数学科向け機械学習入門

2018 年 8 月 31 日

目次

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。

3 章まで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。4 章はその限りではない。

2　機械学習入門

2.1　問題設定

2.1.1　大枠設定

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

$\mathcal{X}(:= \mathbb{R}^d)$ を特徴量空間、$\mathcal{Y}(:= \mathbb{R}^m)$ 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。

写像 $f : \mathcal{X} \to \mathcal{Y}$ のうち、その場で解析の対象とするものの集合となる可分ヒルベルト空間を H と置く。（これを「選択されたモデル」と言い、その具体例は後述する）

データ $D := (x_i, y_i)_{i=1}^n$[*1]から、$y_i = f(x_i) + \epsilon_i$($\epsilon_i$ は $\mathcal{Y}$ 上に値をとる何らかのノイズ。多くはガウシアンノイズ) という状況を仮定し、実際に裏で動いている関数 f に対する解析を行っていくのが機械学習である。往々にして、新たに与えられた $x \in \mathcal{X}$ に対して、$y \in \mathcal{Y}$ の値（もしくはその分布）を推定していくことになる。

2.1.2　頻度論とベイズ論

頻度論では、何らかの良い $f \in H$ を選択し、その f をもって、新たな x に対して $\hat{y} = f(x)$ という形で推測を行う。

ベイズ論と呼ばれる手法では、まず H 上の確率測度 $dp_H(f)$ を仮定する。そして尤度関数によって修正された同じく H 上の確率測度 $dp_H(f|D)$ を考える。

$$dp_H(f|D) \propto L(D|f) dp_H(f) \tag{1}$$

ただし、$L(D|f)$ は「真の写像が f だったときのデータ D の尤度」である。

最尤推定では、$L(D|f)$ が最大になる f を真の f と考える。MAP 推定と呼ばれる手法では、H が有限次元ユークリッド空間と同型で、$dp_H(f|D), dp_H(f)$ がルベーグ測度に対して絶対連続である場合に、ラドンニコディム導関数 $dp_H(f|D)/dh$ が最大になるような f を真の f と考える。

ベイズ推定では新たな特徴量 x に対して、$\mathcal{Y}$ 上の確率密度 $p_{\mathcal{Y}}(y|x)$ を次のように定義する。

$$p_{\mathcal{Y}}(y|x) := \int_H p_{\mathcal{Y}}(y|x, f) dp_H(f|D) \tag{2}$$

$p_{\mathcal{Y}}(y|x, f)$ は「x, f を固定したときの $\mathcal{Y}$ の確率密度」で、ノイズ ϵ がガウス型である場合、これは正規分布となる。

式を見れば容易にわかる通り、$dp_{\mathcal{Y}}(y|x)$ とは、x が与えられた時の y の分布を決定づける測度であり、「予測分布」と呼ばれる。実用上は、x が与えられた時の y の推定値は、$\hat{y} := argmax_y p(y|x)$ で決められる。

ベイズ推定の利点は、「予測の自信」を測れるところにある。ノイズがガウス型であれば、同じデータの下で行われる頻度論とベイズ論の結果は一致する。しかし頻度論は一つの $\hat{y}$ が推定値であることしかわからない

[*1] 実際には x と y は確率変数と考えたほうが、数理統計学的には自然である。ϵ_i を x_i とは独立とすることで、ノイズの立ち位置を明確にできる。しかし特にそれ以外にご利益がなく、ランダム測度などの難解な議論を持ち出さねばならず、表記が非常に複雑怪奇になっていまうため、特に指定がなければ本書ではデータは与えられた定数であると考える。後述の生成測度の概念を用いた解析の際は、この仮定は外される

が、ベイズ推定では $p_{\mathcal{Y}}(y|x)$ を計算できるので、例えば「真の y が $\hat{y} \pm 0.001$ に収まる可能性は何 % か？」という疑問にも答えを出すことができる。

今後は、特に誤解の恐れがなければ連続濃度な標本空間上の確率測度をすべて dp と書き、標本空間を明記しない。また、dp がルベーグ測度に対して絶対連続であるとき、p でその確率密度関数を表すものとする。

