データサイエンス・シリーズ 6

データ学習アルゴリズム

(渡辺澄夫著, 共立出版 2001)

発表日: 平成 15 年 4 月 23 日
担当者: 岩崎 唯史
担当箇所: 第 1 章 学習と確率
1.1 学習とは
1.2 確率変数と情報科学
1.2.1 離散値をとる確率変数
1.2.2 連続値をとる確率変数
1.2.3 確率変数の変換
1.2.4 平均と分散

第 1 章 学習と確率

1.1 学習とは

本書の目的

“知識” を持たない「学習者 $=$ コンピュータ」が「先生 (教師) $=$ 人間」のすること (情報処理, 運動,...) を模倣 ($=$ 学習) する数理的基礎・枠組を紹介 (図 1.1).

- 人間が何気なく行なっていることは, コンピュータにとって実現困難.
- 「教師なし学習 (学習者が自ら概念を生成)」は「教師あり学習」と数理的に等価な問題.
- 抽象化した「学習」の概念は, 研究分野 (情報科学, 知能科学,...) や時代 (見た目, 流行,...) に依らない.

■学習モデルの推移

1980 年半ば以前:
学習者の応答から内部の構造が特定できるモデル.

1980 年半ば以降:
外部からは隠れた部分を持ち, 学習者の表面的な振舞だけからは内側は分からないモデル (本書で中心的に紹介).

■確率論の必要性

　確率論は学習という事柄と本質的な部分を共有. 「学習という問題を考えること」は「確率を考えること」とおそらく等価.*

(*) 「学習」とは目的に合うように確率分布を偏らせること.

1.2 確率変数と情報科学

1.2.1 離散値をとる確率変数

■確率変数と確率関数

要素の数が高々可算である集合 Ω の中を確率的にばらつく変数 X があり，X が Ω のある集合 ω_i になる確率が $p(\omega_i)$ のとき，X を集合 Ω 上の確率変数とよび，$p(\cdot)$ を確率関数とよぶ．ただし，$p(\omega_i)$ は次式を満たす必要がある．

$$p(\omega_i) \geq 0, \quad \sum_{\omega_i \in \Omega} p(\omega_i) = 1.$$

■試行と実現値

確率変数 X をばらつかせて値をとることを試行といい，試行によって得られる値を実現値という．

1.2.2 連続値をとる確率変数

■確率変数と確率関数

M 次元ユークリッド空間 R^M 上を確率的にばらつく変数 $\boldsymbol{X} = (X_1, X_2, \ldots, X_M)$ を考える．R^M の任意の開集合 A について，$\boldsymbol{X}$ が A の中の値をとる確率 $P(A)$ が

$$p(A) = \int_A p(\boldsymbol{x}) d\boldsymbol{x} = \int_A p(x_1, x_2, \ldots, x_M) dx_1 dx_2 \ldots dx_M,$$

となる関数 $p(\boldsymbol{x})$ が存在するとき，$p(\boldsymbol{x})$ を $\boldsymbol{X}$ の確率密度関数 (密度関数) という．また $\boldsymbol{X}$ を密度関数 $p(\boldsymbol{x})$ をもつ確率変数という．ただし，$p(\boldsymbol{x})$ は次式を満たす必要がある．

$$p(\boldsymbol{x}) \geq 0, \quad \int_{R^M} p(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

※ $p(\boldsymbol{x})d\boldsymbol{x}$ は $x_i < X_i < x_i + dx_i$ $(i = 1, 2, \ldots, M)$ となる確率を表す．$p(\boldsymbol{x})$ 自体は確率ではないことに注意．

◆ 例 3: 正規分布

S を $M \times M$ の実数値の正定値行列とする．R^M 上の確率変数 $\boldsymbol{X}$ の密度関数が

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det S}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})\cdot S^{-1}(\boldsymbol{x}-\boldsymbol{m})\right),$$

で与えられるとき，$\boldsymbol{X}$ を平均 $\boldsymbol{m}$，共分散行列 S をもつ正規分布に従う確率変数という．

◆ 例 4: 一様分布

R^1 上の開集合 $(-a, a)$ $(a > 0)$ を考える．この開集合の中の確率変数 X の密度関数が一様分布

$$p(x) = \begin{cases} \frac{1}{2a} & |x| < a, \\ 0 & |x| \geq a, \end{cases}$$

で与えられるとき，X を一様乱数という．

◆ 例 5: 分配関数と統計物理学

R^M 上のエネルギー関数 (ハミルトニアン) $H(x)$ をもつ物理学的な系が「温度 $1/\beta(=k_{\mathrm{B}}T)$ の平衡状態にある」とき, x で指定される微視的状態が実現する確率密度関数 (状態密度関数) は

$$p(\boldsymbol{x}) = \frac{1}{Z}\exp(-\beta H(\boldsymbol{x})), \quad Z = \int \exp(-\beta H(\boldsymbol{x}))d\boldsymbol{x},$$

で与えられる. ここで Z は分配関数 (状態和) と呼ばれ, 平衡状態にある系の性質を知る上で重要な関数である.

