Chương 3

Xác suất và lý thuyết thông tin

Chương này mô tả lý thuyết xác suất và lý thuyết thông tin.

Lý thuyết xác suất là một công cụ toán học để biểu diễn các phát biểu không chắc chắn. Nó cung cấp một phương tiện để định lượng tính không chắc chắn và các tiên đề để suy diễn các phát biểu không chắc chắn mới. Trong các ứng dụng trí tuệ nhân tạo, ta sử dụng lý thuyết xác suất theo hai cách chính. Thứ nhất, các quy luật của xác suất cho ta biết cách các hệ thống trí tuệ nhân tạo nên suy luận, vì vậy ta thiết kế các thuật toán để tính toán hoặc xấp xỉ các biểu thức khác nhau

được suy ra từ lý thuyết xác suất. Thứ hai, ta có thể sử dụng xác suất và thống kê để phân tích lý thuyết về hành vi của các hệ thống trí tuệ nhân tạo được đề xuất.

Lý thuyết xác suất là một công cụ nền tảng của nhiều lĩnh vực khoa học và kỹ thuật. Chương này sẽ giúp người đọc có nền tảng chủ yếu về kỹ thuật phần mềm và ít tiếp xúc với lý thuyết xác suất có thể hiểu được nội dung trong cuốn sách.

Trong khi lý thuyết xác suất cho phép ta đưa ra các phát biểu không chắc chắn và suy luận trong điều kiện có sự không chắc chắn, thì lý thuyết thông tin cho phép ta định lượng mức độ không chắc chắn bởi một phân phối xác suất.

Nếu bạn đã quen thuộc với lý thuyết xác suất và lý thuyết thông tin, bạn có thể bỏ qua toàn bộ chương này, ngoại trừ Mục 3.14, phần mô tả các biểu đồ được sử dụng để mô tả các mô hình xác suất có cấu trúc trong học máy. Nếu bạn hoàn toàn chưa có kinh nghiệm với các chủ đề này, chương này sẽ đủ để bạn thực hiện thành công các dự án nghiên cứu về học sâu, nhưng bạn nên tham khảo thêm tài liệu khác, chẳng hạn như *Probability Theory: The Logic of Science* (Jaynes, 2003, [6]).

3.1 Tại sao lại là xác suất?

Nhiều nhánh của khoa học máy tính chủ yếu xử lý các thực thể hoàn toàn mang tính xác định và chắc chắn. Một lập trình viên thường có thể yên tâm giả định rằng CPU sẽ thực thi mỗi lệnh máy một cách hoàn hảo. Lỗi phần cứng có xảy ra, nhưng đủ hiếm để hầu hết các ứng dụng phần mềm không cần phải được thiết kế để xử lý chúng. Vì nhiều nhà khoa học máy tính và kỹ sư phần mềm làm việc trong môi trường tương đối rõ ràng và chắc chắn, nên có thể gây ngạc nhiên khi học máy lại sử dụng nhiều lý thuyết xác suất.

Điều này là do học máy luôn phải xử lý các đại lượng không chắc chắn, và đôi khi cũng cần xử lý các đại lượng ngẫu nhiên (không mang tính xác định). Tính không chắc chắn và tính ngẫu nhiên có thể phát sinh từ nhiều nguồn khác nhau. Các nhà nghiên cứu đã đưa ra những lập luận thuyết phục về việc lượng hóa tính không chắc chắn bằng cách sử dụng xác suất ít nhất là từ những năm 1980. Nhiều lập luận được trình bày ở đây là tóm tắt hoặc lấy cảm hứng từ *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Pearl 1988, [7]).

Hầu như mọi hoạt động đều đòi hỏi một khả năng nào đó để suy luận trong điều kiện không chắc chắn. Thực tế, ngoài những mệnh đề toán học đúng theo định nghĩa, rất khó để nghĩ đến bất kỳ mệnh đề nào hoàn toàn đúng hoặc bất kỳ sự kiện nào chắc chắn sẽ xảy ra.

