

数学科向け機械学習入門（体験版）

2018 年 2 月 27 日

1 はじめに

世は空前の人工知能ブーム。猫も杓子も AI だディープラーニングだと、なんでも機械学習に任せてしまえばいいという考えが蔓延している。

機械学習エンジニアの待遇の良さや需要の高さを耳にする機会は多いことだろう。かつて数学徒の目指す職種としてありがちだったクオンツやアクチュアリーと違い、スーツを着ることを強要されるのが少ないという点も魅力的だ。そうなるに「自分も機械学習を学んで AI エンジニアになり、ガッポガッポ大儲け。タワーマンションの上層階でたくさん女侍らして、高級ワインを飲むんじゃあ！」という思考になり、機械学習の本に手が伸びる数学徒も少なくないことかと思う。

しかし、勉強をし始めた多くの数学徒は匙を投げてしまう。その理由の大半が「あまりに数学的に適当すぎて読んでいられない」というものだ。

その気持ちはよくわかる。かく言う私もそうだった。世間には「一般向け」と称した、あまりにも粗雑な機械学習の入門書が溢れかえっており、数学的にまともな文献は少数派だ。もちろん一部存在はするが、とてもじゃないが初学者が読んで役立てることはあまりに難しい。

数学的に適当、とは具体的にどういうことなのか。これは大きく分けて 3 つあるように思う。

1. 確率測度と確率密度関数が区別されていない
2. 確率変数であるものと確率変数でないものが区別されていない
3. 目の前の関数がどこからどこへの写像かわからない

ある程度わかってくるとなんとなく忖度して雰囲気を読んでいくことができるようになるが、厳密に読んでいく訓練を受けてきた機械学習初学者の数学徒にはなかなか酷な話だ。

そのため、本書の前半ではなるべく数学科の解析学の授業で習う用語、流儀を用いて、機械学習の基礎を書くことを心掛けた。本書の内容を理解すれば、そう苦勞せず一般向けの機械学習資料を読んでいくことができるはずだ。

前半部分の前提知識は数学科学部レベルの関数解析学、確率論、数理統計学とする。後半はその限りではない。

体験版追記：本書は体験版ということで、ページ数も限られるのもあり、具体的な手法についてあまり詳細な解説はできていない。特にベイズを用いた推定の具体的なやり方についてはほとんど解説できなかったと言っている。興味のある人は自力で調べてみてほしい。本書を読みこなせた人なら難なく学べるはずだ。

2 機械学習入門

2.1 問題設定

2.1.1 大枠設定

まずおおむねの完備な確率空間 (Ω, \mathcal{F}, P) を設定する。

特に表記がなければ、すべての可測空間の σ 加法族はボレル集合族、線形空間に対応する体は実数体であるものとする。

線形距離空間 \mathcal{X} を特徴量空間、線形距離空間 \mathcal{Y} をラベル空間と呼び、それぞれの元 x, y を「特徴量」「ラベル」と呼ぶ。

ここで、データから何らかの「よい」写像 $f: \mathcal{X} \rightarrow \mathcal{Y}$ を構成するのが「機械学習」である。機械学習とは、関数近似の機構なのである。

ここで、独立同分布な確率変数列 $\mathcal{D} := \{D_i(\omega)\}_{i=1}^n$ を考える。各 D_i は $D_i: \Omega \rightarrow \mathcal{X} \times \mathcal{Y}$ となる可測関数である。この確率変数列 $\mathcal{D} (= \{D_i\}_{i=1}^n)$ を入力データと呼び、確率変数 D_i の \mathcal{X}, \mathcal{Y} への射影をそれぞれ X_i, Y_i と表記する。

増大情報系 $\mathcal{F}_0 \subset \mathcal{F}_1 (\subset \mathcal{F})$ を考える。 \mathcal{F}_0 はデータ観測前の保持情報、 \mathcal{F}_1 はデータ観測後の保持情報である。すなわち

$$\mathcal{F}_1 := \mathcal{F}_0 \vee \sigma([D_i]_{i=1}^n) \quad (1)$$

連続写像 $f: \mathcal{X} \rightarrow \mathcal{Y}$ の満たす集合全体を \mathcal{Z} と表記する。モデルに対応した \mathcal{Z} の部分集合 (詳細は後述) を \mathcal{Z}_M と書く。

連続汎関数 $F: \mathcal{Z} \rightarrow \mathbb{R}$ の集合を \mathcal{Z}^* と書く。 \mathcal{Z}^* の位相は弱位相であるとし、 \mathcal{F}_1 -可測な確率変数 $F^*: \Omega \rightarrow \mathcal{Z}^*$ によって定義される評価関数 $F_{\mathcal{D}} := F^*(\omega)$ に対して

$$\hat{f} := \operatorname{argmin}_{f \in \mathcal{Z}_M} F_{\mathcal{D}}(f) \quad (2)$$

を満たす $\hat{f} \in \mathcal{Z}_M$ を見つける。これが機械学習における大枠の問題設定である。

2.1.2 パラメータを通した \hat{f} の構成方法

パラメータを導入する。パラメータの集合となる2つの線形位相空間を Θ_0, Θ と表記する。写像 $\theta^*: \Theta \rightarrow \mathcal{Z}$ はあらかじめ一つ固定しておき、確率変数 $\hat{\theta}_0(\omega): \Omega \rightarrow \Theta_0, \hat{\theta}(\omega): \Omega \rightarrow \Theta$ を考える。実際に機械学習問題を解くにあたっては、この $\hat{\theta}$ を構成することを通して関数 \hat{f} を構成する。今後は、分かりやすくするために $\theta^*(\theta)$ を $f(\cdot, \theta)$ と表記する。

