Tìm chân lý từ dữ liệu thô: suy diễn chủ đề và hình học [1]

Ngày nay chúng ta quen dần với sự hiện diện của các thuật toán học máy thông minh. Chúng tự động rà quét qua các nguồn thông tin khổng lồ để truy tìm những nội dung ẩn sau rừng dữ liệu thô sơ và hỗn độn. Một số thuật toán lấy ở đầu vào hàng vạn, triệu trang văn bản tin tức trên mạng, ở đầu ra chúng cho ta biết những bài viết về giáo dục, thể thao, hay nghệ thuật. Hoặc những chủ đề tinh hơn, về thi cử, về doping trong thể thao, về một phong cách mới nổi của nghệ thuật đương đại. Cũng những thuật toán này, khi được áp vào bộ mã di truyền của các cá thể trong một và nhiều quần thể loài người, đã góp phần giúp các nhà khoa học kiến tạo những giả thuyết lý thú về nguồn gốc và sự pha trộn tổ tiên của loài người trong lịch sử xa xôi.

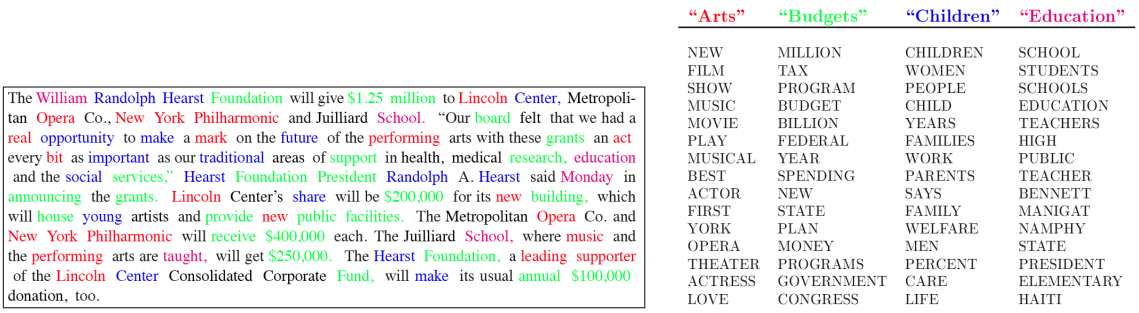
Những thuật toán kể trên dựa vào một nền tảng chung là một lớp mô hình thống kê, gọi là mô hình chủ đề (*topic models*). Đây hiện là một công cụ phổ biến và hữu hiệu trong việc tự động tìm tòi nội dung các văn bản, không chỉ văn bản mà cả nhiều dạng dữ liệu khác, lấy từ kho những bức ảnh chụp, thư viện điện tử, bộ mã di truyền, mạng xã hội. Mô hình thống kê là một công cụ toán học đặc biệt để mô tả motif của dữ liệu: từ đó cho ta các phương pháp tìm tòi suy diễn về các *motif*, hay *pattern* ẩn sau các tần suất về thống kê của dữ liệu thô.

Trong bài viết này tôi sẽ giới thiệu và tập trung vào một vài khía cạnh toán học căn bản của lớp mô hình chủ đề và các phương pháp thống kê kéo theo. Đặc biệt, có một sự liên hệ mật thiết giữa mô hình chủ đề và hình học lồi, giúp cho ta hiểu rõ hơn bản chất của thuật toán học máy đề cập ở trên. Nó cũng giúp ta tạo ra các thuật toán suy diễn mới, hiệu quả hơn.

1 Mô hình chủ đề

Mô hình *Phân bổ ẩn Dirichlet* (tiếng Anh: *Latent Dirichlet Allocation* (LDA)) được David Blei, Andrew Ng và Michael Jordan thuộc Đại học California, Berkeley giới thiệu vào năm 2001 (Blei et al, 2003). Nó đóng một vai trò đặc biệt quan trọng trong ngành học máy và trí tuệ nhân tạo. Một mô hình tương tự, về mặt toán học, được phát kiến độc lập bởi một nhóm tác giả khác thuộc ĐH Oxford, Jonathan Pritchard, Matthew Stephens và Peter Donnelly, cũng có một vị trí quan trọng trong ngành di truyền quần thể (Pritchard et al, 2000). Hai bài báo này gộp lại hiện có trên 30,000 trích dẫn trên Google Scholar, cho thấy sự ảnh hưởng đáng kinh ngạc của

[1] Tác giả: Nguyễn Xuân Long, Đại học Michigan, USA. Bài báo đã được đăng trên tạp chí Pi của hội Toán học Việt Nam, số tháng 5, 2017.