Có ba nguồn gốc có thể gây ra tính không chắc chắn:

1. Tính ngẫu nhiên vốn có trong hệ thống được mô hình hóa. Ví dụ, hầu hết các cách diễn giải về cơ học lượng tử mô tả động lực của các hạt hạ nguyên tử là mang tính xác suất. Ta cũng có thể tạo ra các kịch bản lý thuyết mà ta giả định có động lực ngẫu nhiên, chẳng hạn như một trò chơi bài giả định, trong đó ta cho rằng các lá bài được xáo trộn ngẫu nhiên thực sự.

2. Sự quan sát không đầy đủ. Ngay cả các hệ thống mang tính xác định cũng có thể xuất hiện yếu tố ngẫu nhiên khi ta không thể quan sát tất cả các biến số điều khiển hành vi của hệ thống. Ví dụ, trong bài toán Monty Hall, một thí sinh trong chương trình trò chơi được yêu cầu chọn giữa ba cánh cửa và giành giải thưởng đằng sau cánh cửa được chọn. Hai cánh cửa dẫn đến một con dê, trong khi cánh cửa thứ ba dẫn đến một chiếc xe hơi. Kết quả dựa trên lựa chọn của thí sinh là mang tính xác định, nhưng từ góc nhìn của thí sinh, kết quả là không chắc chắn.

3. Mô hình hóa không đầy đủ. Khi sử dụng một mô hình phải loại bỏ một phần thông tin đã quan sát được, thông tin bị loại bỏ sẽ dẫn đến sự không chắc chắn trong dự đoán của mô hình. Ví dụ, giả sử ta chế tạo một robot có thể quan sát chính xác vị trí của mọi vật thể xung quanh. Nếu robot phân chia không gian thành các ô rời rạc khi dự đoán vị trí tương lai của các vật thể này, thì việc rời rạc hóa sẽ khiến robot ngay lập tức trở nên không chắc chắn về vị trí chính xác của các vật thể: mỗi vật thể có thể nằm ở bất kỳ đâu trong ô rời rạc mà nó chiếm giữ theo như được quan sát.

Trong nhiều trường hợp, sử dụng một quy tắc đơn giản nhưng không chắc chắn lại thực tế hơn là sử dụng một quy tắc phức tạp nhưng chắc chắn, ngay cả khi quy tắc thực sự mang tính xác định và hệ thống mô hình của ta có khả năng đáp ứng quy tắc phức tạp. Ví dụ, quy tắc đơn giản “Hầu hết các loài chim đều bay” rất dễ phát triển và có tính ứng dụng rộng rãi, trong khi một quy tắc có dạng “Chim bay, ngoại trừ chim non chưa học bay, chim bệnh hoặc bị thương đã mất khả năng bay, các loài chim không bay như chim cánh cụt, đà điểu châu Phi và chim kiwi. . . ” sẽ tốn kém để phát triển, duy trì và truyền đạt, và sau tất cả những nỗ lực này, quy tắc đó vẫn rất mong manh và dễ thất bại.

Mặc dù rõ ràng là ta cần một phương tiện để biểu diễn và suy luận về tính không chắc chắn, nhưng không phải ngay lập tức ta có thể nhận ra rằng lý thuyết xác suất cung cấp tất cả các công cụ mà ta mong muốn cho các ứng dụng trí tuệ

