Khám phá đặc trưng

Khám phá đặc trưng

Trong một bộ dữ liệu, một đặc trưng thường được biểu diễn dưới dạng dữ liệu thô, nó là một thuộc tính có thể đo lường được và thường được mô tả bởi một cột trong tập dữ liệu. Tuy nhiên chúng ta thường gặp khó khăn khi nhìn vào một bảng dữ liệu chứa rất nhiều con số bởi vì chúng ta khó có thể hình dung và hiểu được dữ liệu chỉ dưới dạng số thông thường. Bên cạnh đó, việc sử dụng các mô hình học máy khi không có hiểu biết nhất định về bộ dữ liệu khiến cho công sức của chúng ta trở nên kém hữu dụng và chúng ta khó có thể tìm ra một mô hình phù hợp với bài toán đang cần giải quyết. Do vậy, ta cần hiểu được các đặc trưng và phần nào mối quan hệ giữa chúng.

Việc phân tích khám phá dữ liệu này là rất có ích bởi nó góp phần vào phát hiện lỗi sớm trong tập dữ liệu, kiểm tra tính đúng đắn các giả định, xác định mối quan hệ giữa các biến đặc trưng, đánh giá hướng và mức độ ảnh hưởng của các biến đặc trưng lên biến nhãn, và lựa chọn một mô hình phù hợp với các mối quan hệ giữa các biến đặc trưng và biến nhãn, ...

Phân tích khám phá dữ liệu thường bao gồm cả phân tích khám phá phi hình ảnh và phân tích khám phá hình ảnh. Phân tích khám phá phi hình ảnh là việc sử dụng các thông số thống kê để đem lại cho các nhà phân tích những hiểu biết nhất định về bộ dữ liệu từ các con số đó. Cũng như vậy, phân tích khám phá dữ liệu hình ảnh cung cấp thông tin về bộ dữ liệu tới các nhà phân tích nhưng theo một cách trực quan hóa thông qua dạng hình ảnh. Đối với tư duy con người, hình ảnh luôn là công cụ vô cùng mạnh mẽ để truyền tải thông tin một cách hiệu quả nhất. Chúng ta sẽ cùng nhau tìm hiểu về cả hai cách phân tích khám phá dữ liệu trên trong bài viết này, qua đó giúp chúng ta có được các kỹ năng cần thiết để khai phá được lượng thông tin cần thiết trước khi lựa chọn các mô hình phù hợp với bài toán thực tế của chúng ta.

Phân tích khám phá phi hình ảnh

Phân tích khám phá phi hình ảnh (non-graphical exploring data analysis) là bước đầu tiên khi bạn bắt đầu phân tích dữ liệu như một phần của phương pháp phân tích dữ liệu nói chung. Bước phân tích dữ liệu sơ bộ này tập trung vào bốn điểm chính sau:

- Các độ đo về xu hướng tập trung: Giá trị trung bình, trung vị, mốt, …
- Các độ đo về độ phân tán của phân phối: Phương sai, độ lệch chuẩn, …
- Hình dạng của phân phối
- Sự tồn tại của các ngoại vi, ngoại lệ

Phân tích khám phá phi hình ảnh có thể được theo dõi hoặc tham gia trực tiếp vào quá trình phân tích khám phá hình ảnh.

Các độ đo về xu hướng tập trung

Các độ đo về xu hướng tập trung là một số các thông số thống kê cơ bản và hữu ích. Chúng tóm tắt một mẫu hoặc một tập hợp bởi một giá trị đơn điển hình. Hai độ đo phổ biến nhất cho dữ liệu dạng số là giá trị trung bình và trung vị.

- Trung bình (Mean): Giá trị trung bình trên toàn bộ các điểm dữ liệu
- Trung vị (Median): Giá trị tại điểm dữ liệu chia bộ dữ liệu thành hai phần bằng nhau
- Mốt (Mode): Giá trị phổ biến nhất trong tập dữ liệu

Trung bình (Mean)

Trung bình, hay còn gọi là trung bình cộng, trung bình số học, đưa ra một “giá trị kỳ vọng” của bộ dữ liệu.

Trung bình mẫu, được viết dưới dạng $\bar{X}$, được tính bằng cách lấy tổng của các quan sát chia cho số lượng các mẫu quan sát:

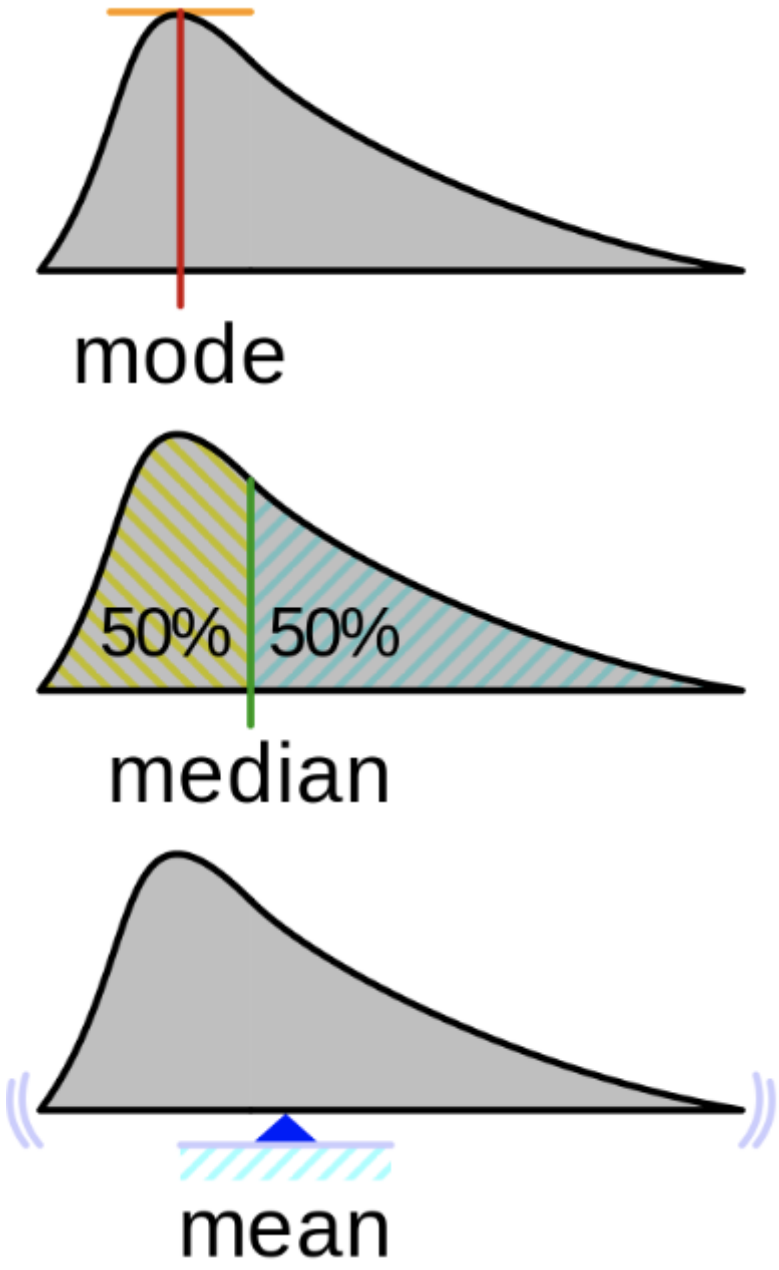
Giá trị trung bình của không gian mẫu, được viết dưới dạng μ, tương tự như trung bình mẫu nhưng được tính trên toàn bộ không gian dữ liệu (còn gọi là quần thể dữ liệu).