Hình 1: Hình bên trái là ví dụ của một trong nhiều văn bản đầu vào của thuật toán. Hình bên phải là đầu ra của thuật toán, cho biết những chủ đề chính của văn bản ở bên trái. Mỗi chủ đề được minh hoạ theo từng cột bằng các từ vựng có tần suất xuất hiện cao (nguồn: Blei et al (2003)).

chúng trong những địa hạt hoàn toàn cách biệt, từ khoa học đến công nghệ. Chỉ với kiến thức toán sơ cấp, chúng ta có thể mô tả kỹ thuật tiên tiến này một cách khá dễ dàng.

Trước khi mô tả LDA, cần có đôi lời về sự cần thiết của mô hình xác suất trong suy diễn thống kê. Lấy $\boldsymbol{x}$ làm ký hiệu cho đầu vào của một thuật toán, còn $\boldsymbol{\theta}$ là đầu ra. Một cách cụ thể $\boldsymbol{x}$ là biểu diễn của tập các văn bản, còn $\boldsymbol{\theta}$ là biểu diễn cho các chủ đề chúng ta muốn chắt lọc từ nội dung của dữ liệu $\boldsymbol{x}$. $\boldsymbol{x}$ lấy giá trị ở trong một không gian mà chúng ta sẽ định nghĩa một cách hình thức. Tương tự như vậy với $\boldsymbol{\theta}$. Để có thể suy diễn được chủ đề $\boldsymbol{\theta}$ từ dữ liệu thô $\boldsymbol{x}$, chúng ta cần phải thiết lập được một sự liên hệ giữa $\boldsymbol{x}$ và $\boldsymbol{\theta}$. Sự liên hệ ấy sẽ được mô tả bằng một phân bố xác suất, ký hiệu bởi $p(x|\boldsymbol{\theta})$, mà trong đó $\boldsymbol{\theta}$ đóng vai trò là một tham số của phân phối cho biến ngẫu nhiên $\boldsymbol{x}$.

Trong thảo luận của chúng ta về sự thiết lập của mô hình, dữ liệu được xem là hiện sinh (realization) của một biến ngẫu nhiên $\boldsymbol{x}$, còn chủ đề $\boldsymbol{\theta}$ là một tham số cho phân phối của dữ liệu. Việc sử dụng ngôn ngữ xác suất để xác lập mô hình toán học cho dữ liệu được xem là hết sức tự nhiên, vì dữ liệu thường có yếu tố *bất định*: mỗi lần thu thập dữ liệu mới ta thường nhận một giá trị khác nhau, cho dù quy luật sinh ra dữ liệu có thể được xác định bởi cùng một phân bố "chân lý". Dẫu ta không bao giờ biết chắc chắn được phân bố chân lý, bằng những hiểu biết từ dữ liệu thực tế chúng ta có thể tìm một lớp phân bố xấp xỉ khả dĩ cho phân bố chân lý. Lớp phân bố ấy sẽ được tham số hoá bởi $\boldsymbol{\theta}$. Bằng dữ liệu thực, chúng ta sẽ tìm giá trị tốt nhất có thể cho $\boldsymbol{\theta}$.

Mô hình sinh $p(\boldsymbol{x}|\boldsymbol{\theta})$ Bây giờ chúng ta đã sẵn sàng với việc thiết lập mô hình chủ đề một cách chặt chẽ, còn gọi là *mô hình sinh* cho dữ liệu. Đó chẳng qua là phân bố xác suất $p(x|\boldsymbol{\theta})$. Giả sử V là bộ từ điển gồm có d phần tử (từ vựng) được đánh số bằng $V = \{1,\ldots,d\}$. Mỗi văn bản được xem là một hiện sinh (realization) của biến ngẫu nhiên $\boldsymbol{x}$. Không mất tính tổng quát, giả dụ văn bản dài n từ, ta viết $\boldsymbol{x} = (x_1,\ldots,x_n) \in V^n$.

