Đồ án – Phân lớp

Multivariate Statistical Applied

1 tháng 3, 2025

Tóm tắt nội dung

Phân loại (Classification) là một trong những bài toán quan trọng trong học máy (Machine Learning) và thống kê, với mục tiêu phân nhóm các đối tượng dựa trên các đặc trưng (features) đã biết. Trong đó, Phân tích phân biệt tuyến tính (Linear Discriminant Analysis - LDA) là một phương pháp phổ biến, giúp tìm kiếm một không gian chiếu tối ưu để phân tách các nhóm dữ liệu. LDA không chỉ giúp cải thiện khả năng phân loại mà còn hỗ trợ giảm chiều dữ liệu, giúp tối ưu hóa hiệu suất tính toán và tránh hiện tượng quá khớp (overfitting).

Trong bài toán phân loại hai lớp, LDA có thể được hiểu như Fisher's Linear Discriminant (FLD), trong đó mục tiêu là tìm một vector chiếu tối ưu để tối đa hóa độ phân biệt giữa hai lớp bằng cách tối đa hóa phương sai giữa các lớp và tối thiểu hóa phương sai trong từng lớp. Tuy nhiên, LDA không chỉ giới hạn trong bài toán hai lớp mà còn có thể mở rộng để phân biệt nhiều lớp bằng cách tìm tập hợp các vector chiếu, giúp tối ưu hóa khả năng phân tách giữa nhiều nhóm dữ liệu trong không gian đặc trưng.

Trong nghiên cứu này, chúng tôi tiến hành thực nghiệm với một case study thực tế, áp dụng LDA trên các tập dữ liệu có nhãn để đánh giá hiệu suất của phương pháp này. Chúng tôi sử dụng các chỉ số đánh giá như độ chính xác (Accuracy), độ nhạy (Sensitivity), độ đặc hiệu (Specificity) và F1-score để đo lường khả năng phân loại của mô hình. Kết quả thực nghiệm cho thấy rằng LDA hoạt động tốt khi dữ liệu có phân phối chuẩn và ma trận hiệp phương sai đồng nhất giữa các lớp, nhưng có thể bị suy giảm hiệu suất khi các giả định này không được thỏa mãn.

Bên cạnh đánh giá lý thuyết và thực nghiệm, chúng tôi cũng phân tích các ứng dụng thực tế của LDA trong nhiều lĩnh vực khác nhau như nhận dạng khuôn mặt, chẩn đoán y khoa, tài chính và sinh học. Với khả năng giảm chiều dữ liệu hiệu quả trong khi vẫn bảo toàn thông tin quan trọng, LDA trở thành một công cụ mạnh mẽ trong việc trích xuất đặc trưng và phân loại dữ liệu có cấu trúc phức tạp.

Những phát hiện từ nghiên cứu này giúp cung cấp một cái nhìn toàn diện về vai trò của LDA trong bài toán phân loại, từ bài toán phân biệt hai lớp đơn giản đến bài toán phân biệt nhiều lớp, cùng với các ứng dụng thực tế của nó.

Từ khóa: Phân loại (Classification), Phân tích phân biệt tuyến tính (LDA), Fisher's Linear Discriminant (FLD), Học máy (Machine Learning), Giảm chiều dữ liệu (Dimensionality Reduction), Ứng dụng LDA.

Mục lục

1 Giới thiệu

1.1 Phân biệt (Discrimination) và Phân lớp (Classification)

1.2 Ý nghĩa về khoa học

1.3 Ý nghĩa về ứng dụng

1.4 Phát biểu bài toán

- **Giả sử:**
  - Có g nhóm khác nhau, ký hiệu $G = \{G_1, G_2, \ldots, G_g\}$. Mỗi nhóm có một phân phối xác suất riêng với trung bình μ_G và ma trận hiệp phương sai Σ_G.
  - Một quan sát x có d đặc trưng và thuộc về một trong các nhóm G.
- **Đầu vào:**
  - Tập dữ liệu gồm n mẫu, mỗi mẫu có d đặc trưng:

$$X = \{x_1, x_2, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$$

  - Nhãn tương ứng với các mẫu:

$$Y = \{y_1, y_2, \ldots, y_n\}, \quad y_i \in \{G_1, G_2, \ldots, G_g\}$$

- **Đầu ra:** Nhóm G^* tối ưu cho một quan sát mới x, xác định theo:

$$G^* = \arg\max_G P(G|x)$$

  hoặc theo hàm phân biệt (discriminant function):

$$G^* = \arg\max_G \delta_G(x)$$

- **Mục tiêu của bài toán:** là xây dựng một hàm phân loại f để gán một quan sát x có d đặc trưng vào một trong g nhóm (classes) sao cho tối ưu được độ chính xác phân loại.

  Cụ thể, cần tìm một ánh xạ:

$$f : \mathbb{R}^d \to \{G_1, G_2, \ldots, G_g\}$$

  sao cho với mỗi quan sát x, nhóm dự đoán $G^* = f(x)$ là nhóm có xác suất hậu nghiệm cao nhất hoặc hàm phân biệt tối đa.

1.4.1 Cấu trúc chung (Framework)

Hình 1: Mô hình cấu trúc chung của bài toán phân lớp

1.4.2 Đóng góp

Một câu hỏi quan trọng là làm thế nào để xác định hàm phân biệt ($\delta_G(x)$). Một trong những phương pháp phổ biến để giải quyết vấn đề này là Phân tích Phân biệt Tuyến tính (LDA). Tương tự như Phân tích Thành phần Chính (PCA), LDA chiếu toàn bộ tập dữ liệu lên một siêu phẳng có số chiều nhỏ hơn,

với mục tiêu phân tách dữ liệu một cách rõ ràng hơn. Hàm phân biệt $\delta_G(x)$ trong LDA có dạng tuyến tính, cho phép phân tách các lớp bằng cách tìm ra hướng chiếu tối ưu nhất.