今後本書では、 θ_0 をハイパーパラメータ、 θ をパラメータと呼び、単にパラメータといった場合は後者のみを指すものとする。ハイパーパラメータはパラメータの数や後述の事前分布の分散など、パラメータにまつわる数値を決定するものである。

(ここで、聡明な読者は「 $\mathcal{Z}_M = \operatorname{Im}(\theta^*)$ でいいのでは？」と思うかもしれない。しかし、これではバيز的な機械学習において非常に困る事態が起こってしまう。 $\operatorname{Im}(\theta^*) \subset \mathcal{Z}_M$ は常に成り立つが、逆は頻度論でしか成り立たない)

$\hat{\theta}$ の \mathcal{F}_0 -条件付き確率、 \mathcal{F}_1 -条件付き確率をそれぞれ事前確率、事後確率と呼ぶ。またそれらの確率測度に Θ 上のルベーグ測度とのラドンニコディム導関数が存在すれば $p(\theta), p(\theta|\mathcal{D})$ と書き、事前分布、事後分布と

呼ぶ。

2.1.3 問題設定

ここでいよいよ統一的な問題設定を定義する。

特徴量空間 \mathcal{X} 、ラベル空間 \mathcal{Y} 、データ \mathcal{D} 、評価関数 $F_{\mathcal{D}}$ 、事前分布 $p(\theta)$ 、事後分布 $p(\theta|\mathcal{D})$ 、超事前分布 $p(\theta_0)$ 、パラメータ $\hat{\theta}$ の可測性、パラメータと f の関係 $\theta^*(\theta)$ をまとめて $(\mathcal{X}, \mathcal{Y}, \mathcal{D}, F_{\mathcal{D}}, p(\theta), p(\theta|\mathcal{D}), p(\theta_0), \hat{\theta} - \text{mesurable}, \theta^*(\theta))$ と表記し、これを今後「機械学習問題」と呼ぶことにする。

これができるだけ「よい」関数になるよう調整したい。このとき \hat{f} が実現できる範囲の集合こそが \mathcal{Z}_M であり、一般には「仮説集合」と呼ばれる。(ちなみに、 $\hat{\theta}_0, \hat{\theta}$ が共に \mathcal{F}_1 -可測であれば、これは $Im(\theta^*)$ と一致する)

また、 $\hat{\theta}_0$ が \mathcal{F}_0 -可測で $\hat{\theta}$ が \mathcal{F}_1 -可測である場合を頻度論的機械学習問題と呼び、 $\hat{\theta}_0$ が \mathcal{F}_0 -可測だが $\hat{\theta}$ が \mathcal{F}_1 -可測でない場合をベイズ論的機械学習問題と呼ぶ。ベイズ論の解説は次章にて。

また、 $\hat{\theta}_0, \hat{\theta}$ が共に \mathcal{F}_1 -可測でない場合を階層ベイズ論的機械学習問題と呼ぶが、本書では扱わない。つまり $\hat{\theta}_0$ はすべてデータ観測前の段階であらかじめ決まっている定数であるものとする。

2.2 代表的なモデル

いよいよここから機械学習における代表的なモデルについて説明する。ここまでの問題設定が難解だった分、ここからはかなりあっさり進んでいくことができる。

解説するモデルは大きく分けて3つ、「線形回帰」「カーネル回帰」「ニューラルネットワーク」である。カーネル回帰とニューラルネットワークは後の章で詳細な数学的研究について解説するため、この章では触りだけで済ませる。

このいずれに対してもある関数を決め、最終的に構成される関数はそれとパラメータの積の和とするのが普通である。

$$\text{線形回帰 (基底関数)} \quad \{\phi_i : \mathcal{X} \rightarrow \mathcal{Y}\}_{i=1}^n (N \text{ はハイパーパラメータとして定める}) \quad (3)$$

$$\text{カーネル回帰 (カーネル関数)} \quad k : \mathcal{X}^2 \rightarrow \mathcal{Y} \quad (4)$$

$$\text{ニューラルネットワーク (活性化関数)} \quad \eta : \mathbb{R} \rightarrow \mathbb{C} \quad (5)$$

2.2.1 線形回帰

最も基本的なモデルである。

写像 $\theta^* : \Theta \rightarrow \mathcal{Z}$ は次のように定義する。

$$\theta^*(\theta)(x) := \sum_{j=1}^d \sum_{k=1}^N \theta_{kj} \phi_k(x_j) \quad (6)$$

ただし、 θ_{kj} は、行列 $\theta \in \mathbb{R}^{N \times d}$ の kj 成分である。

そして、入力データ \mathcal{D} に対して、評価関数 $F_{\mathcal{D}} : \mathcal{Z} \rightarrow \mathbb{R}$ は次のように定義する。

$$F_{\mathcal{D}}(f) := \frac{1}{2} \sum_{i=1}^n \|Y_i - f(X_i)\|_{\mathcal{Y}}^2 \quad (7)$$

これを最小化するような行列 $\theta \in \mathbb{R}^{N \times d}$ を求めればよい。ただし、 $\|\cdot\|_{\mathcal{Y}}$ は \mathcal{Y} 上の距離である。