統計物理学:
微視的な系に確率統計の方法を適用し, (その集団としての) 巨視的な系の物理的性質を調べる. 無限自由度 $(M \to \infty)$ の場合には様々な数学的手法が利用できる.

例. 多くの原子分子 $(M \sim 10^{23})$ から成る系全体の振舞い.

◆ 例 6: 音声情報科学や画像情報科学

音声処理: 音声信号を確率変数と考えその密度関数を推定.
画像処理: 文字画像を確率変数と考えその密度関数を推定.

われわれが日常的に行なっていることをコンピュータに行なわせることは難 (確率変数の自由度の高さが問題).

◆ [注 2] 選択公理とルベーグ積分論

選択公理と矛盾しないように確率が定義できるか?
体積 (リーマン積分) を集合論的に一般化 $\rightarrow$ ルベーグ積分 (測度)

◆ **離散集合上の値をとる確率変数を連続値上の確率変数をして扱う**

一般化された密度関数： $\bar{p}(\boldsymbol{x}) = p(\boldsymbol{x}) + \sum_{i=1}^{\infty} p_i \delta(\boldsymbol{x} - \boldsymbol{x}_i).$

デルタ関数： $\delta(x - x_i) = \begin{cases} \infty & x = x_i, \\ 0 & x \neq x_i, \end{cases}$

$$\int_{-\infty}^{\infty} \delta(x - x_i) dx = 1.$$

開集合 A の中に値をとる確率： $\bar{p}(A) = \int_A p(\boldsymbol{x}) d\boldsymbol{x} + \sum_{\boldsymbol{x}_i \in A} p_i.$

◆ **[注 3] 超関数とフーリエ変換 (デルタ関数の積分表示)**

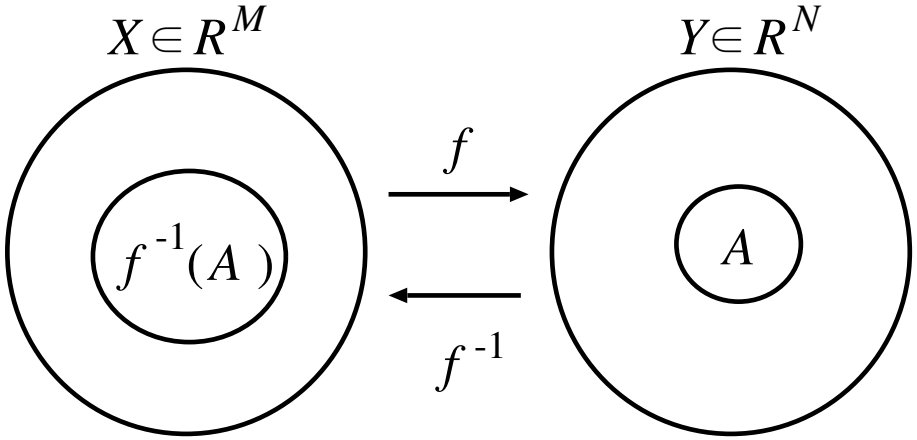
$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ikx} dk.$$

1.2.3 確率変数の変換

R^M 上の確率変数 $\boldsymbol{X}$ が密度関数 $p(\boldsymbol{x})$ をもつとする．$R^M \to R^N$ への連続な関数 $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_N(\boldsymbol{x}))$ が与えられたとき，R^N 上に値をとる確率変数 $\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X})$ の確率密度 $q(\boldsymbol{y})$ は，任意の開集合 $A \subset R^N$ に対して

$$\int_{\boldsymbol{y} \in A} q(\boldsymbol{y}) d\boldsymbol{y} = \int_{\boldsymbol{x} \in \boldsymbol{f}^{-1}(A)} p(\boldsymbol{x}) d\boldsymbol{x},$$

によって定まる．ただし，$\boldsymbol{f}^{-1}(A) = \{\boldsymbol{x} \in R^M; \boldsymbol{f}(\boldsymbol{x}) \in A\}$.