Giả sử rằng có k chủ đề ẩn khác nhau có thể dùng để mô tả nội dung tập các văn bản ta có trong tay. k chủ đề được biểu diễn bằng $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)$, tương ứng với k phần tử nằm trong đơn hình xác suất

$$\Delta^{d-1} := \{u \in [0,1]^d | u_1 + \ldots + u_d = 1\}.$$

Mỗi chủ đề $\theta_i \in \Delta^{d-1}$ tương ứng với một tần suất nhất định của các từ vựng trong bộ từ điển. Ví dụ: một chủ đề về "giáo dục" sẽ có nhiều từ với tần suất xuất hiện cao, như "trường", "lớp", "học sinh", "học phí", v.v. Một chủ đề về "nghệ thuật" sẽ có những từ khác với tần suất lớn, như "nghệ sĩ", "mới", "biểu diễn", "tình yêu", v.v. (xem minh hoạ ở Hình 1). Thật khó tưởng tượng một văn bản chỉ đề cập đến một chủ đề duy nhất. Thực tế hơn, mỗi văn bản bao gồm sự pha trộn

của nhiều chủ đề cùng một lúc, dẫu mức độ pha trộn có thể nhiều ít khác nhau. (Còn ở ứng dụng trong sinh học phân tử, mỗi cá thể trong một cộng đồng có thể là kết quả của sự pha trộn các nguồn gốc khác nhau về mã di truyền, từ châu Phi, Á hay Âu).

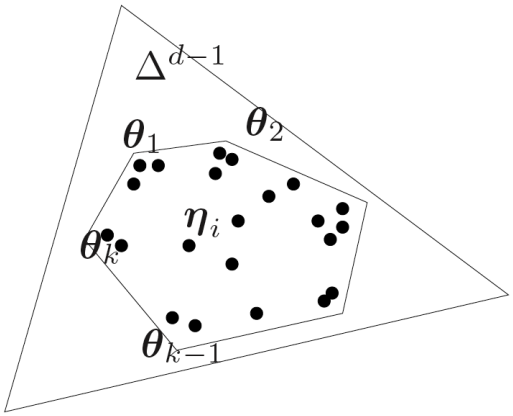
Trực quan trên gợi cho ta một khái niệm hình học rất tự nhiên: điểm ngẫu nhiên nằm trong *tập bao lồi* của các chủ đề: $G := \text{conv}\{\theta_1, \ldots, \theta_k\} \subset \Delta^{d-1}$. Điểm ngẫu nhiên này, với ký hiệu $\boldsymbol{\eta}$, được xác lập bởi công thức

$$\boldsymbol{\eta} = \beta_1\theta_1 + \ldots + \beta_k\theta_k,$$

trong đó $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ là một biến ngẫu nhiên lấy giá trị trong đơn hình xác suất Δ^{k-1}. Chúng ta sẽ giả dụ rằng $\boldsymbol{\beta}$ tuân theo phân phối Dirichlet trong Δ^{k-1}, phân bố này có hàm mật độ như sau, cho mỗi $\boldsymbol{\beta} \in \Delta^{k-1}$,

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \frac{\Gamma(\sum_{j=1}^k \gamma_j)}{\prod_{j=1}^k \Gamma(\gamma_j)} \prod_{j=1}^k \beta_j^{\gamma_j - 1}.$$

Trong công thức trên đây, Γ biểu thị hàm số Gamma, còn $\gamma_1, \ldots, \gamma_k$ là những tham số của mật độ Dirichlet. Định nghĩa trên đây cho (gián tiếp) một phân phối đối với biến ngẫu nhiên $\boldsymbol{\eta}$ lấy giá trị ở trong đa diện lồi G. Ta có thể viết ra một cách tường minh phân phối này, ký hiệu bởi $\mathrm{d}P(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\gamma})$, bằng công thức chuyển biến (điều này không cần thiết cho bài viết).



Để ý rằng tham số $\gamma_1, \ldots, \gamma_k$ điều khiển mật độ của biến ngẫu nhiên $\boldsymbol{\eta} \in G$. Ví dụ, nếu như $\gamma_1 = \ldots = \gamma_k = 1$, $p(\cdot|\boldsymbol{\gamma})$ trở thành hằng số. Do đó khi $k < d$, nếu $\theta_1, \ldots, \theta_k$ là k đỉnh của đa diện lồi G, biến ngẫu nhiên $\boldsymbol{\eta}$ được định nghĩa như ở trên sẽ tuân theo phân bố *đồng nhất* trong lòng đa điện G. Nếu $\gamma_1 = \ldots = \gamma_k \ll 1$, phân phối của $\boldsymbol{\eta}$ tập trung phần lớn khối (mass) ở sát các mặt biên của G. Ngược lại nếu $\gamma_1 = \ldots = \gamma_k \gg 1$, phần lớn khối tập trung ở sâu bên trong lòng đa diện G.