ところで、聡明な読者はこう疑問に思ったかもしれない。「 N を際限なく大きくとれば、どんなデータに対しても完璧に近似できるのでは？」と

その予測は正しい。確かに $n < N$ であれば、基底関数がよほど歪でない限り損失関数はかならず 0 にできてしまう。しかし、そのようにして作った関数が、未知のデータに対して力を発揮できるだろうか（未知のデータへの予測性能を汎化性能という）

もちろんそんなことはない。強引な近似を行えば、データのノイズにもフィッティングしてしまい、とてもじゃないが未知のデータに対する予測性能などもてるはずもない（これを過学習という）

かといって、表現能力の低い近似だとろくな予測性能を発揮できない（これを未学習という）

未学習を無理に解決すると過学習を起こし、過学習を無理に解決すると未学習を起こす。（これを、バイアスとバリエーションのトレードオフという）

では、その過学習と未学習の折り合いをつけるにはどうすればよいか。その代表例が「正則化」と呼ばれる手法である。正則化は、簡単に言うと「近似の複雑さにペナルティを与える」手法だ。基本的には次のように定義する。

$$F_D(f) := \frac{1}{2} \sum_{i=1}^n \|Y_i - f(X_i)\|_{\mathcal{Y}}^2 + \lambda \sum_k \sum_j |\theta_{kj}|^l \quad (8)$$

ただし、 l は自然数 1 もしくは 2 で、 $l = 1$ のときこれを「Lasso 回帰」、 $l = 2$ のときこれを「Ridge 回帰」という。 λ は θ_0 の実数部分となるハイパーパラメータで、0.01 や 0.001 くらいの数字がセットされる場合が多い。

Lasso 回帰にはパラメータが疎になりやすい（値が 0 になるパラメータがたくさん存在する可能性が高い）、また Ridge 回帰は数学的な扱いが楽で解析的に最適化できる、という特徴があり、どちらも一長一短である。

機械学習問題としては、 $(\mathbb{R}^d, \mathbb{C}, \mathcal{D}, \frac{1}{2} \sum_{i=1}^n \|Y_i - f(X_i)\|_{\mathcal{Y}}^2 + \lambda \sum_k \sum_j |\theta_{kj}|^l, -, -, (N, \lambda), \mathcal{F}_1\text{-mesurable}, \sum_{j=1}^d \sum_{k=1}^N \theta_{kj} \phi_k(x_j))$

2.2.2 カーネル回帰

一昔前に流行ったサポートベクターマシンなどにつながっていく方法である。

カーネル関数 $k: \mathcal{X}^2 \rightarrow \mathcal{Y}$ を決めておく。ハイパーパラメータは正則化項の係数 λ のみ。パラメータは $\theta := \{a_i\}_{i=1}^n, \Theta := \mathbb{R}^n$ という形で定める。

データ数をこれまでと同じく n とおいて

$$f(x) := \sum_{i=1}^n a_i k(x, X_i) \quad (9)$$

という形であらわされる関数に対してパラメータを決定する。

カーネル回帰でも正則化は同様に行う。すなわち評価関数は

$$F_D(f) := \sum_{i=1}^n \|Y_i - f(X_i)\|_{\mathcal{Y}} + \sum_{i=1}^n \|a_i\|^l \quad (10)$$

k の満たしている条件などの詳細は 4 章にて。

2.2.3 ニューラルネットワーク

最近流行りの手法である。ここではシャロウニューラルネットワークについて解説する。

ハイパーパラメータは中間層ノード数、ステップ数、学習率の3つである。これらをそれぞれ $N \in \mathbb{N}$, $Step \in \mathbb{N}$, $\epsilon \in \mathbb{R}$ と表記する。すなわち $\Theta_0 := \mathbb{N}^2 \times \mathbb{R}$

パラメータに関しては、 $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, $C \in \mathbb{C}$ と表記する。今後 a, b の存在する空間 \mathbb{R}^{d+1} のことを \mathbb{Y}^{d+1} と書く。この a, b, C の組を N と同数用意するため、 $\Theta := (\mathbb{Y}^{d+1} \times \mathbb{C})^N$ となる。

$$f(x) := \sum_{j=1}^N C_j \eta(a_j \cdot x - b_j) \quad (11)$$

ただし、 $a_j \cdot x$ はユークリッド内積を表す。

損失関数はいつもと同じで

$$F_{\mathcal{D}}(f) := \frac{1}{2} \sum_{i=1}^n \|f(X_i) - Y_i\|_{\mathcal{Y}}^2 \quad (12)$$

機械学習問題としては $(\mathbb{R}^d, \mathbb{C}, \mathcal{D}, \frac{1}{2} \sum_{i=1}^n \|Y_i - f(X_i)\|_{\mathcal{Y}}^2, -, -, (N, Step, \alpha), \mathcal{F}_1 - measurable, F_{\mathcal{D}}(f) := \sum_{i=1}^n \|f(X_i) - Y_i\|_{\mathcal{Y}}^2)$

このニューラルネットワーク、学習結果を解析的に計算するのはまず不可能である。（実はノード数が無限大なら解析的に計算する方法がある。その理論と、それを用いた手法については5章で解説する）そのため、実際にはパラメータによって生まれる勾配からパラメータを更新していく。

すなわち、 m ステップ目のあるパラメータ θ_j^m を θ_j^{m+1} に更新する式は

$$\theta_j^{m+1} := \theta_j^m - \epsilon \nabla_{\theta_j} \frac{\partial F_{\mathcal{D}}(f)}{\partial \theta_j} \Big|_{\theta_j^m} \quad (13)$$