Trung vị (Median)

Cũng được biết như là phần trăm thứ 50, giá trị trung vị của mẫu là giá trị ở giữa khi các quan sát được viết ra một cách có thứ tự.

Giá trị trung vị của một quần thể là phần trăm thứ 50 của toàn bộ quần thể đó.

So sánh giữa trung bình và trung vị: Nhìn chung, trung bình và trung vị đều cho biết giá trị "điển hình" hoặc "trung tâm" của một phân phối. Với các phân phối đối xứng như là phân phối chuẩn, giá trị trung bình bằng với giá trị trung vị. Tuy nhiên, giá trị trung bình được sử dụng nhiều hơn trong các thủ tục thống kê cũng như các mô hình toán học. Trong khi giá trị trung vị lại có nhiều hơn khả năng được áp dụng để chống lại các giá trị ngoại vi.



Hình 1. Các giá trị trung tâm cơ bản (từ Wikipedia)

Các độ đo về độ phân tán của phân phối

Độ phân tán (Variability)

Độ phân tán là một trong những chủ đề cơ bản của thống kê. Nhìn chung, không phải mọi điểm trên một phân phối đều có chung một giá trị xét trên một biến mà ta quan tâm. Các giá trị biểu hiện xu hướng tập trung như là trung bình và trung vị là không hoàn thiện bởi chúng không cho biết độ phân tán của dữ liệu xung quanh trung tâm. Tất cả các suy luận thống kê để phải tính đến độ phân tán của dữ liệu.

Độ lệch (Deviation)

Độ lệch của một quan sát là khoảng cách từ điểm đó tới giá trị trung bình của tập dữ liệu. Một cách công thức hóa, độ lệch của quan sát thứ *i* được cho dưới dạng: *Xi-X*.

Giá trị trung bình của các độ lệch luôn bằng 0 bởi tổng của tất cả các độ lệch trên toàn bộ tập dữ liệu bằng 0.

Phương sai (Variance)

Phương sai mẫu, kí hiệu là *s2*, là một thước đo độ phân tán thường được sử dụng. Nó được tính bằng trung bình của bình phương các độ lệch.

$$s^2 = Phương\ sai = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

Phương sai của quần thể, kí hiệu là *σ2*, được tính tương tự phương sai mẫu nhưng tính trên toàn bộ phân phối. Do đó, *σ2* sẽ bằng trung bình của các bình phương độ lệch trên toàn bộ phân phối.

Độ lệch chuẩn (Standard Deviation)

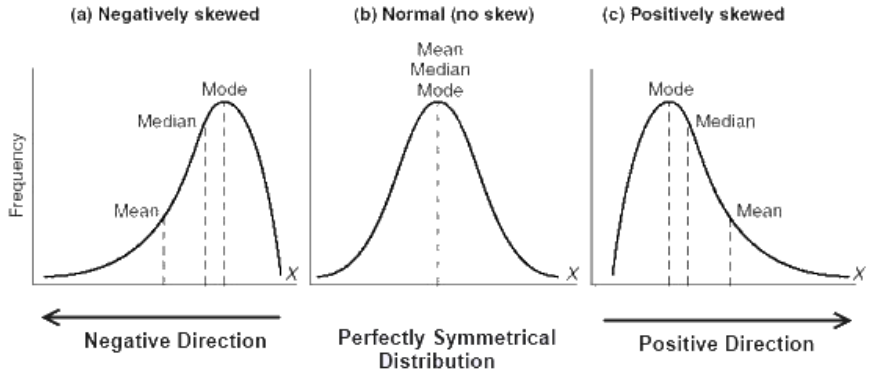
Độ lệch chuẩn được tính bằng căn bậc hai của phương sai. Điều này có thể được hiểu rằng độ lệch chuẩn gần như là khoảng cách trung bình từ các quan sát đến giá trị trung bình của phân phối.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

Hình dạng của phân phối

Các giá trị biểu thị xu hướng tập trung và độ trải rộng là không đủ để ta hình dung một phân phối. Bởi lẽ các phân phối có cùng giá trị trung bình và độ lệch chuẩn vẫn có thể có các hình dạng phân phối rất khác nhau. Người ta định nghĩa một số hình dạng cơ bản của các phân phối như sau:

- Phân phối đối xứng
  - Phân phối hình chuông hoặc phân phối chuẩn
  - Phân phối đồng dạng (uniform)
- Phân phối xiên
  - Phân phối xiên trái (skewed to the left hoặc positively skewed)
  - Phân phối xiên phải (skewed to the right hoặc negatively skewed)
- Phân phối thể dạng
  - Unimodal
  - Bimodal



Hình 2. Các hình dạng của các phân phối cơ bản

Sự tồn tại của các ngoại vi, ngoại lệ (outliers)

Một ngoại vi hay ngoại lệ là một quan sát nằm "rất xa" các giá trị khác trong mẫu ngẫu nhiên của phân phối. Khái niệm "rất xa" tùy thuộc vào việc các nhà phân tích dữ liệu định nghĩa các giá trị "bình thường" ra sao trong ngữ cảnh của vấn đề. Các giá trị ngoại vi có thể biểu thị cho các điểm dữ liệu nhiễu hay một sự kiện hiếm gặp trong biến ngẫu nhiên mà ta cần xem xét cẩn thận để hiểu lý do vì sao chúng xuất hiện trong mẫu dữ liệu và chúng có ảnh hưởng đến vấn đề chúng ta đang giải quyết hay không. Các phân tích nên được thử nghiệm trên cả tập dữ liệu chứa và không chứa các ngoại vi, và nếu không có sự lý giải hợp lý nào về sự tồn tại của chúng trước khi ta loại trừ chúng thì kết quả của các phân tích này cần được báo cáo và xem xét kỹ lưỡng.