Chúng ta đã gần xong việc định nghĩa cho phân bố của dữ liệu $\boldsymbol{x}$. Mỗi văn bản sẽ có tương ứng một tần suất $\boldsymbol{\eta}$ kể trên. Khi biết $\boldsymbol{\eta}$, văn bản $\boldsymbol{x}$ được coi là một dãy các từ vựng trong V. Như đã định ở trên, văn bản n từ được viết thành $\boldsymbol{x} = (x_1, \ldots, x_n) \in V^n$. Mỗi x_i là một biến ngẫu nhiên độc lập tuân theo phân phối phân loại: tung một con xúc xắc d mặt sao cho xác suất của x_i lấy giá trị mặt j sẽ là η_j, $j = 1, \ldots, d$. Công thức hàm khối xác suất cho $\boldsymbol{x}$, khi biết $\boldsymbol{\eta}$:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \prod_{i=1}^n \prod_{j=1}^d \eta_i^{1(x_i=j)}. \qquad (1)$$

Ở trên, $1(A) := 1$ nếu biểu hiện A đúng, và 0 nếu A sai.

Tóm lại, nếu đã biết giá trị các tham số $\theta_1, \ldots, \theta_k$, cũng như $\gamma_1, \ldots, \gamma_k$ được cho trước, mô hình sinh dữ liệu được định nghĩa bởi hàm mật độ sau đây

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \int p(\boldsymbol{x}|\boldsymbol{\eta})\, \mathrm{d}P(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\gamma}). \qquad (2)$$

Ở công thức trên đây, $\mathrm{d}P(\boldsymbol{\eta}|\boldsymbol{\theta}, \boldsymbol{\gamma})$ là ký hiệu cho phân bố của biến ngẫu nhiên $\boldsymbol{\eta}$.

2 Suy diễn thống kê đối với chủ đề và hình học lồi

Trên đây chúng ta đã định nghĩa mô hình Phân bổ ẩn Dirichlet thông qua phân bố xác suất $p(\boldsymbol{x}|\boldsymbol{\theta})$. Giả sử rằng chúng ta có trong tay m văn bản, chúng sẽ được xem là m hiện

sinh độc lập của biến ngẫu nhiên $\boldsymbol{x}$, và được ký hiệu bởi $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$. Mỗi văn bản có đúng n từ. Như vậy, lượng dữ liệu thô là mn từ trong m văn bản. Bài toán thống kê là phải tìm ra các chủ đề $\boldsymbol{\theta}$ được dùng để xác định phân bố sinh ra tập m văn bản.

Bài toán 1. *Cho trước số chủ đề k và các tham số $\gamma_1, \ldots, \gamma_k > 0$. Với các văn bản $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ là m hiện sinh độc lập của phân phối $p(\boldsymbol{x}|\boldsymbol{\theta})$, hãy tìm các chủ đề $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.*

Chúng ta không thể tìm ra chính xác các chủ đề $\boldsymbol{\theta}$, chỉ có thể ước lượng chúng. Nhưng ta kỳ vọng rằng nếu số lượng dữ liệu m, n càng lớn thì các ước lượng đối với $\boldsymbol{\theta}$ càng tiến gần đến giá trị "chân lý" của chúng. Mặt khác, đây cũng không phải là vấn đề đơn giản về mặt tính toán, mà phải cần những hỗ trợ về thuật toán đủ mạnh: đó là các thuật toán "học máy" lấy đầu vào là tập m văn bản, và đầu ra chính là các ước lượng về "chủ đề" như tôi đưa ra ở đầu bài viết.

Để hiểu bản chất các thuật toán đó, cần quay về một bài toán dễ hơn trong hình học lồi.

Ước lượng các đỉnh của một đa diện Có một bài toán kinh điển trong hình học lồi: làm thế nào để ước lượng được các đỉnh của một đa diện lồi G nếu như chúng ta quan sát được m hiện sinh độc lập rút ra theo phân bố đồng nhất từ trong lòng của G. Bài toán sau đây khái quát hơn một chút, ở chỗ sử dụng một phân bố Dirichlet với một vector tham số $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ đã được cho trước.

