

Bài 8-THỐNG KÊ NHIỀU CHIỀU

Trong các chương trước chúng ta đã nghiên cứu các vấn đề liên quan đến một biến định lượng như nhật độ, thống kê, ước lượng, kiểm định giá trị trung bình, kiểm định phương sai, so sánh hai trung bình, so sánh nhiều trung bình (phân tích phương sai).

Khi có nhiều biến thì để hiểu người học phải có sự hiểu biết sâu hơn về toán học đặc biệt là các vấn đề trình bày trong đại số tuyến tính như không gian vectơ, ánh xạ tuyến tính, dạng toàn phương, giá trị riêng và vectơ riêng $v, v, \dots$

Sau đây là một số phần được trình bày trong thống kê nhiều chiều

1-Thống kê mô tả

Giả thiết thường đưa ra là k biến phân phối chuẩn nhiều chiều (Multivariate Normal distribution) $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ là vectơ trung bình (kỳ vọng), $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai. Từ ma trận hiệp phương sai có thể tìm được ma trận tương quan $\boldsymbol{\rho}$.

Nếu lấy mẫu quan sát gồm n vectơ ngẫu nhiên trong không gian k chiều thì tính được vectơ trung bình cộng và ma trận hiệp phương sai mẫu $\mathbf{S}$.

Việc nghiên cứu phân phối của và phân phối của $\mathbf{S}$ (thường gọi là phân phối Wishart) là sự mở rộng của bài toán nghiên cứu phân phối của trung bình cộng và phương sai mẫu s^2 trong trường hợp một biến chuẩn $N(\mu, \sigma^2)$.

Việc tìm các ước lượng của vectơ $\boldsymbol{\mu}$ và ma trận $\boldsymbol{\Sigma}$ và nghiên cứu các tính chất của các ước lượng đó là sự mở rộng của bài toán ước lượng μ và σ^2 đối với biến chuẩn $N(\mu, \sigma^2)$.

Việc tìm miền tin cậy (thường gọi là elipsoit tin cậy) của vectơ $\boldsymbol{\mu}$ là sự mở rộng của bài toán tìm khoảng tin cậy đối với trung bình μ của một biến chuẩn.

Việc so sánh 2 vectơ trung bình $\boldsymbol{\mu}_1$ và $\boldsymbol{\mu}_2$ là sự mở rộng của bài toán so sánh 2 trung bình μ_1 và μ_2 của một biến chuẩn trên 2 tổng thể. Ở đây cũng phân chia thành so sánh khi lấy mẫu độc lập và so sánh khi lấy mẫu theo cặp.

Việc so sánh nhiều vectơ trung bình được trình bày trong phần phân tích phương sai một nhân tố nhiều chiều (One way Manova) và là sự mở rộng của bài toán phân tích

phương sai một nhân tố (One way Anova) đối với một nhân tố có nhiều mức. Sau phân tích phương sai là so sánh các trung bình của các mức của nhân tố với rất nhiều tiêu chuẩn (Test) so sánh. Có thể mở rộng sang phân tích phương sai 2 nhân tố (Two way multivariate analysis of variance).

2-Hồi quy bội tuyến tính nhiều chiều (Multivariate Linear regression models)

Phần này trình bày lại bài toán hồi quy bội tuyến tính và hồi quy đa thức đối với một biến phụ thuộc y với cách nhìn của thống kê nhiều chiều. Tiếp theo là sự mở rộng bài toán tương quan và hồi quy tuyến tính đối với một biến (một chiều) sang hồi quy bội tuyến tính nhiều chiều với các nội dung như khảo sát mô hình, cách tính các hệ số hồi quy, tìm phân phối của các hệ số hồi quy, dự báo . . .

3-Phân tích thành phần chính (Principal components)

Có thể nhìn phương pháp thành phần chính dưới 2 góc độ:

+ Giảm số chiều để có hình ảnh trông thấy được(Data reduction)

Đám mây quan sát gồm n điểm trong không gian k chiều. Với $k > 3$ chúng ta không nhìn thấy đám mây. Để có một hình ảnh trông thấy được phải chọn một hệ tọa độ trục giao mới trong không gian k chiều sao cho hình chiếu của n điểm trên trục thứ nhất (thành phần chính 1) có biến động (phương sai) lớn nhất (so với mọi đường thẳng - trục - trong không gian k chiều), trục thứ hai (thành phần chính thứ hai) có biến động lớn nhất trong mọi trục vuông góc với trục thứ nhất, tiếp theo là trục thứ ba (thành phần chính thứ 3) vuông góc với mặt phẳng của 2 trục đầu. . .

