

機械学習の整理帳

tomixy

2025 年 7 月 30 日

目次

第 I 部	機械学習の概念	3
第 1 章	学習の原理	4
	意思決定のプロセス	4
	モデルとパラメータによる学習	5
	パラメータ数とモデルの表現力	6
	汎化能力	6
	過学習	7
第 2 章	学習の手法	10
	教師あり学習	10
	教師あり学習の流れ	11
	教師なし学習	11
	強化学習	12

第 II 部 教師あり学習の構成要素	13
第 3 章 損失関数	14
損失関数の設計	14
分類器の-marginと損失関数	15
分類で使う損失関数 : 0/1 損失関数	16

第 I 部

機械学習の概念

第 1 章

学習の原理



意思決定のプロセス

経験に基づいて意思決定を行うために人間が用いるプロセスは記憶・定式化・予測フレームワークと呼ばれ、次の 3 つのステップで構成されている。

1. 記憶：過去の同じような状況を思い出す
2. 定式化：全般的なルールを定式化する
3. 予測：このルールを使って将来起こるかもしれないことを予測する


コンピュータに「記憶・定式化・予測」フレームワークを使わせることで、コンピュータに私たちと同じように考えさせることができる。

1. 記憶：巨大なデータテーブルを調べる
2. 定式化：さまざまなルールや式を調べてデータに最適なモデルを作成する
3. 予測：モデルを使って未来（未知）のデータについて予測を行う



モデルとパラメータによる学習


コンピュータはデータを使って**モデル** (model) を構築するという方法で問題を解く。

 **モデル** データを表すルールが集まりであり、予測を行うために使うことができる

モデルは、対象の問題で獲得したい分類器や予測器、生成器などであり、入力（具体的なデータ）から出力（予測結果）を計算する関数とみなすことができる。

パラメータ

モデルの挙動を調整する「つまみ」となる変数を**パラメータ** (parameter) という。
パラメータ θ によって関数の挙動が決まることを、関数が θ によって「特徴づけられた」という。


 **パラメトリックモデル** パラメータによって特徴づけられたモデル

パラメータ θ によって挙動が決まるモデルを $f(x; \theta)$ と表記する。


「;」以降の変数は、この関数の入力ではないことを表している。

学習の定義

モデルのパラメータを調整して、最適なモデルを構築することを**学習**という。

 **学習** 「データから**学習**する」とは、データからモデルの最適なパラメータを推定すること

人間の学習も、脳内にある神経回路のパラメータ（シナプスの重みなど）を調整して実現されている。



パラメータ数とモデルの表現力

モデルは、成り立つかもしれない**仮説**を表すものであり、パラメータの値によって、異なる仮説を表現することができる。

つまり、パラメータの推定は、複数存在する仮説の中から一つを選択することとみなすことができる。

たとえば、モデルのパラメータ θ が $\{-1, 0, 1\}$ のいずれかの値をとる場合は、

- $\theta = -1$ の場合のモデルが表す仮説
- $\theta = 0$ の場合のモデルが表す仮説
- $\theta = 1$ の場合のモデルが表す仮説

の中から選択しているとみなせる。

さまざまな仮説の中から選ぶことができる場合、「モデルの**表現力**が高い」という。

単純には、パラメータ数が多いモデルほど仮説数が多く、表現力が高いといえる。



汎化能力

計算機は多くの情報を誤りなく大量に記憶することができる。

そのため、起きうる事象を十分網羅できるようにデータを用意できれば、わざわざモデルを作らなくても、過去の似たような値をそのまま使えばよいのではないかと考えることもできる。

学習時のデータをすべてそのまま記憶し、それを予測時に利用するアプローチを**丸暗記 (memorization)** という。

しかし、世の中の多くの問題では、すべてのケースを前もって列挙したり、それらを記憶しておくことはできない。

特に、入力値が**高次元データ**（画像、音声、言語、時系列、etc.）や**連続値**である場合、すべての事例を網羅することは不可能である。

丸暗記できない場合は機械学習が有効

機械学習は、有限の「訓練データ」を用いて、無限ともいえる「未知のデータ」に対してもうまく動くようなモデルを作る手法といえる。

未知のデータに対してもうまく動く能力を**汎化能力**（**generalization ability**）という。

機械学習では、単に訓練データでうまくいくようなモデルを見つけるだけでなく、「未知のデータでどれだけうまくいくか」を表す汎化能力をどのように獲得するかが重要な課題となる。