Việc phát hiện ngoại vi sẽ được dễ dàng nhìn thấy với dữ liệu liên tục thông qua địa đồ phân tán sẽ trình bày ở phần sau. Còn trong phạm vi của mục này, ta xét tới việc tính toán khoảng trải giữa (InterQuantile Range – IQR) và sử dụng tích của các tứ phân vị để định nghĩa được các giá trị như thế nào sẽ là các ngoại vi. Khoảng trải giữa được tính bằng khoảng cách từ tứ phân vị thứ nhất (Q1) đến tứ phân vị thứ ba (Q3). Nhắc lại rằng, các giá trị trong khoảng tứ phân vị thứ nhất chiếm 25% phân phối, trong khoảng tứ phân vị thứ hai chiếm 50% phân phối và trong khoảng tứ phân vị thứ ba chiếm 75% phân phối.

Các ngoại vi tiềm năng sẽ nằm ngoài khoảng giá trị sau:

$$[Q1 – (1.5 \times IQR), Q3 + (1.5 \times IQR)],$$

và các ngoại vi khả năng cao sẽ nằm ngoài khoảng giá trị sau:

$$[Q1 – (3 \times IQR), Q3 + (3 \times IQR)].$$

Phân tích khám phá hình ảnh

Một trong những thứ có ích nhất mà bạn có thể làm để hiểu dữ liệu của bạn đó là thể hiện chúng trong một định dạng nhất định của hình ảnh. Hình ảnh hóa dữ liệu cho phép bạn tương tác với chúng, phân tích chúng trong một cách thức đơn giản và tường minh. Ngoài ra, hình ảnh hóa dữ liệu cũng giúp ta nhận ra những mẫu mới trong dữ liệu, và cũng có thể khiến cho dữ liệu phức tạp của ta trở nên dễ hiểu hơn. Bộ não của chúng ta hoạt động hiệu quả hơn với các dữ liệu hình dạng, màu sắc, chiều dài của các biểu đồ hoặc các đồ thị bởi chúng trực quan hơn rất nhiều so với dữ liệu dạng bảng tính.

Hình 3. Hình ảnh hóa dữ liệu

Ví dụ minh họa ở trên là rất rõ ràng, bức ảnh đã thể hiện khéo léo các thành phần của món ăn và số lượng mỗi thành phần. Chìa khóa ở đây là bộ não chúng ta sử dụng năng lượng thụ động để hiểu các thông tin dạng hình ảnh. Nếu bạn phải dành hàng giờ, hàng ngày, thậm chí hàng tuần để phân tích dữ liệu dạng bảng tính, thì khi phân tích dữ liệu bằng hình ảnh, bạn sẽ nhận ra điều gì là có giá trị trong dữ liệu của bạn một cách nhanh chóng hơn rất nhiều.

Các bộ dữ liệu mẫu

Để hiểu và dễ dàng áp dụng các phương pháp khai phá dữ liệu vào thực tế hơn, chúng ta cần phải lựa chọn một bộ dữ liệu và làm việc trên nó. Trong mục này, tôi lựa chọn hai bộ dữ liệu là Thành tích của Học Sinh và bộ dữ liệu về hoa Iris có sẵn trong thư viện MatPlotLib trong Python. Bộ dữ liệu Thành tích của Học sinh là một bản rút gọn của bộ dữ liệu về học sinh của UCI. Bộ dữ liệu này khảo sát thành tích của học sinh trung học cơ sở của hai trường tại Bồ Đào Nha. Các thuộc tính của dữ liệu gồm có các lớp học của học sinh, các thuộc tính liên quan tới nhân khẩu học, xã hội và trường học. Bộ dữ liệu được thu thập bởi các báo cáo và khảo sát của hai trường. Các bạn có thể đọc thông tin chi tiết về bộ dữ liệu tại đây: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Biểu đồ Histogram

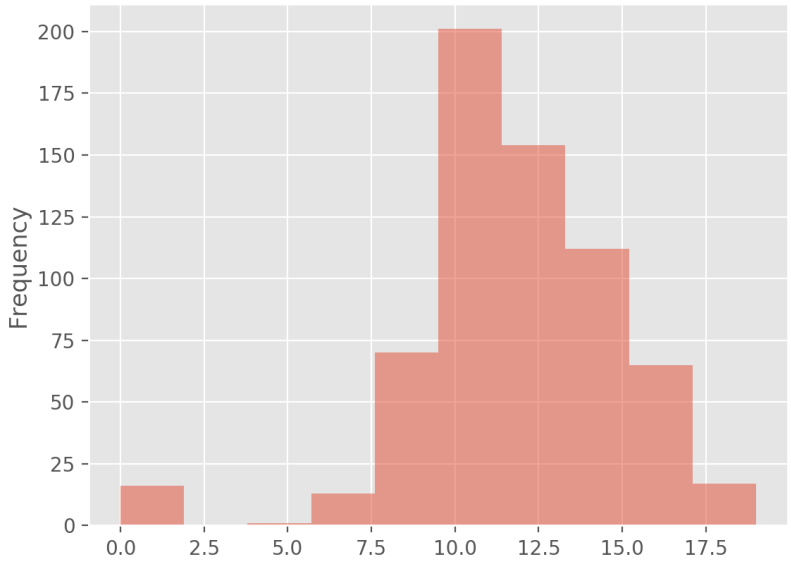
Histogram là một trong Bảy Công cụ Cơ bản về Chất lượng. Nó là một kỹ thuật đồ họa giúp xác định những yếu tố hữu ích cho việc giải quyết các vấn đề của bạn. Biểu đồ giúp bạn thấy phân phối của một đặc trưng trong tập dữ liệu. Chúng cũng thể hiện rằng bộ dữ liệu có bao nhiêu mẫu và chúng được phân bố như thế nào trong không gian của bộ dữ liệu.