Vì thế, nghiên cứu này đóng góp vào lĩnh vực phân loại dữ liệu bằng cách cung cấp một đánh giá toàn diện về Phân tích Phân biệt Tuyến tính (LDA) từ cả góc độ lý thuyết và thực nghiệm. Cụ thể, các đóng góp chính của nghiên cứu bao gồm:

1. **Phân tch chi tiết về LDA trong bài toán phân loại:**

   Nghiên cứu trình bày một cách hệ thống nguyên lý hoạt động của LDA, bao gồm cách tiếp cận Fisher's Linear Discriminant (FLD) trong bài toán phân loại hai lớp và cách mở rộng phương pháp này cho nhiều lớp dữ liệu. Chúng tôi thảo luận các giả định quan trọng của LDA, bao gồm tính phân phối chuẩn của dữ liệu và tính đồng nhất của ma trận hiệp phương sai giữa các lớp.

2. **Đánh giá thực nghiệm hiệu suất của LDA:**

   Chúng tôi thực hiện các thí nghiệm trên các tập dữ liệu có nhãn để đánh giá hiệu suất phân loại của LDA trong các điều kiện khác nhau. Các chỉ số đánh giá như ..... được sử dụng để định lượng hiệu quả của mô hình. Kết quả thực nghiệm cho thấy LDA có hiệu suất tốt khi dữ liệu tuân theo các giả định lý thuyết, nhưng có thể bị suy giảm khi vi phạm các giả định này.

3. **Phân tích ứng dụng thực tế của LDA:**

   Nghiên cứu và áp dụng ứng dụng thực tế của LDA trên tập dữ liệu tiêu chuẩn để đánh giá tính ứng dụng và hiệu suất của phương pháp này. Chúng tôi nhấn mạnh vai trò của LDA trong việc trích xuất đặc trưng và giảm chiều dữ liệu, đặc biệt trong các bài toán có số lượng đặc trưng lớn và yêu cầu tối ưu hóa tính toán.

2 Các công trình nghiên cứu liên quan

2.1 Phân tích phân biệt tuyến tính (Linear Discriminant Analysis)

2.1.1 Phát biểu bài toán LDA tổng quát

Cho tập dữ liệu gồm N điểm $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ trong không gian d-chiều. Các điểm dữ liệu được chia thành K lớp, với mỗi lớp k chứa N_k điểm. Ký hiệu tập chỉ số của lớp k là:

$$\mathbf{C}_k = \{n \mid n \in \text{lớp } k\}, \quad k = 1, 2, ..., K.$$

Đầu ra

Ma trận chiếu $W \in \mathbb{R}^{d \times (K-1)}$, trong đó mỗi cột là một vector chiếu $\mathbf{w}_i$. Dữ liệu sau khi chiếu xuống không gian mới:

$$\mathbf{y}_n = W^T \mathbf{x}_n, \quad 1 \leq n \leq N.$$

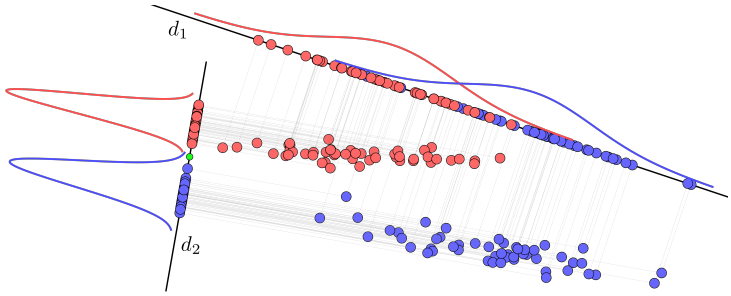
Với $\mathbf{y}_n$ thuộc không gian $(K-1)$-chiều sao cho các điểm dữ liệu thuộc các lớp khác nhau được phân tách tốt nhất trong không gian mới này.

2.1.2 Ý tưởng giải quyết bài toán

Để tiếp cận bài toán phân lớp nhiều lớp, trước hết xem xét bài toán đơn giản nhất: phân loại hai lớp.

Bài toán LDA hai lớp

Xét bài toán phân loại hai lớp (lớp C_1 - màu đỏ và lớp C_2 - màu xanh).



Hình 2: Trực quan hóa dữ liệu

Giả sử:

- Dữ liệu của hai lớp có phân phối chuẩn.
- Giá trị kỳ vọng của mỗi lớp lần lượt là $\mathbf{m}_1$ và $\mathbf{m}_2$.
- Độ phân tán của dữ liệu trong từng lớp được đo bằng phương sai s_1^2 và s_2^2.

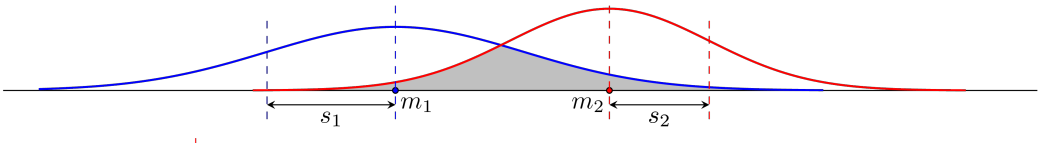
Bài toán đưa về thành: Tìm một hướng chiếu $\mathbf{w}$ sao cho dữ liệu của hai lớp tách biệt rõ ràng nhất.