過学習

訓練データではうまくいっているのに、訓練時には見なかった未知のデータではうまくいかない状態を**過学習**（**overfitting**）という。

過学習が起こる原理

たくさんのデータを集めれば、仮説が本当に成立するかどうかを高い確率で検証することができる。

データ数に対して仮説数が多い場合、正しい仮説よりも、たまたま成り立ってしまう誤った仮説が含まれる可能性が高くなる。これが過学習が起こる場合である。

- 検証する仮説数が少ない場合：事例をすべて満たしている仮説が本当に成り立つ可能性が高い
- 検証する仮説数が多い場合：事例をすべて満たしている仮説も、たまたま成り立っているだけの可能性が高い

機械学習は、多くの仮説（パラメータがある値をとる時のモデル）の中から、多くの訓練事例を説明する仮説がどれかを探す問題ともいえる。

このとき、過学習は、たまたま多くの訓練事例でうまくいくようなモデルが見つかってしまうことで起こる。

過学習を防ぐ原理

訓練データ数に比べて、検証する仮説数が少ない状況であれば、見つかった仮説がたまたま成り立ったものでなく、実際に関係がある可能性が高くなる。

そのため、過学習を防止するには、

$$\text{訓練データ数} > \text{検証する仮説数}$$

という状況を目指すことが有効となる。

過学習の防止策：訓練データを増やす


訓練データを増やすことができれば、多くの事例で仮説を検証できるので、たまたま仮説が成り立つ可能性を小さくすることができる。

しかし、訓練データを増やすことは困難である場合も多い。

そこで、訓練データに意味を変えないような変換を加えて、人工的にデータを水増しする **データオーグメンテーション**（**data augmentation**）という手法が有効となる。

過学習の防止策：仮説数を少なくする

訓練データ数が同じであれば、その中で仮説数を少なくすることも過学習の防止として有効である。

 **オッカムの剃刀** ある事柄を説明するためには、必要以上に多くを仮定すべきでない

仮説数を少なくするには、単純にはモデルのパラメータ数を少なくすればよい。

モデルに制約を加えることで、可能な限り単純なモデルを使うことで、仮説数を少なくする

ことができる。

しかし、過学習を抑える別の仕組みがある場合は、必ずしもパラメータ数が少ない方が汎化するとは限らない。

たとえば、ニューラルネットワーク (neural network) は、パラメータ数が膨大であるにも関わらず、高い汎化能力を持っている。

第 2 章

学習の手法



教師あり学習

教師あり学習（**supervised learning**）では、入力 x と推定したい出力 y からなるペア (x, y) を訓練データとして利用し、



入力 x から望ましい出力 y を予測できるような

モデル $y = f(x; \theta)$ を学習すること



を目標とする。

訓練データは、**教師ありデータ**や**学習データ**と呼ばれることもある。

$$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

学習と推論

教師あり学習は、**学習**（**training**）と**推論**（**inference**）という 2 つのフェーズに分けられる。

学習フェーズでは、訓練データをうまく推定できるように、すなわち

$$y_i = f(x_i; \theta)$$

となるように、モデルのパラメータを調整していく。

推論フェーズでは、学習によって得られたパラメータ $\hat{\theta}$ を使ったモデル $f(x; \hat{\theta})$ を使い、新しいテストデータ \tilde{x} の出力を

$$\tilde{y} = f(\tilde{x}; \hat{\theta})$$

として求める。



教師あり学習の流れ

教師あり学習では、**訓練データ**、**モデル**、**損失関数**、**目的関数**、**最適化**をそれぞれ設計して組み合わせることで、学習を実現する。

1. **訓練データ**を用意する： $(x_i, y_i)_{i=1}^n$
2. 学習対象の**モデル**を用意する： $y = f(x; \theta)$
3. **損失関数**を設計する： $l(y, y')$
4. **目的関数**を導出する： $L(\theta) = \sum_i l(y_i, f(x_i; \theta)) + R(\theta)$
5. **最適化**問題を解く（**学習**）： $\theta^* = \underset{\theta}{\operatorname{argmin}} L(\theta)$
6. 学習して得られたモデルを**評価**する