Chiều đám mây quan sát (n điểm quan sát) lên mặt phẳng của thành phần chính 1 và thành phần chính 2 sẽ được hình ảnh gần đúng tốt nhất (trung thành nhất) của đám mây quan sát. Dựa trên hình ảnh 2 chiều này để phân tích đám mây quan sát, các phân tích đó được bổ sung bởi hình chiếu trên mặt phẳng thành phần chính 1 – thành phần chính 3 và hình chiếu trên mặt phẳng thành phần chính 2 – thành phần chính 3.

+ Coi phương pháp thành phần chính là một trong nhiều phương pháp phân tích nhân tố (Factor analysis).

Phương pháp phân tích nhân tố cho là tuy có k biến nhưng chúng không độc lập, quan hệ giữa chúng, thể hiện qua ma trận phương sai S , được lý giải là do chúng chung nhau một số ít nhân tố (Factor). Cần tìm ra các nhân tố chung đó để có thể tái hiện lại ma trận hiệp phương sai S .

4-Phân tích chính tắc (Canonical Correlation analysis)

Khi có 2 nhóm biến chúng ta có thể lấy 1 cặp gồm 1 biến của nhóm 1 và một biến của nhóm 2. Tìm cặp có cho hệ số tương quan lớn nhất trong tất cả các cặp có thể tìm được. Cặp biến đó được gọi là cặp biến chính tắc 1. Tiếp theo tìm cặp biến có hệ số tương quan lớn nhất trong số các cặp biến không tương quan (uncorrelated) với cặp đầu và gọi đó là cặp biến chính tắc thứ 2 v. v . . .

Có thể rút gọn việc khảo sát mối quan hệ giữa 2 nhóm biến về việc khảo sát một số ít cặp biến chính tắc.

5- Phân tích phân biệt và bài toán xếp loại (Discrimination and classification)

Nếu đám mây quan sát bao gồm r nhóm khác nhau thì có thể tìm cách tách biệt (phân biệt) chúng bằng một số hàm gọi là hàm phân biệt (Discriminant function). Hay dùng nhất là các hàm phân biệt tuyến tính (Còn gọi là hàm phân biệt Fisher - linear discriminant function). Căn cứ vào giá trị của các hàm này để phân biệt nhóm này với nhóm khác.

Bây giờ nếu có một quan sát mới thì nên xếp nó vào nhóm nào trong r nhóm nói trên. Bài toán này được gọi chung là bài toán xếp loại (Classification).

Có nhiều phương pháp khác nhau với những tiêu chuẩn khác nhau để xếp loại. Nhưng nếu đã có các hàm phân biệt thì có thể dùng các giá trị của các hàm đó tại điểm quan sát mới này để xếp loại.

6- Phân tích chùm (Cluster analysis)

Có n điểm quan sát, có thể ghép các điểm lại thành một số nhóm hay không?

Vấn đề này gọi chung là phân tích chùm. Có rất nhiều phương pháp nhưng hay dùng nhất là ghép thành cây (Hierarchical cluster analysis). Coi các điểm như những chiếc lá, các lá gần nhau sẽ ghép lại thành nhánh con, các nhánh con gần nhau sẽ ghép lại thành cành nhỏ, các cành nhỏ gần nhau sẽ ghép lại thành cành to, các cành to sẽ ghép lại thành cây.

Có 2 giai đoạn:

Đối với 2 điểm (2 lá) phải định nghĩa **khoảng cách giữa 2 điểm** để sau đó tìm 2 điểm (2 lá) gần nhau nhất. Có rất nhiều định nghĩa khoảng cách giữa 2 điểm đối với các biến định lượng và biến định tính.

Khi đã ghép các điểm (lá) thành nhánh và sau đó thành cành thì mỗi nhánh, mỗi cành là một nhóm điểm (lá), phải định nghĩa **khoảng cách giữa 2 nhóm (2 cành)**. Cũng có rất nhiều định nghĩa khoảng cách giữa 2 nhóm.

Việc chọn khoảng cách giữa 2 điểm và khoảng cách giữa 2 nhóm dẫn đến các cách ghép nhóm thành cây rất khác nhau.

Phân tích chùm được dùng rất rộng rãi trong sinh học và đem lại các cách phân loại mới trong các ngành học liên quan đến sinh học.