nhân tạo. Lý thuyết xác suất ban đầu được phát triển để phân tích tần suất của các sự kiện. Rất dễ thấy lý thuyết xác suất có thể được sử dụng để nghiên cứu các phép thử như rút một quân bài cụ thể trong trò chơi poker. Những loại phép thử này thường có thể lặp lại. Khi ta nói rằng một kết quả có xác suất xảy ra là p, điều đó có nghĩa là nếu ta lặp lại phép thử (ví dụ, rút một quân bài) vô số lần, thì có một tỉ lệ p của các lần thử sẽ dẫn đến kết quả đó. Kiểu suy luận này dường như không phải ngay lập tức có thể áp dụng cho các mệnh đề không lặp lại. Nếu một bác sĩ phân tích bệnh nhân và nói rằng bệnh nhân có 40% khả năng bị cúm, điều này có ý nghĩa rất khác – ta không thể tạo ra vô số bản sao của bệnh nhân, và cũng không có lý do gì để tin rằng các bản sao khác nhau của bệnh nhân sẽ có cùng triệu chứng nhưng lại có các tình trạng nền khác nhau. Trong trường hợp bác sĩ chẩn đoán bệnh nhân, ta sử dụng xác suất để biểu diễn **mức độ tin tưởng**, với 1 biểu thị sự chắc chắn tuyệt đối rằng bệnh nhân bị cúm và 0 biểu thị sự chắc chắn tuyệt đối rằng bệnh nhân không bị cúm. Loại xác suất đầu tiên, liên quan trực tiếp đến tỉ lệ xảy ra của các sự kiện, được gọi là **xác suất theo tần suất**, trong khi loại sau, liên quan đến mức độ chắc chắn định tính, được gọi là **xác suất Bayes**.

Nếu ta liệt kê một số thuộc tính mà ta mong đợi lý luận theo lẽ thường về tính không chắc chắn phải có, thì cách duy nhất để thỏa mãn các thuộc tính đó là xem xác suất Bayes hoạt động giống hệt như xác suất theo tần suất. Ví dụ, nếu ta muốn tính xác suất một người chơi sẽ thắng trong trò chơi poker với một bộ bài nhất định, ta sử dụng chính xác cùng các công thức như khi tính xác suất một bệnh nhân mắc một bệnh nào đó dựa trên các triệu chứng nhất định. Để biết thêm chi tiết về lý do tại sao một tập hợp nhỏ các giả định theo lẽ thường lại dẫn đến việc cùng một hệ tiên đề phải kiểm soát cả hai loại xác suất, xem *The Foundations of Mathematics and other Logical Essays* (Ramsey, 1926, [8]).

Xác suất có thể được xem như sự mở rộng của logic để xử lý tính không chắc chắn. Logic cung cấp một tập hợp các quy tắc hình thức để xác định mệnh đề nào được suy ra là đúng hay sai, với giả định rằng một tập hợp các mệnh đề khác là đúng hoặc sai. Lý thuyết xác suất cung cấp một tập hợp các quy tắc hình thức để xác định khả năng một mệnh đề là hợp lý, dựa trên tính hợp lý của các mệnh đề khác.

3.2 Biến ngẫu nhiên

Một 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 chữ hoa, và các giá trị mà nó có thể nhận bằng các chữ thường. Ví dụ, x_1 và x_2 đều là các giá trị mà biến ngẫu nhiên X có thể nhận. Đối với các biến có giá trị là véctơ, ta sẽ viết biến ngẫu nhiên là $\boldsymbol{X}$ và một trong các giá trị của nó là $\boldsymbol{x}$. Bản thân biến ngẫu nhiên chỉ là mô tả các trạng thái có thể xảy ra; nó phải được kết hợp với một phân phối xác suất để chỉ rõ mức độ khả năng của mỗi trạng thái này.

Biến ngẫu nhiên có thể là rời rạc hoặc liên tục. Một biến ngẫu nhiên rời rạc là biến có một số hữu hạn hoặc vô hạn đếm được các trạng thái. Lưu ý rằng các trạng thái này không nhất thiết phải là các số nguyên; chúng cũng có thể là các trạng thái được đặt tên mà không được coi là có giá trị số. Một biến ngẫu nhiên liên tục nhận giá trị trong cả một khoảng số thực.

3.3 Phân phối xác suất

Phân phối xác suất là một mô tả về mức độ khả năng mà một biến ngẫu nhiên hoặc tập hợp các biến ngẫu nhiên có thể nhận từng trạng thái khả dĩ của chúng. Cách ta mô tả các phân phối xác suất phụ thuộc vào việc các biến đó là rời rạc hay liên tục.