Trước khi tìm lời giải, ta xét các trường hợp về cách dữ liệu bị phân tán khi chiếu lên một trục bất kỳ.

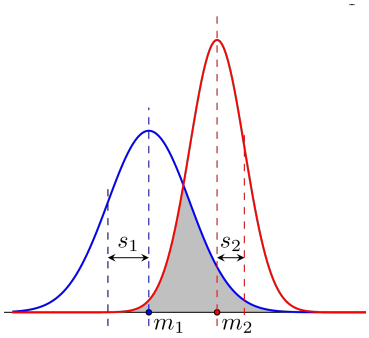
Trường hợp 1: Hai lớp dữ liệu có phương sai lớn

Khi dữ liệu trong mỗi lớp có phương sai lớn, các điểm dữ liệu bị trải rộng và có phần chồng lấn lớn.

Trường hợp 2: Phương sai nhỏ nhưng khoảng cách giữa trung tâm hai lớp quá gần



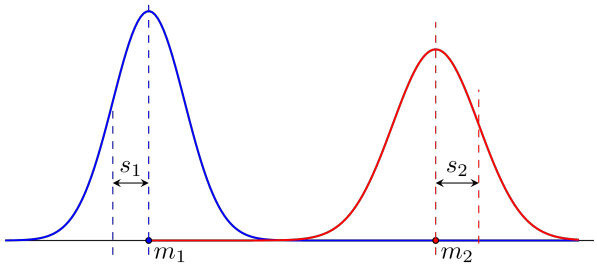
Hình 3: Trường hợp 1: Phương sai lớn làm dữ liệu chồng lấn nhiều



Hình 4: Trường hợp 2: Trung tâm hai lớp quá gần nhau

Khi phương sai nhỏ, dữ liệu trong từng lớp ít phân tán, nhưng nếu trung tâm hai lớp $\mathbf{m}_1$ và $\mathbf{m}_2$ gần nhau, phần chồng lấn giữa hai lớp vẫn lớn.

Trường hợp 3: Phương sai nhỏ và khoảng cách giữa hai trung tâm đủ lớn



Hình 5: Trường hợp 3: Phân bố tốt nhất để phân lớp

Khi dữ liệu trong mỗi lớp tập trung (phương sai nhỏ) và hai trung tâm cách xa nhau, phần chồng lấn giữa hai lớp được giảm đáng kể.

Ý tưởng bài toán LDA

Từ ba trường hợp trên, để có một hướng chiếu tối ưu cho phân loại, ta cần:

- Giảm phương sai trong mỗi lớp, giúp dữ liệu của từng lớp ít phân tán hơn.
- Tăng khoảng cách giữa trung tâm hai lớp, giúp hai lớp cách xa nhau hơn, giảm phần chồng lấn.

Tổng quát hóa lên bài toán nhiều lớp

Khi mở rộng lên bài toán với K lớp, ý tưởng vẫn giữ nguyên: tìm một không gian chiếu sao cho tỉ số giữa phương sai giữa các lớp và phương sai trong từng lớp là lớn nhất. Thay vì chỉ dùng một vector chiếu $\mathbf{w}$ như bài toán hai lớp, LDA tổng quát sử dụng ma trận chiếu W với $K-1$ hướng chiếu tối ưu. Toán học đằng sau việc tìm các hướng chiếu này dựa trên việc tối ưu hóa tỉ số giữa ma trận tán sắc giữa các lớp S_B và ma trận tán sắc trong lớp S_W. Việc tìm nghiệm của bài toán tương đương với việc tính các vector riêng của ma trận $S_W^{-1}S_B$.

2.2 Phân biệt cực đại hợp lý (Maximum Likelihood Discriminant)

2.2.1 Giới thiệu

Phân biệt cực đại hợp lý (Maximum Likelihood Discrimination - MLD) là một phương pháp phân loại dựa trên nguyên lý xác suất. Ý tưởng chính là ước lượng xác suất để một điểm dữ liệu thuộc về một lớp cụ thể, sau đó gán nhãn cho điểm đó theo lớp có xác suất lớn nhất.

2.2.2 Cơ sở toán học

Phương pháp này dựa trên các khái niệm xác suất và thống kê, đặc biệt là xác suất tiên nghiệm, xác suất có điều kiện và ước lượng tham số theo nguyên lý hợp lý cực đại (MLE). Để hiểu rõ hơn, trước tiên ta cần xây dựng nền tảng toán học quan trọng.

Xác suất tiên nghiệm và xác suất có điều kiện
Mỗi lớp dữ liệu C_k có một xác suất tiên nghiệm $P(C_k)$, phản ánh tỷ lệ xuất hiện của lớp đó trong tổng thể dữ liệu. Để xác định lớp của một điểm dữ liệu, ta cần xét xác suất có điều kiện $P(\mathbf{x}|C_k)$. Theo định lý Bayes:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})} \tag{2.2.1}$$

với $P(\mathbf{x})$ là xác suất quan sát dữ liệu:

$$P(\mathbf{x}) = \sum_j P(\mathbf{x}|C_j)P(C_j) \tag{2.2.2}$$

Để quyết định một điểm thuộc về lớp nào, ta chỉ cần so sánh giá trị xác suất hậu nghiệm $P(C_k|\mathbf{x})$ cho từng lớp và chọn lớp có xác suất lớn nhất. LDA dựa trên giả định rằng xác suất có điều kiện $P(\mathbf{x}|C_k)$ tuân theo phân phối chuẩn, giúp đơn giản hóa mô hình và xác định được tiêu chí phân tách tối ưu.

