Chương 3
Xác suất và lý thuyết thông tin

Trong chương này, ta sẽ tìm hiểu về lý thuyết xác suất và lý thuyết thông tin.

Lý thuyết xác suất là nền tảng toán học để biểu diễn các sự kiện bất định. Nó cho ta một phương thức để lượng hóa tính bất định cùng với các tiên đề để suy diễn ra các sự kiện bất định mới. Trong các ứng dụng AI, các lý thuyết xác suất thường được sử dụng để giải quyết hai vấn đề chính. Thứ nhất, các luật xác suất

cho ta biết các hệ thống AI nên suy diễn như thế nào, từ đó, ta có thể thiết kế các thuật toán để tính hoặc xấp xỉ các biểu thức xác suất. Thứ hai, ta có thể sử dụng xác suất và thống kê để phân tích về mặt lý thuyết hành vi của các hệ thống AI.

Lý thuyết xác suất là công cụ cốt lõi trong nhiều ngành khoa học và kĩ thuật. Chúng tôi viết chương này với mục đích giúp những bạn có kiến thức nền tảng chủ yếu về công nghệ phần mềm, ít tiếp xúc với lý thuyết xác suất, có thể hiểu được những gì trình bày trong cuốn sách này.

Trong khi lý thuyết xác suất cho phép ta suy luận ra các sự kiện mới và suy luận ngay cả khi có sự bất định, lý thuyết thông tin lại cung cấp thêm cho ta một cách thức để lượng hóa sự bất định thể hiện thông qua một phân phối xác suất.

Nếu bạn đã làm quen với lý thuyết xác suất và lý thuyết thông tin trước đây có thể bỏ qua chương này, ngoại trừ mục 3.14. Trong mục 3.14, chúng tôi mô tả các đồ thị dùng để biểu diễn các mô hình có cấu trúc sử dụng trong ML. Nếu bạn hoàn toàn không có chút kiến thức nào về hai mảng này trước đây, thì kiến thức trong chương này có lẽ đủ để bạn thực hiện các dự án nghiên cứu về DL, tuy nhiên chúng tôi cũng khuyến khích bạn tham khảo thêm nguồn khác, chẳng hạn như [Jaynes, 2003].

3.1 Tại sao cần xác suất?

Nhiều nhánh của khoa học máy tính chủ yếu nghiên cứu các đối tượng hoàn toàn xác định và chắc chắn. Một lập trình viên thường giả định rằng sẽ không có lỗi xảy ra mỗi khi CPU thực thi một dòng lệnh. Các lỗi phần cứng thực tế có xảy ra, nhưng rất hiếm, và hầu hết các ứng dụng phần mềm không cần phải quan tâm đến yếu tố này. Thế nên người ta có thể cảm thấy ngạc nhiên khi biết rằng ML lại sử dụng rất nhiều lý thuyết xác suất, trong khi các nhà khoa học máy tính và các kĩ sư phần mềm lại làm việc trong một môi trường tương đối "sạch" và ổn định.

Lĩnh vực ML luôn phải đương đầu với các đại lượng bất định (*uncertain*), và đôi khi là các đại lượng ngẫu nhiên (*stochastic*). Tính bất định và tính ngẫu nhiên có thể đến từ nhiều nguồn khác nhau. Từ cuối những năm 1980, các nhà nghiên cứu có thể nói là thành công trong các lý luận sử dụng xác suất để lượng hóa

tính bất ổn. Nhiều lý luận chúng tôi trình bày ở đây được tổng kết hoặc được lấy cảm hứng từ [Pearl, 1988].

Gần như tất cả các hoạt động đều cần một chút khả năng suy luận nếu chúng có tính bất định. Trên thực tế, ngoài những phát biểu toán học là luôn đúng theo định nghĩa, rất khó để hình dung ra bất kỳ một mệnh đề nào là đúng tuyệt đối, hoặc bất kỳ sự kiện nào được đảm bảo chắc chắn sẽ xảy ra.

Sự bất định thường đến từ ba nguồn chính:

1. **Tính ngẫu nhiên vốn có trong hệ thống được mô hình hóa**. Ví dụ, hầu hết các cách diễn giải của cơ học lượng tử cho rằng động lượng của các hạt hạ nguyên tử mang tính xác suất. ta cũng có thể tạo ra các tình huống lý thuyết mà ở đó, ta gán cho nó bản chất xác suất. Ví dụ như trong trò chơi các quân bài, ta thường giả định rằng các quân bài đều được xáo trộn một cách thực sự ngẫu nhiên.

2. **Quan sát không đầy đủ**. Tất cả các hệ thống xác định đều trông giống ngẫu nhiên khi ta không thể quan sát được tất cả các biến số quyết định hành vi của hệ thống đó. Ví dụ, trong bài toán Monty Hall, người chơi sẽ được đề nghị chọn một cánh cửa có phần thưởng trong ba cánh cửa có sẵn. Đằng sau hai trong ba cánh cửa đó là một con dê, và đằng sau cánh cửa còn lại là một chiếc ô tô. Rõ ràng một khi người chơi đã chọn thì phần thưởng đằng sau cánh cửa được chọn đó là xác định, nhưng từ góc độ của người chơi, phần thưởng đằng sau cánh cửa lại là một sự kiện bất định.

3. **Mô hình hóa không đầy đủ**. Khi ta sử dụng một mô hình mà trong mô hình đó ta phải bỏ đi một số thông tin quan sát được, những thông tin bị bỏ đi đó tạo ra sự bất định trong các kết quả dự đoán của mô hình. Ví dụ, giả sử ta tạo ra một robot có khả năng quan sát chính xác vị trí của mọi vật thể xung quanh nó. Nếu con robot đó thực hiện rời rạc hóa không gian trong quá trình nó dự đoán vị trí trong tương lai của các vật xung quanh, sự rời rạc hóa đó lập tức khiến con robot trở nên không chắc chắn về vị trí chính xác của các vật thể: mỗi vật thể có thể ở bất kì vị trí nào bên trong mỗi ô rời rạc mà con robot quan sát được.

Trong nhiều trường hợp, sử dụng các quy luật chứa yếu tố bất định nhưng đơn giản sẽ phù hợp với thực tế hơn là các quy luật tất định nhưng phức tạp, ngay cả khi các quy luật của hệ thống thực sự là tất định và trong hệ thống của ta về bản

chất có bao gồm một luật phức tạp. Ví dụ, một luật đơn giản như "hầu hết chim có thể bay" không những dễ phát triển mà còn hữu dụng hơn là luật kiểu như "Chim có thể bay, ngoại trừ những con non chưa học bay, những con bị ốm hay bị thương đến nỗi bị mất khả năng bay, những loài chim không biết bay như đà điểu châu Úc, đà điểu Châu phi và chim Kiwi..". Các luật phức tạp như thế thường tốn nhiều chi phí để phát triển, khó bảo trì và truyền đạt, và mặc dù tốn rất nhiều công sức, các luật này vẫn không bền và đôi khi không thể áp dụng được.

Tuy rõ ràng là ta cần một cách để biểu diễn và suy luận về sự bất định, nhưng khó có thể biết được liệu lý thuyết xác suất có cung cấp đầy đủ các công cụ ta cần để phát triển các ứng dụng trí tuệ nhân tạo hay không. Lý thuyết xác suất ban đầu được phát triển để phân tích tần suất của các sự kiện. Không khó để thấy rằng lý thuyết xác suất có thể được ứng dụng để nghiên cứu các sự kiện kiểu như rút được các quân bài thuộc cùng một bộ trong trò poker. Những sự kiện kiểu như vậy thường lặp đi lặp lại. Khi ta nói rằng một sự kiện có một xác suất xảy ra, điều đó có nghĩa là nếu ta lặp lại phép thử đó rất nhiều lần, tỉ lệ mà sự kiện đó lặp lại chính là xác suất xảy ra của sự kiện đó. Cách suy luận này có vẻ như không thể áp dụng được cho các sự kiện không thể lặp lại. Chẳng hạn, khi một bác sĩ thực hiện chuẩn đoán với một bệnh nhân, và kết luận rằng bệnh nhân đó có 40% khả năng bị cảm cúm, ý nghĩa của điều này lại rất khác, ta không thể tạo ra rất nhiều phiên bản của cùng người bệnh đó, có cùng triệu chứng đó nhưng với các điều kiện gây ra triệu chứng khác nhau. Trong trường hợp bác sĩ khám bệnh này, xác suất biểu diễn *mức độ tin cậy* với 1 mang ý nghĩa rằng người bệnh chắc chắn bị cúm, và 0 mang ý nghĩa người bệnh chắc chắn không bị cúm. Dạng xác suất đầu tiên, liên quan tới tần suất một sự kiện xảy ra, được gọi là *frequentist probability*, trong khi đó, dạng xác suất còn lại liên quan tới lượng hóa mức độ bất định, được gọi là *Bayesian probability*.

Nếu liệt kê một vài tính chất mà ta cho rằng các lập luận thông thường về tính bất định cần phải có, thì cách duy nhất để thỏa mãn các tính chất đó là coi xác suất Bayes như là xác suất tần xuất. Ví dụ, nếu ta tính xác suất một người chơi thắng trong trò poker, biết rằng người đó đang cầm một bộ các quân bài xác định, ta sẽ dùng cùng một công thức tính như khi ta tính xác suất một bệnh nhân bị bệnh nào đó khi biết các triệu chứng. Để biết rõ hơn tại sao một tập nhỏ các giả thiết thông thường lại suy ra tính chất của cả hai hướng xác suất có cùng một tập các tiên đề, xem [Ramsey, 1926].

Xác suất có thể được coi là sự mở rộng của logic để lập luận về sự bất định. Logic cung cấp một tập các luật chuẩn để xác định mệnh đề nào đúng hoặc sai khi biết trước một tập mệnh đề khác là đúng hay sai. Còn lý thuyết xác suất, theo một cách linh hoạt hơn, lại cung cấp một tập các luật chuẩn để xác định **khả năng** đúng hay sai của một mệnh đề, khi biết trước **khả năng** đúng hay sai của một tập mệnh đề khác.

