Duda & Hart
3.2 MAXIMUM LIKELIHOOD ESTIMATION

3.2.1 The General Principle

作成者: おおつかたく

December 11, 2025

目次

Purpose and Background

目的と背景

- パターン認識では事前確率や条件付き密度を正確に知ることは難しいため、得られたサンプルから確率やパラメータを推定する必要がある。推定法として最尤推定法とベイズ推定がある。
- 今回は教師あり学習における「最尤推定法」の基本原理を解説する。

前提の話

前提条件

- サンプル集合をクラスごとに分類したとし、c 個のサンプル集合 $\mathcal{X}_1, \ldots, \mathcal{X}_c$ があるとする。ここで、$\mathcal{X}_i$ 内のサンプルは、確率法則 $p(\vec{x}|\omega_i)$ に従って独立に抽出されたものである。
- 我々は、$p(\vec{x}|\omega_i)$ が既知の parametric form（パラメータで記述できる形）を持っていると仮定する。よって、$p(\vec{x}|\omega_i)$ はパラメータベクトル θ_i の値によって一意に決定される。

前提条件（続き）

- 例えば、$p(\vec{x}|\omega_i) \sim N(\mu_i, \Sigma_i)$（正規分布）であるかもしれず、この場合、$\theta_i$ の成分には μ_i と Σ_i 両方の成分が含まれることになる。
- そこで、$p(\vec{x}|\omega_i)$ の θ_i への依存性を明示するために、$p(\vec{x}|\omega_i)$ を $p(\vec{x}|\omega_i, \theta_i)$ と書くことにする。
  - $p(\vec{x}|\omega_i; \theta_i)$ と表記することがある。
  - 厳密に言えば、$p(\vec{x}|\omega_i, \theta_i)$ という表記は θ_i が確率変数であることを暗示する（条件付き確率に見える）。
  - ここでは最尤推定法による解析において、θ_i を通常のパラメータとして扱う。

問題提起・単純化

問題提起・単純化

- 我々の問題は、サンプルから提供される情報を用いて、未知のパラメータベクトル $\theta_1, \ldots, \theta_c$ の良い推定値を得ることである。
- この問題の扱いを単純化するために、「$\mathcal{X}_i$ 内のサンプルは、$i \neq j$ である場合の θ_j については何の情報も与えない」と仮定
- すなわち、「異なるクラスのパラメータは関数的に独立している」と仮定する。（すべてのクラスのサンプルが同じ共分散行列を共有している場合などこれが当てはまらない場合もある。）

問題提起・単純化（続き）

- これにより、各クラスを個別に扱うことができ、クラスの区別を示す記号（添字）を削除して表記を簡略化することができる。
- したがって、この仮定の下では、次のような形式の c 個の独立した問題が存在することになる
  - 確率法則 $p(\vec{x}|\theta)$ に従って独立に抽出されたサンプル集合 $\mathcal{X}$ を用いて、未知のパラメータベクトル θ を推定せよ。

尤度

尤度

- $\mathcal{X}$ が n 個のサンプルを含んでいるとし、$\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\}$ とする。このとき、サンプルは独立に抽出されたので、以下が成り立つ。

$$p(\mathcal{X}|\theta) = \prod_{k=1}^{n} p(\vec{x}_k|\theta) \tag{1}$$

- θ の関数として見たとき、$p(\mathcal{X}|\theta)$ はサンプル集合に対する θ の尤度（likelihood）と呼ばれる。

尤度（続き）

- θ の最尤推定値（maximum likelihood estimate）とは、定義により、この $p(\mathcal{X}|\theta)$ を最大化する値 $\hat{\theta}$ のことである（10 ページ図参照）。
- 直感的には、これは実際に観測されたサンプルと、ある意味で最もよく合致する θ の値に対応している。