3.3.1 Biến rời rạc và hàm trọng số xác suất

Một phân phối xác suất trên các biến rời rạc có thể được mô tả bằng **hàm trọng số xác suất** (probability mass function – PMF). Ta thường ký hiệu các hàm trọng số xác suất bằng chữ p viết thường. Thường thì ta gán mỗi biến ngẫu nhiên với một hàm trọng số xác suất khác nhau nên trong ngữ cảnh nếu có nhiều biến ngẫu nhiên thì hàm trọng số xác suất của mỗi biến ngẫu nhiên được có thêm ký hiệu biến ngẫu nhiên ở vị trí chỉ số dưới, ví dụ p_X để chỉ hàm trọng số xác suất của biến ngẫu nhiên X.

Hàm trọng số xác suất ánh xạ mỗi trạng thái của biến ngẫu nhiên với xác suất để biến ngẫu nhiên đó đạt tới trạng thái này. Nếu x là một trạng thái của biến ngẫu nhiên X, ta viết $x \in X$. Xác suất $p(x)$ để $X = x$ được ký hiệu là $P(X = x)$, hoặc $P(x)$ nếu trong ngữ cảnh không có biến ngẫu nhiên nào khác ngoài X – ký hiệu này trùng khớp và có ý nghĩa như $p(x)$. Trong các phát biểu với giá trị x tổng quát

của biến ngẫu nhiên X, ta cũng thường viết là $P(X)$. Xác suất này bằng 1 chỉ ra rằng $X = x$ là chắc chắn và xác suất bằng 0 chỉ ra rằng $X = x$ là không thể. Đôi khi ta định nghĩa biến trước, sau đó sử dụng ký hiệu $\sim$ để chỉ rõ biến đó tuân theo phân phối nào: $X \sim p(x)$.

Các hàm trọng số xác suất có thể tác động lên nhiều biến cùng một lúc. Một phân phối xác suất như vậy trên nhiều biến được gọi là **phân phối xác suất đồng thời**. Giá trị $p(x, y)$ biểu thị xác suất để $X = x$ và $Y = y$ đồng thời xảy ra, ký hiệu $P(X = x, Y = y)$. Để đơn giản, ta cũng có thể viết $P(x, y)$, và trong các phát biểu với giá trị x, y tổng quát, ta viết $P(X, Y)$.

Để là một hàm trọng số xác suất trên một biến ngẫu nhiên X, một hàm p phải thỏa mãn các tính chất sau:

- Miền xác định của p phải là tập hợp tất cả các trạng thái có thể có của X.
- $\forall x \in X,\ 0 \leq p(x) \leq 1$. Một sự kiện không thể xảy ra có xác suất bằng 0 và không có trạng thái nào có xác suất nhỏ hơn thế. Tương tự, một sự kiện chắc chắn xảy ra có xác suất bằng 1 và không có trạng thái nào có khả năng xảy ra lớn hơn thế.
- $\sum_{x \in X} p(x) = 1$. Ta gọi tính chất này là **tính chuẩn hóa**. Nếu không có tính chất này, ta có thể thu được xác suất lớn hơn 1 khi tính xác suất của một trong nhiều sự kiện xảy ra.

Ví dụ, xét biến ngẫu nhiên rời rạc X có k trạng thái khác nhau. Ta có thể đặt một phân phối đều trên X – nghĩa là làm cho mỗi trạng thái của nó có xác suất như nhau – bằng cách đặt hàm trọng số xác suất của nó là

$$P(X = x_i) = \frac{1}{k} \tag{3.1}$$

với mọi i. Ta có thể thấy rằng điều này đáp ứng các yêu cầu của một hàm trọng số xác suất. Giá trị $\frac{1}{k}$ là số dương vì k là một số nguyên dương. Ta cũng thấy rằng

$$\sum_i P(X = x_i) = \sum_i \frac{1}{k} = \frac{1}{k} \times k = 1, \tag{3.2}$$

nên phân phối đã thực sự 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 các phân phối xác suất bằng cách sử dụng **hàm mật độ xác suất** (probability density function – PDF) thay vì hàm trọng số xác suất. Để là một hàm mật độ xác suất, một hàm p phải thỏa mãn các tính chất sau:

- Miền xác định của p phải là tập hợp tất cả các trạng thái có thể có của X.
- $\forall x \in X,\ p(x) \geq 0$. Lưu ý rằng không nhất thiết $p(x) \leq 1$.
- $\int p(x)\, dx = 1$, trong đó miền lấy tích phân ở đây không được ghi thì mặc định là tập hợp tất cả các trạng thái có thể có của X.