3.2 Biến ngẫu nhiên

Biến ngẫu nhiên là một biến có thể nhận các giá trị khác nhau một cách ngẫu nhiên. Ta thường ký hiệu biến ngẫu nhiên bằng một ký tự in thường, không nghiêng và các giá trị mà nó có thể nhận bằng các chỉ số là các ký tự in thường. Ví dụ, x_1 và x_2 là hai giá trị mà biến ngẫu nhiên x có thể nhận. Ta sẽ ký hiệu một vector các biến ngẫu nhiên là $\mathbf{x}$ và một giá trị cụ thể của vector ngẫu nhiên đó là $\boldsymbol{x}$. Bản thân một biến ngẫu nhiên chỉ là một cách mô tả tập các trạng thái có thể tồn tại; biến ngẫu nhiên đó phải đi kèm với một phân phối xác suất quyết định khả năng các trạng thái đó xuất hiện.

Các biến ngẫu nhiên có thể là liên tục hay rời rạc. Một biến rời rạc chỉ có thể nhận giá trị trong tập hữu hạn hoặc vô hạn **đếm được**. Chú ý rằng các giá trị của biến ngẫu nhiên rời rạc không nhất thiết phải là số nguyên; chúng có thể chỉ đơn thuần là tập các trạng thái được đặt tên và không có trị số tương ứng. Một biến ngẫu nhiên liên tục sẽ nhận các giá trị thực.

3.3 Phân phối xác suất

Phân phối xác suất mô tả khả năng một biến hay một tập biến ngẫu nhiên nhận một giá trị nào đó. Cách biểu diễn một phân phối lại phụ thuộc vào các biến ngẫu nhiên đó là rời rạc hay liên tục.

3.3.1 Biến rời rạc và hàm khối xác suất

Phân phối xác suất trên các *biến rời rạc* có thể được mô tả bởi *hàm khối xác suất* (*probability mass function -- PMF*). Ta thường kí hiệu các hàm khối xác suất bởi một chữ cái in hoa P. Thông thường, mỗi biến ngẫu nhiên sẽ có một hàm

khối xác suất tương ứng, và bạn phải tự suy ra biến nào được gán với hàm PMF nào dựa trên kí hiệu của biến đó, thay vì dựa trên tên của hàm PFM; $P(x)$ thường sẽ khác với $P(y)$.

PMF ánh xạ mỗi giá trị của một biến ngẫu nhiên tới xác suất mà biến ngẫu nhiên mang giá trị này. Xác suất $\mathrm{x} = x$ được kí hiệu là $P(x)$, với giá trị 1 có nghĩa là chắc chắn $\mathrm{x} = x$ và 0 có nghĩa là $\mathrm{x} = x$ không thể xảy ra. Đôi khi để phân biệt PMF nào đang được sử dụng, ta sẽ viết rõ ràng tên của biến ngẫu nhiên đó: $P(\mathrm{x} = x)$. Nhưng đôi khi ta sẽ định nghĩa biến ngẫu nhiên trước, sau đó sử dụng kí hiệu $\sim$ để chỉ định phân phối tương ứng: $\mathrm{x} \sim P(\mathrm{x})$.

Các hàm khối xác suất có thể áp dụng cho nhiều biến ngẫu nhiên cùng lúc. Một phân phối xác suất cho nhiều biến được gọi là một *đồng phân phối xác suất* (*joint probability distribution*). $P(\mathrm{x} = x, \mathrm{y} = y)$ là xác suất đồng thời xảy ra $\mathrm{x} = x$ và $\mathrm{y} = y$. Để đơn giản, ta có thể viết gọn thành $P(x, y)$ thay cho $P(\mathrm{x} = x, \mathrm{y} = y)$.

Để P có thể là một PMF cho một biến ngẫu nhiên x, nó phải thỏa mãn các tính chất sau:

- Tập xác định của P phải là tập tất cả các giá trị mà x có thể nhận.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. Một sự kiện không thể xảy ra sẽ có xác suất bằng 0, và không sự kiện nào có thể có xác suất nhỏ hơn 0. Tương tự, một sự kiện chắc chắn xảy ra sẽ có xác suất bằng 1, và không sự kiện nào có xác suất xảy ra lớn hơn 1.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. Ta gọi tính chất này là *tính được chuẩn hóa* (*being normalized*). Nếu không có tính chất này, ta có thể gặp phải các giá trị xác suất lớn hơn 1 khi tính toán xác suất để xảy ra một sự kiện trong tập các sự kiện khả dĩ.

Chẳng hạn, xét một biến ngẫu nhiên rời rạc x có thể nhận k giá trị khác nhau. ta có thể chỉ định cho nó một *phân phối đều* (*uniform distribution*)—tức là gán cho mỗi giá trị trong tập giá trị một xác suất như nhau—bằng cách thiết lập hàm PMF của nó như sau:

$$P(\mathrm{x} = x_i) = \frac{1}{k} \qquad (3.1)$$

với mọi i. Ta có thể thấy rằng hàm này thỏa mãn tất cả các điều kiện đặt ra ở trên cho một hàm khối xác suất. Giá trị $\frac{1}{k}$ dương vì k dương. Ta cũng thấy:

$$\sum_i P(\mathrm{x} = x_i) = \sum_i \frac{1}{k} = 1 \qquad (3.2)$$

Do đó, phân phối này cũng thỏa mãn *tính được chuẩn hóa*.

3.3.2 Biến liên tục và hàm mật độ xác suất

Khi làm việc với các biến ngẫu nhiên liên tục, ta mô tả phân phối xác suất bằng cách sử dụng *hàm mật độ xác suất* (probability density function - PDF) thay vì hàm khối xác suất. Để được xem là một hàm mật độ xác suất, hàm số p phải thỏa mãn các tính chất sau:

- Miền giá trị của p phải là tập tất cà các trạng thái có thể có của x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Lưu ý, ta không yêu cầu $p(x) \leq 1$.
- $\int p(x)dx = 1$.

Hàm mật độ xác suất $p(x)$ không trực tiếp cho ta biết xác suất của một trạng thái cụ thể; thay vào đó, xác suất rơi vào phần vi phân có độ lớn δx, và bằng $p(x)\delta x$.

Ta có thể lấy tích phân hàm mật độ xác suất để tìm ra khối xác suất thực tế của một tập các điểm. Cụ thể, xác suất để x nằm trong tập $\mathbb{S}$ được tính bằng cách lấy tích phân $p(x)$ trên toàn tập đó. Chẳng hạn, trong trường hợp đơn biến, xác suất để x nằm trong khoảng $[a, b]$ được tính bởi $\int_{[a,b]} p(x)dx$.

Để tìm hiểu một ví dụ về hàm mật độ xác suất tương ứng với một mật độ xác suất cụ thể cho biến ngẫu nhiên liên tục, ta xét một phân phối đều trên một khoảng số thực. Ta có thể thực hiện với hàm số $u(x; a, b)$, với a, b là hai đầu mút của khoảng, với $b > a$. Dấu ";" là viết tắt của cụm "tham số hóa bởi"; ta xét x là đối số của hàm, trong khi đó a, b là các tham số định nghĩa hàm. Để chắc chắn rằng không có khối xác suất nào nằm ngoài khoảng $[a, b]$, ta giả sử $u(x; a, b) = 0$ với mọi $x \notin [a, b]$. Trong khoảng $[a, b]$, $u(x; a, b) = \frac{1}{b-a}$. Có thể thấy rằng biểu thức này không âm ở mọi giá trị của x. Thêm vào đó, nó có tích phân bằng 1. Ta thường ký hiệu rằng x tuân theo phân phối đều trên khoảng $[a, b]$ bằng cách viết $\mathrm{x} \sim U(a, b)$.

3.4 Xác suất biên

Đôi khi ta đã biết phân phối xác suất trên một tập các biến, và muốn xác định phân phối xác suất trên một tập con của chúng. Phân phối xác suất trên tập con được gọi là *phân phối xác suất biên* (*marginal probability distribution*).

Ví dụ, giả sử cho trước các biến ngẫu nhiên rời rạc x, y, và $P(\mathrm{x}, \mathrm{y})$. Ta có thể tìm $P(\mathrm{x})$ bằng *quy tắc cộng* (*sum rule*):

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y) \qquad (3.3)$$

Cái tên "xác suất biên" bắt nguồn từ quá trình tính toán các xác suất biên khi thực hiện trên giấy. Giá trị của $P(\mathrm{x}, \mathrm{y})$ được viết theo các ô lưới, các giá trị khác nhau của x được viết theo hàng và các giá trị khác nhau của y được viết theo cột, ta tính tổng mỗi dòng của ô lưới rồi viết $P(x)$ vào biên (lề) của tờ giấy, bên phải mỗi hàng.

> ND: Giả sử ta gieo 2 cục xúc xắc, xác suất nhận được kết quả trong tổng số 36 khả năng sẽ như trong bảng. Biên bên phải và bên dưới là các xác suất biên

i \ j	1	2	3	4	5	6	$Px(i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
3	1/36	1/36	1/36	1/36	1/36	1/36	1/6
4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
5	1/36	1/36	1/36	1/36	1/36	1/36	1/6
6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$Py(j)$	1/6	1/6	1/6	1/6	1/6	1/6	

Đối với biến liên tục, ta dùng phép tích phân thay vì tính tổng:

$$p(x) = \int p(x, y)dy \qquad (3.4)$$

3.5 Xác suất có điều kiện

Trong nhiều trường hợp, ta quan tâm tới xác suất của một vài sự kiện, khi đã biết một vài sự kiện khác xảy ra trước đó, được gọi là *xác suất có điều kiện* (conditional probability). Ta ký hiệu xác suất có điều kiện $\mathrm{y} = y$ khi biết $\mathrm{x} = x$ là $P(\mathrm{y} = y|\mathrm{x} = x)$. Xác suất có điều kiện này có thể được tình bằng công thức:

$$P(\mathrm{y} = y|\mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)} \qquad (3.5)$$

Xác suất có điều kiện chỉ xác định khi $P(\mathrm{x} = x) > 0$. Dễ hiểu, ta không thể tính xác suất có điều kiện với một sự kiện không bao giờ xảy ra.