Bài toán 2. *Cho trước $\boldsymbol{\gamma} \in \mathbb{R}^k_+$. Giả sử $\eta^{(1)}, \ldots, \eta^{(m)}$ là m hiện sinh độc lập của một biến ngẫu nhiên $\boldsymbol{\eta}$ lấy giá trị trong G như đã định nghĩa ở mục trước. Hãy đưa ra một ước lượng tốt nhất cho đa diện G.*

"Tốt nhất" được đo bằng một metric cụ thể, như Hausdorff: với hai đa diện G và G' bất kỳ trong một không gian chung $\mathbb{R}^d$

$$d_H(G, G') = \min\{\epsilon \geqslant 0 | G \subset G'_\epsilon, G' \subset G_\epsilon\}$$

trong đó $G_\epsilon = \{\theta + \nu | \theta \in G, \nu \in \mathbb{R}^d, \|\nu\|_2 \leqslant \epsilon\}$.

Có một ước lượng rất tự nhiên và kinh điển: lấy hình bao lồi của các hiện sinh $\eta^{(1)}, \ldots, \eta^{(m)}$,

$$\hat{G} := \text{conv}(\eta^{(1)}, \ldots, \eta^{(m)}).$$

Bỏ qua chi tiết $\hat{G}$ không nhất thiết có k đỉnh, ước lượng này có một đảm bảo cần thiết: có thể chứng minh được rằng, nếu $m \to \infty$, $d_H(G, \hat{G}) \to 0$ trong xác suất. Kết quả này là một dạng luật số lớn trong hình học lồi. Ngoài ra còn có một số kết quả tinh hơn về tốc độ hội tụ, và về phân bố tiệm cận (Reitzner, 2005; Bárány & Vu, 2007).

Tuy khá giống nhau, bài toán suy diễn về chủ đề (Bài toán 1) của chúng ta phức tạp hơn Bài toán 2 ở một số khía cạnh. Thứ nhất, ở Bài toán 1 ta không có trong tay các hiện sinh của vector tần suất $\eta^{(1)}, \ldots, \eta^{(m)}$, thay vào đó là các hiện sinh $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ ($\boldsymbol{x}$ liên hệ với η qua phương trình (1)). Thứ hai, vector tham số γ có thể bao gồm các giá trị khác với 1, và khác nhau (để phản ánh xác thực hơn dữ liệu thực tế). Nói cách khác, phân phối trong lòng đa diện lồi G thường không bao giờ là đồng nhất, tuy đó là một giả thuyết khá phổ biến trong hình học lồi cổ điển.

Ước lượng bằng cực đại của hàm khả dĩ Trong thống kê toán có một phương pháp ước lượng rất mạnh và hữu dụng, dựa vào phép lấy cực đại của hàm khả dĩ của Fisher. Hàm *khả dĩ* (likelihood function) là một hàm số đối với tham số $\boldsymbol{\theta}$ của (mật độ) phân bố $p(\boldsymbol{x}|\boldsymbol{\theta})$. Ước lượng cực đại của hàm khả dĩ, viết tắt bằng MLE, khi đã được cho các dữ liệu $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ được định nghĩa như sau:

$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}) \qquad (3)$$

trong đó hàm khả dĩ $p(\boldsymbol{x}^{(i)}|\cdot)$ được cho bởi công thức (2).

Định lý sau cho một bảo đảm quan trọng đối với ước lượng MLE cho lớp mô hình chủ đề:

Định lý 2.1. *Giả sử G có k đỉnh $\theta_1, \ldots, \theta_k$, và $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ là ước lượng MLE định nghĩa bởi* (3). *Viết $\hat{G} := conv(\hat{\theta}_1, \ldots, \hat{\theta}_k)$. Nếu m và n cùng tiến tới vô cùng, trong đó $\log\log m \leqslant \log n = o(m)$, ta có*

$$d_H(\hat{G}, G) = O_P\left(\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}\right)^{1/(2(q+\alpha))}.$$

Trên đây, $q = \min\{(k-1), (d-1)\}$, α là một hằng số dương phụ thuộc vào $\boldsymbol{\gamma}$, O_P có nghĩa là "hằng số cận trên trong xác suất", được xác định đối với phân phối $p(\boldsymbol{x}|\boldsymbol{\theta})$.

