MACHINE LEARNING

Các kỹ thuật xây dựng đặc trưng

TS. Trần Quang Quý

Khái niệm

Trong học máy và thống kê, trích chọn đặc trưng (hay còn gọi bằng nhiều cụm từ như trích chọn đặc tính, lựa chọn đặc trưng, lựa chọn thuộc tính, chọn lựa thuộc tính, chọn lọc đặc trưng, tiếng Anh: *feature selection*, có thể là *variable selection*, *attribute selection* hay *variable subset selection*), là một quá trình chọn lọc một tập con chứa các thuộc tính liên quan để sử dụng trong quá trình xây dựng mô hình. Các kỹ thuật trích chọn đặc trưng được dùng cho một số lý do:

- Đơn giản hóa các mô hình để giúp các nhà nghiên cứu/người dùng diễn dịch dễ dàng hơn.
- Giảm thời gian huấn luyện.
- Tránh lời nguyền chiều (curse of dimensionality).
- Tăng cường tổng quát hóa bằng cách giảm sự quá khớp.

Curse of dimensionality

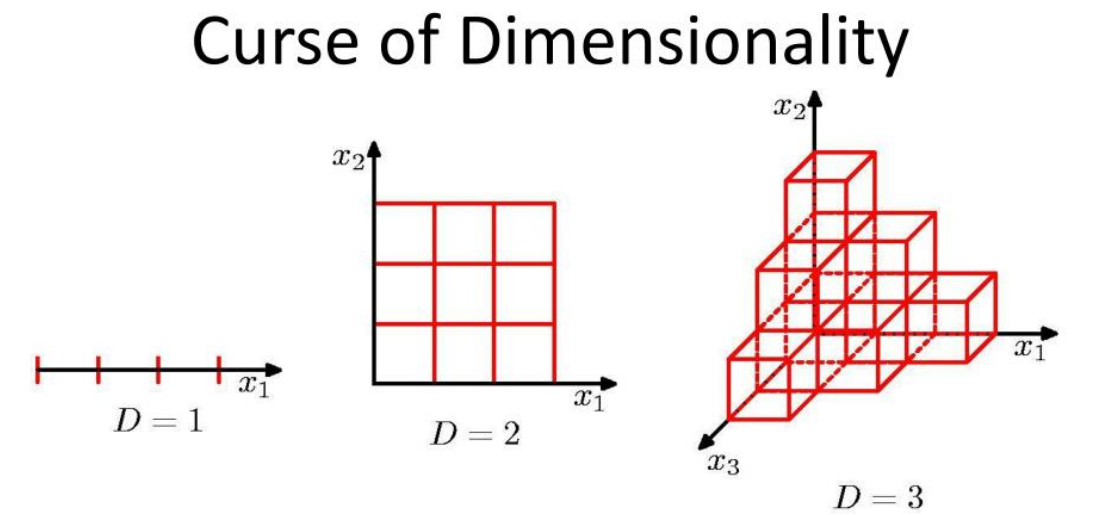

Figure 1: Lời nguyền chiều dữ liệu

Đặc trưng dữ liệu

Khái niệm

- Mỗi một điểm dữ liệu trong một mô hình học máy thường được biểu diễn bằng một véc tơ được gọi là *véc tơ đặc trưng.* Trong cùng một mô hình, các véc tơ đặc trưng của các điểm thường có kích thước khác nhau. Tuy nhiên dữ liệu trong thực tế thường ở dạng thô với kích thước khác nhau hoặc kích thước như nhau nhưng số chiều quá lớn gây trở ngại trong việc lưu trữ. Vì vậy, việc lựa chọn, tính toán đặc trưng phù hợp cho mỗi bài toán là một bước rất quan trọng.
- Quá trình trích trọn đặc trưng được hiểu là chỉnh sửa dữ liệu thô về dữ liệu dạng chuẩn, loại bỏ đi những dữ liệu nhiễu (noise). Dữ liệu chuẩn này phải đảm bảo giữ được những thông tin đặc trưng của dữ liệu thô ban đầu. Ngoài ra chúng ta còn phải thiết kế những phép biến đổi để có những đặc trưng phù hợp cho bài toán.