Một điều quan trọng là ta không nên nhầm lẫn giữa *xác suất có điều kiện* với việc tính toán (dự đoán) hệ quả của hành động cho trước. *Xác suất có điều kiện* một người nói tiếng Đức khi biết trước rằng họ đến từ Đức là khá cao, nhưng nếu chọn ngẫu nhiên một người và dạy người này nói tiếng Đức thì cũng không thay đổi được việc họ đến từ đâu. Tính toán chuỗi hệ quả của một hành động được gọi là *truy vấn can thiệp* (intervention query). Truy vấn can thiệp thuộc về lĩnh vực *mô hình nhân quả* (causal modeling), nằm ngoài phạm vi đề cập của cuốn sách này.

3.6 Quy tắc dây chuyền của xác suất có điều kiện

Bất kỳ đồng phân phối xác suất nào trên nhiều biến ngẫu nhiên cũng có thể được phân tách thành các phân phối có điều kiện chỉ trên một biến:

$$P(\mathrm{x}^{(1)}, ..., \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, ..., \mathrm{x}^{(i-1)}) \tag{3.6}$$

Tính chất này còn được gọi là *quy tắc dây chuyền* (*chain rule*), hoặc *quy tắc nhân* (*product rule*) của xác suất. Có thể suy ngay ra công thức trên từ định nghĩa của xác suất có điều kiện trong phương trình (3.5).

Ví dụ, áp dụng định nghĩa hai lần, ta có:

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} | \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} | \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} | \mathrm{b}, \mathrm{c}) P(\mathrm{b} | \mathrm{c}) P(\mathrm{c}) \end{aligned}$$

3.7 Độc lập và độc lập có điều kiện

Hai biến ngẫu nhiên x, y được coi là *độc lập* nếu phân phối xác suất của chúng có thể được biểu diễn bằng tích của hai nhân tử, một chỉ dựa trên x và một chỉ dựa trên y:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x) p(\mathrm{y} = y) \tag{3.7}$$

Hai biến ngẫu nhiên x và y được xem là *độc lập có điều kiện* với một biến ngẫu nhiên z cho trước nếu phân phối xác suất có điều kiện trên cả x và y có thể được tách theo cách sau đây cho mọi giá trị của z:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, x \in \mathrm{x}, p(\mathrm{x} = x, \mathrm{y} = y|\mathrm{z} = z) = p(\mathrm{x} = x|\mathrm{z} = z)p(\mathrm{y} = y|\mathrm{z} = z) \tag{3.8}$$

Ta có thể ký hiệu độc lập và độc lập có điều kiện một cách súc tích như sau: $\mathrm{x} \perp \mathrm{y}$ nghĩa là x độc lập với y, trong khi đó $\mathrm{x} \perp \mathrm{y}|\mathrm{z}$ nghĩa là x và y độc lập có điều kiện khi biết z.

3.8 Kỳ vọng, phương sai, hiệp phương sai

Kỳ vọng (expectation), hoặc *giá trị kỳ vọng* (expected value) của hàm số $f(x)$ tương ứng với phân phối xác suất $P(\mathrm{x})$ là trung bình, hoặc giá trị trung bình của f với x lấy ra từ P. Đối với biến rời rạc, kỳ vọng có thể được tính bằng tổng:

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \sum_x P(x)f(x) \tag{3.9}$$

còn đối với biến liên tục, kỳ vọng được tính bằng tích phân:

$$\mathbb{E}_{\mathrm{x} \sim p}[f(x)] = \int p(x)f(x)dx \tag{3.10}$$

Trong trường hợp phân phối được xác định rõ ràng, để đơn giản hơn, ta có thể chỉ viết tên của biến ngẫu nhiên đang cần tính kỳ vọng, ví như ký hiệu ở trên có thể viết là $\mathbb{E}_{\mathrm{x}}[f(x)]$. Nếu biến ngẫu nhiên mà ta đang lấy kỳ vọng cũng rõ ràng (không sợ nhầm lẫn trong ngữ cảnh sử dụng), ta có thể bỏ qua phần chỉ số dưới, và chỉ viết $\mathbb{E}[f(x)]$. Mặc định, ta hiểu rằng $\mathbb{E}[\cdot]$ là trung bình trên giá trị của tất cả các biến ngẫu nhiên trong ngoặc vuông. Tương tự, nếu không sợ nhầm lẫn, ta có thể bỏ qua dấu ngoặc vuông.

Kỳ vọng là hàm tuyến tính, chẳng hạn,

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{x}}[f(x)] + \beta \mathbb{E}_{\mathrm{x}}[g(x)] \tag{3.11}$$

với α, β không phụ thuộc vào x.

Phương sai (variance) cho ta biết các giá trị của hàm số của một biến ngẫu nhiên x có thể dao động đến mức nào khi ta lấy mẫu các giá trị x khác nhau từ

phân phối xác suất của nó:

$$\mathrm{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \qquad (3.12)$$

Khi giá trị phương sai nhỏ, thì các giá trị của $f(x)$ tụ lại gần giá trị kỳ vọng. Căn bậc hai của phương sai được gọi là *độ lệch chuẩn* (standard deviation).

Hiệp phương sai (covariance) cho ta biết một chút về mức độ liên quan tuyến tính giữa hai giá trị với nhau, cũng như tỉ lệ giữa hai biến này với nhau:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])] \qquad (3.13)$$

Giá trị tuyệt đối của hiệp phương sai lớn có nghĩa là giá trị của các biến ngẫu nhiên có mức độ thay đổi lớn và đều xa giá trị trung bình tương ứng cùng lúc. Nếu dấu của hiệp phương sai là dương, thì cả hai biến có xu hướng cùng nhận giá trị tương đối cao. Nếu dấu của hiệp phương sai là âm, thì một biến có xu hướng đạt giá trị tương đói cao trong khi biến kia có xu hướng đạt giá trị tương đối thấp và ngược lại. Những thước đo khác như là *độ tương quan* (correlation) sẽ chuẩn hóa sự đóng góp của mỗi biến vào hiệp phương sai, dùng để đo lường mức độ liên quan giữa các biến, hơn là độ lớn của các biến riêng lẻ.

Hai khái niệm *hiệp phương sai* và *phụ thuộc* có liên quan với nhau nhưng hoàn toàn khác nhau. Ta nói rằng chúng có liên quan với nhau là bởi hai biến độc lập có hiệp phương sai bằng 0, và hai biến có hiệp phương sai khác 0 thì phụ thuộc với nhau. Dù vậy, độc lập là một đặc tính tách biệt với hiệp phương sai. Hai biến có hiệp phương sai bằng 0, thì chắc chắn không có phụ thuộc tuyến tính với nhau. Tính độc lập là một giả thiết mạnh hơn so với hiệp phương sai bằng 0, bởi vì tính độc lập loại trừ cả các mối quan hệ phi tuyến (giữa 2 biến). Hoàn toàn có có khả năng hai biến phụ thuộc (phi tuyến) nhưng có hiệp phương sai bằng 0. Ví dụ, giả sử đầu tiên ta lấy mẫu số thực x từ một phân phối đều liên tục trên khoảng $[-1, 1]$. Tiếp theo, ta lấy mẫu biến ngẫu nhiên s. Với xác suất $\frac{1}{2}$, ta chọn $s = 1$. Trong các trường hợp khác, ta chọn $s = -1$. Ta có thể sinh biến ngẫu nhiên y bằng cách gán $y = sx$. Rõ ràng là x và y không hề độc lập với nhau, bởi x hoàn toàn xác định độ lớn của y. Và dù vậy thì, $\mathrm{Cov}(x, y) = 0$

Ma trận hiệp phương sai (*covariance matrix*) của vector ngẫu nhiên $\mathbf{x} \in \mathbb{R}^n$ là ma trận kích cỡ $n \times n$ thỏa mãn:

$$\mathrm{Cov}(\mathrm{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j) \qquad (3.14)$$

Các phần tử trên đường chéo chính của ma trận hiệp phương sai chính là phương sai:

$$\text{Cov}(\text{x}_i, \text{x}_i) = \text{Var}(\text{x}_i) \tag{3.15}$$

3.9 Các phân phối xác suất thông dụng

3.9.1 Phân phối Bernoulli

Phân phối **Bernoulli** là phân phối xác suất rời rạc của một biến ngẫu nhiên nhị phân. Nó được đặc trưng bởi duy nhất một tham số $\phi \in [0, 1]$, cũng là xác suất để biến ngẫu nhiên có giá trị bằng 1. Một số tính chất của phân phối Bernoulli:

$$P(\text{x} = 1) = \phi \tag{3.16}$$
$$P(\text{x} = 0) = 1 - \phi \tag{3.17}$$
$$P(\text{x} = x) = \phi^x(1 - \phi)^{1-x} \tag{3.18}$$
$$\mathbb{E}_\text{x}[\text{x}] = \phi \tag{3.19}$$
$$\text{Var}_\text{x}[\text{x}] = \phi(1 - \phi) \tag{3.20}$$

3.9.2 Phân phối Multinoulli

Phân phối *multinoulli* (đôi khi còn được gọi là *phân phối phạm trù*) là dạng tổng quát của phân phối **Bernoulli**, trên một đơn biến rời rạc với k trạng thái khác nhau, trong đó k là hữu hạn.[1] Phân phối phạm trù được tham số hoá bởi một vector $\mathbf{p} \in [0, 1]^{k-1}$, với p_i là xác suất của trạng thái thứ i. Xác suất của trạng thái thứ k được xác định bởi $1 - \mathbf{1}^T\mathbf{p}$. Lưu ý rằng ta phải có ràng buộc $\mathbf{1}^T\mathbf{p} \leq 1$. Các phân phối phạm trù thường được sử dụng để chỉ phân phối theo phạm trù các đối tượng, do đó ta thường không giả định trạng thái 1 có giá trị số là 1, vân vân. Vậy nên, việc tính toán kỳ vọng và phương sai của các biến ngẫu nhiên trong phân phối phạm trù thường là không cần thiết.

