

# 目次

第 1 章	機械学習の枠組み	2
1.1	枠組み	2
1.2	損失と相対エントロピー	3
1.3	勾配降下法	4
1.4	学習アルゴリズムの性能評価	4
第 2 章	スパース性	5
2.1	スパース性とは	5
第 3 章	バンディット問題と強化学習	6
第 4 章	Deep Learning	7
4.1	形式的定義	7
4.2	TCNN	7
参考文献		8

## 第 1 章

# 機械学習の枠組み

まさか情報「科学」らしくなっていることに驚愕している。「学習」は統計的には「推定」「予測」と同義であるが、特に「機械」が行うアルゴリズム的行為を表現する際に使う、concept with an attitude である。

### 1.1 枠組み

機械の構成の際に、モデルの選択など、人間により恣意的に挿入された偏りを帰納バイアスという。帰納バイアスが弱いことを「万能」という。逆に特徴量エンジニアリングや CNN などは、ドメイン知識に基づいて適切な帰納バイアスを導入することで狭い分野に対する高い性能を目指す。

定義 1.1.1 (model).

- (1) 観測によって得られる情報をデータという。情報理論の対象。データ生成分布を背後に想定することで、推測統計学が始まる。
  - (a) 仮説のパラメータを推定するために直接用いられるデータを学習データ (training data) または訓練データという。
  - (b) 正則化パラメータの調整など、学習結果の性能を検証するためのデータを検証データ (validation data) という。
  - (c) 学習データと検証データを併せて、観測によってすでに人類の手元に得られたデータを観測データ (observed data) という。単にデータといったときはこれを指す。
  - (d) 予測精度を評価するためのデータをテストデータという。いまだに人類に観測されていないデータを使う場合もある。
  - (e) データが入力と出力の組  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  で表されるとき、入力データと出力データという。 $|\mathcal{Y}| < \infty$  を満たす出力空間の元をラベルという。 $|\mathcal{Y}| = 2$  を満たすときのデータ  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  を 2 値データという。
- (2) 入力空間から出力空間への関数の族  $\mathcal{F} \subset \text{Map}(\mathcal{X}, \mathcal{Y})$  を仮説集合 (hypothesis set) または統計モデルという。 $|\mathcal{Y}| < \infty$  であるとき、仮説  $f \in \mathcal{F}$  を判別器 (classifier) といい、これを定める関数  $\mathcal{X} \rightarrow \mathbb{R}$  を判別関数 (discriminant function) という。
- (3) 学習データの集合から仮説集合への関数  $\mathcal{X} \times \mathcal{Y} \ni D \rightarrow \mathcal{F}$  を学習アルゴリズムという。
- (4) 学習アルゴリズムと実際のデータの差を評価する関数を損失関数という。

例 1.1.2.  $|\mathcal{Y}| < \infty$  であるときは判別問題、 $\mathcal{Y} = \mathbb{R}$  であるときは回帰問題という。それぞれ、0-1 損失や二乗損失などの損失関数が使われる。 $(x, x', y) \in \mathcal{X}^2 \times \mathcal{Y}$ ,  $\mathcal{Y} = \mathbb{Z}/2\mathbb{Z}$  と定式化される問題は、 $\mathcal{X}$  に順序が誘導されるため、ランキング問題という。

定義 1.1.3. データ  $D \subset \mathcal{X} \times \mathcal{Y}$  に対して、新たなデータを生み出すことを考え、 $\mathcal{X} \times \mathcal{Y}$  上の確率分布、または、 $\mathcal{F} \subset \text{Map}(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \times \mathcal{Y})$  生成モデルという。

## 1.2 損失と相対エントロピー

丸暗記を過学習と呼ぶ．この組み合わせ論的な解を超えることを考える．そこに神が宿る．そのプロセスは，相対エントロピーの（適切な）最小化として完全に特徴づけられる．

情報幾何学は，その一般化として  $C^*$ -代数による量子確率論による定義と理論展開もある．Fisher 情報量は統計多様体の曲率を定め，曲率が大きい時は「大きく変化する」が，これが Cramer-Rao の不等式である．

定義 1.2.1 (over fitting, regularization).

