

Duda & Hart

Chapter 4 NONPARAMETRIC TECHNIQUES

4.1 INTRODUCTION

4.2 DENSITY ESTIMATION

作成者: おおつかたく

February 8, 2026

Purpose and Background

4.1 INTRODUCTION

4.2 DENSITY ESTIMATION

Conclusion

Purpose and Background

目的と背景

- 前章では、基礎となる確率密度関数の形式が既知であるという仮定の下で、教師あり学習 (supervised learning) を扱った
- 本章では、基礎となる密度の形式が既知であると仮定せずに使用できる、ノンパラメトリック (非パラメトリック) な手法について検討する。

4.1 INTRODUCTION

イントロダクション

- 一般的なパラメトリック（パラメータで記述できる）な形式が、実際に遭遇する密度関数に適合することは稀なのでほとんどのパターン認識の応用において、確率密度関数の形式が既知であるという仮定は疑わしい。
- 特に、古典的なパラメトリック密度関数はすべて単峰性（極大値を1つしか持たない）だが、実際の問題は多峰性*（複数の山がある状態）の密度を含んでいることが多い。

*多峰性密度の例: 音声認識における音素の分布、画像認識における物体の位置分布など、マルチモーダルなデータセットのこと。

ノンパラメトリック手法の種類

1. 密度関数の推定: サンプルパターンから $p(\vec{x}|\omega_i)$ を推定し、設計に利用する。
2. 事後確率の直接推定: 密度推定を経由せず、直接 $P(\omega_i|\vec{x})$ を推定する。(例: 最隣接法)
3. 特徴空間の変換: パラメータ手法を適用しやすくするために空間を変換する。(例: フィッシャーの線形判別)

4.2 DENSITY ESTIMATION

未知の確率密度関数を推定する

- 基本原理: ベクトル \vec{x} がある領域 \mathcal{R} に収まる確率 P は、密度関数 $p(\vec{x}')$ をその領域で積分した値になる。

$$P = \int_{\mathcal{R}} p(\vec{x}') d\vec{x}' \quad (1)$$

- したがって、 P は密度関数 $p(\vec{x})$ の平滑化された、あるいは平均化されたものであり、この確率 P を推定することによって、この平滑化された p の値を推定することができる。

推定式の導出

- n 個のサンプル $\vec{x}_1, \dots, \vec{x}_n$ が確率法則 $p(\vec{x})$ に従って独立に抽出されたと仮定する。
- 明らかに、これら n 個のうち k 個が領域 \mathcal{R} に入る確率は、以下の二項分布の法則で与えられる。

$$P_k = \binom{n}{k} P^k (1 - P)^{n-k}.$$

そして、 k の期待値は以下となる。

$$E[k] = nP. \quad (2)$$

- この k に関する二項分布は平均付近に非常に鋭いピークを持つため、比率 k/n は確率 P （平滑化された密度関数）の非常に良い推定量になると期待される。
- ここで、 $p(\vec{x})$ が連続であり、かつ領域 \mathcal{R} が非常に小さく、その内部で p がほとんど変化しないと仮定すると、以下のようなになる。

$$\int_{\mathcal{R}} p(\vec{x}') d\vec{x}' \approx p(\vec{x}) V. \quad (3)$$

推定式の完成

$$\int_{\mathcal{R}} p(\vec{x}') d\vec{x}' \approx p(\vec{x})V. \quad (3)$$

- ここで、 \vec{x} は領域 \mathcal{R} 内の点であり、 V は \mathcal{R} によって囲まれる体積である。(1)、(2)、(3) を組み合わせると、 $p(\vec{x})$ に対する以下の推定式に到達する。

$$p(\vec{x}) \approx \frac{k/n}{V}. \quad (4)$$

実用上の課題

体積 V を固定してサンプルを増やしていけば、比率 k/n は期待通り（確率的に）収束するが、得られるのはあくまで $p(\vec{x})$ の空間平均値の推定にすぎない。

$$\frac{P}{V} = \frac{\int_{\mathcal{R}} p(\vec{x}') d\vec{x}'}{\int_{\mathcal{R}} d\vec{x}'}.$$

サンプル数 n が限られている場合、体積 V を小さくしすぎると領域内にサンプルが含まれなくなり、推定値が 0 または無限大になってしまう。そのため、ある程度の k/n の分散（バラつき）と $p(\vec{x})$ の平均化（解像度の低下）を受け入れる必要がある。

どのように回避するか

- もし無制限の数のサンプルが利用可能であるならば、 \vec{x} における密度を推定するために、 \vec{x} を含む領域の列 $\mathcal{R}_1, \mathcal{R}_2, \dots$ を形成する。
- 最初の領域は1つのサンプルで、2番目は2つのサンプルで、といった具合に使用する。
- V_n を \mathcal{R}_n の体積、 k_n を \mathcal{R}_n に落ちるサンプルの数、 $p_n(\vec{x})$ を $p(\vec{x})$ に対する n 番目の推定値と仮定する：

$$p_n(\vec{x}) = \frac{k_n/n}{V_n}. \quad (5)$$

収束の条件

$p_n(\vec{x})$ が $p(\vec{x})$ に収束するためには、以下の3つの条件が必要であると考えられる。

1. $\lim_{n \rightarrow \infty} V_n = 0$ (領域が収縮すること)
2. $\lim_{n \rightarrow \infty} k_n = \infty$ (領域内のサンプル数が無限に増えること)
3. $\lim_{n \rightarrow \infty} k_n/n = 0$ (全サンプルに対する割合は無視できるほど小さいこと)

収束の条件 詳細

1. 領域が収縮かつ p が \vec{x} において連続であれば、空間平均 P/V が $p(\vec{x})$ に収束
2. 頻度比率が確率 P に（確率的に）収束
3. 式 (5) で与えられる $p_n(\vec{x})$ がそもそも収束するために必要（たとえ膨大な数のサンプルが最終的に小さな領域 \mathcal{R}_n 内に落ちたとしても、それらは全サンプル数に対して無視できるほど小さな割合にしかならないことを意味する）

代表的なアプローチ 1

これまでの条件を満たす領域の列を得るための一般的な方法

- Parzen-window method(パルゼン窓法、カーネル密度推定): 体積 V_n を n のある関数 (例えば $V_n = 1/\sqrt{n}$ など) として指定することにより、初期領域を縮小させていく方法。この場合、確率変数 k_n および k_n/n が適切に振る舞うこと、より正確には $p_n(x)$ が $p(x)$ に収束することを示す必要がある。

代表的なアプローチ 2

これまでの条件を満たす領域の列を得るための一般的な方法

- k_n -nearest-neighbor estimation(k 近傍法) : k_n を n のある関数 (例えば $k_n = \sqrt{n}$ など) として指定する方法。
ここでは、体積 V_n は \vec{x} の近傍の k_n 個のサンプルを囲むまで拡大される。

これらの手法は両方とも実際に収束するが、有限サンプルでの挙動について意味のある言及をすることは困難。

Conclusion

- ノンパラメトリック手法は、密度関数の形式を仮定せずに、観測データから直接、密度関数や事後確率を推定するもの。
- 密度推定は領域内のサンプル数と体積の比に基づく近似であり、代表的手法に領域を制御する「カーネル密度推定」とサンプル数を制御する「k-NN」がある。