Mô hình chung các bài toán học máy

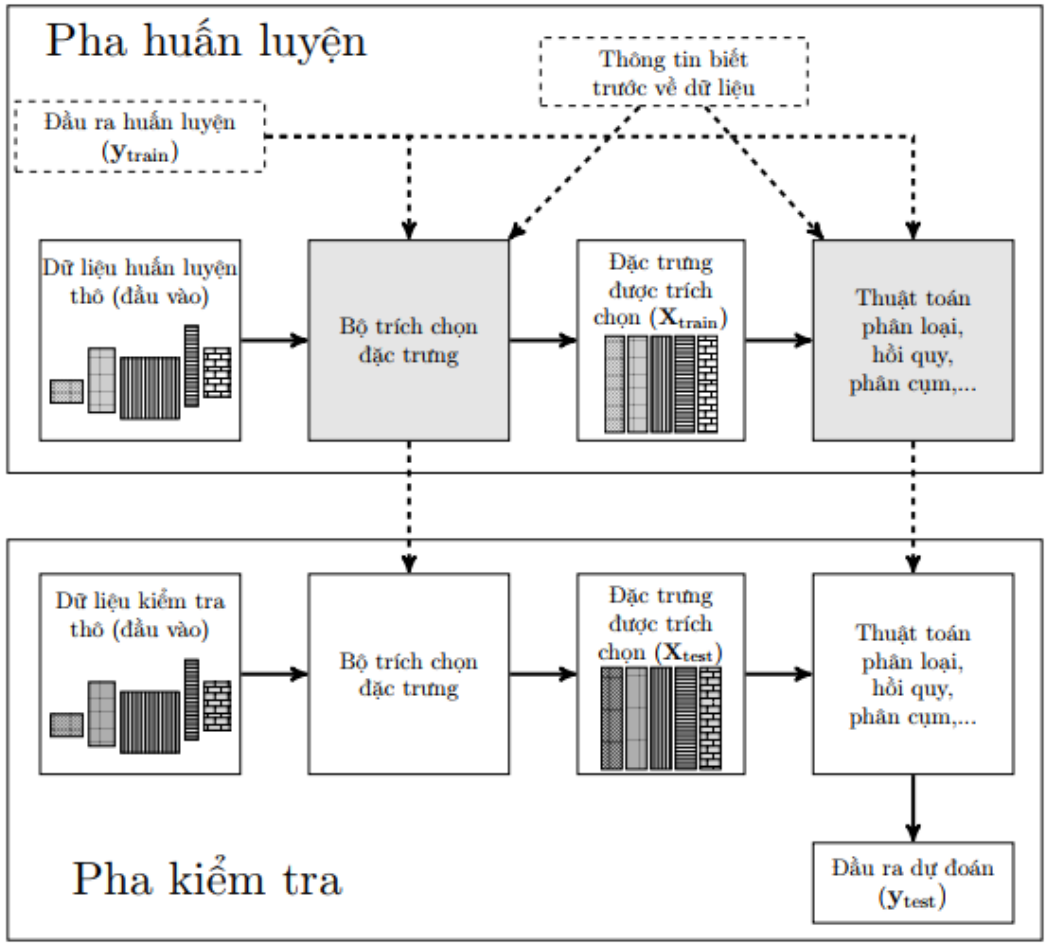

Figure 2: Mô hình chung các bài toán học máy

Mô hình chung các bài toán học máy

Phần lớn các mô hình học máy có thể được minh họa như hình trên. Có hai pha lớn trong mỗi bài toán học máy là:

- Pha huấn luyện
- Pha kiểm tra

Pha huấn luyện xây dựng mô hình dựa trên dữ liệu huấn luyện. Dữ liệu kiểm tra được sử dụng để đánh giá hiệu quả của mô hình.

Một số kỹ thuật trích chọn đặc trưng

1. Trực tiếp lấy dữ liệu (raw data)
2. Lựa chọn các biến số phù hợp (feature selection)
3. Giảm chiều dữ liệu (dimensionality reduction)
4. Túi từ (Bag-of-words)

Raw Data

Khái niệm

Với bài toán phân loại chữ số viết tay trong bộ cơ sở dữ liệu MNIST, mỗi bức ảnh có số chiều là 28 pixel x 28 pixel (tất nhiên việc *crop* và chỉnh sửa mỗi bức ảnh đã được thực hiện từ trước rồi, đó đã là một phần của feature engineering rồi). Một cách đơn giản thường được dùng là *kéo dài* ma trận 28x28 này để được 1 vector có số chiều 784. Trong cách này, các cột (hoặc hàng) của ma trận ảnh được đặt chồng lên (hoặc cạnh nhau) để được 1 vector dài. Vector dài này được trực tiếp sử dụng làm feature đưa vào các bộ classifier/clustering/regression/... Lúc này, giá trị của mỗi pixel ảnh được coi là một feature.