◆ 例 7: 正規分布と χ^2 分布

$$\textbf{正規分布: } p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

に従う R^1 **上の確率変数** X **を考える．** $R^1 \to R^1$ **への変換** $Y = f(X) = X^2$ **によって得られる** R^1 **上の確率変数** Y **の密度関数を** $q(y)$ **とおく．このとき**

$$\int_0^{a^2} q(y)dy \equiv \int_{-a}^{a} p(x)dx, \ \ \textbf{for}^{\forall} a \geq 0.$$

$$\textbf{自由度 1 の } \chi^2 \textbf{ 分布: } q(x) = \begin{cases} \frac{1}{\sqrt{x}} p(\sqrt{x}) = \frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{x}{2}\right), & x \geq 0 \\ 0 & x < 0 \end{cases}$$

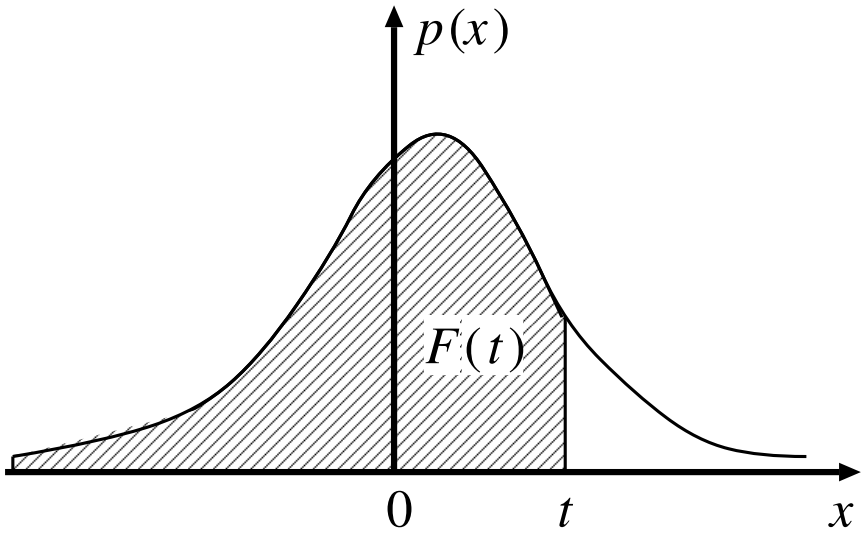
$$\textbf{(参考) 自由度 } n \textbf{ の } \chi^2 \textbf{ 分布: } q(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \ \ x \geq 0$$

$$\Gamma(m) = \int_0^{\infty} e^{-x} x^{m-1} dx, \ \ m > 0.$$

◆ 累積分布関数

確率変数 X の密度関数 $p(x)$ に対する累積分布関数

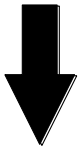
$$F(t) = \int_{-\infty}^{t} p(x)dx.$$



$$0 \leq F(t) \leq 1, \;\; F(-\infty) = 0, \;\; F(\infty) = 1, \;\; \frac{dF(t)}{dt} = p(t).$$

◆ **例 8: コンピュータ上で所望の密度関数に従う疑似乱数を生成**

区間 (0,1) の一様疑似乱数 U (生成アルゴリズム様々あり)



確率変数の変換: $X = F^{-1}(U)$

所望の密度関数に従う疑似乱数 X

◆ **例 9: 正規分布に従う疑似乱数の生成**

区間 (0,1) の互いに独立な一様疑似乱数を U_1, U_2 とする．このとき, 変数変換 $(U_1, U_2) \to (X_1, X_2)$

$$
\begin{aligned}
X_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2), \\
X_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2),
\end{aligned}
$$

によって生成される X_1, X_2 は $\mathrm{N}(0,\,1)$ の正規分布に従う疑似乱数.

◆ 例 10: 確率変数の重ね合わせと畳込み

確率変数 $X = (X_1, X_2)$ の密度関数を $p(x_1, x_2)$ とする．このとき，2 変数の重ね合わせで定義される確率変数 $Y = X_1 + X_2$ の密度関数 $q(y)$ は

$$q(y) = \int p(x_1, y - x_1) dx_1 = \int p(y - x_2, x_2) dx_2.$$

◆ 例 11: 画像識別と特徴量

画像に関する (生の) 確率変数を $X \in R^M$ とする．

良い画像認識システムを作る

カテゴリ内のばらつきが少なく (文字変形の影響が小さい)，カテゴリ間の重なりが少ない (識別能力が高い) $X \to Y \in R^N$ $(M > N)$ への変換関数 $Y = f(X)$ を見付ける (特徴を上手に抽出)．

変換後の確率変数 $f(X)$ を特徴量という．

1.2.4 平均と分散

R^M 上の確率変数 X が密度関数 $p(\boldsymbol{x})$ をもつとする．$R^M \to R^1$ への関数 f が与えられたとき, 確率変数 $f(X)$ の平均を次式で定義.

$$E(f(\boldsymbol{X})) = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x},$$

確率変数 X の平均ベクトル, 共分散行列の定義.