Như chúng ta đã biết, dữ liệu bao gồm hai loại là dữ liệu liên tục và dữ liệu dạng nhóm. Histogram chỉ thực sự có ý nghĩa với dữ liệu dạng nhóm. Nếu dữ liệu của bạn thuộc dạng liên tục, chúng

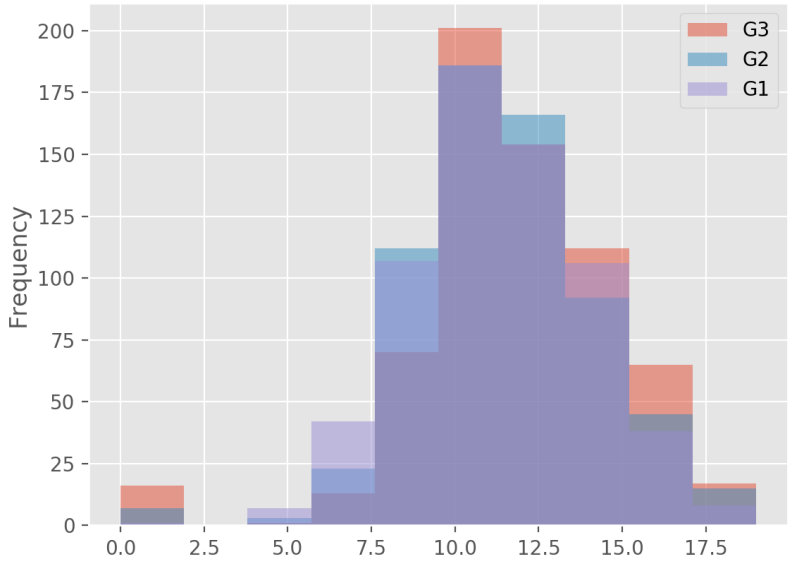
phải được ngăn chia thành các khoảng giá trị (gọi là bin) hoặc rời rạc hóa bởi việc biến đổi sang dữ liệu dạng nhóm bằng cách gom nhóm những giá trị tương tự lại với nhau. Để làm điều này, toàn bộ các giá trị của đặc trưng được chia vào một chuỗi các khoảng giá trị, thường là liên tục, có độ dài bằng nhau và không chồng chéo lên nhau. Các khoảng này sẽ trở thành các nhóm của đặc trưng dạng đó. Tiếp đó, số lượng các giá trị rơi vào mỗi khoảng được gọi là tần suất của khoảng (nhóm) đó.

Để biểu diễn một histogram, ta sẽ sử dụng bộ dữ liệu về học sinh kết hợp với thư viên MatPlotLib trong ngôn ngữ lập trình Python. Phương thức *.plot.hist()* có thể xử lý cả dữ liệu dạng nhóm và rời rạc hoá giữ liệu liên tục trong trường hợp cần thiết.

```
import pandas as pd
from matplotlib import pyplot as plt

plt.style.use('ggplot') # Look pretty

student_dataset = pd.read_csv("students.csv", index_col=0)

my_series = student_dataset.G3
my_dataframe = student_dataset[['G3', 'G2', 'G1']]

my_series.plot.hist(alpha=0.5)
plt.show()
my_dataframe.plot.hist(alpha=0.5)
plt.show()
```



Hình 4. Histogram của đặc trưng G3 .



Hình 5. Histogram của 3 đặc trưng G3, G2, G1.

Nếu bạn quan tâm tới tỉ lệ phần trăm trên mỗi bin hơn là tần suất của chúng, bạn có thể truyền vào tham số *normed*=True, kết quả của bạn sẽ được chuyển thành dạng phần trăm. Ngoài ra, bạn có thể đọc thêm tài liệu online của MatPlotLib để sử dụng histogram một cách hiệu quả hơn.

Việc biết rằng một đặc trưng được phân bố như thế nào trên khắp bộ dữ liệu là điều tương đối hữu ích, vì một số mô hình học máy chỉ hoạt động tốt khi phân bố của dữ liệu là dạng phân phối chuẩn (Gaussian). Với những mô hình như thế, nếu việc khám phá dữ liệu của chúng ta với một histogram chỉ ra rằng dữ liệu của chúng ta bị lệch thì bạn cũng không cần quá lo lắng. Vì thực tế có một số kỹ thuật biến đổi dữ liệu có thể hợp thức hóa dữ liệu bị lệch để đưa vào các mô hình theo phân phối chuẩn như thế.

Địa đồ phân tán (Scatter plot)

Tương tự như biểu đồ histogram, địa đồ phân tán (scatter plot) cũng là một trong Bảy Công cụ Cơ bản về Chất lượng. Hãy thêm nó vào thành một trong những kỹ năng của bạn để bạn có nhiều lựa chọn hơn trong việc phân tích dữ liệu. Hãy bắt đầu với địa đồ 2 biến (dữ liệu gồm 2 đặc trưng).

Địa đồ phân tán 2 chiều (2d scatter plot)

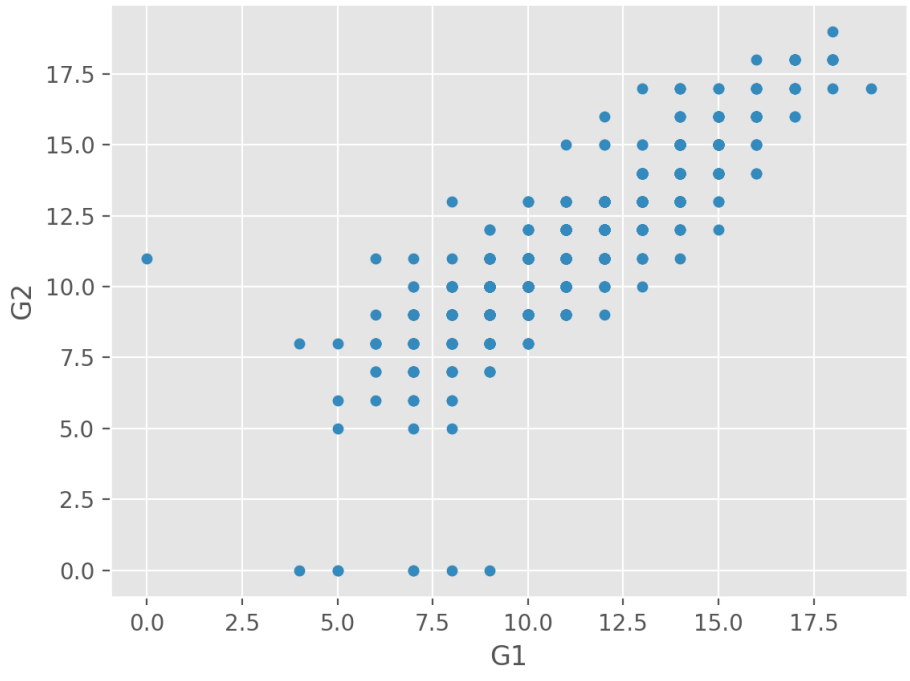
Địa đồ phân tán 2 chiều được sử dụng để kiểm tra trực quan xem liệu có sự tương quan nào giữa hai đặc trưng được lập biểu đồ hay không. Cả hai trục của một địa đồ phân tán 2 chiều đều biểu diễn cho những đặc trưng số và phân biệt. Chúng không phải là đặc trưng liên tục, nhưng chúng ít nhất phải có thứ tự vì mỗi bản ghi trong tập dữ liệu sẽ được phát họa thành một điểm với vị trí của nó dọc theo các trục và tương ứng với các giá trị của thuộc tính. Nếu không có thứ tự, vị trí của các điểm dữ liệu sẽ không có nhiều ý nghĩa.