尤度（続き）

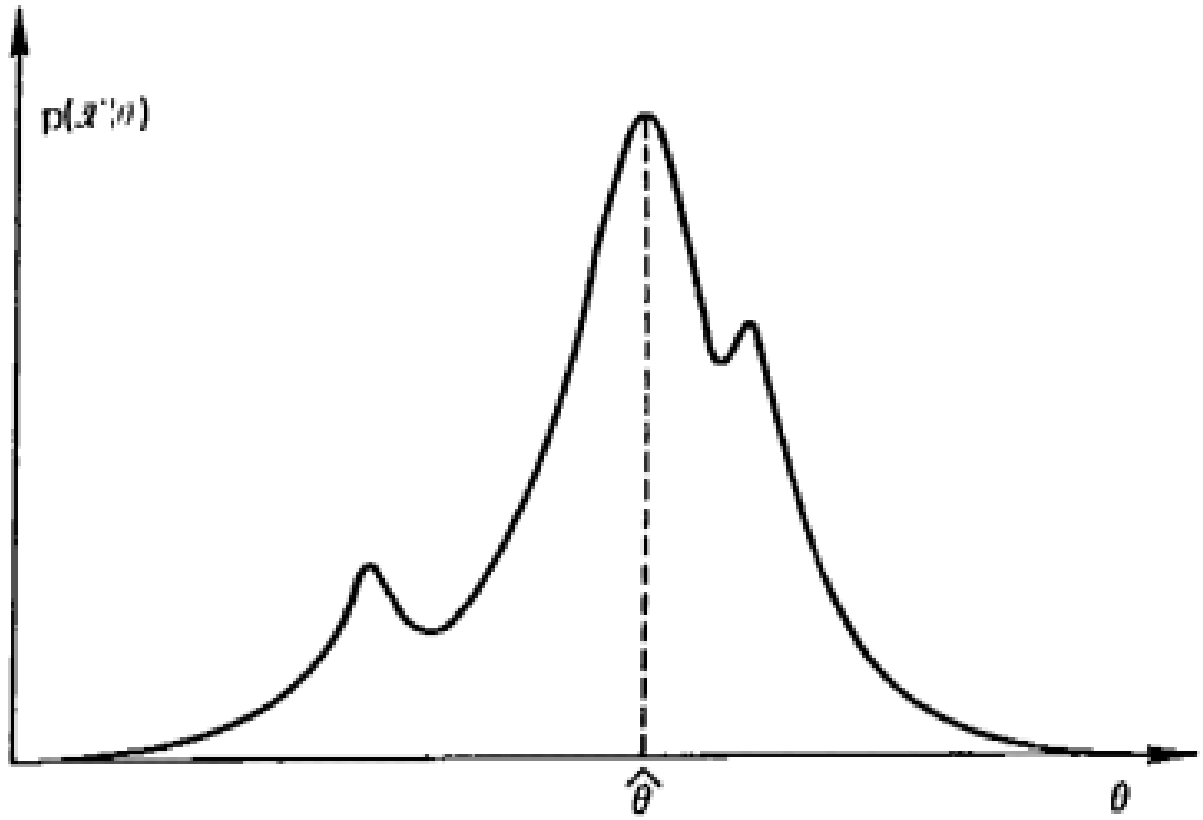

FIGURE 3.1. The maximum likelihood estimate for a parameter θ.

対数尤度（logarithm of the likelihood）

対数尤度

- 解析的な目的のためには、通常、尤度そのものよりも対数尤度（logarithm of the likelihood）を扱う方が簡単である。
- 対数は単調増加関数であるため、対数尤度を最大化する $\hat{\theta}$ は、尤度そのものも最大化する。
- もし $p(\mathcal{X}|\theta)$ が振る舞いの良い（滑らかな）、θ の微分可能な関数であれば、$\hat{\theta}$ は微分積分の標準的な手法によって求めることができる。

対数尤度（続き）

θ を p 成分ベクトル $\theta = (\theta_1, \ldots, \theta_p)^t$ とし、∇_θ を勾配演算子とすると,

$$\nabla_\theta \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}. \tag{2}$$

$l(\theta)$ を対数尤度関数とすると、

$$l(\theta) \equiv \ln p(\mathcal{X}|\theta). \tag{3}$$

対数尤度（続き）

すると、以下が成り立つ

$$l(\theta) = \sum_{k=1}^{n} \ln p(\vec{x}_k|\theta). \tag{4}$$

$$\frac{\partial l}{\partial \vec{\theta}} = \sum_{k=1}^{n} \frac{\partial}{\partial \vec{\theta}} \ln p(\vec{x}_k|\theta). \quad (i = 1, \ldots, p) \tag{5}$$

対数尤度（続き）

- したがって、θ の最尤推定値に対する必要条件のセットは、p 個の方程式 $\nabla_\theta l = 0$ のセットから得ることができる。
- 勾配演算子を用いず表すと、

$$\begin{cases} \dfrac{\partial l}{\partial \theta_1} = 0 \\ \dfrac{\partial l}{\partial \theta_2} = 0 \\ \vdots \\ \dfrac{\partial l}{\partial \theta_p} = 0 \end{cases} \tag{6}$$

Conclusion

まとめ

- 観測データが得られる確率（尤度）を最大化するパラメータこそが最適であると定義し、対数尤度の勾配を用いてその値を導出する方法の基本原理を説明した。

functionally independent

- 異なるクラスのパラメータは関数的に独立している（functionally independent）とは、一言で言えば、あるクラスのパラメータを推定する際に、他のクラスのデータやパラメータを一切気にする必要がないということ。

functionally independent（続き）

- 例えば、「男性クラス (ω_1)」と「女性クラス (ω_2)」の身長の分布（正規分布）を作るとする。
- 男性の平均身長 (μ_1) を計算するのに、女性のデータは一切関係ない。女性の分散 (σ_2^2) を計算するのに、男性のデータは必要ない。
- よって「男性だけのデータで男性のパラメータを決める」「女性だけのデータで女性のパラメータを決める」という、完全に別々の作業として処理できる。

functionally independent（続き）

- つまり、全体の尤度関数（最大化したい式）が、$L(\theta_1, \theta_2, \dots) = L_1(\theta_1) \times L_2(\theta_2) \times \dots$ のように、個別の項の掛け算にきれいに分解できるということ。

ふるまいの良い関数（well-behaved function）

- 「$p(\mathcal{X}|\theta)$ が振る舞いの良い（滑らかな）、θ の微分可能な関数」という表現は、数学的に厳密な定義というよりは、「微積分を使って最大値を求めるテクニック（勾配法など）が使えるような、都合の良い形をしている」ということを意味している。
- 「これから紹介する $\nabla_\theta l = 0$ （偏微分してゼロと置く）という便利な解法を使うために、グラフがギザギザしたり途切れたりしていない、滑らかな形であると仮定しますよ」という、数学的な「お断り」の意。

尤度

- 確率（Probability）「ルール（モデル）が決まっているときに、データが出る可能性」例：サイコロを振って「3」が出る確率は 1/6
- 尤度（Likelihood）「データが出たときに、そのルール（モデル）がどれくらい信頼できるか」例：サイコロを 10 回振ったら「3」が 3 回出たとする。そのサイコロが「公平（全部 1/6）」な場合、どれくらいありそう？もし「3 が出やすいサイコロ」だとしたら、その仮説の方がもっと自然？

確率と尤度の違い

- 確率は「ルールが決まっていて、そのルールのもとで何が起きやすいか」
- 尤度は「実際に起きたことを見て、そのルールはどれくらい信頼できるか」