

機械学習入門

。

2017 年 10 月 18 日

1 確率論の基礎の復習

この章では、機械学習の数学的定式化に必要な確率論の基礎中の基礎を復習する。

目的はあくまで機械学習への応用であるため、細々とした議論は行わず、機械学習を定式化するために必要な最低限の項目だけ解説する

1. 確率空間と確率変数

標本空間 Ω に対し、次を満たす部分集合族 \mathcal{F} を σ 加法族と呼ぶ

$$1. \phi, \Omega \in \mathcal{F} \quad (1)$$

$$2. A \in \mathcal{F} \rightarrow A^c \in \mathcal{F} \quad (2)$$

$$3. \text{可算個の } \Omega \text{ の部分集合 } A_1, A_2, \dots, \text{があり、任意の } n \text{ に対して } A_n \in \mathcal{F} \text{ なら、} \cup_{n=1}^{\infty} A_n \in \mathcal{F} \quad (3)$$

写像 $P : \mathcal{F} \rightarrow [0, 1]$ が次を満たすとき、 P を確率測度という

$$1. P(\Omega) = 1 \quad (4)$$

$$2. \text{互いに素な可算個の集合 } A_1, A_2, \dots, \in \mathcal{F} \text{ に対し、} P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) \quad (5)$$

この (Ω, \mathcal{F}, P) の組を、確率空間と呼ぶ

$\mathcal{B}(\mathbb{R})$ で、 \mathbb{R} の開集合をすべて含む最小の σ 加法族を表し、ボレル集合族と呼ぶ。

写像 $X : \Omega \rightarrow \mathbb{R}$ が次の条件を満たすとき、(1 次元実) 確率変数であるという

$$\text{任意の } B \in \mathcal{B}(\mathbb{R}) \text{ に対して、} X^{-1}(B) \in \mathcal{F} \quad (6)$$

ここで、 $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P \circ X^{-1})$ は確率空間となる

確率変数 X に対して、分布関数 $F_X : \mathbb{R} \rightarrow [0, 1]$ を次のように定義する

$$F_X(x) := P(X \leq x) \quad (7)$$

今回扱う確率変数は、その確率変数に対応する確率測度 $P \circ X^{-1}$ が \mathbb{R} 上のボレル測度 μ に対して絶対連続であるものとし、ボレル測度とのラドンニコディム導関数を確率密度関数と呼ぶ

すなわち

$$f_X(x) := \frac{d(P \circ X^{-1})(x)}{d\mu(x)} \quad (8)$$

Ω の部分集合族 \mathcal{G} が、 σ 加法族の条件を満たし、さらに $\mathcal{G} \subset \mathcal{F}$ が成り立つなら、 \mathcal{G} を部分 σ 加法族という。
 Ω の部分集合族 \mathcal{H} に対して、 $\sigma(\mathcal{H})$ で \mathcal{H} を含む最小の σ -加法族を表すものとする。
また、確率変数 Y に対して、 $\sigma(Y)$ で Y を可測にする最小の σ -加法族を表すとする。

2. 独立性と条件付き期待値

集合 $A, B \in \mathcal{F}$ が独立であるとは、 $P(A \cup B) = P(A)P(B)$ が成り立つことである

部分 σ -加法族 $\mathcal{G}_1, \mathcal{G}_2$ が独立であるとは、任意の集合 $A_1 (\in \mathcal{G}_1), A_2 (\in \mathcal{G}_2)$ が独立となることである

確率変数 X, Y が独立であるとは、 σ -加法族 $\sigma(X), \sigma(Y)$ が独立となることである

集合 $A \in \mathcal{F}$ の、 $B \in \mathcal{F}$ に対する条件付き確率を次のように定義する

2 機械学習入門

ここからは本格的に機械学習の内容を学んでいく。

工学への応用であるため、数学的な厳密性に関しては数学書に比べればかなり適当なものになるがご容赦願いたい。

1. クラス分け

ここでは母集団の濃度は可算無限であるとする。(応用上確実に有限なのだが、限りなく大きい場合を扱う) 考察に用いたい各データの数値を、特徴量ベクトル \mathbf{x} として分類したいクラスの集合を C と置く

$\mathbf{x} \in K^d$ (K^d は位相空間。 K は \mathbb{R} や \mathbb{N} の部分集合など様々で、具体的な中身は分析対象による) (9)

$y \in C := [c_1, c_2, \dots, c_n]$ (c_i は各クラス名) (10)

これに対して、 \mathbf{x} と y が紐づけられた濃度 $m := |\mathcal{D}|$ のデータ $\mathcal{D} := [(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)]$ が与えられた上で、分類関数 $\hat{y}: K^d \rightarrow C$ を見繕いたい。

この写像がどういった場合に優れていると定義するかは、状況による。(後々具体例を書く)