Có thể là một sự tương quan cùng chiều hoặc ngược chiều (cùng tăng hoặc một tăng một giảm) giữa các đặc trưng được lập biểu đồ. Kiểu tương quan có thể được đánh giá thông qua xu hướng đường chéo tổng thể của các điểm dữ liệu.

Các sự tương quan cùng chiều hoặc ngược chiều có thể biểu thị cho một mối quan hệ tuyến tính hoặc không tuyến tính. Nếu một đường thẳng có thể được vẽ chạy qua địa đồ phân tán và hầu hết các điểm có xu hướng gần và vây quanh đường thẳng này thì chúng ta có thể khẳng định tương đối chắc chắn rằng có một mối quan hệ tuyến tính giữa các đặc trưng được lập biểu đồ. Tương tự, nếu một đường cong có thể được vẽ chạy qua các điểm của địa đồ thì đó dường như là một mối quan hệ phi tuyến tính. Nếu không phải là một đường thẳng và cũng không phải là một đường cong phù hợp với hình tổng thể của phân bố thì khả năng cao là không có sự tương quan hay mối liên hệ nào giữa các đặc trưng được lập biểu đồ, hay ít nhất là ta chưa đủ thông tin để xác định điều đó.

Hãy bắt đầu bằng đoạn code để phát họa địa đồ phân tán của bộ dữ liệu về học sinh. Ở đây, tôi lựa chọn xây dựng địa đồ phân tán của hai thuộc tính là G1 và G2.

```
import pandas as pd

from matplotlib import pyplot as plt

plt.style.use('ggplot') # Look pretty

student_dataset = pd.read_csv("students.csv", index_col=0)

student_dataset.plot.scatter(x='G1', y='G2')

plt.show()
```



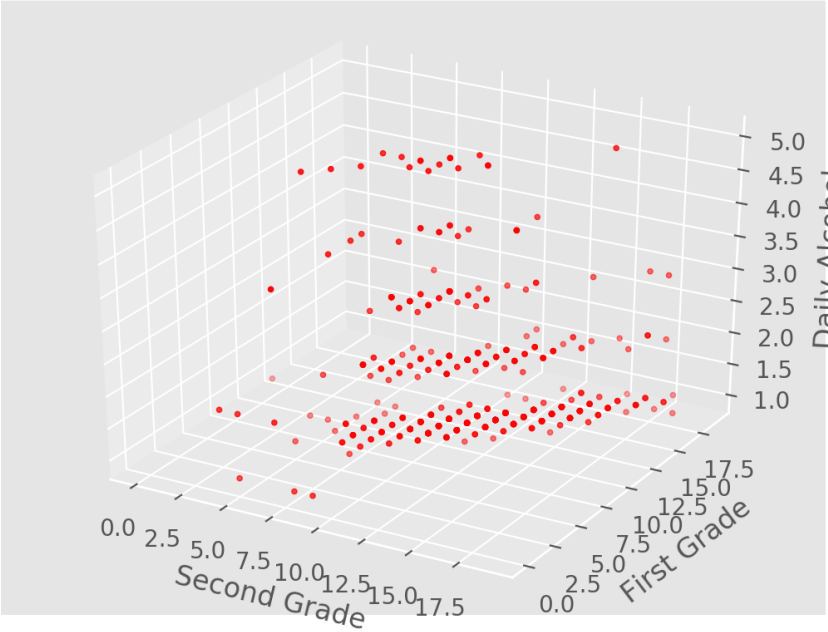
Hình 6. Địa đồ phân tán của bộ dữ liệu thành tích học sinh dựa trên hai thuộc tính G1 và G2

Nhìn vào địa đồ ta có thể thấy rằng giữa đặc trưng điểm kiểm tra đầu kì và điểm kiểm tra thứ hai trong kỳ dường như có sự tương quan cùng chiều và tồn tại quan hệ tuyến tính ở đây, mặc dù vẫn còn một số điểm ngoại lệ tách biệt khỏi khối này. Ta có thể thấy rằng, không ai trong số các học sinh này đạt điểm tốt ở kỳ thi đầu tiên (G1) mà lại bị điểm thấp ở kỳ thi thứ hai (G2).

Địa đồ phân tán 3 chiều (3D Scatter Plot)

Cũng giống như địa đồ phân tán 2 chiều, địa đồ phân tán 3 chiều cũng cho ta thấy mối tương quan giữa 3 biến được lập biểu đồ. Các bạn có thể tham khảo đoạn code bên dưới để biết cách lập địa đồ phân tán 3 chiều.

```
import matplotlib

import matplotlib.pyplot as plt

from mpl_toolkits.mplot3d import Axes3D

import pandas as pd

matplotlib.style.use('ggplot') # Look Pretty

# If the above line throws an error, use plt.style.use('ggplot') instead

student_dataset = pd.read_csv("students.csv", index_col=0)

fig = plt.figure()

ax = fig.add_subplot(111, projection='3d')

ax.set_xlabel('Second Grade')

ax.set_ylabel('First Grade')

ax.set_zlabel('Daily Alcohol')

ax.scatter(student_dataset.G1, student_dataset.G2, student_dataset['Dalc'], c='r', marker='.')

plt.show()
```



Hình 7. Địa đồ phân tán 3 chiều của bộ dữ liệu Học sinh

Nhìn vào địa đồ, chúng ta có thể thấy một số điều như sau:

- Vẫn tồn tại một mối tương quan tuyến tính, cùng chiều giữa điểm thi lần đầu và điểm thi lần thứ hai. Nhìn chung là điểm thi lần đầu càng cao thì điểm thi lần hai càng cao.
- Có một số lượng tương đối lớn các học sinh uống rượu hàng ngày.
- Một học sinh càng uống nhiều rượu hàng ngày thì điểm số của học sinh đó càng có xu hướng tệ.
- Sự tiêu thụ rượu hàng ngày không thực sự là một đặc trưng gây nên việc bỏ thi (được điểm 0) ở kỳ thi thứ hai.