Hàm mật độ xác suất
Với giả định là dữ liệu trong mỗi lớp tuân theo phân phối chuẩn (Gaussian):

$$P(\mathbf{x}|C_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right) \tag{2.2.3}$$

trong đó μ_k là vector trung bình và Σ_k là ma trận hiệp phương sai.

Giả định phân phối chuẩn giúp đơn giản hóa bài toán phân loại vì nó cho phép ta mô hình hóa dữ liệu theo một mô hình toán học rõ ràng. Nếu các lớp dữ liệu thực sự tuân theo phân phối chuẩn với cùng một ma trận hiệp phương sai, thì ranh giới phân loại sẽ là một đường thẳng hoặc mặt phẳng (trong không gian nhiều chiều). Điều này giải thích vì sao LDA là một bộ phân loại tuyến tính.

Ước lượng hợp lý cực đại (Maximum Likelihood Estimation - MLE)
Trong bài toán phân loại, ta cần ước lượng các tham số của phân phối xác suất $P(\mathbf{x}|C_k)$ để có thể tính được xác suất hậu nghiệm $P(C_k|\mathbf{x})$. Một trong những phương pháp phổ biến nhất để ước lượng tham số là phương pháp hợp lý cực đại (MLE), nhằm tìm các giá trị tham số tối ưu sao cho xác suất quan sát dữ liệu là lớn nhất.

Giả sử dữ liệu trong mỗi lớp C_k tuân theo phân phối chuẩn $\mathcal{N}(\mu_k, \Sigma_k)$, ta thực hiện **Ước lượng tham số của phân phối chuẩn**

Cho tập dữ liệu $X = \{x_1, x_2, ..., x_n\}$ được lấy mẫu từ một phân phối chuẩn một chiều $\mathcal{N}(\mu, \sigma^2)$, ta cần tìm giá trị ước lượng của các tham số μ (trung bình) và σ (độ lệch chuẩn).

Hàm mật độ xác suất (PDF) của một điểm dữ liệu là:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.2.4}$$

Khi có nhiều quan sát độc lập $X = \{x_1, x_2, \ldots, x_n\}$, hàm hợp lý là tích các xác suất của từng điểm dữ liệu:

$$L(\mu,\sigma|X) = \prod_{i=1}^{n} f(x_i|\mu,\sigma) \tag{2.2.5}$$

Thay biểu thức $f(x|\mu,\sigma)$ vào, ta có:

$$L(\mu,\sigma|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{2.2.6}$$

Vì việc làm việc với tích nhiều số mũ có thể phức tạp, ta thường lấy log của hàm hợp lý (log-likelihood):

$$\log L(\mu,\sigma|X) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{2.2.7}$$

Tìm ước lượng MLE cho μ

Lấy đạo hàm của log-likelihood theo μ:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Đặt bằng 0 để tìm giá trị cực đại:

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

Suy ra:

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2.8}$$

Như vậy, giá trị ước lượng hợp lý cực đại của μ chính là trung bình mẫu.

Tìm ước lượng MLE cho σ

Lấy đạo hàm của log-likelihood theo σ:

$$\frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Đặt bằng 0:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Nhân cả hai vế với σ^3, ta thu được:

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Suy ra:

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2.2.9}$$

Tức là ước lượng hợp lý cực đại của phương sai là trung bình của bình phương độ lệch so với trung bình mẫu. Điều này khác với ước lượng không chệch của phương sai (dùng mẫu) là $\frac{1}{n-1}\sum(x_i - \mu)^2$.

Trong LDA, ta cần ước lượng các tham số μ_k và Σ_k của từng lớp C_k. Công thức trên cho thấy rằng μ_k có thể ước lượng bằng trung bình của các điểm thuộc lớp đó, và Σ_k có thể tính bằng trung bình ma trận hiệp phương sai của từng điểm với trung tâm lớp. Nếu giả định rằng tất cả các lớp có cùng ma trận hiệp phương sai $\Sigma_k = \Sigma$, ta có thể ước lượng nó bằng cách lấy trung bình của các Σ_k. Khi các tham số được ước lượng bằng MLE, ta có thể áp dụng định lý Bayes để tính toán xác suất hậu nghiệm và xác định miền quyết định cho từng điểm dữ liệu.

2.2.3 Quy tắc Phân biệt cực đại hợp lý

Trong bài toán phân loại, mục tiêu của phân biệt cực đại hợp lý (Maximum Likelihood Discriminant Rule - MLD) là gán nhãn cho một điểm dữ liệu $\mathbf{x}$ vào nhóm C_k sao cho xác suất có điều kiện $P(\mathbf{x}|C_k)$ đạt giá trị lớn nhất. Điều này đồng nghĩa với việc chọn nhóm có xác suất hậu nghiệm lớn nhất theo định lý Bayes:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})}.$$

Do $P(\mathbf{x})$ là hằng số không phụ thuộc vào nhóm C_k, nên quy tắc phân loại có thể được đơn giản hóa thành:

$$C^* = \arg\max_{C_k} P(\mathbf{x}|C_k)P(C_k). \tag{2.2.10}$$

Xác định miền quyết định

Một điểm dữ liệu $\mathbf{x}$ sẽ được gán vào nhóm C_k nếu:

$$P(\mathbf{x}|C_k)P(C_k) > P(\mathbf{x}|C_j)P(C_j), \quad \forall j \neq k. \tag{2.2.11}$$

Nếu giả định mỗi nhóm có phân phối chuẩn với cùng ma trận hiệp phương sai Σ, thì tiêu chí phân loại có thể được viết dưới dạng hàm quyết định tuyến tính:

$$d_k(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log P(C_k). \tag{2.2.12}$$

Điểm $\mathbf{x}$ sẽ được gán vào nhóm C_k nếu $d_k(\mathbf{x}) > d_j(\mathbf{x})$ với mọi $j \neq k$.