- (1) 訓練データは有限であるから，組み合わせ論的な解は存在する．このように，訓練データのみには使えないアルゴリズムを，過学習という．対して，理想的な状況を汎化という．
- (2) 損失関数にペナルティ項などの追加の構造を付加することにより，過学習や ill-posed problem をとこうとする手法群を正則化という．

注 1.2.2. 深層学習では，特殊な構造を持ったレイヤーを用意すること自体が正則化の代わりをする．この現象を implicit regularization という．これは深層学習の汎化性能を説明する可能性がある．

定義 1.2.3 (predictive loss, empirical loss).  $l(\hat{y}, y)$  を損失関数とする．これに対して，仮説毎に定まる損失関数の値の平均値を損失  $\mathcal{F} \rightarrow \mathbb{R}$  という． $h \in \mathcal{F}$  を仮説とし，真の分布を  $P$ ，観測データが定める経験分布を  $\mathbb{P}_n$  とする．

- (1)  $R(h) := E[l(h(X), Y)]$  を予測損失という．
- (2)  $\hat{R}(h) := E_{\mathbb{P}}[l(h(X), Y)]$  を経験損失という．

要諦 1.2.4. 学習した仮説の精度を評価するにあたって，経験損失と予測損失との間の違いを見積もることが肝要となる．

補題 1.2.5. 経験損失は，予測損失の不偏推定量である．

定義 1.2.6 (relative entropy / KL divergence).

- (1)  $H$  をヒルベルト空間とすると，可観測量は  $C^*$ -部分代数  $\mathcal{A} \subset B(H)$  をなす．
- (2) 状態とは，可観測量  $A \in \mathcal{A}$  に対して期待値  $\langle A \rangle \in \mathbb{C}$  を定める線型汎関数  $\langle - \rangle := \langle \psi | - | \psi \rangle : B(H) \rightarrow \mathbb{C}$  をいう．これは Riesz の表現定理より，ある元  $\psi \in H$  が存在して，これが定める内積から定まる．
- (3)  $C^*$ -代数と状態の組を量子確率空間といい，これは確率空間の一般化となっている．
- (4) エントロピーの概念を，2つの状態の間に定義される距離として一般化したものを相対エントロピーまたは **Kullback-Leibler divergence** という．具体的には，可測空間  $X$  上の2つの確率測度  $P, Q$  について， $Q$  は  $P$  に関して絶対連続であるとき，相対エントロピーを

$$S(Q|P) := \int_X \log \frac{dQ}{dP} dP$$

と定める．相対エントロピーは非対称であることを除いて距離を定める．

定義 1.2.7. 判別問題の損失関数  $l$  について，

- (1)  $\inf_{h \in \text{Meas}(\mathcal{C}, \mathcal{Y})} R(h)$  を **Bayes 誤差** という．
- (2) 下限を達成する可測関数がモデル  $\mathcal{F}$  の元であるとき，これを **Bayes 規則** という．

## 1.3 勾配降下法

相対エントロピーの最小化アルゴリズムには、勾配降下法 / 最急降下法がある。ここからは数値解析の領域である。しかし、これは損失関数が下に凸である場合にしか保証されない。しかし、深層学習では、ほとんどすべてのモデルが勾配降下法で訓練されていて、これが何故かうまく行く。

## 1.4 学習アルゴリズムの性能評価

## 第 2 章

# スパース性

正則化によってスパース性を誘導する方法は、直接スパース性を仮定する方法に比べてより柔軟であるだけでなく、凸最適化問題に帰着するため、計算量の上で大きなメリットがあります。

### 2.1 スパース性とは

定義 2.1.1. 実際の説明変数は次元に比べて少数であるという仮定のことをいう。さらに一般化すれば、低ランク行列の推定問題をいう。

例 2.1.2. ゲノム変異からの予測，時空間データの解析，推薦システムなど。自然言語処理ではデータは素になるかもしれないが，統計的な優位性は生じない。

## 第 3 章

# バンディット問題と強化学習

強化学習は意思決定モデルをデータから、報酬を最大するように学習する機械学習の一分野である。

## 第 4 章

# Deep Learning

### 4.1 形式的定義

### 4.2 TCNN

CNN はデータの位相構造にマッチするため，強力な性能を発揮する．そこで，この成功を一般化させたい．TCNN のレイヤーの位相構造は，Topological Data Analysis の知見が用いられている．

## 参考文献

- [1] Ephy Love "Topological Deep Learning" (2021)
- [2] 富岡亮太『スパース性に基づく機械学習』(2015)
- [3] 金森敬文『統計的学習理論』(2015)
- [4] 本多淳也，中村篤祥『バンディット問題の理論とアルゴリズム』(2016)
- [5] 持橋大地，大羽成征『ガウス過程と機械学習』(2019)
- [6] 森村哲郎『強化学習』(2019)