Hình ảnh hóa dữ liệu nhiều chiều

Địa đồ phân tán rất hiệu quả trong việc truyền tải dữ liệu bằng việc ánh xạ một đặc trưng sang một chiều cụ thể trong không gian, nơi mà trực giác của chúng ta dễ dàng hiểu được. Tuy nhiên, khả năng của chúng ta bị giới hạn bởi chỗ chúng ta chỉ dễ dàng cảm nhận và hiểu một hình ảnh trong không gian không lớn hơn 3 chiều mà thôi. Nhưng trong thực tế thì các bộ dữ liệu trong thế giới thực thường có khoảng 10 đặc trưng trở lên. Cá biệt có những bộ dữ liệu có thể có 10.000 đặc trưng. Vậy làm thế nào để ta có thể hình ảnh hóa dữ liệu khi chúng ta có một bộ dữ liệu chứa nhiều hơn ba chiều?

Hệ tọa độ song song (Parallel Coordinate)

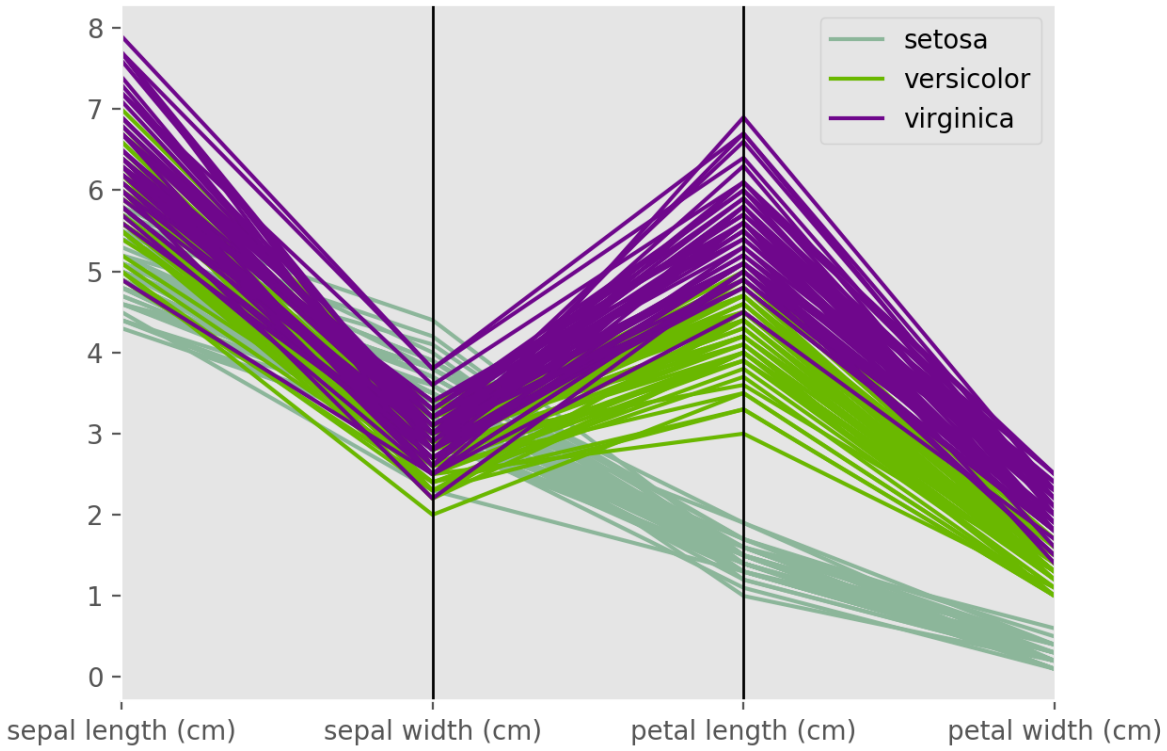
Hệ tọa độ song song tương đối giống với địa độ phân tán trong khía cạnh mỗi trục tọa độ ánh xạ tới một miền đặc trưng số. Nhưng thay vì có các trục tọa độ được căn chỉnh sao cho chúng vuông góc với nhau từng đôi một thì hệ tọa độ song song sẽ đặt các trục tọa độ một cách song song theo chiều dọc. Hay nói cách khác, một hệ tọa độ song song là một tập các trục biểu thị cho các đặc trưng được bố trí một cách song song theo chiều dọc.

Mỗi quan sát được vẽ đồ thị là một hình gồm nhiều đường (sau đây gọi là polyline), là một chuỗi kết nối các đường phân đoạn. Điểm kết nối giữa các đường này nằm ở các trục tọa độ. Vì mỗi trục ánh xạ tới một miền đặc trưng số nên mỗi polyline mô tả toàn bộ giá trị của một quan sát (điểm dữ liệu) theo các đặc trưng.

Hệ tọa độ song song là một kĩ thuật biểu đồ hữu dụng mà bạn sẽ muốn thêm vào các kĩ năng của mình. Chúng là một kĩ thuật hình ảnh hóa dữ liệu nhiều chiều (lớn hơn 3 chiều) bởi vì chúng cho phép bạn dễ dàng quan sát các điểm dữ liệu với nhiều hơn 3 trường bằng cách thêm các trục tọa độ vào hệ tọa độ song song tương ứng với các đặc trưng của dữ liệu. Tuy nhiên ở một vài khía cạnh, nó sẽ khó hiểu do số lượng các trục quá nhiều hoặc do số lượng điểm dữ liệu quá lớn. Nếu dữ liệu của bạn có hơn 10 đặc trưng, hệ tọa độ song song có thể không giúp ích được nhiều cho bạn.

