機械学習原論

2018 年 2 月 13 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。そうなると「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で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、そう苦労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、微分方程式論とする。

後半では高度な数学を用いた機械学習研究について解説する。こちらでは普通に院レベル以上の解析学を用いる。

2　機械学習入門

2.1　問題設定

2.1.1　大枠設定

まずおおもとの確率空間 $(\Omega, \mathcal{F}, P)$ 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。

線形位相空間 $\mathcal{X}$ を特徴量空間、線形位相空間 $\mathcal{Y}$ 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、独立同分布な確率変数列 $\mathcal{D} \coloneqq \{D_i(\omega)\}_{i=1}^n$ を考える。各 D_i は $D_i \colon \Omega \to \mathcal{X} \times \mathcal{Y}$ 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

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

連続写像 $f \colon \mathcal{X} \to \mathcal{Y}$ 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 $F \colon \mathcal{Z} \to \mathbb{R}$ 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 $F^* \colon \Omega \to \mathcal{Z}^*$ によって定義される評価関数 $F_{\mathcal{D}} \coloneqq F^*(\omega)$ に対して

$$\hat{f} \coloneqq argmin_{f \in \mathcal{Z}_M} F_{\mathcal{D}}(f) \tag{2}$$

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 $\theta^* \colon \Theta_0 \times \Theta \to \mathcal{Z}$ はあらかじめ一つ固定しておき、確率変数 $\hat{\theta}_0(\omega) \colon \Omega \to \Theta_0, \hat{\theta}(\omega) \colon \Omega \to \Theta_1$ を考える。実際に機械学習問題を解くにあたっては、この $\hat{\theta}_0, \hat{\theta}$ 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 $\theta^*(\theta_0, \theta)$ を $f(\cdot, \theta_0, \theta)$ 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。

(ここで、聡明な読者は「$\mathcal{Z}_M = Im(\theta^*)$ でいいのでは？」と思うかもしれない。しかし、これではベイズ的な機械学習おいて非常に困る事態が起こってしまう。$Im(\theta^*) \subset \mathcal{Z}_M$ 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。また、$\hat{\theta}$ の事前分布を超事前分布と呼ぶ

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 $(\mathcal{X}, \mathcal{Y}, \mathcal{D}, F_{\mathcal{D}}, p(\theta), p(\theta|\mathcal{D}), p(\theta_0), \hat{\theta} - mesurable)$ と表記し、これを今後「機械学習問題」と呼ぶことにする。

機械学習問題において、$\hat{f}$ は次のように計算できる。

$$\hat{f}(x) := E[f(x, \hat{\theta}_0, \hat{\theta})|\mathcal{F}_1] \tag{3}$$

厳密に言えば、これが最適な $\hat{f}$ となるように $\hat{\theta}$ を構成するのが機械学習である。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。

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 ベイズの定理と共役事前分布

この項では、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。

$$\hat{f}(x) = \int_{\Theta} f(x, \theta_0, \theta) p(\theta|\mathcal{D}) d\theta \tag{4}$$

となるわけだが、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

2.2 項ではこの 2 つの手法について解説する。

定理 2.1. 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

$$p(\theta|\mathcal{D}) = \frac{1}{W} p(\theta) \Pi_{i=1}^{n} p(D_i|\theta) \tag{5}$$

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。

2.2.1　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」

2.2.2　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。

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] ルベーグ積分入門-使うための基礎と応用-, 吉田伸生 (2006)