Xác suất phân loại sai

Với hai nhóm $J = 2$, xác suất phân loại sai có thể được tính như sau:

$$p_{21} = P(X \in R_2|1) = \int_{R_2} f_1(x)dx.$$

$$p_{12} = P(X \in R_1|2) = \int_{R_1} f_2(x)dx.$$

Tổng xác suất phân loại sai được tính bằng:

$$P_{error} = 1 - \sum_k P(C_k) \int_{R_k} P(\mathbf{x}|C_k)d\mathbf{x}. \tag{2.2.13}$$

Xác suất phân loại sai phụ thuộc vào độ chồng lấn giữa các phân phối xác suất có điều kiện $P(\mathbf{x}|C_k)$. Nếu các nhóm có phân phối gần nhau, xác suất phân loại sai sẽ cao hơn.

Mở rộng Phân biệt cực đại hợp lý với Chi phí phân loại sai

Trong nhiều ứng dụng, các lỗi phân loại có thể có chi phí khác nhau. Khi đó, ta sử dụng ma trận chi phí $C(i|j)$ để biểu diễn chi phí khi phân loại mẫu thuộc nhóm C_j thành nhóm C_i. Chi phí kỳ vọng được tính bằng:

$$R(C_i|\mathbf{x}) = \sum_j C(i|j)P(C_j|\mathbf{x}). \tag{2.2.14}$$

Quy tắc phân loại tối ưu để giảm chi phí kỳ vọng là chọn nhóm C_k sao cho:

$$C^* = \arg\min_{C_k} R(C_k|\mathbf{x}). \tag{2.2.15}$$

2.2.4 Ưu điểm & Nhược điểm của Maximum Likelihood Discrimination (MLD)

Ưu điểm:

- **Tính chính xác cao với cỡ mẫu lớn**: Khi số lượng dữ liệu huấn luyện n tăng, MLD đạt được tính không chệch suy rộng và tiệm cận quyết định tối ưu theo tiêu chí Bayes.
- **Cơ sở toán học vững chắc**: MLD tối đa hóa xác suất hậu nghiệm để ra quyết định phân loại, giúp đảm bảo tính tối ưu thống kê.
- **Áp dụng được cho nhiều loại mô hình xác suất**: Có thể sử dụng với nhiều loại phân phối khác nhau như Gaussian, Bernoulli, Poisson,...
- **Hàm quyết định đơn giản với giả định phù hợp**: Nếu dữ liệu tuân theo phân phối chuẩn với cùng ma trận hiệp phương sai Σ, giúp giảm chi phí tính toán trong phân loại.

Nhược điểm:

- **Phụ thuộc vào giả định phân phối dữ liệu**: Nếu dữ liệu không tuân theo phân phối giả định, hiệu suất của MLD bị suy giảm nghiêm trọng.
- **Độ phức tạp tính toán cao với dữ liệu nhiều chiều**: Khi số chiều của dữ liệu lớn, cần ước lượng ma trận hiệp phương sai Σ và nghịch đảo của nó, gây tốn kém về mặt tính toán.
- **Hiệu suất kém với cỡ mẫu nhỏ**: Khi số lượng mẫu nhỏ, ước lượng của MLD có thể bị chệch. Có thể sử dụng các phương pháp như ước lượng Bayes hoặc ước lượng hợp lý cực đại có điều chỉnh (Regularized MLE) để khắc phục.
- **Không phù hợp với phân phối không chính quy**: Nếu dữ liệu có phân phối bất thường hoặc phụ thuộc phi tuyến phức tạp, MLD hoạt động kém hiệu quả so với các phương pháp phi tham số khác.

2.3 Quy tắc phân biệt Bayes (Bayes Discriminant Rule)

2.4 Quy tắc phân biệt tuyến tính của Fisher (Fisher's Linear Discriminant Rule)

2.5 So sánh

Phương pháp	Nguyên lý	Ưu điểm	Hạn chế
Phân biệt cực đại hợp lý (Maximum Likelihood Discriminant)	Ước lượng tham số bằng phương pháp cực đại hợp lý (MLE).	Không cần thông tin tiên nghiệm, dựa hoàn toàn vào dữ liệu quan sát.	Cần dữ liệu lớn để ước lượng chính xác.
Quy tắc phân biệt Bayes (Bayes Discriminant Rule)	Áp dụng định lý Bayes để tính xác suất hậu nghiệm.	Tận dụng được thông tin tiên nghiệm, có thể cải thiện độ chính xác.	Cần biết xác suất tiên nghiệm, nếu ước lượng sai sẽ ảnh hưởng đến kết quả.
Quy tắc phân biệt tuyến tính của Fisher (Fisher's Linear Discriminant)	Tìm mặt phẳng phân tách tuyến tính tối ưu giữa các nhóm.	Không yêu cầu giả định về phân phối, tính toán đơn giản.	Không hoạt động tốt nếu dữ liệu phi tuyến.