Hệ tọa độ song song là tương đối hữu ích trong trường hợp các điểm dữ liệu có tính chất tương tự nhau, khi đó các polyline sẽ có xu hướng hội tụ nhau. Để dựng được một hệ tọa độ song song với Pandas và MatPlotLib, bạn phải chỉ định một đặc trưng để nhóm theo (có thể không phải là đặc trưng số). Điều này dẫn đến việc mỗi giá trị riêng biệt của một đặc trưng được vẽ bởi một màu duy nhất khi dựng biểu đồ. Qua đó, ta dễ dàng quan sát được các nhóm điểm dữ liệu có tính chất tương tự. Dưới đây là một ví dụ về hệ tọa độ song song sử dụng bộ dữ liệu hoa Iris của SciKit-Learn:

```
from sklearn.datasets import load_iris

from pandas.tools.plotting import parallel_coordinates

import pandas as pd

import matplotlib.pyplot as plt

import matplotlib

# Look pretty...

matplotlib.style.use('ggplot')

# If the above line throws an error, use plt.style.use('ggplot') instead


# Load up SKLearn's Iris Dataset into a Pandas Dataframe

data = load_iris()

df = pd.DataFrame(data.data, columns=data.feature_names)

df['target_names'] = [data.target_names[i] for i in data.target]


# Parallel Coordinates Start Here:

plt.figure()

parallel_coordinates(df, 'target_names')

plt.show()
```



Hình 8. Hệ tọa độ song song của dữ liệu hoa Iris

Giao diện hệ tọa độ song song của Pandas rất dễ để sử dụng, tuy nhiên bạn cũng cần phải cẩn trọng khi sử dụng nó. Bởi lẽ nó chỉ hỗ trợ một tỉ lệ cho tất cả các trục. Nếu bạn có một số đặc trưng có tỉ lệ nhỏ và một số khác lại có tỉ lệ lớn, thì bạn sẽ phải đối mặt với một hệ tọa độ rất khó nhìn. Để ứng phó với điều này, bạn có thể lựa chọn một số cách như sau trước khi lập địa đồ:

- Chuẩn hóa dữ liệu của bạn trước khi vẽ biểu đồ về cùng một phạm vi
- Thay đổ tỉ lệ ban đầu sang tỉ lệ log
- Tạo ra một số hệ tọa độ song song khác nhau. Mỗi hệ tọa độ này biểu thị cho một số đặc trưng có tỉ lệ miền giá trị tương tự nhau

Đường cong Andrews (Andrew's curve)

Một biểu đồ Andrews, hay còn được gọi là đường cong Andrews, giúp bạn hình ảnh hóa dữ liệu nhiều chiều, dữ liệu đa biến bằng thiết lập mỗi quan sát của tập dữ liệu thành một đường cong. Các giá trị đặc trưng của một quan sát làm thành các hệ số của đường cong, nên những đường cong có các đặc tính giống nhau có xu hướng nhóm lại gần nhau hơn. Do đó, đường cong Andrew có một vài ứng dụng trong việc phát hiện ngoại vi.

Cũng giống như hệ tọa độ song song, mọi đặc trưng được vẽ biểu đồ phải là đặc trưng số vì phương trình đường cong được lập theo triển khai Fourier. Cụ thể, giả sử ta có một giá trị *x* là một điểm dữ liệu *d* chiều thuộc **R***d*. Ta có thể biểu diễn dữ liệu nhiều chiều *x* dưới dạng *x = {x1, x2,…, xd}*. Để hình ảnh hóa chúng, biểu đồ Andrews định nghĩa một chuỗi Fourier hữu hạn:

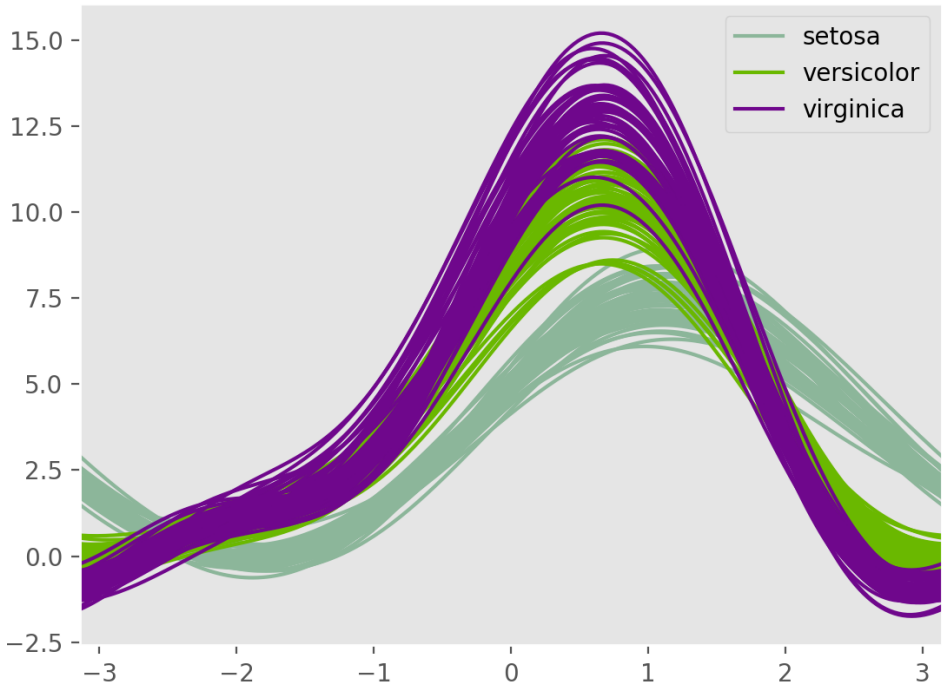
$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 sin(t) + x_3 cos(t) + x_4 sin(2t) + x_5 cos(2t) + \dots$$

Hàm này sau đó được vẽ đồ thị trong khoảng giá trị -π < t < π. Do vậy, mỗi điểm dữ liệu có thể được xem dưới dạng một đường cong trong khoảng (-π; π). Công thức này có thể được hiểu dưới dạng một phép chiều của điểm dữ liệu lên véc-tơ:

$$V = (\frac{1}{\sqrt{2}}, sin(t), cos(t), sin(2t), cos(2t), \dots)$$

Nếu có một cấu trúc nào đó tồn tại trong dữ liệu thì ta sẽ có thể nhìn thấy được qua đường cong Andrews của dữ liệu. Chúng ta tiếp tục lập biểu đồ trên bộ dữ liệu Iris để thấy rõ hơn nhận định này:

```
from sklearn.datasets import load_iris
from pandas.tools.plotting import andrews_curves

import pandas as pd
import matplotlib.pyplot as plt
import matplotlib

# Look pretty...
matplotlib.style.use('ggplot')
# If the above line throws an error, use plt.style.use('ggplot') instead

# Load up SKLearn's Iris Dataset into a Pandas Dataframe
data = load_iris()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target_names'] = [data.target_names[i] for i in data.target]

# Andrews Curves Start Here:
plt.figure()
andrews_curves(df, 'target_names')
plt.show()
```



