TRƯỜNG ĐẠI HỌC BÁCH KHOA TP. HCM - ĐHQG TP. HCM

o0o



Ứng dụng phương pháp LDA trong việc trích xuất đặc trưng

MÔN HỌC: HỌC MÁY VÀ ỨNG DỤNG

GVHD: Dương Tuấn Anh

Sinh viên thực hiện

Phạm Văn Thành – 1814029

Phan Đình Khánh – em không nhớ.

TP HỒ CHÍ MINH, THÁNG 03/2018

Mục lục

[I. Giới thiệu 3](#_Toc71398036)

[II. Fisher LDA 3](#_Toc71398037)

[III. Thí nghiệm 4](#_Toc71398038)

[1. Mô tả tập dữ liệu 4](#_Toc71398039)

[2. Minh họa kết quả 4](#_Toc71398040)

[IV. Kết luận 5](#_Toc71398041)

[V. Tài liệu tham khảo 5](#_Toc71398042)

1. **Giới thiệu**

Phân tích phân biệt tuyến tính Fisher (Fisher LDA) là một kỹ thuật giảm kích thước đơn giản nhưng hiệu quả để phân loại nhị phân. Nó chiếu một tập dữ liệu vào một không gian chiều thấp hơn và duy trì khả năng phân tách lớp tốt. Tuy nhiên, hiệu suất của Fisher LDA quá nhạy cảm với một số ngoại loai.   
Trong phần tiếp theo, chúng tôi giới thiệu về Fisher LDA. Phần III nói về tập dữ liệu và minh họa chúng trên ngôn ngữ lập trình Python. Cuối cùng chúng ta kết luận và đánh giá thuật toán.

1. **Fisher LDA**

Đối với một phân biệt tuyến tính được đặc trưng bởi , chúng tôi tìm cách thu được một giá trị vô hướng y bằng cách chiếu các mẫu lên một đường . Trong số tất cả các đường thẳng có thể, chúng tôi muốn chọn một dòng tối đa hóa khả năng phân tách của các đại lượng vô hướng. Fisher đề xuất tối đa hóa sự khác biệt giữa giá trị trung bình của các mẫu dự kiến, được chuẩn hóa bằng thước đo độ phân tán bên trong lớp, trong đó lớp phân tán bên trong (within-class scatter) được định nghĩa là tổng phương sai của cụm hoặc các mẫu được chiếu. Do đó, mức độ phân biệt đối xử được đo lường bằng tỷ lệ phân biệt đối xử Fisher

**(1)**

Trong đó:

và biểu thị giá trị trung bình và ma ma trận hiệp phương sai của các trường hợp dương

và biểu thị giá trị trung bình và ma ma trận hiệp phương sai của cho các trường hợp âm,

được định nghĩa là ma trận hiệp phương sai giữa các lớp,

được định nghĩa là ma trận hiệp phương sai trong lớp.

Đạo hàm và giải phương trình (1), chúng ta nhận được

**.**

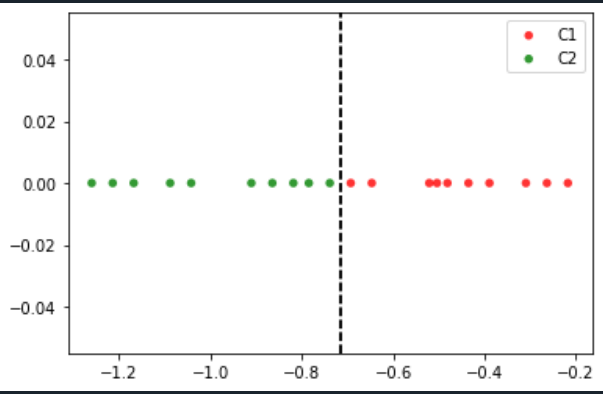
Bằng cách giải quyết vấn đề giá trị riêng tổng quát này cho kết quả

(tức là chiếu nó tới vector riêng có giá trị riêng lớn nhất ).

1. **Thí nghiệm**
2. **Mô tả tập dữ liệu**

Để minh họa thuật toán trên, ta sử dụng tập dữ liệu với 10 mẫu được gán nhãn 1 và 10 mẫu được gãn nhãn là 2. Để trực quan, chúng ta chiếu kết quả thu được lên không gian 2 chiều.

1. **Minh họa kết quả**



Hình 1: Chiếu dữ liệu mẫu lên 2 hướng.

1. **Kết luận**

* LDA là một phương pháp giảm chiều dữ liệu có sử dụng thông tin về label của dữ liệu. LDA là một thuật toán học có giám sát.
* Ý tưởng cơ bản của LDA là tìm một không gian mới với số chiều nhỏ hơn không gian ban đầu sao cho hình chiếu của các điểm trong cùng 1 lớp lên không gian mới này là gần nhau trong khi hình chiếu của các điểm của các classes khác nhau là khác nhau.
* LDA có giả sử ngầm rằng dữ liệu của các classes đều tuân theo phân phối chuẩn và các ma trận hiệp phương sai của các lớp là gần nhau.
* LDA hoạt động rất tốt nếu các classes là khả phân tách tuyến tính, tuy nhiên, chất lượng mô hình giảm đi rõ rệt nếu các lớp là không khả phân tách tuyến tính. Điều này dễ hiểu vì khi đó, chiếu dữ liệu lên phương nào thì cũng bị chồng lần, và việc tách biệt không thể thực hiện được như ở không gian ban đầu.
* Mặc dù có hạn chế, ý tưởng về small within-class và large between-class được sử dụng rất nhiều trong các mô hình classification khác.

1. **Tài liệu tham khảo**