すなわち、損失関数を θ の関数ととらえて、各 θ_j ごとに傾きの方向大きさに応じてそちら方面に摺り寄せる。これを Step 回繰り返す。

ニューラルネットはこの3つの手法の中でもっとも表現能力が高く、過学習もしやすいし、極値にもハマリやすい。正則化もパラメータが明らかに等価に扱えるものではないため、そのままノルムを計算してもうまくいかない。

過学習や極値にハマるリスクを抑える方法はドロップアウトや確率的勾配法などいろいろとあるが、本書では解説しない。

3 ベイズ学習

この章では、ベイズ論的機械学習についてもう少し掘り下げる。ベイズを用いるため、 $\hat{\theta}_0$ は \mathcal{F}_1 -可測だが、 $\hat{\theta}$ は \mathcal{F}_1 -可測ではない。ベイズ論に深入りすると一生機械学習の議論に戻ってこれなくなるため、簡単に入口だけを解説する。

3.1 ベイズの定理と共役事前分布

ベイズの手法を用いる時も、普通に前章と同じく $\theta^*(\theta) := f$ を決めておく。ただし、問題は θ が時刻 1 で可測になっていない点にある。そのため

$$\hat{f}(x) = \int_{\Theta} f(x, \theta) p(\theta | \mathcal{D}) d\theta \quad (14)$$

とする（要するに \mathcal{F}_1 条件付き期待値）。果たしてこんなものを解析的に計算するなどということが、本当にできるのだろうか。

もちろん一般には現実的なリソースで計算できるものではない。実際にこれを近似計算する方法のうち代表的なものは 2 つあり、

1. MCMC を利用する
2. 共役事前分布を利用する
3. 変分推論

2.2 項ではこの 1.2 の手法について解説する。

定理 3.1. ベイズの定理

$\hat{\theta} = \theta$ である時のデータ D_i の尤度を $p(D_i | \theta)$ と表記する。

このとき、任意の $\theta \in \Theta$ に対して、ある定数 W が存在し、次の式が成り立つ

$$p(\theta | \mathcal{D}) = \frac{1}{W} p(\theta) \prod_{i=1}^n p(D_i | \theta) \quad (15)$$

この W を解析的に表現するのは簡単だが、実際にパソコン上で求めるとなると非常に難しい。ここでの 2 つの手法は、どちらもこの定数を計算せずに済む方法である。

3.1.1 マルコフ連鎖モンテカルロ法

まず、基本的な考え方として「比がわかればその具体的な値を知る必要はない」というものがある。

例えば、箱の中にたくさんの青いボールと赤いボールが入っており、赤いボールを引けば 100 万円もらえるが、青いボールを引くとなにももらえないという状況があったとする。このとき、引く人間の関心ごとは「赤いボールと青いボールの個数の比は」であって、それさえわかれば「赤いボールと青いボールがそれぞれ何個入っているか」は気にならないはずだ。そして、ボールを一個引き出したときにもらえる金額の期待値を求めたいときにはどうすればいいか。最も手っ取り早いのは、「適当に大量のボールを取り出して、その中の赤と青の比を見る」ことであろう。

このモンテカルロ法はそういった発想に基づいている。サンプリングされる θ の確率分布が $p(\theta|\mathcal{D})$ に近くなるような方法で、多数の θ をサンプリングしてくるのである。

$\{\theta_1, \theta_2, \dots, \theta_N\}$ と N 個の θ を事後分布に従うようサンプリングしてきたとする。そうしたら、平均値 $\frac{1}{N} \sum_{i=1}^N f(x, \theta_i)$ は、 $N \rightarrow \infty$ により $\int_{\Theta} f(x, \theta) p(\theta|\mathcal{D}) d\theta$ に \mathcal{X} 上ほとんど至るところ確率 1 で収束する。

事後分布からの具体的なサンプリング方法については、今回は説明しない。

3.1.2 共役事前分布

実は、事前分布と尤度関数の間にある関係を仮定すれば、MCMC のような複雑な手法を使うよりも圧倒的に簡単な方法で事後分布を直接計算できる。実際には分析対象がよほどアレなデータでない限り、こちらの方法を用いることが多い。

事前分布と尤度関数が「ガンマ分布-ポアソン分布」など、特定の組み合わせに該当すれば、事後分布は事前分布のパラメータを変更したものになる。

3.2 誤差の見積もり

ベイズの特徴としては、「判断誤差を見積もれる」という点がある。

関数 $f(x)$ の代わりに $\mathcal{X} \times \mathcal{Y}$ 上の密度関数 $p: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ を推定したい。上記の計算と同じように

$$\hat{p}(y|x) = \int_{\Theta} p(y|x, \theta) p(\theta|\mathcal{D}) d\theta \quad (16)$$

という計算で x に対する周辺分布が計算できる。ただし、 $p(y|x)$ は $D = (X, Y)$ に対して $X = x$ が成り立つときの条件付き密度。 $p(y|x, \theta)$ も同様に定義される。

3.3 変分推論

変分推論とは、 Θ 上のわかりやすい密度関数 $q(\theta)$ を、なるべく $p(\theta|\mathcal{D})$ に似せよう、という手法である。実現可能な q の集合を \mathcal{Q} とおき。

$$\hat{q} := \operatorname{argmin}_{q \in \mathcal{Q}} KL(q||p) \quad (17)$$

によって q を求める。

$KL(q||p)$ は、カルバックライブラー情報量と呼ばれる基準で、 p と q の間の、ある種の「距離」を表す。

ベイズの測度論を用いた厳密な定式化は、主に渡辺澄夫氏によって研究されている。渡辺氏の理論では平気な顔して代数幾何が登場するので、興味がある人はぜひ調べてみよう。ここではそのような領域にまでは立ち入らない。情報量基準による議論を微分幾何学でゴリ押しする情報幾何学もおすすめだ。