確率空間 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ があったとする。ここでの標本空間は上記の母集団と異なることに注意。

確率空間上の可測関数 $X: \tilde{\Omega} \rightarrow K^d \times C$ について考えたい。ただし、値域となる空間の σ -加法族は K が可算集合ならば $\mathcal{B}(K^d) \times 2^C$ とする。

データ \mathcal{D} の入力は、 m 個の $\Omega \rightarrow K^d \times C$ な確率変数として捉えることができる ((Ω, \mathcal{F}, P) が存在するものとする)

すなわち、入力データ \mathcal{D} に対して、 $K^d \times C$ - 値確率変数となるような独立同分布な確率変数列 $[X_i]_{i=1}^m$ が

$$X_i = \mathcal{D}_i = (\mathbf{x}_i, y_i) \quad (11)$$

という値になるものと考えればよい

ここで、データ観測前の所持情報を \mathcal{F}_0 , データ観測後の所持情報を \mathcal{F}_1 と置く。これはどちらも \mathcal{F} の部分 σ -加法族で、 $\mathcal{F}_0 \subset \mathcal{F}_1$ であれば、これらは増大情報系の条件を満たし。

$$\mathcal{F}_1 = \mathcal{F}_0 \vee \sigma([X_i]_{i=1}^m) \quad (12)$$

とおける

余談:入力データを確率変数列で捉えるのは、数学的に厳密でありながら難解な数学の概念を持ち出さずに済む方法である。

これは、数学上の都合だけかというそうではなく、 \mathcal{F}_0 と \mathcal{F}_1 の間に $\mathcal{F}_{\frac{l}{m}}, l \in [1, 2, \dots, m]$ という部分 σ -加法族があると考え、 $\mathcal{F}_{\frac{l}{m}} := \mathcal{F}_0 \vee \sigma([X_i]_{i=1}^l)$ と置けば「通常データは1つずつ取り込まれていき、解析者の保持する情報が増えていく」という実際の状況にも対応させることができる

写像 $\hat{y} : K^d \rightarrow C$ の満たす集合を Y とおく。 Y の位相は後述の F を連続関数にする最弱の位相と定義する

入力データ \mathcal{D} に対して、最適な写像 \hat{y} を導き出したい。

ただし、 Y の任意の元を取ってこれるかと言えばそうではなく、ある程度制約をつけないと役に立たない（オーバーフィット、具体例を交えて後述）

選択されたモデル M (詳細は後述) に対して、 $Y_M (\subset Y)$ でモデルに対応する分類関数の集合とし、この上で最適化を行っていく。

見つけたい最適な分類関数は、このように書ける

$$\tilde{y} := \operatorname{argmax}_{y \in Y_M} [F_{\mathcal{D}}(y)] \quad (13)$$

ただし、 $F_{\mathcal{D}}$ はモデルと入力データから定まる $F_{\mathcal{D}} : Y \rightarrow \mathbb{R}$ の汎関数である（ Y は線形空間ではないので厳密には汎関数ではないが、「引数関数の関数」という意味でここでは「汎関数」と呼ぶものとする）

ベイズ統計

データ \mathcal{D} を観測する前の C 上の密度関数を事前分布と呼び、観測後の密度関数は事後確率という。

X_C で C への射影、 X_K で K^d への射影を表すものとし、 $\tilde{\mathcal{D}} := \{\omega; [X(\omega)]_{i=1}^m = \mathcal{D}(\omega)\}$ とおく

$$f_{X_C}^{\theta_0}(c) : \text{事前確率} \quad (14)$$

$$f_{X_C}^{\theta_1}(c | \{\omega; [X(\omega)]_{i=1}^m = \mathcal{D}(\omega)\}) : \text{事後確率} \quad (15)$$

ただし、 $f_X(x|A)$, $A \in \mathcal{F}$ は条件付き分布とし、確率密度関数 $f_X(x)$ に対して、 $\omega \in A$ という情報がある上で $X(\omega)$ の密度関数を表すものとする。すなわち

$$f_{X_C}^{\theta_1}(c | \tilde{\mathcal{D}}) := f_{X_C(\omega) | \omega \in \tilde{\mathcal{D}}}^{\theta_1}(c) P(\tilde{\mathcal{D}}) \quad (16)$$

ベイズの定理：

事後分布は事前分布とデータの尤度の積に比例する

すなわち、ある定数 L が存在し、任意の $c_i \in C$ に対して

$$f_{X_C}^{\theta_1}(c_i | \tilde{\mathcal{D}}) = L f_{X_C}^{\theta_0}(c_i) f_{X_K}^{\theta_0}(x | \{\omega; X_C(\omega) = c_i\}) \quad (17)$$

参考文献

<http://www2.itc.kansai-u.ac.jp/~afujioka/2014/ig/141112ig.pdf>