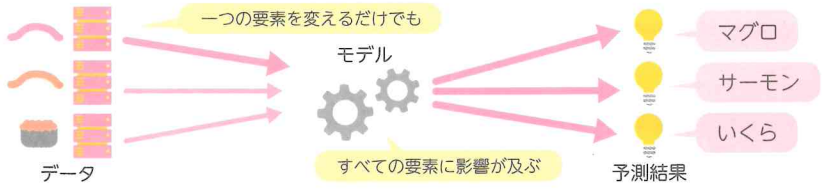
22 フィードバックループ

機械学習システムで注意しなければならないのは、システムの振る舞いを完全には制御できない点です。モデルを随時更新するようなシステムではフィードバックループが起こり、予期せぬ動作を引き起こす場合もあります。

機械学習を使ったシステムの落とし穴

機械学習を使ったシステムには大きな落とし穴があります。それは、**コードの書き方だけではシステムの振る舞いが規定できない**点です。機械学習がデータを必要とする以上、システムの振る舞いはデータに大きく依存してしまいます。そのため、もし誤りを含んだデータをモデルが学習してしまうと、モデルの出力が意図しないものになってしまう可能性があるのです。また、機械学習システムでは、システムのうち何かしら一つの要素を変更すると、他のすべての要素も変わってしまう（Changing Anything Changes Everything, CACE）、いわば「あちらを立てればこちらが立たない」ケースがあるのもしばしば問題となります。たとえば、寿司の画像を読み込ませ、寿司ネタを判別するモデルを作成したとします。そのうち、特定のネタ（たとえばマグロ）の判別精度がよくない場合には、機械学習モデルのパラメータをいじったり、マグロの画像データを追加したりします。これによってマグロの判別精度がよくなったとしても、他のネタの判別精度が保たれることは保証できません。他のネタは判別が難しくなることも十分考えられます。機械学習ではモデルの中身がブラックボックスとなるため、振る舞いを監視することが重要です。

■ Changing Anything Changes Everything



フィードバックループ

観測された最新のデータに基づいて随時モデルを更新していくような機械学習システムでは、システムの使用開始前にその振る舞いを予測するのが難しい場合があります。特に気を付けたいのが、フィードバックループです。**フィードバックループ**とは、システムの振る舞いが環境に影響を及ぼし、次に観測するデータが環境から影響を受けて変化してしまう現象です。

この際、システムの振る舞いの変化が急であったり頻繁に起こったりする場合は、振る舞いの変化の検出は比較的かんたんです。一方、システムの振る舞いが徐々に変わっていったり、モデルの更新の頻度が低い場合には、振る舞いの変化に気づくのが遅れる場合があります。

直接的なフィードバックループの例としては、予測警備があげられます。予測警備とは、過去の犯罪のデータをモデルに学習させ、犯罪が多く起こると予測される場所を重点的に警備する警備の方法です。警察は犯罪の起こる場所を重点的にパトロールするため、その場所での検挙件数は多くなります。これによってさらに犯罪のデータが蓄積されていき、その場所での警備はさらに強化されていきます。これは、確証バイアス（仮説を実証する情報ばかりを集め、反例を集めようとしない傾向のこと）の自動化にほかなりません。

■ 直接的なフィードバックループ