4 再生核ヒルベルト空間とカーネル法

この章では、前章で簡単にしか触れられなかったカーネル法の、もっと詳細な理論的背景の入り口を解説する。

4.1 再生核ヒルベルト空間

定義 4.1. カーネル関数

写像 $k: \mathcal{X}^2 \rightarrow \mathcal{Y}$ が対称非負定値関数であるとは、任意の $x_1, x_2, \dots, x_n \in \mathcal{X}$ に対して

$$k(x_i, x_j) = k(x_j, x_i) \quad (18)$$

$$\text{任意の } c_1, c_2, \dots, c_n \in \mathbb{R} \text{ に対して、} \sum_{i=1}^n \sum_{j=1}^n c_i c_j k(x_i, x_j) \geq 0 \quad (19)$$

が成り立つことである。この時、 k を \mathcal{X} 上のカーネル関数であるという。

定義 4.2. 再生核ヒルベルト空間

写像 $\mathcal{X} \rightarrow \mathcal{Y}$ 全体の部分集合でヒルベルト空間となるもの \mathcal{H} が、再生核 k を持つ再生核ヒルベルト空間であるとは、次が満たされることを言う。

$$\text{任意の } x \in \mathcal{X} \text{ に対して、} k(\cdot, x) \in \mathcal{H} \quad (20)$$

$$\text{任意の } x \in \mathcal{X}, f \in \mathcal{H} \text{ に対して、} \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad (21)$$

厳密に言えば、対称非負定置関数に対して、それを核とする再生核ヒルベルト空間が一意に存在する。

4.2 カーネル表現定理

上記の設定を踏襲したうえで、 $f \in \mathcal{H}, b \in \mathcal{Y}$ と置く。前々章の問題設定でいうと $\mathcal{Z}_M = \mathcal{H} + \mathcal{Y}$ と書ける。ここで損失関数は一般には次のように表記できる。

$$F_{\mathcal{D}}(f + b) := L(f + b; \mathcal{D}) + \Psi(\|f\|_{\mathcal{H}}) \quad (22)$$

ただし、 $L: \mathcal{Z}_M \rightarrow \mathbb{R}, \Psi: \mathbb{R} \rightarrow \mathbb{R}$ であり、 Ψ は単調非減少関数である。このとき、次の定理が成り立つ。

定理 4.1. カーネル表現定理

$f + b \in \mathcal{Z}_M$ とする。このとき、 $F_{\mathcal{D}}$ を最小化するような $\hat{f} \in \mathcal{H}, \hat{b} \in \mathcal{Y}$ が存在すれば、そのような $f \in \mathcal{H}$ の中に

$$f(x) = \sum_{i=1}^n \hat{a}_i k(x, X_i) \quad (23)$$

と表記できるものが存在する。

ただし、 $\{\hat{a}_i\}_{i=1}^n \in \Theta$ である。

5 ニューラルネットワークの積分表現理論

この章では、「中間層のノード数が無限大であれば学習結果を解析的に決定できる」という理論に基づく、ニューラルネットのアルゴリズムについて簡単に解説する。

中間層のノード数が無限大だった場合、ニューラルネットの計算はある関数 $T(a, b)$ を用いて次のように書ける。

$$f(x) = \int_{\mathbb{Y}^{d+1}} T(a, b) \eta(a \cdot x - b) da db \quad (24)$$

近似したい関数 $f: \mathbb{R}^d \rightarrow \mathbb{C}$ に対して、リッジレット関数 ψ と活性化関数 η が適当な条件を満たす場合、リッジレット作用素と呼ばれる線形作用素 \mathcal{R}_ψ を用いて

$$f(x) = \int_{\mathbb{Y}^{d+1}} (\mathcal{R}_\psi f)(a, b) \eta(a \cdot x - b) da db \quad (25)$$

が \mathcal{X} 上ほとんどいたるところで成り立つ。つまりこのリッジレット作用素によって、学習結果であるはずの $T(a, b)$ を解析的に計算できてしまう。

もちろん実際には有限のノードしか使えないので近似が必要だが、この理論を用いて初期値を決定すると、精度、速度ともに向上がみられる。

5.1 活性化関数がある場合

まずは $\eta \in L^\infty(\mathbb{R} \rightarrow \mathbb{C})$ が成り立つ場合について解説する。

定義 5.1. リッジレット変換

$f \in L^1(\mathbb{R}^d \rightarrow \mathbb{C})$ のリッジレット関数 $\psi \in L^\infty(\mathbb{R} \rightarrow \mathbb{C})$ によるリッジレット変換は、次のように定義される。

$$(\mathcal{R}_\psi f)(a, b) := \int_{\mathbb{R}^d} f(x) \overline{\psi(a \cdot x - b)} |a|^s dx \quad (26)$$

ただし、 s は非負実数。0 か 1 を使うことが多い。

このリッジレット作用素の双対作用素はこのようになる。

定義 5.2. 双対リッジレット変換関数 $T \in L^1(\mathbb{Y}^{+K} \rightarrow \mathbb{C}; |a|^{-s} da db)$ の、活性化関数 $\eta \in L^\infty(\mathbb{R} \rightarrow \mathbb{C})$ による双対リッジレット変換は

$$(\mathcal{R}_\psi^* T)(x) := \int_{\mathbb{Y}} T(a, b) \eta(a \cdot x - b) |a|^{-s} da db \quad (27)$$