Feature selection

Khái niệm

Giả sử rằng các điểm dữ liệu có số features khác nhau (do kích thước dữ liệu khác nhau hay do một số feature mà điểm dữ liệu này có nhưng điểm dữ liệu kia lại không thu thập được), và số lượng features là cực lớn. Chúng ta cần *chọn* ra một số lượng nhỏ hơn các feature phù hợp với bài toán

Dimensionality reduction

Khái niệm

- Phương pháp đơn giản nhất trong giảm chiều dữ liệu là dùng phép chiếu ngẫu nhiên (random projection). Trong đó sử dụng một ma trận chiếu (projection matrix) d chiều, với $d < n$ (n là chiều của dữ liệu gốc), ma trận chiếu này nhân với từng điểm dữ liệu gốc để được dữ liệu có véc tơ có số chiều thấp hơn.
- Giả sử dữ liệu ban đầu là một vector $\mathbf{x} \in \mathbb{R}^D$, $\mathbf{A}$ là một ma trận trong $R^{d \times D}$ và $\mathbf{z} = \mathbf{A}\mathbf{x} \in \mathbb{R}^d$. Nếu $d < D$, ta thu được một vector với số chiều nhỏ hơn. Đây là kỹ thuật khá phổ biến trong giảm chiều dữ liệu.

Bag of words

Khái niệm

- Giả sử chúng ta có bài toán phân loại tin rác. Ta thấy rằng nếu một tin có chứa các từ *khuyến mại, giảm giá, trúng thưởng, miễn phí, quà tặng, tri ân, ...* thì nhiều khả năng đó là một tin nhắn rác. Vậy phương pháp đơn giản nhất là *đếm* xem trong tin đó có bao nhiêu từ thuộc vào các từ trên, nếu nhiều hơn 1 ngưỡng nào đó thì ta quyết định đó là tin rác.
- Khái niệm túi đựng từ chính là đếm số lần xuất hiện của các từ xuất hiện trong câu thông qua một dạng từ điển các từ thu thập được.

Ví dụ về túi đựng từ

Giả sử chúng ta có hai văn bản đơn giản:

```
(1) John likes to watch movies. Mary likes movies too.
```

và

```
(2) John also likes to watch football games.
```

Dựa trên hai văn bản này, ta có danh sách các từ được sử dụng, được gọi là *từ điển* với 10 *từ* như sau:

```
["John", "likes", "to", "watch", "movies", "also", "football", "games", "Mary", "too"]
```

Với mỗi văn bản, ta sẽ tạo ra một vector đặc trưng có số chiều bằng 10, mỗi phần tử đại diện cho số từ tương ứng xuất hiện trong văn bản đó. Với hai văn bản trên, ta sẽ có hai vector đặc trưng là:

```
(1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]
```

Figure 3: Túi đựng từ

Một vài lưu ý với túi đựng từ

- Với những ứng dụng thực tế, *từ điển* có nhiều hơn 10 từ rất nhiều, có thể đến một trăm nghìn hoặc cả triệu, như vậy vector đặc trưng thu được sẽ rất *dài*. Một văn bản chỉ có 1 câu, và 1 tiểu thuyết nghìn trang đều được biểu diễn bằng các vector có số chiều bằng 100 nghìn hoặc 1 triệu.
- Có rất nhiều từ trong từ điển không xuất hiện trong một văn bản. Như vậy các vector đặc trưng thu được thường có rất nhiều phần tử bằng 0. Các vector có nhiều phần tử bằng 0 được gọi là *sparse vector* (véc tơ thưa).
- Nhược điểm lớn nhất của BoW là nó không mang thông tin về thứ tự của các từ. Cũng như sự liên kết giữa các câu, các đoạn văn trong văn bản. Ví dụ, ba câu sau đây: "*Em yêu anh không?*", "*Em không yêu anh*", và "*Không, (nhưng) anh yêu em*" khi được trích chọn đặc trưng bằng BoW sẽ cho ra ba vector giống hệt nhau, mặc dù ý nghĩa khác hẳn nhau.