Một hàm mật độ xác suất $p(x)$ không cung cấp xác suất của một trạng thái cụ thể một cách trực tiếp, thay vào đó là xác suất rơi vào một vùng vô cùng nhỏ có thể tích δx được cho bởi $p(x)\,\delta x$.

Ta có thể tích phân hàm mật độ để tìm xác suất thực tế của một tập hợp các điểm. Cụ thể, xác suất để X nằm trong một tập nào đó được cho bởi tích phân của $p(x)$ trên tập hợp đó. Ví dụ trong trường hợp một biến, xác suất để X nằm trong khoảng $[a, b]$ được cho bởi $\int_{[a,b]} p(x)\, dx$ hay $\int_a^b p(x)\, dx$.

Ví dụ về hàm mật độ xác suất tương ứng với một mật độ xác suất cụ thể trên một biến ngẫu nhiên liên tục, hãy xét phân phối đều trên một khoảng của tập số thực. Ta có thể thực hiện điều này bằng một hàm $u(x; a, b)$, trong đó a và b là các điểm đầu và cuối của khoảng, với $a < b$. Ký hiệu ";" có nghĩa là "được tham số hóa bởi"; ta coi x là đối số của hàm, trong khi a và b là các tham số xác định hàm. Để đảm bảo rằng không có trọng số xác suất nào ngoài khoảng, ta nói $u(x; a, b) = 0$ với mọi $x \notin [a, b]$. Trong khoảng $[a, b]$, $u(x; a, b) = \dfrac{1}{b - a}$. Ta có thể thấy rằng hàm này không âm ở mọi nơi. Ngoài ra, nó có tích phân bằng 1. Ta thường ký hiệu X tuân theo phân phối đều trên $[a, b]$ bằng cách viết $X \sim U(a, b)$.

3.4 Xác suất biên

Đôi khi ta biết phân phối xác suất trên một tập các biến và muốn biết phân phối xác suất chỉ trên một tập con của chúng. Phân phối xác suất trên tập con này được gọi là **phân phối xác suất biên**.

Ví dụ, giả sử ta có các biến ngẫu nhiên rời rạc X và Y, và ta biết $P(X, Y)$. Ta có thể tìm $P(X)$ bằng **quy tắc tổng**:

$$\forall x \in X, \quad P(X = x) = \sum_y P(X = x, Y = y). \tag{3.3}$$

Tên gọi "xác suất biên" xuất phát từ quá trình tính xác suất biên trên giấy. Khi các giá trị của $P(x, y)$ được viết trong một lưới với các giá trị khác nhau của x trong các hàng và các giá trị khác nhau của y trong các cột, thì một cách tự nhiên, ta cộng dồn theo một hàng của lưới, ta tính được và viết $P(x)$ ở lề giấy ngay bên phải của hàng đó.

Đối với biến liên tục, ta cần lấy tích phân thay vì tính tổng:

$$p(x) = \int p(x, y)\, dy. \tag{3.4}$$

3.5 Xác suất có điều kiện

Trong nhiều trường hợp, ta quan tâm đến xác suất của một sự kiện nào đó, với điều kiện một sự kiện khác đã xảy ra. Xác xuất này được gọi là xác suất có điều kiện. Ta ký hiệu xác suất có điều kiện để $Y = y$ khi biết $X = x$ là $P\left(Y = y \mid X = x\right)$. Xác suất có điều kiện này có thể được tính bằng công thức

$$P\left(Y = y \mid X = x\right) = \frac{P(Y = y, X = x)}{P(X = x)}. \tag{3.5}$$

Xác suất có điều kiện chỉ được định nghĩa khi $P(X = x) > 0$. Chúng ta không thể tính xác suất có điều kiện dựa trên một sự kiện không bao giờ xảy ra.