- **Nhận xét:**
  - Nếu dữ liệu có phân phối chuẩn và đủ lớn, phương pháp phân biệt cực đại hợp lý sẽ được sử dụng để ước lượng tham số tối ưu.
  - Nếu cần tận dụng thông tin tiên nghiệm, quy tắc phân biệt Bayes sẽ được sử dụng để tối ưu hóa phân loại theo xác suất hậu nghiệm.
  - Nếu không muốn giả định về phân phối dữ liệu, quy tắc phân biệt tuyến tính của Fisher sẽ được sử dụng để tìm mặt phẳng phân tách tuyến tính tốt nhất.
- **Kết luận:** Quy tắc phân biệt tuyến tính của Fisher là một bước tiến giúp giảm yêu cầu về giả định phân phối dữ liệu, trong khi quy tắc phân biệt Bayes cải thiện phương pháp phân biệt cực đại hợp lý bằng cách sử dụng xác suất tiên nghiệm.

3 Phương pháp: Linear Discriminant Analysis (LDA)

3.1 LDA với hai lớp (Fisher's Discriminant Rule)

3.1.1 Bài toán LDA hai lớp

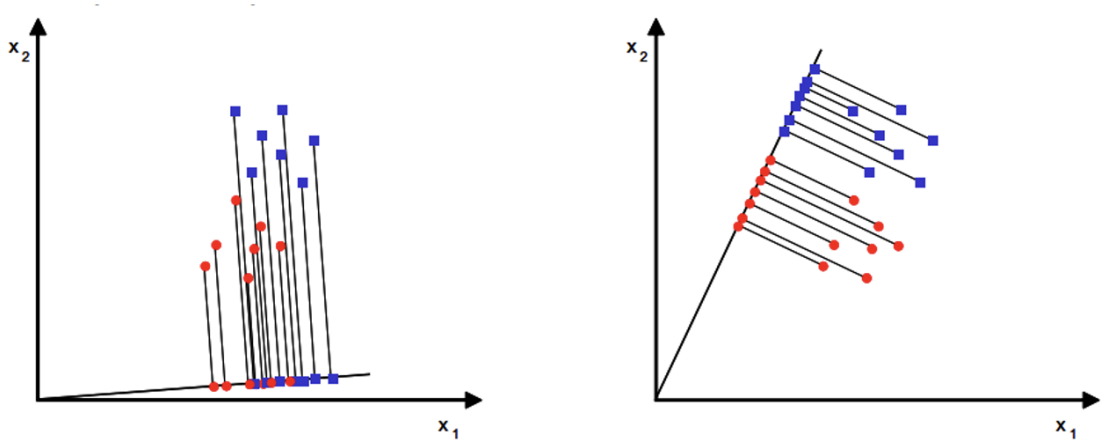
Giả sử có N điểm dữ liệu $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ trong không gian d-chiều, với N_1 điểm thuộc lớp thứ nhất và $N_2 = N - N_1$ điểm thuộc lớp thứ hai.

Tập chỉ số các điểm thuộc từng lớp:

- $\mathbf{C}_1 = \{n \mid 1 \leq n \leq N_1\}$: là tập chỉ số của các điểm thuộc lớp thứ nhất.
- $\mathbf{C}_2 = \{m \mid N_1 + 1 \leq m \leq N\}$: là tập chỉ số của các điểm thuộc lớp thứ hai.

Mục tiêu là tìm một vector hệ số $\mathbf{w}$ để chiếu dữ liệu lên một trục sao cho khả năng phân biệt giữa các lớp là tốt nhất. Giá trị sau khi chiếu của mỗi điểm là:

$$y_n = \mathbf{w}^T \mathbf{x}_n, \quad 1 \leq n \leq N$$



Hình 6: Chiếu tập dữ liệu lên các đường thẳng khác nhau thì khả năng phân biệt giữa các lớp cũng khác nhau.

Ta thấy phép chiếu bên trái đã có sự tối đa hóa về việc phân chia dữ liệu độc lập nhau. Để có thể tìm được một vector chiếu tốt nhất, ta cần xác định một quy tắc (rule) để phân chia dữ liệu.

Như đã phân tích trong Mục 2, hai lớp được xem là phân biệt tốt (discriminative) nếu khoảng cách giữa chúng lớn, tức là phương sai giữa lớp (between-class variance) lớn, và dữ liệu trong từng lớp có mức độ đồng nhất cao, tức là phương sai trong lớp (within-class variance) nhỏ.

Từ đó, phương pháp Linear Discriminant Analysis (LDA) có thể được hiểu là một thuật toán nhằm tìm một phép chiếu tối ưu sao cho tỷ lệ giữa phương sai giữa lớp và phương sai trong lớp đạt giá trị lớn nhất. Mục tiêu của phương pháp này là tối ưu hóa khả năng phân biệt giữa các lớp trong không gian đặc trưng.

Để cực đại hóa tỷ lệ giữa phương sai giữa lớp và phương sai trong lớp, cần xác định một hàm mục tiêu phù hợp. Hàm mục tiêu này được xây dựng dựa trên việc tìm cực trị của tỷ số giữa phương sai giữa lớp và phương sai trong lớp nhằm mô hình hóa và tối ưu hóa quá trình phân tách giữa các lớp.

3.1.2 Xây dựng hàm mục tiêu

1. **Phương sai giữa lớp (Between-Class Variance)**

   Gọi μ_i là vector trung bình của các mẫu thuộc lớp $\mathbf{C}_i$ trong không gian ban đầu $\mathbb{R}^d$, được tính bằng:

   $$\mu_i = \frac{1}{N_i} \sum_{x_n \in \mathbf{C}_i} \mathbf{x}_n, \quad i = 1, 2$$

   trong đó:

   - N_i là số lượng điểm dữ liệu thuộc lớp $\mathbf{C}_i$.
   - $\mathbf{x}_n$ là một điểm dữ liệu thuộc lớp $\mathbf{C}_i$.
   - $\boldsymbol{\mu}_i$ là vector trung bình của lớp $\mathbf{C}_i$ trong không gian đặc trưng $\mathbb{R}^d$.