Ví dụ về BoW

1. Sử dụng Véc tơ histogram của ảnh trong Computer Vision
2. Xét cửa sổ ảnh (patch)

Cụ thể hơn có thể xem tại đây: BoW and Computer Vision

Chuẩn hóa véc tơ đặc trưng

Các điểm dữ liệu đôi khi được đo đạc với những đơn vị khác nhau, m và feet chẳng hạn. Hoặc có hai thành phần (của vector dữ liệu) chênh lệch nhau quá lớn, một thành phần có khoảng giá trị từ 0 đến 1000, thành phần kia chỉ có khoảng giá trị từ 0 đến 1 chẳng hạn. Lúc này, chúng ta cần chuẩn hóa dữ liệu trước khi thực hiện các bước tiếp theo.

Một vài phương pháp chuẩn hóa thường dùng:

1. Rescaling (Thay đổi lại kích thước)
2. Standardization (Chuẩn hóa giá trị)
3. Scaling to unit length (Chia tỷ lệ theo chiều dài đơn vị)

Rescaling

Khái niệm

Phương pháp đơn giản nhất là đưa tất cả các thành phần về cùng một khoảng, [0,1] hoặc [-1,1]. Nếu muốn đưa một thành phần về khoảng [0,1], công thức sẽ là:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

trong đó x là giá trị ban đầu, x' là giá trị sau khi chuẩn hóa, $min(x), max(x)$ được tính trên toàn bộ dữ liệu training data ở cùng một thành phần. Việc này được thực hiện trên từng thành phần của véc tơ dữ liệu $\mathbb{x}$.

Standardization

Khái niệm

Một phương pháp nữa cũng hay được sử dụng là giả sử mỗi thành phần đều có phân phối chuẩn với kỳ vọng là 0 và phương sai là 1. Khi đó, công thức chuẩn hóa sẽ là:

$$x' = \frac{x - \bar{x}}{\sigma}$$

với $\bar{x}, \sigma$ lần lượt là kỳ vọng và phương sai của thành phần đó trên toàn bộ training data.

Scaling to unit length

Khái niệm

Một lựa chọn khác nữa cũng được sử dụng rộng rãi là chuẩn hóa các thành phần của mỗi vector dữ liệu sao cho toàn bộ vector có độ lớn (Euclid, tức norm 2) bằng 1. Việc này có thể được thực hiện bằng:

$$\mathbf{x}' = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$$

Lý thuyết về chuẩn (norm)

Khái niệm

- Trong không gian một chiều, việc đo khoảng cách giữa hai điểm đã rất quen thuộc: lấy trị tuyệt đối của hiệu giữa hai giá trị đó. Trong không gian hai chiều, tức mặt phẳng, chúng ta thường dùng khoảng cách Euclid để đo khoảng cách giữa hai điểm. Khoảng cách này chính là cái chúng ta thường nói bằng ngôn ngữ thông thường là đường chim bay. Đôi khi, để đi từ một điểm này tới một điểm kia, con người chúng ta không thể đi bằng đường chim bay được mà còn phụ thuộc vào việc đường đi nối giữa hai điểm có dạng như thế nào nữa.
- Việc đo khoảng cách giữa hai điểm dữ liệu nhiều chiều, tức hai vector, là rất cần thiết trong Machine Learning. Chúng ta cần đánh giá xem điểm nào là điểm gần nhất của một điểm khác; chúng ta cũng cần đánh giá xem độ chính xác của việc ước lượng; và trong rất nhiều ví dụ khác nữa.

Một số chuẩn thường dùng

Giả sử các vectors $\mathbf{x} = \lfloor x_1; x_2; ...; x_n \rfloor, \mathbf{y} = \lfloor y_1; y_2; ...; y_n \rfloor$. Nhận thấy rằng khoảng cách Euclid chính là một norm, norm này thường được gọi là norm 2:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + ... x_n^2}$$

Với p là một số không nhỏ hơn 1 bất kỳ, hàm số sau đây:

$$\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + ... |x_n|^p\right)^{\frac{1}{p}}$$

được chứng minh thỏa mãn ba điều kiện bên trên, và được gọi là norm p.

