20 能動学習

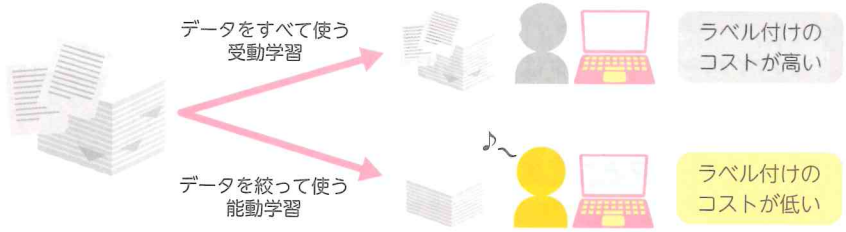
機械学習において教師あり学習を行うには、教師データとなる大量のラベル付きデータが必要です。一般に、教師データ用のラベル付けには時間がかかりますが、能動学習を使うと予測精度を悪化させずに、ラベル付けするデータを少なくできます。

ラベル付きデータの作成は煩雑

機械学習（特に教師あり学習）を行うためには大量のラベル（正解）付きデータが必要ですが、ラベル付けの作業は煩雑です。そのため、教師データをやみくもに作って学習（受動学習）するのではなく、教師データの数を絞って学習する**能動学習**を採用すると効率的です。

そもそもラベル付けにおける効率化の重要性がイメージできない、という方のために、例を挙げましょう。ここでは、「ポケモン」のキャラクターがそれぞれ何というキャラクターなのか、機械学習で判定したいとします。画像には正解の情報が付与されていないため、画像一つ一つの正解を人が判定し、その上で教師データを作成します。その際、判定する人は当然、ポケモンの全キャラクターの見た目を正確に把握していなければなりません。また、正解を入力する際はキーボードなどを使いますが、800種類以上あるポケモンの判定を行うには1つのキーに1体のキャラクターを対応させるだけでは足りません。結果として、キーボードを何回も複雑に打つ必要があり、非常に手間がかかります。効率的に教師データを作成することは重要なのです。

■ラベル付けのコストを低減する能動学習



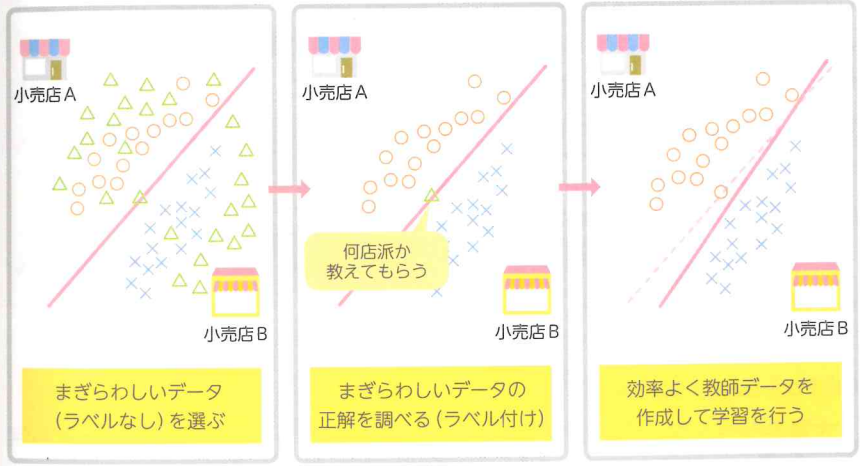
ラベルを選ぶ基準

すでに確認したように、教師データを効率的に作成するには、大量のデータからラベル付けすべきデータを厳選する必要があります。しかし、何を基準に厳選すればよいのでしょうか。

答えは、**まぎらわしいデータ**のラベルを作成することです。明らかに区別がつく大量のデータよりも、区別がまぎらわしい少数のデータのラベルを作成して学習したほうが精度向上につながるためです。このことは、人の学習に置き換えてもすんなりとイメージできるでしょう。

これらを踏まえた上で、Section02でも使ったA店派とB店派の例を使って説明します。下図のように、A店派とB店派（ラベル付きデータ）の分布のほか、何店派なのか不明（ラベルなしデータ）の分布がわかっているとします。ラベル付きデータのみで派閥の境界線を書いたのが下図左です。効率的に境界線の精度を上げるには、派閥不明の家庭のうち、一番まぎらわしい（境界線上に最も近い）家庭を選んで、どちらの派閥に属するのかを聞く（ラベルを付ける）のがよいでしょう。明らかに区別できる家庭（A店あるいはB店に近いと一目でわかる家庭）を選んでラベルを付けても、境界線の精度はほとんど上がりません。

■教師データは「まぎらわしい」データのラベルを選ぶ



○：A店派（ラベルあり）　×：B店派（ラベルあり）　△：派閥不明（ラベルなし）