> [1] "Multinoulli" là một thuật ngữ mới xuất hiện gần đây, được đặt tên bởi Gustavo Lacerda và trở nên phổ biến từ Murphy (2012). Phân phối phạm trù là một trường hợp đặc biệt của *phân phối đa thức* (multinomial distribution). Một phân phối đa thức là phân phối của các véc-tơ trong $0, ..., n^k$, thể hiện số lần mỗi k phạm trù xuất hiện sau n phép lấy mẫu từ một phân phối đa thức. Có nhiều tài liệu sử dụng thuật ngữ "multinomial" khi đề cập đến các

> phân phối phạm trù mà không chỉ rõ rằng chúng chỉ là một trường hợp con của phân phối đa thức khi $n = 1$.

Phân phối Bernoulli và phân phối phạm trù là đủ để mô tả bất kỳ phân phối nào trên miền xác định của chúng. Lý do không hẳn là vì hai phân phối này thực sự mạnh, mà vì miền xác định của chúng đơn giản; chúng mô tả các trạng thái rời rạc, loại trạng thái mà ta có thể liệt kê tất cả. Đối với các biến liên tục, số trạng thái là vô hạn không đếm được, do đó bất kỳ phân phối nào được mô tả bởi một lượng nhỏ tham số buộc phải có những giới hạn nghiêm ngặt.

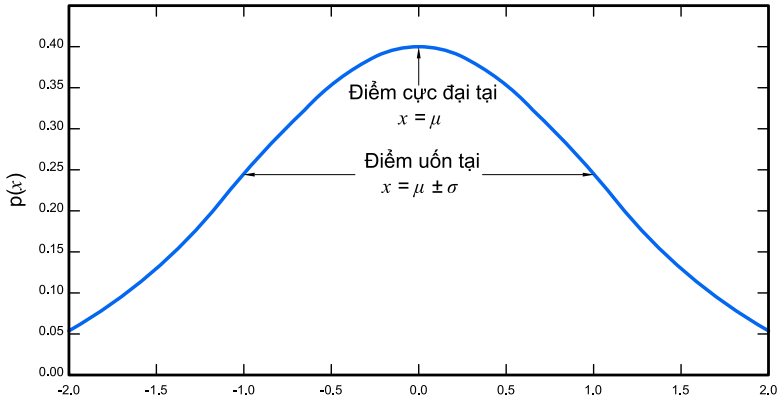
3.9.3 Phân phối Gauss

Phân phối được sử dụng phổ biến trên tập số thực là *phân phối chuẩn*, thường được biết đến với tên gọi *phân phối Gauss*:

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \qquad (3.21)$$

Hình 3.1 biểu diễn hàm mật độ xác suất của một phân phối chuẩn.

Phân phối chuẩn có hai tham số: $\mu \in \mathbb{R}$ và $\sigma \in (0, \infty)$. Tham số μ chỉ ra tọa độ đỉnh trung tâm. Đây cũng chính là giá trị trung bình của phân phối: $\mathbb{E}[\mathrm{x}] = \mu$. Độ lệch chuẩn của phân phối được ký hiệu là σ và phương sai σ^2.



Hình 3.1: Phân phối chuẩn. Đồ thị của phân phối chuẩn $\mathcal{N}(x; \mu, \sigma)$ trông giống như một "đường cong hình chuông" cổ điển, với toạ độ x ở đỉnh trung tâm được xác định bởi μ, và độ rộng của "chuông" phụ thuộc vào σ. Trong hình trên, chúng tôi mô tả một phân phối chuẩn tắc với $\mu = 0$ và $\sigma = 1$.

Khi định giá trị hàm mật độ xác suất của phân phối chuẩn, ta cần bình phương và nghịch đảo giá trị σ. Nhưng khi cần định giá trị đó một cách thường xuyên với

nhiều giá trị σ khác nhau, một cách tham số hoá hiệu quả hơn đó là sử dụng một tham số $\beta \in (0, \infty)$ để kiểm soát *độ nét* (*precision*), hay phương sai nghịch đảo, của phân phối:

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x-\mu)^2\right) \tag{3.22}$$

Phân phối chuẩn là lựa chọn phổ biến và hợp lý cho nhiều ứng dụng. Đặc biệt là trong trường hợp ta không có tri thức tiên nghiệm về dạng phân phối mà các số thực tuân theo, phân phối chuẩn là một lựa chọn mặc định tốt vì hai lý do chính sau.

Thứ nhất, nhiều phần phối ta cần tìm thực sự rất gần với các phân phối chuẩn. *Định lý giới hạn trung tâm* chỉ ra rằng trong một số trường hợp nhất định, phân phối của tổng của nhiều biến ngẫu nhiên độc lập là phân phối xấp xỉ chuẩn. Có nghĩa là trong thực tế, nhiễu với phân phối chuẩn là một mô hình khá thành công cho các hệ thống phức tạp, ngay cả khi hệ thống đó có thể được tách thành các thành phần có hành vi mang tính cấu trúc.

Thứ hai, trong số tất cả các phân phối xác suất có cùng phương sai, phân phối chuẩn mã hoá một lượng lớn nhất sự bất định đối với các số thực. Do đó, ta có thể coi phân phối chuẩn là phân phối đưa vào ít tri thức tiên nghiệm nhất trong mô hình. Chứng minh đầy đủ cho ý tưởng này đòi hỏi nhiều công cụ toán học và ta sẽ tìm hiểu đầy đủ hơn trong phần 19.4.2.

Trường hợp tổng quát hơn của phân phối chuẩn trong không gian $\mathbb{R}^n$ được gọi là *phân phối chuẩn đa biến* (multivariate normal distribution). Phân phối này được tham số hoá bởi một ma trận đối xứng xác định dương $\boldsymbol{\Sigma}$:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3.23}$$

Trong đó, tham số $\boldsymbol{\mu}$ biểu thị vector giá trị trung bình của phân phối và $\boldsymbol{\Sigma}$ là ma trận hiệp phương sai của phân phối. Giống như đối với phân phối một chiều, khi cần định giá trị hàm mật độ xác suất nhiều lần trên nhiều giá trị khác nhau của các tham số, ma trận hiệp phương sai không phải là cách tham số hóa hiệu quả về mặt tính toán, bởi vì ta phải nghịch đảo $\boldsymbol{\Sigma}$ để tính. Thay vào đó, ta sử dụng *ma trận độ nét* (precision matrix) $\boldsymbol{\beta}$ để đánh giá hàm mật độ xác suất:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3.24}$$

Ma trận hiệp phương sai thường được chọn là ma trận đường chéo. Ngoài ra, có một phiên bản đơn giản hơn nữa là phân phối Gauss *đẳng hướng* (*isotropic*), với ma trận hiệp phương sai tỉ lệ với ma trận đơn vị.

3.9.4 Phân phối mũ và phân phối Laplace

Khi làm việc với DL, ta thường muốn có một phân phối xác suất với một đỉnh nhọn tại $x = 0$. *Phân phối mũ* (*exponential distribution*) có thể đáp ứng điều đó:

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x) \qquad (3.25)$$

Phân phối mũ sử dụng một hàm chỉ thị $\mathbf{1}_{x \geq 0}$ để chỉ định xác suất bằng 0 cho tất cả các giá trị âm của x.

Ngoài ra, có một phân phối xác suất liên quan chặt chẽ với phân phối mũ là *phân phối Laplace*, cho phép ta đặt một đỉnh nhọn của khối xác suất tại một điểm μ tuỳ ý:

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right) \qquad (3.26)$$

3.9.5 Phân phối Dirac và phân phối thực nghiệm

Đôi khi, ta sẽ cần toàn bộ khối lượng của một phân phối xác suất được gom lại quanh một điểm duy nhất. Điều này có thể được thực hiện bằng cách định nghĩa một hàm mật độ xác suất sử dụng *hàm delta Dirac*, $\delta(x)$:

$$p(x) = \delta(x - \mu) \qquad (3.27)$$

Hàm delta Dirac được định nghĩa sao cho nó có giá trị bằng 0 ở mọi điểm ngoại trừ 0, nhưng lại có tích phân toàn phần trên miền xác định bằng 1. Hàm delta Dirac không phải là một hàm số thông thường, với mỗi đầu vào x cho ra một giá trị thực tương ứng; thay vào đó, nó là một loại đối tượng toán học khác mang tên *hàm tổng quát hoá* được định nghĩa theo các tính chất riêng khi lấy tích phân. Ta có thể coi hàm delta Dirac như một điểm giới hạn của chuỗi các hàm số có khối hẹp dần khi tiến tới 0.

Bằng cách định nghĩa $p(x) = \delta(x - \mu)$, ta đạt được một đỉnh vô cùng cao và hẹp của khối xác suất tại $x = \mu$.

Ta thường sử dụng phân phối delta Dirac như một thành phần của một *phân phối thực nghiệm* (*empirical distribution*),

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}) \qquad (3.28)$$

Phân phối này gán một khối lượng $\frac{1}{m}$ cho mỗi điểm trong tập m các điểm mẫu $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}$. Phân phối delta Dirac chỉ cần thiết khi ta cần định nghĩa phân phối thực nghiệm trên các biến liên tục. Đối với biến rời rạc thì đơn giản hơn: một phân phối thực nghiệm có thể được coi như một phân phối phạm trù, với xác suất gán cho mỗi giá trị đầu vào đơn giản là bằng với *tần số thực nghiệm* (empirical frequency) của giá trị đó trong tập huấn luyện.