   Sau khi dữ liệu được chiếu lên phương $\mathbf{w}$, trung bình mẫu trong không gian mới (không gian 1 chiều) được ký hiệu là m_i:

   $$m_i = \frac{1}{N_i} \sum_{x_n \in \mathbf{C}_i} y_n$$

   Thay $y_n = \mathbf{w}^T \mathbf{x}_n$, ta có:

   $$m_i = \frac{1}{N_i} \sum_{x_n \in \mathbf{C}_i} \mathbf{w}^T \mathbf{x}_n$$

   Sử dụng tính chất tuyến tính của tích vô hướng:

   $$m_i = \mathbf{w}^T \left(\frac{1}{N_i} \sum_{x_n \in \mathbf{C}_i} \mathbf{x}_n \right) = \mathbf{w}^T \mu_i$$

   $$\implies (m_1 - m_2) = w^T \mu_1 - w^T \mu_2 = w^T (\mu_1 - \mu_2)$$

   Xét đẳng thức:

   $$(a^T b)^2 = (a^T b)(a^T b) = a^T b b^T a \quad (3.1)$$

   với a, b là hai véc tơ cùng chiều bất kỳ.

   Khi đó, phương sai giữa lớp (between-class variance) là:

   $$(m_1 - m_2)^2 = [w^T (\mu_1 - \mu_2)]^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w$$

   (Áp dụng đẳng thức ma trận 3.1)

   Đặt $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$, khi đó phương sai giữa lớp trở thành:

   $$(m_1 - m_2)^2 = \mathbf{w}^T S_B \mathbf{w}$$

2. **Phương sai trong lớp (Within-Class Variance)**

   Phương sai trong lớp (within-class variance) được định nghĩa là tổng phương sai của hai lớp $s_1^2 + s_2^2$.

   Các phương sai trong lớp (within-class variances) được định nghĩa là:

$$
\begin{aligned}
s_k^2 &= \sum_{n \in \mathbf{C}_1} (y_n - m_k)^2, \quad k = 1, 2, \dots \\
&= \sum_{n \in \mathbf{C}_1} \left(\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mu_k\right)^2 \\
&= \sum_{n \in \mathbf{C}_1} \left(\mathbf{w}^T (\mathbf{x}_n - \mu_k)\right)^2
\end{aligned}
$$

- **Lưu ý:** Các phương sai trong lớp ở đây không được tính theo cách trung bình như phương sai thông thường. Mức độ ảnh hưởng của mỗi phương sai trong lớp phụ thuộc vào số lượng điểm dữ liệu trong lớp đó. Lớp có nhiều điểm dữ liệu hơn sẽ đóng góp nhiều hơn vào tổng phương sai trong lớp. Do đó, thay vì lấy trung bình, ta giữ nguyên tổng phương sai trong lớp bằng cách nhân phương sai với số lượng điểm dữ liệu trong lớp. Ta có thể hiểu phương sai trong lớp thực chất là phương sai thông thường nhân với số lượng điểm dữ liệu trong lớp đó.

Vậy phương sai trong lớp khi đó:

$$
\begin{aligned}
s_1^2 + s_2^2 &= \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} (y_n - m_k)^2, \quad k = 1, 2, \dots \\
&= \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} \left(\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mu_k\right)^2 \\
&= \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} \left(\mathbf{w}^T (\mathbf{x}_n - \mu_k)\right)^2
\end{aligned}
$$

Áp dụng đẳng thức ma trận 3.1

$$
\begin{aligned}
s_1^2 + s_2^2 &= \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} \left(\mathbf{w}^T (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \mathbf{w}\right) \\
&= \mathbf{w}^T \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} \left((\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T\right) \mathbf{w}
\end{aligned}
$$

Đặt $S_w = \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$ thì phương sai trong lớp sẽ trở thành:

$$
s_1^2 + s_2^2 = \mathbf{w}^T S_w \mathbf{w}
$$

3. **Hàm mục tiêu**

LDA là thuật toán đi tìm giá trị lớn nhất của hàm mục tiêu:

$$
J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}
$$

Bây giờ, bài toán tối ưu cho LDA trở thành:

$$
J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \quad \text{và} \quad \mathbf{w} = argmax \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}
$$

với:

- $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$

- $S_w = \sum_{k=1}^{2} \sum_{n \in \mathbf{C}_1} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$

3.1.3 Bài toán tối ưu hoá

Để tìm giá trị lớn nhất của hàm mục tiêu, ta thực hiện đạo hàm $J(w)$ và cho kết quả bằng 0.