[Note 1: それぞれの章へのリンクを貼る]



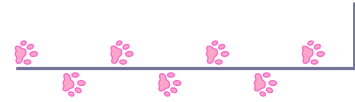
教師なし学習

教師なし学習 (**unsupervised learning**) は、教師（正解）がつけられていないデータ

$$D = \{x_1, \dots, x_n\}$$

を利用した学習であり、

データの特徴を捉え、データの最適な表現や
データ間の関係を獲得すること



を目標とする。

たとえば、教師なし学習では分類する目標がデータとして与えられないため、画像分類は学習できない。

代わりに、画像をどのようにデータとして表現できれば、後続のタスクがうまく処理できるのかを学習する。

教師なし学習の代表例

[Todo 1: book: ディープラーニングを支える技術 p63~]



強化学習

[Todo 2: book: ディープラーニングを支える技術 p66~]



第 II 部

教師あり学習の構成要素

第 3 章

損失関数



損失関数の設計

損失関数 (loss function) は、訓練データで与えられる正解に対し、



予測がどれだけ間違っているのか



を表す関数である。

コスト関数 (cost function) や誤差関数 (error function) と呼ばれることもある。

損失関数は、入力 \mathbf{x} 、正解の出力 \mathbf{y} 、モデルパラメータ θ を引数としてとり、0 以上の値を返す。

$$l(\mathbf{x}, \mathbf{y}; \theta) \geq 0$$

予測が正しければ 0 を返し、予測が間違っていれば正の値を返すようにする。

学習は、損失関数によって表される「現在の予測の間違っている度合い」を最小化するようなパラメータを求める最適化問題を解くことによって実現される。

損失関数の設計の重要性

損失関数は、学習の目的に応じて自由に設定することができる。

どのような損失関数を使うかによって、学習結果のモデルがどのような性質を持つかが決まる。

- 汎化性能
- ノイズに強いかわ弱いかわ
- 平均的な性能が優れているか、最悪の場合の性能が優れているか

たとえば、損失関数が大きく間違っているサンプルを重視するなら、大きな間違いはしないようになる。その反面、訓練データのノイズ（間違ったラベルがある場合など）に弱くなる。

損失関数の微分の形の重要性


損失関数の微分の性質によって、どのような解に収束するかが変わる。



分類器の-marginと損失関数

分類器が高い確信度で予測したにもかかわらず間違えた場合は、損失関数は大きな正の値を返すようにする。

確信度は、margin（margin）によって表される。

 margin 予測結果が境界面からどれだけ離れているか

境界面より離れている（marginが大きい）分類結果は、確信を持って「これは XXX である」と予測していることを表す。

逆に、境界面に近い（marginが小さい）分類結果は、「これはどちらかといえば XXX だが、YYY かもしれない」といったように、確信度が低いことを表している。

marginが大きいにもかかわらず予測が外れている場合は、今の予測や境界面が適切ではないことを示しており、大きな更新が必要となる。



分類で使う損失関数：0/1 損失関数

0/1 損失関数は、モデルによる分類が間違っていたら 1、正しければ 0 を返す関数である。

$$l_{0/1}(\mathbf{x}, y; \theta) = \begin{cases} 0 & \text{if } f(\mathbf{x}; \theta) = y \\ 1 & \text{if } f(\mathbf{x}; \theta) \neq y \end{cases}$$

この損失関数を使えば分類精度を評価できるが、微分がほとんどの位置で 0 になるため、**勾配法**を用いた学習では利用できない。

.....

Zebra Notes

Type	Number
todo	2
note	1