Ta có thể coi phân phối thực nghiệm hình thành từ tập huấn luyện như là cách ta xác định phân phối để lấy mẫu khi ta huấn luyện một mô hình trên tập dữ liệu này. Một quan điểm quan trọng khác về phân phối thực nghiệm đó là phân khối của dữ liệu học có độ hợp lý lớn nhất (maximum likelihood).

3.9.6 Hỗn hợp của các phân phối

Trong thực tế, người ta cũng hay định nghĩa một phân phối xác suất mới bằng cách kết hợp một số phân phối xác suất đơn giản. Cách thức thường xuyên được sử dụng để kết hợp các phân phối với nhau là xây dựng một *phân phối hỗn hợp* (mixture distribution). Một phân phối hỗn hợp được cấu thành từ nhiều phân phối thành phần. Trong mỗi phép thử, việc lựa chọn phân phối thành phần nào để sinh mẫu được xác định bằng cách lấy mẫu thành phần đó từ một phân phối multinoulli (hay phân phối phạm trù):

$$P(\mathrm{x}) = \sum_{i} P(\mathrm{c} = i)P(\mathrm{x}|\mathrm{c} = i) \qquad (3.29)$$

trong đó $P(\mathrm{c})$ là một phân phối phạm trù của các thành phần tạo nên phân phối hỗn hợp. Ta đã tìm hiểu một ví dụ về phân phối hỗn hợp: phân phối thực nghiệm của các biến giá trị thực là một phân phối hỗn hợp với mỗi thành phần Dirac cho một mẫu huấn luyện.

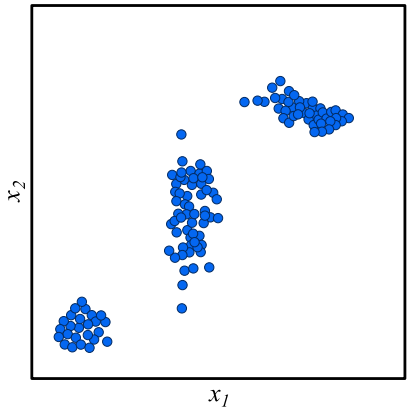
Mô hình hỗn hợp là một chiến lược đơn giản để kết hợp các phân phối xác suất nhằm tạo ra một phân phối phong phú hơn. Ta sẽ khám phá chi tiết hơn về nghệ thuật xây dựng các phân phối xác suất phức tạp từ những phân phối đơn giản trong chương 16.

Mô hình hỗn hợp cũng cho ta có cái nhìn sơ lược về một khái niệm đặc biệt quan trọng sau này - *biến tiềm ẩn* (*latent variable*). Biến tiềm ẩn là một biến ngẫu

nhiên mà ta không thể quan sát trực tiếp. Biến định danh thành phần c của mô hình hỗn hợp phía trên là một ví dụ. Các biến tiềm ẩn có thể có quan hệ với x thông qua phân phối đồng thời (joint distribution), trong trường hợp này, là $P(\mathrm{x}, \mathrm{c}) = P(\mathrm{x}|\mathrm{c})P(\mathrm{c})$. Phân phối $P(\mathrm{c})$ là phân phối của các biến tiềm ẩn và phân phối $P(\mathrm{x}|\mathrm{c})$ thể hiện sự liên hệ giữa biến tiềm ẩn với các biến ta quan sát được. Hai phân phối này xác định hình dạng của phân phối $P(\mathrm{x})$, ngay cả khi ta có thể mô tả $P(\mathrm{x})$ mà không cần tham chiếu tới biến tiềm ẩn. Chúng tôi sẽ thảo luận chi tiết hơn về biến tiềm ẩn ở phần 16.5.

Một mô hình hỗn hợp phổ biến và cực kỳ mạnh là *mô hình Gauss hỗn hợp* (*Gaussian mixture model*), trong đó mỗi thành phần $p(\mathbf{x}|\mathrm{c} = i)$ tuân theo phân phối Gauss, và có tham số kì vọng $\boldsymbol{\mu}^{(i)}$ và hiệp phương sai $\boldsymbol{\Sigma}^{(i)}$ riêng biệt. Một số mô hình hỗn hợp có thể có nhiều ràng buộc hơn nữa. Chẳng hạn, các thành phần có thể có chung hiệp phương sai với nhau thông qua ràng buộc $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}, \forall i$. Giống như trường hợp chỉ có một phân phối Gauss, các mô hình Gauss hỗn hợp có thể có ràng buộc ma trận hiệp phương sai của mỗi thành phần sẽ là ma trận đường chéo hoặc đẳng hướng (*isotropic*).

Ngoài kì vọng và hiệp phương sai, các tham số của một hỗn hợp Gaussian cũng xác định *xác suất tiên nghiệm* (prior probability) $\alpha_i = P(\mathrm{c} = i)$ đối với mỗi thành phần i. Từ "tiên nghiệm" mang ý nghĩa rằng nó thể hiện niềm tin của mô hình đối với c trước khi quan sát $\mathbf{x}$. Ngược lại, $P(\mathrm{c}|\boldsymbol{x})$ được gọi là một *xác suất hậu nghiệm* (posterior probability), bởi nó được tính toán sau khi mô hình đã quan sát $\mathbf{x}$. Một mô hình Gaussian hỗn hợp có thể được coi như một *mô hình xấp xỉ phổ quát* (universal approximator) theo nghĩa nó có thể xấp xỉ một hàm mật độ trơn bất kỳ với sai số tùy ý miễn là số thành phần của nó đủ lớn. Hình 3.2 biểu thị một số phép lấy mẫu từ mô hình Gauss hỗn hợp.



Hình 3.2: Một mẫu lấy từ một mô hình Gauss hỗn hợp. Có ba thành phần trong ví dụ này. Từ trái qua phải, thành phần đầu tiên có ma trận hiệp phương sai đẳng hướng, nghĩa là phương sai của nó có cùng giá trị trên mỗi hướng. Thành phần thứ 2 có ma trận hiệp phương sai chéo, nghĩa là nó có thể kiểm soát phương sai một cách riêng biệt theo hướng của từng trục. Trong ví dụ này, phương sai theo trục x_2 lớn hơn trục x_1. Thành phần thứ 3 có ma trận hiệp phương sai với hạng đầy đủ (full-rank), cho phép nó kiểm soát phương sai một cách riêng biệt theo bất kì hướng nào.

3.10 Những tính chất hữu ích của các hàm thông dụng

Khi làm việc với các phân phối xác suất, đặc biệt là những phân phối xác suất được sử dụng trong các mô hình DL, ta sẽ bắt gặp một số hàm được sử dụng rất thường xuyên.

Một trong số đó là hàm *sigmoid*

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad (3.30)$$

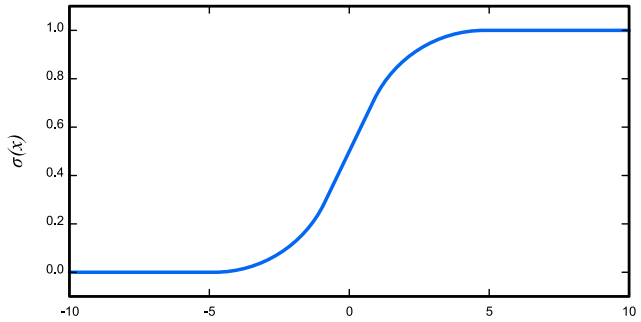
Hàm thường được sử dụng để tính tham số ϕ của một phân phối Bernoulli bởi miền giá trị của nó là $(0, 1)$, nằm trong phạm vi giá trị hợp lệ của tham số ϕ. Hình 3.3 biểu diễn đồ thị của một hàm sigmoid. Hàm sigmoid sẽ **bão hòa**, nghĩa là không bị ảnh hưởng bởi sự thay đổi nhỏ ở đầu vào, khi đối số của nó tiến đến giá trị dương vô cùng hoặc âm vô cùng.

Một hàm số thường được sử dụng khác là **softplus** [Dugas et al., 2001]:

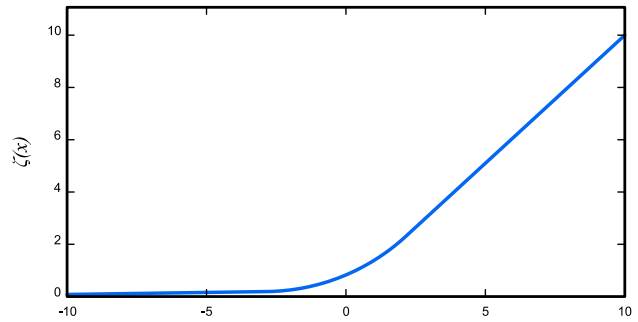
$$\zeta(x) = \log(1 + e^x) \qquad (3.31)$$

Hàm softplus có thể được dùng để tính tham số β và σ của một phân phối chuẩn vì miền giá trị của nó là (0, ∞). Nó cũng thường xuất hiện khi ta làm việc với các biểu thức liên quan đến hàm sigmoid. Cái tên softplus xuất phát từ thực tế rằng nó là phiên bản được làm mượt, hay là một phiên bản "mềm hóa" của

$$x^+ = \max(0, x) \qquad (3.32)$$



> Hình 3.3: Hàm sigmoid



> Hình 3.4: Hàm softplus

Dưới đây là một số tính chất hữu dụng mà bạn đọc được khuyến khích nên học thuộc chúng:

$$\sigma(x) = \frac{e^x}{e^x + e^0} \tag{3.33}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{3.34}$$

$$1 - \sigma(x) = \sigma(-x) \tag{3.35}$$

$$\log \sigma(x) = -\zeta(-x) \tag{3.36}$$

$$\frac{d}{dx}\zeta(x) = \sigma(x) \tag{3.37}$$

$$\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right) \tag{3.38}$$

$$\forall x > 0, \zeta^{-1}(x) = \log(e^x - 1) \tag{3.39}$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy \tag{3.40}$$

$$\zeta(x) - \zeta(-x) = x \tag{3.41}$$

Hàm số $\sigma^{-1}(x)$ còn được gọi là hàm *logit* trong thống kê, nhưng thuật ngữ này ít khi được sử dụng trong ML. Phương trình 3.41 còn cho ta thêm một cách giải

thích cho cái tên "softplus". Hàm softplus là một phiên bản được làm mịn của hàm **phần dương**, $x^+ = \max(0, x)$. Đối ngược với hàm phần dương là hàm **phần âm**, $x^- = \max(0, -x)$. Ta có thể sử dụng $\zeta(-x)$ như là hàm làm mịn cho phần âm. Giống như việc x có thể tìm lại được từ phần dương và phần âm của nó thông qua biểu thức $x^+ - x^- = x$, ta cũng có thể tìm lại x được bằng quan hệ tương tự của $\zeta(x)$ và $\zeta(-x)$ được chỉ ra trong phương trình 3.41.

