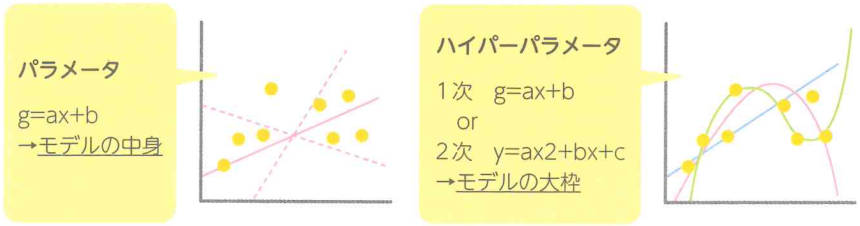
19 ハイパーパラメータとモデルのチューニング

機械学習にも、アルゴリズムの性能を向上させるために人の手でモデルを調整しなくてはならないパラメータがあります。このパラメータを、ハイパーパラメータと呼びます。

ハイパーパラメータ

ハイパーパラメータを理解するにあたり、ここでは多項式を例に解説します。パラメータが直線の傾きや切片など、モデルの中に設定される具体的な値であるのに対し、ハイパーパラメータはモデルを何次式にするのか（直線、二次曲線、三次曲線など）といったモデルの大枠を決める値を意味します。

■ 多項式（直線、二次関数など）の例



ハイパーパラメータが適切でないと、モデルは性能を十分に発揮できません。そのような、性能が十分でない状態によく見られる特徴として「未学習」と「過学習」があります。

未学習とはその名の通り、十分に学習が行われていないことで性能が低い状態を指します。学習データに対する予測や分類の精度が十分に高くない場合、未学習であると言えます。

対して**過学習**とは、学習データに対する精度の向上を重視し過ぎることで、未知のデータに対する精度が下がってしまっている状態を指します。

次ページから、未学習と過学習についてより具体的に見ていきましょう。

未学習と過学習

例として、アルゴリズムで2次元グラフの形（真のモデル）を推測することを考えます。下図の緑線が真のモデルとすると、実際に私たちが取得することができるデータはそこにノイズ（ばらつき）の乗った黄色い点だと考えてください。機械学習では、アルゴリズムがデータを学習することで、真のモデル（緑線）をよく表現できるようなモデル（赤線）を求めます。使うモデルを多項式（1次→直線、2次→2次関数）とすると、この多項式モデルにおけるハイパーパラメータは、次数であると言えます。

下図①のように、次数が1のときには直線となります。しかし真のモデルが曲線であるため、直線では単純過ぎてうまく表現できていないことがわかります。この状態が未学習です。なお、このようにモデルの表現力が足りないことによって、学習データとモデルとの間に生じた誤差のことを、**近似誤差**と呼びます。

さて、次数が1のモデルでは単純過ぎてうまく表現できなかったので、今度は次数を思い切って増やしてみましょう。学習させるのは、次数が9のモデルです。すると下図②のように、学習させたデータ（黄色の点）にぴったりフィットしたモデルを得ることができました。しかし、データのなかった部分は真のモデル（緑線）から大きく外れており、これでは真のモデルをよく表現できているとは言えません。このようなモデルでは、学習データに対する精度は高くなりますが、未知のデータに対する精度は悪くなってしまいます。このように、モデルが過学習してしまったことで、未知のデータ（テストデータ）とモデルとの間に生じた誤差のことを**推定誤差（Validation Loss）**と呼びます。

■ 未学習と過学習