Hình 9. Các đường cong Andrews lập thành từ bộ dữ liệu hoa Iris

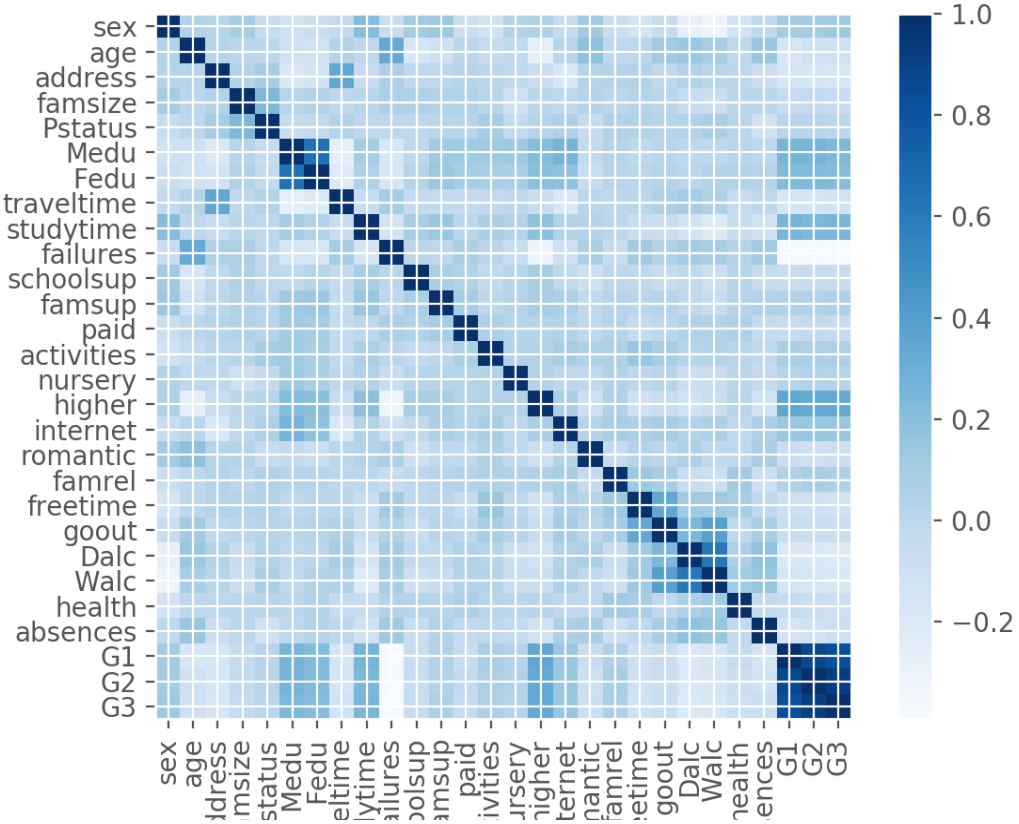
Quan sát các đường cong này ta thấy rằng loài hoa Iris-setosa có các đường cong khá tách rời với các loài hoa còn lại, điều này chứng tỏ dữ liệu về hoa Iris-setosa có những đặc điểm tương đối khác biệt so với hai loài hoa Iris-versicolor và Iris-virginica. Trong khi hai loài hoa này có những đặc điểm tương đối tương tự nhau khi ta nhìn vào đường cong Andrew của chúng. Do đó, nếu không lựa chọn được một mô hình tốt, việc nhầm lẫn giữa hoa Iris-versicolor và hoa Iris-virginica là điều rất dễ xảy ra.

Ma trận tương quan (Correlation Matrix)

Một ma trận tương quan là một bảng chứa các hệ số tương quan giữa các tập đặc trưng. Mỗi đặc trưng trong bảng được tính toán hệ số tương quan với các đặc trưng còn lại trong bảng. Ma trận này là ma trận đối xứng vì độ tương quan giữa đặc trưng *X* và đặc trưng *Y* đương nhiên bằng độ tương quan giữa đặc trưng *Y* và đặc trưng *X*. Ma trận này và các ma trận khác như ma trận hiệp phương sai là rất hữu ích trong việc kiểm tra xem mức độ tương quan giữa hai đặc trưng là như thế nào, và cũng để kiểm tra xem lượng thông tin mà mỗi đặc trưng cung cấp được.

Pandas tính các hệ số tương quan giữa các đặc trưng nằm trong khoảng -1 tới 1. Giá trị 1 thể hiện rằng hai đặc trưng có tính tương quan cùng chiều. Ngược lại, giá trị -1 thể hiện rằng hai đặc trưng có tính tương quan ngược chiều. Các giá trị gần giá trị 0 thể hiện hai đặc trưng có ít tương quan tới nhau. Điều này chỉ ra rằng khi hai giá trị có giá trị tuyệt đối của hệ số tương quan càng lớn thì chúng càng có tính tương quan tới nhau.

```
import pandas as pd

from matplotlib import pyplot as plt

plt.style.use('ggplot') # Look pretty

df = pd.read_csv("students.csv", index_col=0)

print(df.corr())
# code hình ảnh hóa ma trận tương quan
import matplotlib.pyplot as plt

plt.imshow(df.corr(), cmap=plt.cm.Blues, interpolation='nearest')
plt.colorbar()
tick_marks = [i for i in range(len(df.columns))]
plt.xticks(tick_marks, df.columns, rotation='vertical')
plt.yticks(tick_marks, df.columns)

plt.show()
```



Hình 10. Ma trận tương quan giữa các thuộc tính trong bộ dữ liệu học sinh

Kết quả chỉ ra rằng các đặc trưng G1, G2 và G3 có tính tương quan rất lớn với nhau (giống như khi ta đã phân tích ở địa đồ phân tán). Ngoài ra, ta cũng thấy hai đặc trưng là Medu và Fedu cũng khá tương quan tới nhau.

Tổng kết

Việc hiểu dữ liệu thực sự là nền tảng của khoa học dữ liệu. Chúng ta sẽ khó thành công nếu không hiểu được dữ liệu của mình. Bài viết này đã giới thiệu cho bạn cách thức để nhìn vào bộ dữ liệu của bạn bằng việc sử dụng một số kĩ thuật hình ảnh và phi hình ảnh dữ liệu như là các thông số thống kê, histogram, các địa đồ phân tán và các phương pháp hình ảnh hóa dữ liệu có số chiều lớn khác. Hi vọng qua bài viết này, bạn có thể hiểu được các bước và ý nghĩa của chúng trong việc phân tích khám phá dữ liệu.

Tài liệu tham khảo

1, Agresti, A. & Finlay, B., *Statistical Methods for the Social Sciences*, 3th Edition, Prentice Hall, 1997.

2, Howard J. Seltman, *Experimental Design and Analysis*, Chương 4, trang 61-98, 2018

3, Khóa học online *Analysis of Discrete Data,* Đại học PennState, Eberly College of Science

4, Khóa học online *Programming with Python for Data Science*, Edx.com