平均ベクトル: $$\boldsymbol{m} = E(\boldsymbol{X}) = \int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x},$$

共分散行列: $$\begin{aligned} S &= E((\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^T) = E(\boldsymbol{X}\boldsymbol{X}^T) - \boldsymbol{m}\boldsymbol{m}^T \\ &= \int \boldsymbol{x}\boldsymbol{x}^T p(\boldsymbol{x})d\boldsymbol{x} - \int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x} \int \boldsymbol{x}^T p(\boldsymbol{x})d\boldsymbol{x}. \end{aligned}$$

◆ **例 12:**

任意の実数値ベクトル u, v に対して

$$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{u}^T \boldsymbol{v} = \mathrm{tr}(\boldsymbol{v}\boldsymbol{u}^T),$$

が成り立つので, 任意の行列 A に対して次式が成り立つ.

$$\boldsymbol{u} \cdot A\boldsymbol{v} = \boldsymbol{u}^T A\boldsymbol{v} = \mathrm{tr}(A\boldsymbol{v}\boldsymbol{u}^T) = \mathrm{tr}(\boldsymbol{v}\boldsymbol{u}^T A).$$

◆ **例 13:**

確率変数 $\boldsymbol{X} \in R^M$ が平均 $\boldsymbol{m}$, 共分散行列 S の正規分布を密度関数にもつ場合, 任意の行列 A に対して次式が成り立つ.

$$\begin{aligned}
E(\boldsymbol{X} \cdot A\boldsymbol{X}) &= \int (\boldsymbol{x} \cdot A\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \int \mathrm{tr}(A\boldsymbol{x}\boldsymbol{x}^T) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathrm{tr}\left[A \int \boldsymbol{x}\boldsymbol{x}^T p(\boldsymbol{x}) d\boldsymbol{x} \right] \\
&= \mathrm{tr}\left[A \int \left((\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^T + \boldsymbol{m}\boldsymbol{m}^T \right) p(\boldsymbol{x}) d\boldsymbol{x} \right] \\
&= \mathrm{tr}(AS) + \boldsymbol{m} \cdot A\boldsymbol{m}.
\end{aligned}$$

◆ **例 14: 統計力学と熱力学を結ぶ分配関数**

ギブス分布 (例 5 参照): $p(\boldsymbol{x}) = \dfrac{1}{Z(\beta)} \exp(-\beta H(\boldsymbol{x}))$.

以下, 分配関数 $Z(\beta)$ **を用いて表した熱力学的な量.**

自由エネルギー: $F(\beta) \stackrel{\text{def}}{=} -\log Z(\beta) = -\log \left(\int \exp(-\beta H(\boldsymbol{x})) d\boldsymbol{x} \right)$.

エネルギーの平均: $E\{H(\boldsymbol{X})\} \stackrel{\text{def}}{=} \int H(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \dfrac{\partial}{\partial \beta} F(\beta)$.

エネルギーの分散:
$$\begin{aligned} V\{H(\boldsymbol{X})\} &\stackrel{\text{def}}{=} \int (H(\boldsymbol{x}) - E\{H(\boldsymbol{X})\})^2 p(\boldsymbol{x}) d\boldsymbol{x} \\ &= -\frac{\partial^2}{\partial \beta^2} F(\beta). \end{aligned}$$

エントロピー:
$$\begin{aligned} S(\beta) &\stackrel{\text{def}}{=} -E\{\log p(\boldsymbol{X})\} = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x} \\ &= \beta E\{H(\boldsymbol{X})\} - F(\beta). \end{aligned}$$

◆ **例 15:**

H **個の有限集合**

$$\mathcal{C}^H = \{\underbrace{(1,0,0,\ldots,0)}_{H\text{個}}, (0,1,0,\ldots,0), (0,0,0,\ldots,1),\}$$

上の確率変数 $U = (U_1, U_2, \ldots, U_H)$ **を競合的な確率変数という．ここで, 集合** $\{a_h; a_1 + a_2 + \cdots + a_H = 1, a_h > 0, h = 1, 2, \ldots, H\}$ **を考えると, 確率** a_h **で第** h **番目のユニットが** 1**, その他のユニットが** 0 **となる確率変数の確率関数**

$$p(u_1, u_2, \ldots, u_H) = \prod_{h=1}^{H} (a_h)^{u_h}.$$

離散集合 $\mathcal{C}^H$ **上の確率変数** U **を** R^H **上の確率変数として考え, 平均ベクトルを計算.**

$$E(U) = (a_1, a_2, \ldots, a_H).$$