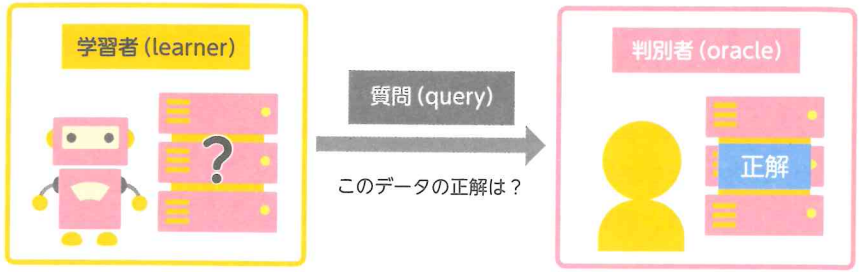
ラベル付きデータの作り方

能動学習では、「**学習者（learner）**」「**判別者（oracle）**」「**質問（query）**」という3つの用語を使って表現します。学習を行う学習者が、データの正解ラベルを知る判別者に対し、正解ラベルを問う質問を行ってラベル付けする、という流れです。なお、学習者は機械学習システム、判別者は人間、質問を行うことはラベル付けをすることと同義です。

■ 3つの用語



流れが理解できたところで、実際にまぎらわしいラベル付きデータを作る代表的な手法として、以下の3つを確認していきましょう。

(1) Membership Query Synthesis

まぎらわしいデータを自ら作り出したあと、判別者に質問を行う方法です。たとえば、手書き数字の画像認識では1と7が似ているため、1と7の中間に見えるような手書き数字の画像を生成し、判別者に正解ラベルを質問します。

(2) Stream-Based Selective Sampling

ラベル付けが済んでいないデータを1つ取り出し、そのデータがまぎらわしければ正解ラベルを質問します。質問しなかったデータは破棄されます。

(3) Pool-Based Samping

大量のラベルなしデータすべてについてまぎらわしさを計算し、もっとも学習に役立つデータの正解ラベルを質問します。

■ ラベル付きデータの作り方