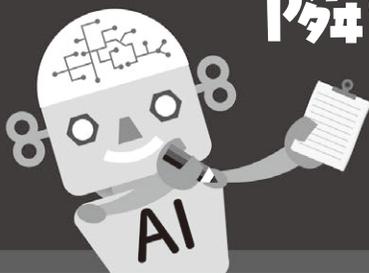
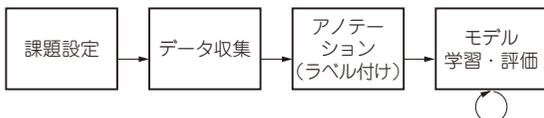


隣のデータサイエンス



第1回 ラベル付けの難しさを提供アプリで体験

萩野 真一



アノテーション後はデータの前処理、説明変数の選択、モデルの選択、パラメータチューニングで精度を上げていく

図1 通常の機械学習のワークフロー



(a) 基準画像 (b) OK (c) NG
基準画像に対して、シミの面積が小さければOK、大きければNG

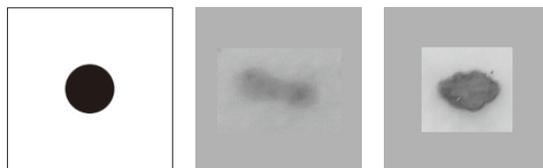
図2 シミのOK/NGの基準

機械学習におけるラベルとは、データを分類するために個々のデータにひもづけられている情報のことです。分類問題のクラス分け、物体検出のバウンディング・ボックスとクラス分け、領域分割のピクセル単位のクラス分けに使います。ラベルを付ける行為はアノテーションと呼ばれます。

ラベルに従って分類されるのですから、ラベル付けが間違っていたら元も子ありません。本稿では、モデルづくりに比べて軽視されがちなラベル付けの重要性を解説します。

ラベル付け間違いは珍しくない

一般的な機械学習の教科書では、ラベルは正しいものとして扱われ、データやラベルの質に関してはあまり触れられていないことが多いようです。図1はよくある機械学習のワーク・フローです(「課題設定」「データ収集」「アノテーション」の順番は前後することがある)。このようにモデルの学習・評価の後に、ラベルを見直すことはほとんどありません。



(a) 基準画像 (b) OK? NG? (c) OK? NG?

このようなデータは判断に迷う
現実の製品異常検査では必ず存在する

図3 OKかNGか判断に迷う例

● 有名なデータセットにも含まれる

実際は論文⁽¹⁾にもあるように、有名なデータセット(アノテーション済みデータ)にも平均3%のラベル誤りがあることが報告されています。特に分類問題で有名なImageNet (LSVRC 2012) の評価セット(validation set)には5.8%の誤りがあると報告されています。通常、公開されているデータセットはラベル誤りがないようにさまざまな工夫がなされているのですが、それでも少量のラベル誤りが含まれているのです。一般的なデータセットにはもっと多くのラベル誤りが含まれていると考えた方がよいでしょう。

● 判断に迷うケースがある

製品異常検査にもいろいろありますが、よくあるのがラインの上を製品が流れ、検査員が目視でOK/NGを判断するものです。歩留まりの関係上、キズ、汚れなどあっても一定以下ならOKとするケースがよくあります。

このような検査で必ず起こるのがOK/NGの判断に迷う、グレーゾーンのデータが存在することです。形がいびつであったり、境界がかすれたりしていると、基準があったとしても判断に迷います。

グレーゾーン判別のために、現場では限度見本を用意し、ガイドラインを作って細かく指示をしている場合があります。それでもすべてのデータで判断が一致することは不可能です。実際に同一画像を複数人で判断する実験を行ってみると、人によって判断が分かれ