3.11 Quy tắc Bayes

Một tình huống thường thấy là khi ta đã có $P(\mathrm{y}|\mathrm{x})$ và muốn tính $P(\mathrm{x}|\mathrm{y})$. Nếu đã biết $P(\mathrm{x})$ thì ta có thể tính giá trị xác suất này bằng *quy tắc Bayes* (Bayes' rule):

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})} \qquad (3.42)$$

Lưu ý rằng dù $P(\mathrm{y})$ xuất hiện trong công thức, nhưng ta không cần biết trước nó vì ta có thể tính qua công thức: $P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x)P(x)$.

Mặc dù quy tắc Bayes có thể dễ dàng suy ra từ định nghĩa của xác suất có điều kiện, nhưng việc biết tên công thức này là cần thiết bởi nhiều tài liệu đề cập đến cái tên này. Nó được đặt theo tên của Reverend Thomas Bayes, người đầu tiên phát hiện ra một trường hợp đặc biệt của công thức này. Phiên bản tổng quát vừa giới thiệu ở trên cũng được độc lập phát hiện bởi Pierre-Simon Laplace.

3.12 Chi tiết kỹ thuật về biến số liên tục

Để hiểu đúng về bản chất các biến ngẫu nhiên liên tục và các hàm mật độ xác suất, ta phải phát triển lý thuyết xác suất từ một nhánh của toán học gọi là *lý thuyết độ đo*. Lý thuyết độ đo nằm ngoài phạm vi của sách này, nhưng ta có thể phác họa một số vấn đề mà lý thuyết độ đo có thể giải quyết.

Trong phần 3.3.2, ta thấy rằng xác suất của một vector giá trị liên tục $\mathbf{x}$ thuộc tập hợp $\mathbb{S}$ được tính bằng tích phân của $p(\boldsymbol{x})$ trên tập $\mathbb{S}$. Tuy nhiên, một vài lựa chọn của tập $\mathbb{S}$ có thể tạo ra nghịch lý. Ví dụ, ta có thể tạo ra hai tập hợp $\mathbb{S}_1$ và $\mathbb{S}_2$ sao cho $p(\mathbf{x} \in \mathbb{S}_1) + p(\mathbf{x} \in \mathbb{S}_2) > 1$ nhưng $\mathbb{S}_1 \cap \mathbb{S}_2 = \emptyset$. Những tập hợp này nói chung

được xây dựng để tận dụng độ chính xác vô hạn của số thực, chẳng hạn, bằng cách tạo ra những tập hợp hình fractan (fractal-shaped) hay những tập hợp là biến đổi tập các số hữu tỉ. (Xem thêm về định lý Banach-Tarski.) Một trong những đóng góp quan trọng của lý thuyết độ đo là nó cung cấp một đặc tính của tập hợp để ta có thể tính xác suất mà không gặp phải các nghịch lý kiểu này. Trong cuốn sách này, ta chỉ tính tích phân trên các tập tương đối đơn giản, do đó khía cạnh này của lý thuyết độ đo không phải là một mối quan tâm lớn.

Đối với mục đích của ta, lý thuyết độ đo chủ yếu dùng để mô tả những định lý được áp dụng cho đa số những điểm trong không gian $\mathbb{R}^n$ ngoại trừ một số trường hợp đặc biệt (corner cases). Thuyết độ đo cung cấp một phương pháp để mô tả một cách chặt chẽ một tập điểm như thế nào là nhỏ không đáng kể. Một tập như vậy được gọi là có **độ đo** 0 (measure zero). Trong cuốn sách này, chúng tôi không mô tả tỉ mỉ khái niệm này. Đối với ta, có thể hiểu một cách trực quan rằng một tập có độ đo 0 sẽ không chiếm dung lượng trong không gian mà ta đo lường. Ví dụ, trong không gian $\mathbb{R}^2$, một đường thẳng có độ đo 0, trong khi một hình đa giác (tính cả phần diễn tích bên trong) có độ đo dương. Tương tự, một điểm có độ đo 0. Hợp của một số đếm được các tập hợp, trong đó mỗi tập hợp là độ đo 0, cũng sẽ có độ đo 0 (vì thế tập hợp tất cả các số hữu tỉ cũng có độ đo 0).

Một thuật ngữ hữu dụng khác xuất phát từ lý thuyết độ đo mang tên *hầu hết mọi nơi* (almost everywhere). Một thuộc tính mà tồn tại ở hầu hết mọi nơi sẽ tồn tại trong toàn bộ không gian, ngoại trừ trên tập hợp có độ đo 0. Bởi vì các ngoại lệ chỉ xảy ra trên một không gian không đáng kể, nên chúng hoàn toàn có thể được bỏ qua trong nhiều trường hợp. Một vài kết quả quan trọng trong lý thuyết xác suất đúng cho mọi giá trị rời rạc nhưng chỉ đúng trong "hầu hết mọi nơi" trong trường hợp liên tục.

Một chi tiết mang tính kỹ thuật khác của các biến liên tục liên quan tới việc xử lý biến ngẫu nhiên liên tục là một hàm tất định của một biến khác. Giả sử ta có hai biến ngẫu nhiên, $\mathbf{x}$ và $\mathbf{y}$, biết rằng $\boldsymbol{y} = g(\boldsymbol{x})$ trong đó g là khả nghịch, liên tục và khả vi. Ta có thể dự đoán rằng $p_y(\boldsymbol{y}) = p_x(g^{-1}(\boldsymbol{y}))$. Thực tế đây là một dự đoán sai.

Xét một ví dụ đơn giản sau, giả sử ta có hai biến ngẫu nhiên vô hướng x và y. Giả sử $\mathrm{y} = \frac{\mathrm{x}}{2}$ và $\mathrm{x} \sim U(0, 1)$. Nếu ta sử dụng luật $p_y(y) = p_x(2y)$ thì p_y sẽ bằng 0 mọi nơi ngoại trừ đoạn $[0, \frac{1}{2}]$, và nó sẽ bằng 1 trên đoạn này. Điều đó có nghĩa là

$$\int p_y(y)dy = \frac{1}{2} \qquad (3.43)$$

vi phạm định nghĩa của một phân phối xác suất, và đây là một sai lầm thường gặp. Vấn đề của cách tiếp cận trên là nó không tính đến sự biến dạng của không gian do hàm g gây ra. Nhắc lại, xác suất của $\boldsymbol{x}$ trên một vùng vi phân nhỏ với kích thước $\delta\boldsymbol{x}$ là $p(\boldsymbol{x})\delta\boldsymbol{x}$. Và bởi vì hàm g có thể mở rộng hoặc thu hẹp không gian, thể tích vùng vi phân bao quanh $\boldsymbol{x}$ trong không gian của $\boldsymbol{x}$ có thể khác thể tích trong không gian của $\boldsymbol{y}$.

Để tìm hiểu cách khắc phục vấn đề này, ta trở lại trường hợp đối với biến vô hướng. Ta cần bảo toàn tính chất:

$$|p_y(g(x))dy| = |p_x(x)dx| \qquad (3.44)$$

Từ đó ta có:

$$p_y(y) = p_x(g^{-1}(y)) \left|\frac{\partial x}{\partial y}\right| \qquad (3.45)$$

Hay tương đương với:

$$p_x(x) = p_y(g(x)) \left|\frac{\partial g(x)}{\partial x}\right| \qquad (3.46)$$

Trong không gian nhiều chiều hơn, đạo hàm được tổng quát hóa thành định thức của *ma trận Jacobi* -- ma trận với $J_{i,j} = \frac{\partial x_i}{\partial y_j}$. Do đó, với vector giá trị thực $\boldsymbol{x}$ và $\boldsymbol{y}$,

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left|\det\left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right| \qquad (3.47)$$

3.13 Lý thuyết thông tin

Lý thuyết thông tin là một nhánh của toán ứng dụng xoay quanh việc xác định lượng thông tin chứa trong trong một tín hiệu. Ban đầu, nó được phát minh trong lĩnh vực điện tín để nghiên cứu việc gửi tin nhắn gồm các chữ cái đơn lẻ qua một kênh truyền nhiễu, chẳng hạn như tryền tin qua đường truyền vô tuyến. Trong lĩnh vực này, lý thuyết thông tin cho ta biết cách thiết kế các đoạn mã tối ưu và tính toán độ dài kì vọng của các tin nhắn được lấy mẫu từ những phân phối xác suất sử dụng nhiều cơ chế mã hoá khác nhau. Đối với ML, ta có thể áp dụng lý thuyết thông tin lên các biến liên tục mà ở đó cách ta diễn giải độ dài các tin nhắn này không áp dụng được. Lý tuyết thông tin là nền tảng cơ bản trong nhiều lĩnh vực kỹ thuật điện và khoa học máy tính. Trong cuốn sách này, chúng tôi chủ

yếu sử dụng một số ý tưởng chính từ lý thuyết thông tin để đặc tả các phân phối xác suất hoặc lượng hóa mức độ tương đồng giữa các phân phối xác suất. Bạn đọc có thể tìm hiểu chi tiết hơn về lý thuyết thông tin ở sách của Cover and Thomas (2006) hoặc MacKay (2003).