Với độc giả chưa có đủ kiến thức về lý thuyết xác suất để hiểu kết quả trên một cách cặn kẽ, chỉ cần biết rằng ước lượng MLE (3) càng tiến gần tới chủ đề chân lý khi lượng dữ liệu càng nhiều. Định lý cho ta biết chặn trên của sai số của ước lượng MLE, khi m và n tăng dần. Hãy tham khảo Nguyen (2015) để biết thêm chi tiết cho cả chặn trên và chặn dưới.

Bayes hay không Bayes Chúng ta tạm dừng một chốc lát, để quay lại với thảo luận về sự cần thiết của mô hình xác suất $p(\boldsymbol{x}|\boldsymbol{\theta})$, một chất keo toán học kết nối dữ liệu quan sát được $\boldsymbol{x}$, với khái niệm "chân lý" ẩn sau đó, $\boldsymbol{\theta}$. Các nhà thống kê đều đồng ý với nhau rằng dữ liệu phải được xem là hiện sinh của một biến ngẫu nhiên, do sự bất định của quan sát. Nhưng có một sự tranh cãi quyết liệt về ý nghĩa của khái niệm $\boldsymbol{\theta}$: nó có bất định hay không? Có hai trường phái đối với câu hỏi này: Các nhà *thống kê tần suất* cho rằng $\boldsymbol{\theta}$ hoàn toàn có thể xác định được. Phương pháp ước lượng cực đại của hàm khả dĩ (MLE) chính là một sản phẩm của các nhà tần suất. Ngược lại, trường phái *Bayes* cho rằng tri thức về $\boldsymbol{\theta}$ cũng bất định, và do đó về mặt toán học, nó cũng phải được biểu diễn bởi một biến ngẫu nhiên. Vì chúng ta không thể quan sát được khái niệm trừu tượng $\boldsymbol{\theta}$, chúng ta không thể xác quyết đúng sai trong tranh cãi này.

Một ưu thế của phương pháp suy diễn Bayes là nó không chỉ cung cấp cho ta một điểm ước lượng cho $\boldsymbol{\theta}$, nó còn cho ta biết mức độ (không) chắc chắn của ước lượng đó, dựa trên những dữ liệu ta thu thập được. Để áp dụng phương pháp Bayes, tham số $\boldsymbol{\theta}$, bởi được xem là biến ngẫu nhiên, phải được gán cho một phân bố trước khi có dữ liệu. Phân phối này được gọi là *phân bố tiên nghiệm*, ký hiệu bởi $\Pi(\boldsymbol{\theta})$, nó là chắt lọc của tri thức mà nhà thống kê có được trước khi anh ta quan sát. Sau khi đã thu thập dữ liệu, m văn bản $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$, kiến thức của nhà thống kê về $\boldsymbol{\theta}$ được tổng kết thành *phân bố hậu nghiệm*. Phân bố hậu nghiệm được xác định bằng công thức Bayes nổi tiếng:

$$\mathrm{d}\Pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\Pi(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\Pi(\boldsymbol{\theta})}. \qquad (4)$$

Lưu ý, cả phân bố tiên nghiệm (a priori) và hậu nghiệm (a posteriori) thường được ký hiệu bởi chữ Π viết hoa, để phân biệt với hằng số π thân thuộc với độc giả tạp chí này. Có thể chứng minh được rằng khi số lượng dữ liệu càng nhiều thì phân bố hậu nghiệm tập trung càng gần về giá trị "chân lý" của $\boldsymbol{\theta}$ (Nguyen, 2015): niềm tin của các nhà thống kê Bayes cũng tiệm cận dần tới chân lý!

3 Thuật toán học chủ đề

Thống kê hiện đại và học máy tập trung đáng kể vào mục tiêu tính toán hiệu quả những đại lượng trọng yếu như ước lượng MLE (3) hay phân bố hậu nghiệm (4). Đây là một thách thức vì những phép tính này nằm trong lớp NP-hard hoặc khó hơn. Trên thực tế, để có độ tin cậy cao các thuật toán phải quét qua một số lớn các văn bản có độ dài khác nhau, con số đó có thể đến hàng vạn, hàng triệu, thậm chí hàng tỷ, như Định lý 2.1 gợi ý. Khi đó, thời gian chạy thuật toán có thể mất hàng ngày hoặc hàng tuần để có được kết quả tương đối chính xác của bài toán tìm cực đại cho hàm khả dĩ.