となる。

このリッジレット関数と活性化関数に対して、次の許容条件を定義する。

定義 5.3. 許容条件関数 f に対して、 \hat{f} で f のフーリエ変換を表すとする。

次の積分 $K_{\eta,\psi}$ が収束し、かつ 0 でないとき、この活性化関数とリッジレット関数の組 (η, ψ) は許容的であるという。

$$K_{\eta,\psi} := (2\pi)^{d-1} \int_{\mathbb{R}} \frac{\overline{\hat{\psi}(\xi)} \hat{\eta}(\xi)}{|\xi|^d} d\xi \quad (28)$$

そして、次の式がニューラルネットによる近似可能性の一つの答えを出す定理である。

定理 5.1. 再構成公式 $f, \hat{f} \in L^1(\mathbb{R}^d \rightarrow \mathbb{C})$ で、 (η, ψ) は許容的な組み合わせであるとする。

$$\mathcal{R}_\eta^* \mathcal{R}_\psi f(x) = K_{\eta,\psi} f(x) \quad (29)$$

が、 \mathcal{X} 上 $a.e$ で成り立つ。

5.2 活性化関数がある有界でない場合

機械学習に取り組んだことのある方は疑問に思ったことだろう。「じゃあ η が Relu や Swish のような、どう見ても L^∞ じゃない場合はどうするのか」と。

実際既存の積分表現理論は、活性化関数の有界性に甘えていた。しかし、Relu が有用である理由の一つに「勾配が消失しない」というものがあり、これは非有界性と切り離すことができない。2015 年に出た論文である [10] によって、ようやく非有界な活性化関数の一部に対しても積分表現理論を行えるようになった。([10] を読めば、非有界な活性化関数を持つニューラルネットを数学的に扱うことがいかに大変で、論文著者が苦労したかよくわかる)

活性化関数のクラスを拡張した代償は、リッジレット関数の範囲である。活性化関数を L^∞ から \mathcal{S}' に広げた代わりに、リッジレット関数がとれる範囲は L^∞ から \mathcal{S} まで縮小している。

それに加え、一定の条件を満たせばリッジレット変換は L^2 に一意的に拡張でき、 L^2 関数に対する再構成公式も定義できる。

この関数解析的な議論を活かしたアルゴリズムについては [5] を参照。

6 数学×機械学習

この章では、ある程度機械学習の経験があり、数学知識も豊富な人向けに、機械学習の数値計算への応用や、逆に数値計算の技法を機械学習に応用する方法について解説する。

6.1 ニューラルネットワークの数値解析への応用

ここでは、PDEs（放物型偏微分方程式）の境界値問題への数値計算に応用する手法について解説する。

PDE の数値解析理論は、現状空間が 4 次元以上だと誤差がひどいことになってしまう。株式の銘柄の数だけ次元が増えるファイナンスなど、SDE や PDE の応用分野ではこれが大いに問題になることも多い。

今回扱う半線形放物型 PDE の境界値問題は次のような形をしている。

$$\frac{\partial u}{\partial t}(t, x) + \frac{1}{2}(\Delta_x u)(t, x) + f(u(t, x), (\nabla_x u)(t, x)) = 0 \quad (30)$$

$$u(T, x) = g(x) \quad (31)$$

ただし、 x の取りうる空間は \mathbb{R}^d , $f: \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$, $g: \mathbb{R}^d \rightarrow \mathbb{R}$ は既知の連続関数であるとし、 $u: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ は $C^{1,2}$ 級とする。

定理 6.1. 非線形ファインマンカツツの公式

W を d 次元標準ブラウン運動。確率空間をそこから生成されるものとし、 Y, Z は適合過程で、 f, g は先ほどと同じで、 $\xi \in \mathbb{R}^d$ とする。

$$Y_t = g(t, \xi + W_T) + \int_t^T f(Y_s, Z_s) ds - \int_t^T \langle Z_t, dW_t \rangle \quad (32)$$

は、上記の半線形熱方程式終端値問題の解 u に対して次を満たす

$$Y_t = u(t, X_t), Z_t = \nabla_x u(t, X_t) \quad (33)$$

ただし、 $X_t := \xi + W_t$

このブラウン運動は、学習の反復 1 回ごとにオイラー丸山法で実装する。今回、このニューラルネットワークを用いた強化学習で学習させる対象は、写像 $Z_t = F_t(X_t)$ 。通常の強化学習での評価関数にあたるものは $Y_t = Q(t, Z_t)$ と定義される。

そしてこの評価関数は、終端時刻 T にて、

$$\psi(\theta) = \|Y_T - g(X_T)\|^2 \quad (34)$$

で評価され、通常の強化学習と同じく更新されていく。

つまり、時刻が N 分割の場合、「毎回 N 回の行動をとった後に一度だけ報酬が与えられる状況での強化学習」と見なして、 X_t に対する最適戦略 Z_t と Y_t を導き出して、それを偏微分方程式の数値解析結果にするという考えである。

通常の Q-学習と違い、状態 X_{t_n} と方策 Z_{t_n} と前時刻での評価 $Y_{t_{n-1}}$ から一意的にこの式で Y_{t_n} が定まるため、その意味では楽だといえる。

$$Y_{t_{n+1}} \approx Y_{t_n} - f(Y_{t_n}, (\nabla_x u)(t_n, \xi + W_{t_n}))(t_{n-2} - t_{n-1}) + \langle (\nabla_x u)(t_n, \xi + W_{t_n}), W_{t_2} - W_{t_1} \rangle \quad (35)$$

