

学習用データセット事典

斉藤 直希

[ご購入はこちら](#)

機械学習にはデータが必要です。特にディープ・ラーニングを行うときは大量に必要となります。すぐ試してみたいのに大量のデータを用意するのは大変な作業です。

TensorFlowやscikit-learnなど、既存の機械学習用ライブラリは、サンプル・データやデータ・リポジトリからデータをダウンロードするためのツールが用意されていることがあります。その場合、当面はそれを利用すればよさそうです。

しかし、そのうち自分でやりたいことが見つかったりすると、データを用途別に用意する必要性が生じてきます。その際に目的に沿ったデータが利用可能であれば大変便利です。

ここでは機械学習などで利用可能な、公開済みのデータセット^{注1}をいくつか紹介します。特に、入手が容易なウェブ上で公開されているものを中心に紹介します。各ウェブ・サイトの規約に従って使用してください。

ウェブに公開されている 学習用データの種類

● 画像や音声やテキスト

機械学習に使用するデータとは、基本的には画像、音声、動画、テキストなど、私たちが日常的に利用する類のデータです。例えばSNSに投稿された画像やカメラで撮影した動画、ウェブ・サイトのテキスト・データなどです。

● トリミングやリサイズを施したもの

そして、機械学習での利用を想定して公開されているデータは、元のデータに対して一手間が加えられた形で提供されることがよく見られます。これは実際に機械学習でデータを利用する手順とも関連しています。

データを利用するとき、収集したデータから不要なデータを除いて必要なデータだけを選択したり、必要な前処理を加えたりします。例えば、ネコの認識を行うためにネコの画像だけを選び出したり、学習のために画像のサイズを同じ大きさにそろえたり、といった

作業があらかじめ行われた上で配布されるということです。

● ラベルを加えたもの

追加の情報が何らかの方法で加えられることもあります^{注2}。追加の情報とは画像データの場合ならそこに写っているものの名前や種類などの情報です。そしてそれがファイル名やフォルダ名の形であったり、データ・ファイルとは別のテキスト・ファイルなどの形であったり、どこかに記録されて学習やテストのときに利用できるようになっています。

● 学習済みモデルが公開されることもある

上記で説明した類のデータとは別に、学習済みのモデル・データが公開される例もあります。これは本来ならデータを元に学習を行うことで作成するはずのモデルをすぐに利用できるように配布するということです。この学習済みモデルは、そのまま利用することもできますし、これを元にオリジナルのモデルを作成することもできます。

(1) 画像データセット

画像は代表的なデータセットで、さまざまなデータセットが公開されています(表1)。中でも「MNIST」は最も有名なデータセットと言ってもよく、ベンチマークのためのデータとして利用される他、大抵の機械学習のツールではこのデータセットを利用したチュートリアルが用意されています。最近では、MNISTの置き換えを意図したデータセットとして「Fashion-MNIST」も公開されています。これはファッ

注1：データセットとは、単に「データを集めたもの」という程度の意味で捉えてもらって構いません。

注2：ちなみに元のデータに何らかの情報を加えることを「アノテーション」、付加した情報のことを「アノテーション」と呼ぶ場合があります。場合によっては同じようなことを指す言葉として「ラベリング」や「ラベル」といった言葉が使われる場合もあります。