Đồ án – Phân lớp

Multivariate Statistical Applied

10 tháng 2, 2025

Tóm tắt nội dung

This is an abstract!

Mục lục

1 Giới thiệu

Trong các bài toán phân tích và tái tổ chức dữ liệu, hay các bài toán phân loại (classification issues), liên quan đến việc nhóm và phân loại các đối tượng (hoặc dữ liệu) dựa trên các đặc trưng (features) của chúng, có thể tiếp cận theo hai góc nhìn.

Thứ nhất, bài toán phân cụm (Clustering) là quá trình phân chia các đối tượng thành các nhóm tự nhiên mà không biết trước nhóm hay nhãn. Việc phân chia này được thực hiện dựa trên việc phân tích cấu trúc của tập dữ liệu và nhóm các đối tượng có sự tương đồng cao vào cùng một nhóm.

Khía cạnh thứ hai là bài toán phân loại (Classification), trong đó có hai quá trình chính: nghiên cứu và xây dựng một hàm phân loại để phân biệt các đối tượng, và gán nhãn cho các nhóm, với nhãn đã biết trước.

Trong bài nghiên cứu này, chúng tôi sẽ tập trung vào bài toán phân lớp theo góc nhìn thứ hai, cụ thể là nghiên cứu phương pháp phân tích phân biệt (Discriminant Analysis) và quá trình phân loại các đối tượng thành các nhóm với nhãn đã biết trước.

1.1 Phân tích phân biệt (Discriminant Rules) và Phân tích phân loại (Classification Rules)

1.2 Phát biểu bài toán

Giả sử có J quần thể (populations), ký hiệu $j = 1, 2, \dots, J$, cần gán một quan sát x vào một trong các nhóm này.

- **Mục tiêu**: Xây dựng một hàm phân loại
  $f : \mathbb{R}^d \to \{1, 2, \dots, J\}$ để dự đoán nhãn y của một quan sát mới $x^* \in \mathbb{R}^d$, gồm hai bước chính:

  1. **Phân biệt các lớp (Discrimination):** Xác định đặc trưng giúp phân biệt các nhóm, từ đó thiết lập quy tắc phân biệt (discriminant rules).

  2. **Phân loại dữ liệu mới (Classification):** Gán nhãn y cho quan sát mới dựa trên các quy tắc đã học.

- **Mục đích**: Tách biệt các nhóm dữ liệu và phân bổ chính xác các quan sát mới vào các nhóm đã biết, đồng thời tối ưu hóa mô hình để giảm sai số phân loại hoặc cực đại hóa xác suất hậu nghiệm $P(y|x)$.

- **Đầu vào**:

  – Một tập dữ liệu huấn luyện gồm n quan sát:

$$\{(x_i, y_i)\}_{i=1}^{n}$$

trong đó:

* $x_i \in \mathbb{R}^d$ là một điểm dữ liệu có d đặc trưng (features).
* $y_i \in \{1, 2, ..., J\}$ là nhãn (label) của x_i, thuộc một trong J nhóm.

– Một quan sát mới $x^* \in \mathbb{R}^d$ cần được phân loại.

- **Đầu ra**:
  - Một mô hình phân loại $f : \mathbb{R}^d \to \{1, 2, ..., J\}$
  - Nhãn dự đoán $y^* = f(x^*)$ cho quan sát mới x^*.
- **Lưu ý**: Mặc dù lý thuyết phân biệt và phân loại có sự khác nhau, trong thực tế, chúng thường kết hợp và hỗ trợ lẫn nhau: mô hình giúp phân biệt nhóm dữ liệu cũng có thể được dùng để phân loại, và ngược lại, một mô hình phân loại tốt thường phản ánh rõ các yếu tố phân biệt giữa các nhóm.

2 Allocation Rules for Known Distributions

3 Discrimination Rules in Practice

4 Bibliography