ニューラルネットは各時刻で用意する。つまりこれは RNN ではなく、単純に N 個のニューラルネットが存在するのみである。時刻が違う中間層同士のつながりはない。

損失関数 $\psi(\theta)$ で計算された結果から、新たなパラメータを決定する。すなわち、学習 m 回目のパラメータを Θ_m とおくと

$$\Theta_{m+1} := \Theta_m - \nabla_{\theta} \psi(\Theta_m) \quad (36)$$

これは要するに確率的勾配法による学習で、ニューラルネットワークに偏微分方程式の正しい関数を近似させようとする取り組みだ。

ただし、普通の確率的勾配法はデータの集合 \mathcal{D} 上の一様測度の元で行われていたが、今回は $C([0, T])$ 上の Wiener 測度の下で確率的勾配法を行っているのである。

「ならば Adam 法などをこの手法に適用してはどうか」「ということは確率過程になるので、抽象 Wiener 空間上の解析を離散化したものと捉えて収束証明ができるのではないか」「カメロン-マルティン部分空間が絡んでくるのではないか」という当然の疑問が浮かんでくる読者もいるだろうが、それは今後の課題である。

6.2 数値解析の深層学習への応用

6.2.1 深層学習とは

これまでの章で扱ったニューラルネットワークは、入力層と隠れ層と出力層が1つずつであったが、ここからは隠れ層が2つ以上のもの、つまり N 層パーセプトロン ($N > 3$) を扱っていく。

近年、結果をあげている深層学習アルゴリズムの層数は圧倒的に深くなっている。数年前は結果を出しているのはせいぜい8層、22層程度だったが、2015年に某大会で152層のニューラルネットワークが成果を上げたのを皮切りに、現在では1000層を超えるネットワークも使われている。

6.2.2 残差学習とは

なぜそのような超深層学習が可能になったのか、そのカギを握るのは、残差学習という概念だ。

これまでの章を読んでもらえばわかる通り、通常のニューラルネットワークは入力 x に対して、学習させた写像 $H(x)$ を直接学習させる。

しかしこの手法は3~6層程度のパーセプトロンなら有効だが、層の数が劇的に増えてくると意味をなさなくなってくる（勾配消失・発散）。それを突破するためにいろいろな方法が考えられたが、どれも中間層の数が数百、数千層単位になる超深層学習には対応できなかった。

そこで登場するのが残差学習である。学習させる対象は、 $H(x)$ ではなく、残差 $F(x) := H(x) - x$ 、これを学習させ、そこに x を加えた値を次の層の入力とする。

ここで、この残差関数 F の、 n 番目の層での値を $F(t_n, \cdot)$ と表記する。また、 n 番目の層の入力を x_{t_n} と表記する。

すると、この残差学習はこう書ける。

$$x_{t_{n+1}} = x_{t_n} + F(t_n, x_{t_n}) \quad (37)$$

これまで数学を学んできた読者ならすでにお気づきだろう。これは常微分方程式のオイラー近似に他ならない。

すなわち、残差学習による超深層ニューラルネットで行われていることは、常微分方程式の離散近似なのである。

近年の主な超深層学習と常微分方程式数値計算の関係を簡単にまとめると

$$ResNet- \quad \quad \quad \text{前進オイラー法} \quad (38)$$

$$RevNet- \quad \quad \quad \text{後退オイラー法} \quad (39)$$

$$PolyNet- \quad \quad \quad \text{前進オイラー法 (連立常微分方程式)} \quad (40)$$

$$FractalNet- \quad \quad \quad \text{2 次ルンゲクッタ法} \quad (41)$$

さらに [5] の著者は、別の常微分方程式の数値計算技法を使うことで、さらなる計算の高速化及び精度の向上を可能にした。

6.2.3 確率的超深層学習と確率微分方程式

ここまで書かれたのは、決定論的な超深層学習である。ここにノイズ付加やドロップアウトなどの確率要素を加えると、必然これらの議論は確率微分方程式の話に突入する。

shake-shake-model

しゃけしゃけモデルとは、その「シェイクシェイク」と 2 度も言うことからわかる通り、まさに計算結果をシェイクするモデルである。

γ_n を (0,1) 上の一様分布に従う確率変数とする。これはブロックごとステップごとに新たに生成される乱数である。そして n ブロック目の入力を X_n と置くと

$$X_{n+1} = X_n + \gamma_n f_1(X_n) + (1 - \gamma_n) f_2(X_n) \quad (42)$$

これは、確率微分方程式の離散近似と見做せ、[7] の著者は別の SDE 数値解析技術を用いて精度の向上を行った。

6.3 確率的勾配法と確率微分方程式

最近、機械学習界限（の極一部）で盛り上がっている考察である。

近年は、ミニバッチ学習のことを確率的勾配法と呼ぶことが多く、この場でもそれに倣う。

通常のバッチ学習において、勾配はどのように決定される。

$$\frac{dC}{d\omega} := \sum_{i=1}^n \frac{dC_i}{d\omega} \quad (43)$$

ただし、 C_i は、データ D_i に対する損失である。

確率的勾配法においては、勾配はこう書ける。

$$\frac{d\hat{C}}{d\omega} := \frac{n}{B} \sum_{j=1}^B \frac{dC_{i_j}}{d\omega} \quad (44)$$

確率的勾配法における 1 ステップごとのパラメータ変化は

$$\Delta\omega = \frac{\epsilon}{n} \left(\frac{dC}{d\omega} + \left(\frac{d\hat{C}}{d\omega} - \frac{dC}{d\omega} \right) \right) \quad (45)$$