Một cách trực quan, ý tưởng cơ bản ẩn sau lý thuyết thông tin là những sự kiện ít có khả năng xảy ra mang lại nhiều thông tin hơn so với những sự kiện có gần như sắp xảy ra. Chẳng hạn, thông điệp "mặt trời mọc vào buổi sáng" không mang lại nhiều thông tin và không cần thiết để gửi, nhưng một thông điệp nói rằng "có nhật thực vào sáng nay" lại rất giàu thông tin.

ta muốn lượng hóa thông tin theo cách thể hiện được những nhận định trên.

- Một sự kiện có nhiều khả năng xảy ra thì chứa đựng ít thông tin, trong trường hợp cực đoan, một sự kiện chắc chắn xảy ra sẽ không mang bất kì thông tin nào.
- Một sự kiện ít có khả năng xảy ra sẽ mang lại nhiều thông tin hơn.
- Các sự kiện độc lập mang thông tin bổ sung. Chẳng hạn, quan sát thấy một đồng xu được tung lên xuất hiện mặt ngửa hai lần sẽ truyền đạt lượng thông tin gấp đôi so với việc quan sát thấy đồng xu được tung lên xuất hiện mặt ngửa một lần.

Để thỏa mãn cả ba tính chất trên, ta định nghĩa một hàm *tự thông tin* (*self-information*) của một sự kiện $\mathrm{x} = x$:

$$I(x) = -\log P(x) \qquad (3.48)$$

Trong cuốn sách này, ta luôn sử dụng log theo nghĩa là logarit tự nhiên cơ số e. Do đó, theo cách ta nghĩa định nghĩa, hàm tự thông tin $I(x)$ sẽ có đơn vị *nat*. Một nat là một lượng thông tin thu được bằng việc quan sát một sự kiện có xác xuất $\frac{1}{e}$. Một số tài liệu khác sử dụng logarit cơ số 2 và đơn vị của nó do đó được gọi là *bit* hoặc *shannon*; thông tin được đo lường trong các bit chỉ là một chuyển đổi tỉ lệ thuận từ thông tin đã được đo lường bởi nat.

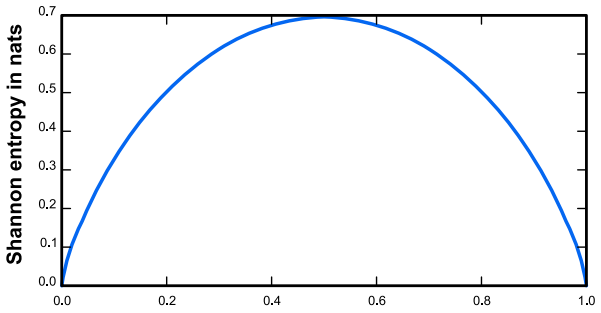
Khi x là liên tục, ta định nghĩa thông tin một cách tương tự, tuy nhiên một số tính chất trong trường hợp rời rạc sẽ bị mất đi. Chẳng hạn, một sự kiện có mật độ đơn vị mang lượng thông tin bằng 0, kể cả khi nó không phải là một sự kiện được đảm bảo sẽ xảy ra.

Hàm tự thông tin chỉ dành cho một kết quả đầu ra duy nhất. Ta có thể lượng hóa sự bất định trong toàn bộ một phân phối xác suất bằng cách sử dụng hàm

Shannon entropy, được kí hiệu là $H(P)$:

$$H(x) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)] \qquad (3.49)$$

Nói cách khác, Shannon entropy của một phân phối là lượng thông tin kì vọng của một sự kiện được lấy ngẫu nhiên ra từ phân phối đó. Nó là cận dưới của số lượng bit (trong trường hợp logarit cơ số 2, với cơ số khác ta sẽ dùng đơn vị khác) trung bình cần thiết để mã hóa các ký tự được lấy ngẫu nhiên ra từ phân phối P. Các phân phối gần như xác định (nghĩa là kết quả gần như chắc chắn xác định) có entropy thấp; các phân phối gần như đồng đều sẽ có entropy cao. Hình 3.5 giải thích rõ hơn về nhận định này. Khi x là liên tục, Shannon entropy còn được biết đến với tên gọi *entropy vi phân* (differential entropy).



> Hình 3.5: Shannon entropy của một biến ngẫu nhiên nhị phân. Đồ thị này chỉ ra lí do các phân phối gần như xác định sẽ có Shannon Entropy thấp, trong khi phân phối gần như đồng đều có Shannon entropy cao. Trên trục hoành, ta biểu diễn p, thể hiện xác suất biến ngẫu nhiên nhị phân đó có giá trị bằng 1. Entropy được tính bởi $(p-1)\log(1-p)$–$p\log p$. Khi p gần với 0, phân phối gần như xác định, bởi biến ngẫu nhiên gần như chắc chắn bằng 0. Điều tương tự xảy ra khi p gần với 1, bởi biến ngẫu nhiên đó gần như chắc chắn bằng 1. Khi $p = 0.5$, entropy đạt cực đại, do phân phối là đồng đều đối với hai kết quả có thể xảy ra.

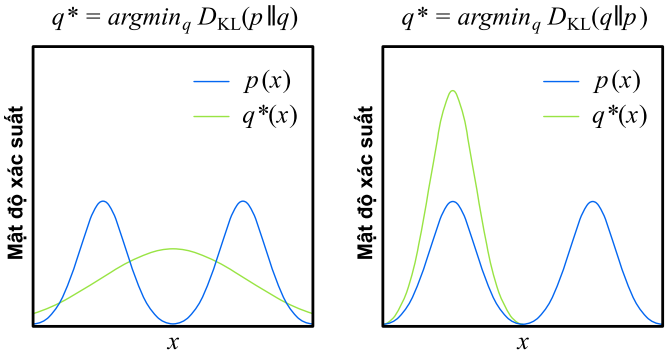
Nếu ta có hai phân phối $P(\mathrm{x})$ và $Q(\mathrm{x})$ riêng biệt của cùng một biến ngẫu nhiên x, ta có thể đo lường sự khác biệt của hai phân phối này bằng cách sử dụng *độ phân kỳ Kullback-Leibler (KL)* (KL divergence):

$$D_{KL}(P||Q) = \mathbb{E}_{\mathrm{x}\sim P}[\log \frac{P(\mathrm{x})}{Q(\mathrm{x})}] = \mathbb{E}_{\mathrm{x}\sim P}[\log P(\mathrm{x}) - \log Q(\mathrm{x})] \qquad (3.50)$$

Trong trường hợp các biến là rời rạc, đây là lượng thông tin thêm (được đo lường bằng các bit nếu ta sử dụng logarit cơ số 2, nhưng trong ML ta thường sử dụng nat và logarit tự nhiên) cần thiết để gửi một thông điệp chứa các kí tự được lấy

ngẫu nhiên ra từ phân phối xác xuất P, khi ta sử dụng một bộ mã được thiết kế để tối thiểu hóa độ dài của thông điệp được lấy ngẫu nhiên ra từ phân phối xác xuất Q.

Độ phân kỳ KL có nhiều tính chất hữu ích, đáng chú ý nhất là nó không âm. Độ phân kỳ KL có giá trị bằng 0 khi và chỉ khi P và Q là cùng một phân phối trong trường hợp các biến là rời rạc, hoặc bằng nhau ở "hầu hết mọi nơi" trong trường hợp các biến liên tục. Bởi độ phân kỳ KL là không âm và được dùng để đo lường sự khác biệt giữa hai phân phối, nó thường được hiểu như để đo một dạng khoảng cách giữa những phân phối này. Nó không phải là một độ đo khoảng cách thực sự bởi vì nó không có tính đối xứng: $D_{KL}(P||Q) \neq D_{KL}(Q||L)$ đối với một số phân phối P và Q. Tính phi đối xứng này đồng nghĩa với có sự khác biệt quan trọng trong quyết định lựa chọn giữa $D_{KL}(P||Q)$ hay $D_{KL}(Q||P)$. Xem hình 3.6 để biết thêm chi tiết.



> Hình 3.6: Độ phân kỳ KL là bất đối xứng. Giả sử ta có một phân phối $p(x)$ và muốn xấp xỉ nó với một phân phối $q(x)$ khác. Ta có thể chọn q sao cho hàm $D_{KL}(p||q)$ hoặc $D_{KL}(q||p)$ nhỏ nhất. Chúng tôi minh hoạ ảnh hưởng của cách làm này bằng cách sử dụng một hỗn hợp của hai phân phối Gauss cho p, và chỉ một phân phối Gauss cho q. Cách chọn phương của khoảng cách KL phụ thuộc vào từng bài toán cụ thể. Một số ứng dụng yêu cầu một cách xấp xỉ có xác suất cao ở bất cứ vị trí nào phân phối thực có xác suất cao, trong khi các ứng dụng khác yêu cầu một cách xấp xỉ hiếm khi đặt xác suất cao ở bất cứ vị trí nào phân phối xác xuất thực có xác suất thấp. Cách chọn phương của độ phân kỳ KL sẽ thể hiện ta ưu tiên lựa chọn nào trong hai lựa chọn trên trong mỗi ứng dụng. (*Bên trái*) Ảnh hưởng khi ta chọn cách cực tiểu hóa $D_{KL}(p||q)$. Trong trường hợp này, ta lựa chọn q có xác suất cao tại những vị trí mà p có xác suất cao. Khi p có nhiều mode, q được chọn để làm mờ chúng, nhằm mục đích đặt một khối lượng xác suất cao lên các mode này. (*Bên phải*) ảnh hưởng khi cực tiểu hóa $D_{KL}(q||p)$.