Vì thế, các nhà thống kê và học máy thực hành chỉ cần tìm các giải pháp xấp xỉ, như phép suy diễn biến phân (variational inference) (Blei et al, 2003). Đối với phương

pháp suy diễn Bayes, phân bố hậu nghiệm cũng được xấp xỉ bằng cách thiết kế một thuật toán mô phỏng chuỗi Markov với thuộc tính ergodic, mà theo đó phân bố dừng của nó chính là phân bố hậu nghiệm cần phải tính. Có một thực trạng khó chịu: phép biến phân xấp xỉ cho ra kết quả rất nhanh, nhưng lại không trúng. Ngược lại, thuật toán mô phỏng chuỗi Markov đưa ra được ước lượng khá trúng, nhưng lại không nhanh.

Sự liên hệ giữa mô hình chủ đề và hình học lồi được mô tả trên đây đã được tận dụng để đưa ra một thuật toán ước lượng chủ đề vừa chính xác, vừa nhanh, trong một số điều kiện nhất định. Vì giới hạn bài vở, tôi chỉ mô tả một cách nôm na ý tưởng này như sau: Giả sử chúng ta biết rằng phần lớn các văn bản tập trung ở gần các đỉnh của đa diện G (nếu như tham số $\gamma_1, \ldots, \gamma_k \ll 1$). Ta có thể áp dụng một thuật toán phân cụm đơn giản để tìm ra trọng tâm của k cụm các văn bản này. Các trọng tâm này cho ta một ước lượng thô khả dĩ về vị trí các đỉnh của G. Để có ước lượng tốt hơn ta cần một số phép "điều chỉnh hình học". Độc giả tò mò có thể tham khảo thêm Yurochkin và Nguyen (2016).

4 Kết luận

Khoa học dữ liệu phát triển mạnh mẽ như hôm nay là nhờ đóng góp về vật chất của công nghệ thông tin, bên cạnh vai trò mang tính nền tảng của thống kê, xác suất nói riêng và toán học nói chung. Công nghệ thông tin mang cho ta nguồn dữ liệu dồi dào, làm nảy sinh một nhu cầu cần suy diễn với những dữ liệu ấy. Thống kê toán cho ta một nền móng để suy diễn một cách hợp lý. Ngược lại sự phát triển của khoa học suy diễn từ dữ liệu, dữ liệu lớn, cũng đưa toán học tới gần với các ứng dụng của xã hội hiện đại hơn bao giờ hết. Vấn đề suy diễn với các cấu trúc phức tạp cũng có tác dụng thúc đẩy sự phát triển nội tại của toán học.

Bài viết này minh hoạ một vài khía cạnh của sự tương tác ấy: chỉ với kiến thức toán sơ cấp chúng ta có thể mô tả được một lớp kỹ thuật tiên tiến của ngành học máy và di truyền quần thể trong hai thập niên đầu thế kỷ 21. Tất nhiên, để sáng tạo ra những kỹ thuật ấy, các tác giả của chúng không chỉ cần có toán học. Họ phải có sự tò mò đối với thực tiễn, có khả năng diễn đạt thành thục bằng mô hình toán học, có sự liều lĩnh để áp mô hình vào dữ liệu thực. Để hiểu thấu đáo và cải thiện các phương pháp suy diễn ấy, để đi đến những bước đột phá trong tương lai, chắc chắn sẽ cần nhiều công cụ toán học mạnh hơn nữa.

Lời cảm ơn Tác giả chân thành cảm ơn GS Ngô Bảo Châu đã khích lệ và góp ý cho bài viết.

5 Tài liệu tham khảo

I. Bárány and V. Vu. Central limit theorems for Gaussian polytopes. *Annals of Probability*, 35:1593–1621, 2007.

D. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21:618–646, 2015.

J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

M. Reitzner. Central limit theorems for random polytopes. *Probability Theory and Related Fields*, 133:483–507, 2005.

M. Yurochkin and X. Nguyen. Geometric Dirichlet means algorithm for topic inference. *Advances in Neural Information Processing Systems (NIPS)* 29:2505-2513, 2016.