と書ける。ここで $\frac{dC}{d\omega}$ は、 \mathcal{F}_t -可測だが、 $\alpha := (\frac{d\hat{C}}{d\omega} - \frac{dC}{d\omega})$ はステップごとに独立な確率変数である。 $E[\alpha] = 0, E[\alpha^2] = n(\frac{n}{B} - 1)F(\omega)$ ($F(\omega)$ は ω 毎に定まる定数) であり、 α は \mathcal{F}_{t+1} -可測であるため、伊藤の表現定理より、ある確率 1 で $[t, t+1]$ 上 L^2 である適過程 $f(s)$ が存在し

$$\alpha = \int_t^{t+1} f(s) dB_s \quad (46)$$

と表記することができる。 α の分散と伊藤積分の等長性より、 $f(s) = \sqrt{n(\frac{n}{B} - 1)F(\omega)}$ が成り立ち、それによって (45) 式は

$$d\omega = \frac{\epsilon}{n} \frac{dC}{d\omega} dt + \frac{\epsilon}{n} f(t) dB_t \quad (47)$$

という確率微分方程式の離散化だと見做すことができる。

だから何だと思うかもしれない。同感だ。実際、確率的勾配法を確率微分方程式と見做すことによる明確な御利益は現時点ではない。([11] ではこれによって最適なミニバッチサイズがバッチサイズに比例することを導き出したと主張しているが、それは α の平均分散を求めれば十分であり、わざわざ SDE の議論に持ち込む意味はない。おそらく論文著者が確率微分方程式を使ってみたかっただけである)

応用数学には予言性が必要だ。つまり、「数学的なこんな概念で記述できた」だけでは意味はなく、その応用先の学問を発展させて初めて価値を持つ。カタストロフィー理論はそれができずに廃れた。

例えば、積分表現理論はそれによってアルゴリズムの改善につながり、またニューラルネットによる近似可能性を証明したから価値がある。カーネルトリックの研究も、それによってカーネル法がうまくいく理由を証明できたことによって存在価値が保証された。6-2 で解説した残差学習を微分方程式の離散化と捉えた研究も、このことにより超深層学習の設計に対して新たな一石を投じたから意味があるのだ。

そのため、この SGD を SDE と捉える考え方にも、現時点ではほぼ価値がない。だが、この考察がなんの意味もない無駄なものだったかという、私はそうは思わない。この先、確率解析学の言葉を用いた議論による発展で、新たなアルゴリズムが開発されたり、「経験的にうまくいっている」だけだったアルゴリズムに収束性の証明が成されたりするかもしれないのだ。

私はここからマリアヴァン解析の議論に持ち込み、渡辺-吉田理論を用いて漸近的な性質を示したりできるのではないかと考えているが、それは今後の課題である。

7 あとがき

如何だったろうか。正直、(原稿締め切りとページ数制限と就活の狭間で慌てふためいたせいで) まだまだ解説しきれていない部分はたくさんある。特に、解説できたのがマルチタスクでない回帰問題のみというのが痛い。これからも加筆修正して github に上げていくが、お前の遅い筆を待つ気はない！という方はぜひとも参考文献リストの書籍や、その他書籍や Qiita 記事などを読んでいってほしい。

ここまで読んでくれた方なら、一体機械学習とは何をやっているのか、機械学習の背後にはどれだけ数学的な議論が隠れているのか、その一端を知ることができているはずだ。

本書の内容を理解した方であれば、きっと一般向けの機械学習書籍を読んでいくこともできるようになっているかと思う。是非とも、機械学習の勉強を進め、積極的に実践経験を積んでいってほしい。この手の分野は自分で手を動かしてナンボだ。本書はそこに辿り着くまでの手助けをしているに過ぎない。

では、読者の数学徒諸君と、機械学習の議論ができる日を楽しみにしながら、いったん筆を置くことにする。本書は常に修正・加筆に取り組んでいるため、意見等があったら気軽にメッセージを送ってほしい。面白い機械学習の数学的研究の資料の紹介なども大歓迎だ。

参考文献

- [1] 確率論 (実教理工学全書), 西尾真紀子 (1978)
- [2] Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series), Kevin P. Murphy (2012)
- [3] ルベーク積分入門-使うための基礎と応用-, 吉田伸生 (2006)
- [4] 関数解析, 黒田成俊 (1980)
- [5] 深層ニューラルネットの積分表現理論, 園田翔 (2017)
- [6] Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations, Weinan E, Jiequn Han, Arnulf Jentzen (2017)
- [7] BEYOND FINITE LAYER NEURAL NETWORKS: BRIDGING DEEP ARCHITECTURES AND NUMERICAL DIFFERENTIAL EQUATIONS, Yiping Lu, Aoxiao Zhong, Quanzheng Li (2017)
- [8] カーネル法の新展開—その理論と応用—, 福水健次 (2012)
- [9] 統計的学習理論, 金森敬文 (2015)
- [10] Neural Network with Unbounded Activation Functions is Universal Approximator, Sho Sonoda, Noboru Murata (2015)
- [11] A BAYESIAN PERSPECTIVE ON GENERALIZATION AND STOCHASTIC GRADIENT DESCENT, Samuel L. Smith, Quoc V. L. (2017)
- [12] Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis, Maxim Raginsky, Alexander Rakhlin, Matus Telgarsky (2017)
- [13] 数理統計学, 吉田朋友 (2006)