2.1.3 分類問題と回帰問題

機械学習には、大きく分けて分類問題と回帰問題の 2 種類に分けられる。

一応これに該当しない分野も存在するが、概ねそれらもこの二種類に帰着できる。

分類問題とは、有限集合 $C := \{c_1, c_2, ..., c_m\}$ に対して、データ $(x_i, c_{j_i})_{i=i}^n$ が与えられた上で (ただし $j_i \in \{1, 2, 3, ..., m\}$) 新たな $x \in \mathcal{X}$ に対して、クラス $c \in C$ を推定する。

このままではこれまでに書いてきた写像 $f : \mathbb{R}^d \to \mathbb{R}^m$ を選ぶという問題に持ち込むことができない。そこで、f は C 上に値をとるのではなく、C の濃度と同じ次元のユークリッド空間 $\mathbb{R}^m$ に値をとる写像であると考える。

頻度論の場合、これまでと同じように構成された写像 $\tilde{f}$ による出力 $\tilde{y} := \tilde{f(x)} \in \mathbb{R}^m$ に対して、ソフトマックス関数 $sm(y)$ で次のように x に対して定まる C 上の確率測度を $\mathcal{Y}$ 上に定義することで、真の写像 $y := sm(\tilde{f}(x))$ を構成する。

$$y_k := f(x)_k := P_C(c_k|x) := sm(\tilde{f}(x)) := \frac{exp[\tilde{y_k}]}{\Sigma_{i=1}^m exp[\tilde{y_i}]} \tag{3}$$

ただし $k \in \{1, 2, ..., m\}$ である

厳密に言えば、このように構成された $P_C(c_k|x)$ が「よい」確率測度となるように f を構成したい。

各入力データは $(x_i, y^{(i)}) := (y_{j_i}^{(i)}$ のみ 1, 他は 0) と考えれば、前項の「入力データは $\mathcal{X} \times \mathcal{Y}$ の部分集合」という前提で語られた問題に帰着できる。

最後に、$argmax_{c_k \in C} P_C(c_k|x)$ により、x に対する推定クラス c_k を決定する。

ベイズ論では予測分布 $p(y|x)$ を考えて、そこから上記のソフトマックス関数にかけてもいいが、[*2]ナイーブベイズ分類機という概念を用いるのが早い。

ベイズの定理を用いて

$$P(c_k|x) := \frac{p(x|c_k)P(c_k)}{p(x)} \tag{4}$$

2.1.4 損失関数と勾配法

2.1.5 過学習と正則化

2.2 具体例

2.2.1 線形回帰 (分類) とロジスティック回帰

2.2.2 カーネル近似

2.2.3 ニューラルネットワーク

[*2] 要するにこれで出てくるのは「予測分布の予測分布」であり、測度論的に厳密に解説するとランダム測度が登場してくる。あまりにややこしいので本書では詳細は語らない。

3　カーネルと再生核ヒルベルト空間

3.1　カーネルトリック

3.2　サポートベクターマシン

4　ニューラルネットワーク

4.1　ニューラルネットの積分表現理論

4.1.1　ニューラルネットの計算

4.1.2　リッジレット変換と積分表現

4.1.3　チコノフ正則化と大域的最適解

4.2　最近のニューラルネットの手法と応用

4.2.1　深層学習

4.2.2　ドロップアウト

4.2.3　CNN

4.2.4　RNN

4.2.5　バッチ正規化

4.2.6　様々な最適化アルゴリズム

ミニバッチ、Adam.RMSprop

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)

[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 forhigh-dimensional parabolic partial differential equationsand backward stochastic differential equations,Weinan E,Jiequn Han,Arnulf Jentzen(2017)

[7] Beyond finite layer neuralnetworks:bridging deep architectures and numerical differential equations,Yiping Lu,Aoxiao Zhong,Quanzheng Li(2017)

[8] カーネル法の新展開ーその理論と応用ー, 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 Murata1(2015)

[11] Pattern Recognition and Machine Learning,Christopher M. Bishop(2006)

[12] A bayesian perspective on generalization and stochastic gradient descent,Samuel.L.Smith , Quoc.V.L(2017)

[13] Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis,Maxim Raginsky,Alexander Rakhlin,Matus Telgarsky(2017)

[14] 数理統計学, 吉田朋友 (2006)