$$
\begin{aligned}
&\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_{\mathbf{w}} \mathbf{w}} \right) = 0 \\
&\Leftrightarrow \left(\mathbf{w}^T S_{\mathbf{w}} \mathbf{w} \right) \frac{d}{d\mathbf{w}} \left(\mathbf{w}^T S_B \mathbf{w} \right) - \left(\mathbf{w}^T S_B \mathbf{w} \right) \frac{d}{d\mathbf{w}} \left(\mathbf{w}^T S_{\mathbf{w}} \mathbf{w} \right) = 0 \\
&\Leftrightarrow \left(\mathbf{w}^T S_{\mathbf{w}} \mathbf{w} \right) 2 S_B \mathbf{w} - \left(\mathbf{w}^T S_B \mathbf{w} \right) 2 S_{\mathbf{w}} \mathbf{w} = 0 \\
&\Leftrightarrow \left(\frac{\mathbf{w}^T S_{\mathbf{w}} \mathbf{w}}{\mathbf{w}^T S_{\mathbf{w}} \mathbf{w}} \right) S_B \mathbf{w} - \left(\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_{\mathbf{w}} \mathbf{w}} \right) S_{\mathbf{w}} \mathbf{w} = 0 \\
&\Leftrightarrow S_B \mathbf{w} - J(\mathbf{w}) S_{\mathbf{w}} \mathbf{w} = 0 \\
&\Leftrightarrow S_{\mathbf{w}}^{-1} S_B \mathbf{w} - J(\mathbf{w}) \mathbf{w} = 0 \\
&\Leftrightarrow S_{\mathbf{w}}^{-1} S_B \mathbf{w} = J(\mathbf{w}) \mathbf{w} \quad (3.2)
\end{aligned}
$$

Vì $J(\mathbf{w})$ là một số vô hướng, nên suy ra $\mathbf{w}$ phải là một vector riêng của ma trận $\mathbf{S}_W^{-1}\mathbf{S}_B$ ứng với một trị riêng nào đó.

Hơn nữa, giá trị của trị riêng này chính bằng $J(\mathbf{w})$. Do đó, để hàm mục tiêu đạt giá trị lớn nhất, ta cần chọn trị riêng lớn nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$. Điều này có nghĩa là $\mathbf{w}$ sẽ là vector riêng ứng với trị riêng lớn nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Vì $\mathbf{w}$ là nghiệm của $\mathbf{w} = \arg\max \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$, nên bất kỳ bội số $k\mathbf{w}$ (với $k \neq 0$) cũng là nghiệm. Điều này là do giá trị hàm mục tiêu chỉ phụ thuộc vào hướng của $\mathbf{w}$, không phụ thuộc vào độ lớn của nó. Do đó, mọi nghiệm thuộc về một đường thẳng trong không gian, không phải là một điểm duy nhất.

Chứng minh:

Giả sử $\mathbf{w}^*$ là nghiệm tối ưu, tức là:

$$
J(\mathbf{w}^*) = \frac{\mathbf{w}^{*T} S_B \mathbf{w}^*}{\mathbf{w}^{*T} S_W \mathbf{w}^*} = \max_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}
$$

Xét bội số $k\mathbf{w}^*$ với $k \neq 0$, ta có:

$$
J(k\mathbf{w}^*) = \frac{(k\mathbf{w}^*)^T S_B (k\mathbf{w}^*)}{(k\mathbf{w}^*)^T S_W (k\mathbf{w}^*)} = \frac{k^2 (\mathbf{w}^*)^T S_B (\mathbf{w}^*)}{k^2 (\mathbf{w}^*)^T S_W (\mathbf{w}^*)} = \frac{(\mathbf{w}^*)^T S_B (\mathbf{w}^*)}{(\mathbf{w}^*)^T S_W (\mathbf{w}^*)} = J(\mathbf{w}) \quad (3.3)
$$

Ở đây ta thấy, khi giải bài toán LDA, hàm mục tiêu $J(\mathbf{w})$ chỉ phụ thuộc vào hướng của $\mathbf{w}$ chứ không phụ thuộc vào độ lớn của nó. Điều này có nghĩa là nếu $\mathbf{w}$ là nghiệm, thì mọi bội số $k\mathbf{w}$ (với $k \neq 0$) cũng là nghiệm.

Để "đóng băng" độ lớn của $\mathbf{w}$ hay chuẩn hóa độ lớn của nó mà không làm thay đổi hướng, ta có thể đặt thêm điều kiện:

$$
(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w} = J(\mathbf{w}) = L \quad (3.4)
$$

với L là trị riêng lớn nhất của $\mathbf{S}_W^{-1}\mathbf{S}_B$.

Điều này đảm bảo rằng khoảng cách giữa trung bình các lớp sau khi chiếu (tính bằng $m_1 - m_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{w}$) đạt đúng mức tối ưu (tương đương với giá trị riêng lớn nhất L). Qua đó, ta có được một

hướng $\mathbf{w}$ đã được chuẩn hóa phù hợp với mục tiêu phân biệt tối ưu.

Bây giờ, chúng ta đi giải bài toán trị riêng (3.2) với điều kiện (3.4).

Với $\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$, thay vào $(*)$:

$$L\mathbf{w} = J(\mathbf{w})\mathbf{w} = \mathbf{S}_W^{-1}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\right]\mathbf{w} = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{w}$$

mà $L = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{w}$ nên:

$$\begin{aligned} & L\mathbf{w} = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{w} \\ \Leftrightarrow \quad & L\mathbf{w} = L\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ \Leftrightarrow \quad & \mathbf{w} = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \end{aligned}$$

Do $\mathbf{w}$ là nghiệm, nên bất kỳ bội số $k\mathbf{w}$ (với $k \neq 0$) cũng là nghiệm (kết quả chứng minh 3.3), nên nghiệm tổng quát của $\mathbf{w}$ là:

$$\mathbf{w} = \alpha\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \qquad \text{với } \alpha \neq 0 \text{ bất kỳ}$$

3.1.4 Phương pháp phân lớp

3.2 LDA với đa lớp

4 Cài đặt và thực nghiệm

5 Kết luận và hướng phát triển

6 Tham khảo