであると言えるでしょう。データベースが取得できる場所としては、たとえば日本政府が行っている統計調査の結果をまとめた「e-Stat」や米国の国勢調査の結果をまとめた「Census」などがあります。

■官公庁や企業が公開しているデータ

e-Start 政府統計の総合窓口（https://www.e-stat.go.jp/）



アメリカ合衆国国勢調査局ホームページ（https://www.census.gov/en.html）