So sánh norm 1 và norm 2

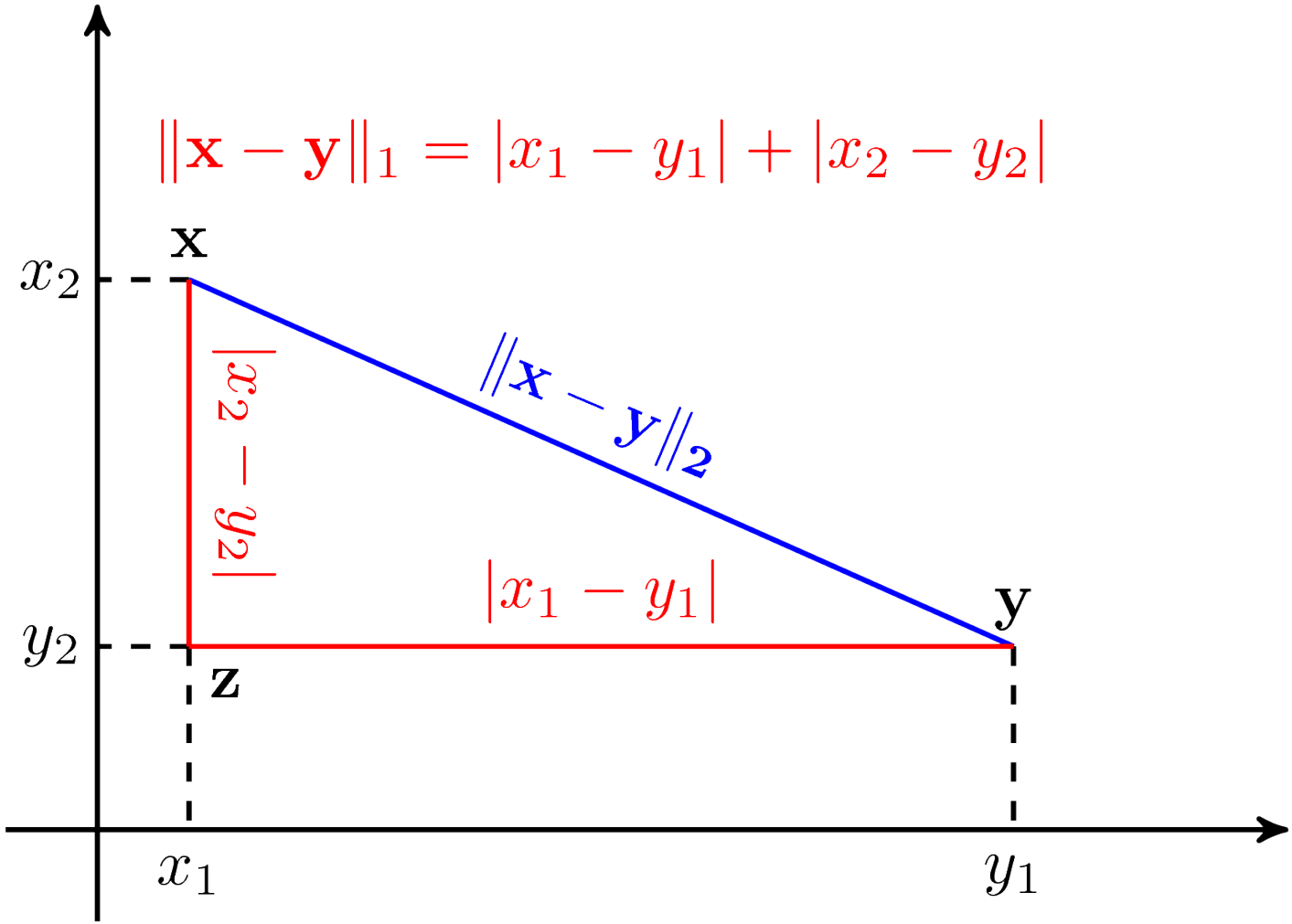

Figure 4: So sánh norm 1 và norm 2 trong không gian 2 chiều

Câu hỏi ôn tập và bài tập

Câu hỏi

1. Trích chọn đặc trưng là gì? Khi nào cần sử dụng trích chọn đặc trưng
2. Trình bày các pha chính khi xây dựng mô hình học máy
3. Trình bày một số kỹ thuật trích chọn đặc trưng phổ biến hiện nay?
4. Trình bày ứng dụng của túi đựng từ (BoW) trong thị giác máy tính (computer vision)

Bài tập thực hành

1. Cài đặt Anaconda và khởi chạy Jupyter Notebook
2. Download tập dữ liệu iris.csv
3. Thực hành bài tập sau trên Jupyter Notebook, lưu file làm trên máy. Link thực hành: Bấm vào đây để làm bài.