> Trong trường hợp này, ta lựa chọn q sao cho có xác suất thấp tại những vị trí mà p có xác suất thấp. Khi p có những mode khác nhau được tách biệt đủ rõ ràng, giống như trong hình, độ phân kỳ KL nhỏ nhất khi ta chọn chỉ một mode, nhằm tránh đặt khối lượng xác suất vào những vùng có xác suất thấp giữa các mode của p. Ở đây, chúng tôi minh hoạ kết quả khi q được chọn để nhắm vào mode bên trái. Ta cũng có thể đạt được giá trị tương đương của khoảng cách KL khi nhắm vào mode bên phải. Nếu các mode không được phân cách bằng một vùng xác suất đủ thấp, thì phương ta chọn này của độ phân kỳ KL cũng sẽ làm mờ các mode.

Một đại lượng có liên quan chặt chẽ với độ phân kỳ KL là *entropy chéo* (cross-entropy) $H(P, Q) = H(P) + D_{KL}(P||Q)$, tương tự như độ phân kỳ KL khi thiếu số hạng bên trái:

$$H(P, Q) = -\mathbb{E}_{\mathrm{x}\sim P} \log Q(x) \qquad (3.51)$$

Tối thiểu hóa hàm cross-entropy đối với Q tương đương với việc tối thiểu hóa độ phân kỳ KL, bởi Q không tham gia vào số hạng bị thiếu.

Khi tính toán những đại lượng này thường xuyên, ta thường gặp phải những biểu thức có dạng $0 \log\ 0$. Theo quy ước trong lý thuyết thông tin, ta có thể coi những biểu thức này như là $\lim_{x\to 0} x\ \log x = 0$

3.14 Các mô hình xác suất có cấu trúc

Các thuật toán ML thường liên quan đến những phân phối xác suất trên một số lượng lớn các biến ngẫu nhiên. Những phân phối xác suất này thường liên quan đến các tương tác trực tiếp giữa một số lượng biến tương đối nhỏ với nhau. Do đó việc sử dụng một hàm đơn lẻ để biểu diễn toàn bộ đồng phân phối xác suất (joint probability distribution) có thể rất kém hiệu quả (cả về mặt thống kê lẫn tính toán).

Thay vì sử dụng một hàm đơn lẻ để biểu diễn phân phối xác suất, ta có thể phân chia một phân phối xác suất thành tích của nhiều thành phần. Chẳng hạn, giả sử ta có ba biến ngẫu nhiên: a, b và c. Giả sử rằng a ảnh hưởng đến giá trị của b, và b ảnh hưởng đến giá trị của c, nhưng a và c là độc lập khi biết b. ta có thể

biểu diễn phân phối xác suất trên cả ba biến bằng tích của các phân phối xác suất trên hai biến:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a})p(\mathrm{b}|\mathrm{a})p(\mathrm{c}|\mathrm{b}) \qquad (3.52)$$

Phép phân tích như trên có thể làm giảm đáng kể số lượng tham số cần thiết để biểu diễn phân phối. Số lượng tham số sử dụng cho mỗi thành phần tỉ lệ với số lượng biến của phần đó theo hàm mũ. Điều đó có nghĩa là ta có thể giảm chi phí biểu diễn một phân phối xuống rất nhiều nếu ta có thể tìm được một cách phân tích nó thành những phân phối nhỏ trên ít biến hơn.

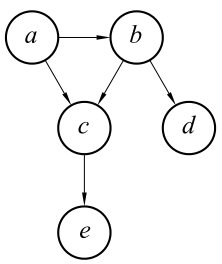
Ta có thể sử dụng đồ thị để minh hoạ cho việc biểu diễn một phân phối thành tích của các thành phần như trên. Ở đây, chúng tôi sử dụng từ "đồ thị" trong lý thuyết đồ thị: là một tập các đỉnh có thể được kết nối với nhau bởi các cạnh. Chúng tôi gọi việc biểu diễn phép phân rã của một xác suất bằng đồ thị là *mô hình xác suất có cấu trúc* (structured probabilistic model) hoặc là *mô hình đồ thị* (graphical model).

Có hai loại mô hình xác suất có cấu trúc chính: có hướng và vô hướng. Cả hai loại mô hình đồ thị này sử dụng một đồ thị $\mathcal{G}$ trong đó mỗi nút (node) trong đồ thị tương ứng với một biến ngẫu nhiên, và một cạnh kết nối hai biến ngẫu nhiên có nghĩa là có thể biển diễn sự tương tác trực tiếp của chúng với nhau bằng phân phối xác suất tương ứng.

Mô hình *có hướng* sử dụng những đồ thị với các cạnh có hướng, và chúng biểu diễn phép phân rã thành các phân phối xác suất có điều kiện, giống như trong ví dụ phía trên. Cụ thể hơn, một mô hình có hướng sẽ có một nút cho mỗi biến x_i trong phân phối, và nút đó tương ứng với phân phối có điều kiện của x_i khi biết các biến cha mẹ của x_i trong đồ thị, kí hiệu là $Pa_{\mathcal{G}}(x_i)$:

$$p(\mathbf{x}) = \prod_i p(\mathrm{x}_i|Pa_{\mathcal{G}}(\mathrm{x}_i)) \qquad (3.53)$$

Hình 3.7 là một ví dụ về một đồ thị có hướng và phép phân rã của phân phối xác suất mà đồ thị đó biểu diễn.



Hình 3.7: Một mô hình đồ thị có hướng đối với các biến ngẫu nhiên $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}$. Đồ thị này tương ứng với phân phối xác suất có phân rã như sau:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b}|\mathrm{a})p(\mathrm{c}|\mathrm{a}, \mathrm{b})p(\mathrm{d}|\mathrm{b})p(\mathrm{e}|\mathrm{c}) \qquad (3.54)$$

Mô hình đồ thị này cho phép ta nhanh chóng nhận thấy một số tính chất của phân phối. Chẳng hạn, a và c có tương tác trực tiếp, nhưng a và e chỉ tương tác gián tiếp thông qua c.

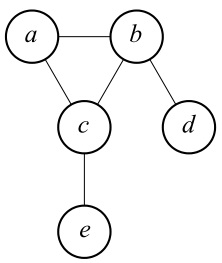
Mô hình *vô hướng* sử dụng đồ thị với các cạnh vô hướng, và chúng biểu diễn phân rã bằng một tập các hàm; không giống như trường hợp có hướng, các hàm này thường không phải là phân phối xác suất. Một tập hợp các nút mà mọi nút trong nó đều có liên kết với nhau trong $\mathcal{G}$ gọi là một *bè* (clique). Mỗi bè $\mathcal{C}^{(i)}$ trong một mô hình vô hướng tương ứng với một thành phần $\phi^{(i)}(\mathcal{C}^{(i)})$. Những thành phần này chỉ là các hàm, không phải phân phối xác suất. Kết quả đầu ra của một thành phần phải là không âm, nhưng không có ràng buộc rằng tổng các thành phần hoặc tích các thành phần là 1 giống như phân phối xác suất.

> ***ND***: Bè (clique) là một khái niệm trong lý thuyết đồ thị để chỉ một đồ thị con đầy đủ: hai đỉnh bất kì đều được nối bằng một cạnh. Xem thêm tại wikipedia.

Xác suất của một cấu hình các biến ngẫu nhiên là **tỉ lệ** với tích của tất cả các hệ số -- cấu hình có tích lớn hơn sẽ có khả năng xảy ra cao hơn. Tất nhiên sẽ có trường hợp tổng của các tích này không bằng 1. Ta chia tích này cho một số hệ số chuẩn hoá Z không đổi, với Z là tổng hoặc tích phân của tất cả các hàm ϕ, để thu được phân phối xác suất đã chuẩn hoá:

$$p(x) = \frac{1}{Z} \prod_i \Phi^i(\mathcal{C}^i) \qquad (3.55)$$

Hình 3.8 cho thấy một ví dụ về đồ thị vô hướng và phép phân rã của phân phối xác suất mà nó biểu diễn.



> Hình 3.8: Một mô hình đồ thị vô hướng đối với các biến ngẫu nhiên $\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}$. Đồ thị này tương ứng với các phân phối xác suất có thể phân rã

như sau:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z}^{1} (\mathrm{a}, \mathrm{b}, \mathrm{c})^{2} (\mathrm{b}, \mathrm{d})^{3} (\mathrm{c}, \mathrm{e}) \qquad (3.56)$$

Mô hình đồ thị này cũng cho phép ta nhanh chóng nhận ra một số tính chất của phân phối giống như trong trường hợp có hướng.

Nên nhớ rằng biểu diễn đồ thị của các phép phân rã này là một ngôn ngữ giúp ta miêu tả các phân phối xác suất. Có hướng hay vô hướng không phải là một thuộc tính của phân phối xác suất; đó là tính chất của một cách mô tả cụ thể của một phân phối xác suất, nhưng mọi phân phối xác suất đều có thể được mô tả theo cả hai cách.

Trong suốt phần I và II của cuốn sách này, chúng tôi sử dụng các mô hình xác suất có cấu trúc đơn thuần như là một ngôn ngữ để biểu diễn casc mối quan hệ xác suất trực tiếp xuất hiện trong các thuật toán ML khác nhau. Bạn đọc không cần thiết phải hiểu sâu hơn về mô hình xác suất có cấu trúc cho đến khi ta thảo luận về các chủ đề nghiên cứu trong phần III, nơi chúng tôi sẽ trình bày các mô hình xác suất có cấu trúc chi tiết hơn nhiều.

Chương này đã tóm lược những khái niệm cơ bản của lý thuyết xác suất có liên quan mật thiết nhất với DL. ta sẽ khám phá một bộ công cụ toán học cơ bản còn lại trong chương tiếp theo.