Điều quan trọng là không nên nhầm lẫn giữa xác suất có điều kiện và việc tính toán điều gì sẽ xảy ra nếu một hành động nào đó được thực hiện. Xác suất có điều kiện để một người đến từ Đức khi biết họ nói tiếng Đức là khá cao, nhưng nếu một người được chọn ngẫu nhiên được dạy nói tiếng Đức, thì quốc gia xuất xứ của họ không thay đổi. Việc tính toán hậu quả của một hành động được gọi là **truy vấn can thiệp**. Truy vấn can thiệp thuộc lĩnh vực **mô hình nhân quả**, mà ta sẽ không tìm hiểu trong cuốn sách này.

3.6 Quy tắc nhân xác suất có điều kiện

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

$$P(X_1, \dots, X_n) = P(X_1) \prod_{i=2}^{n} P\left(X_i \mid X_1, \dots, X_{i-1}\right) \tag{3.6}$$

trong đó ký hiệu X_i để chỉ sự kiện biến ngẫu nhiên X_i đạt giá trị x_i nào đó.

Quan sát này được gọi là **quy tắc chuỗi** hay **quy tắc nhân** xác suất có điều kiện. Nó được suy ra ngay từ định nghĩa của xác suất có điều kiện trong phương trình (3.5). Ví dụ, áp dụng định nghĩa này hai lần, ta được

$$\begin{aligned} P(X, Y, Z) &= P(Y, Z) \cdot P\left(X \mid Y, Z\right) \\ P(Y, Z) &= P(Z) \cdot P\left(Y \mid Z\right) \\ \Rightarrow P(X, Y, Z) &= P\left(X \mid Y, Z\right) \cdot P\left(Y \mid Z\right) \cdot P(Z). \end{aligned}$$

3.7 Độc lập và độc lập có điều kiện

Hai biến ngẫu nhiên X và Y **độc lập** nếu phân phối xác suất của chúng có thể được biểu diễn dưới dạng tích của hai nhân tử, một nhân tử chỉ liên quan đến X và nhân tử kia chỉ liên quan đến Y:

$$\forall x \in X,\ y \in Y, \quad P(X = x, Y = y) = P(X = x) \times P(Y = y). \tag{3.7}$$

Hai biến ngẫu nhiên Y và Y **độc lập có điều kiện** khi biết biến ngẫu nhiên Z nếu phân phối xác suất có điều kiện trên X và Y cũng phân tích theo cách này đối với mọi giá trị của Z:

$$P\left(X = x, Y = y \mid Z = z\right) = P\left(X = x \mid Z = z\right) \times P\left(Y = y \mid Z = z\right) \tag{3.8}$$

với mọi $x \in X,\ y \in Y,\ z \in Z$.

Ta có thể ký hiệu tính độc lập và độc lập có điều kiện như sau: $X \perp Y$ có nghĩa là X và Y độc lập, trong khi $X \perp Y \mid Z$ có nghĩa là X và Y độc lập có điều kiện khi biết Z.

3.8 Kỳ vọng, phương sai và hiệp phương sai

Kỳ vọng hay **giá trị kỳ vọng** của biến ngẫu nhiên X, trong trường hợp rời rạc, với hàm trọng số xác suất $p(x)$, được tính bởi tổng các tích giá trị của X với trọng số xác suất tương ứng:

$$EX = \sum_{x \in X} x\, p(x), \tag{3.9}$$

trong khi đối với biến ngẫu nhiên liên tục, nó được tính bằng phép tính tích phân

$$EX = \int x\, p(x)\, dx \tag{3.10}$$

Cho biến ngẫu nhiên X và một hàm số $f(x)$. Trong lý thuyết xác suất, với điều kiện nhất định của f, thì $f(X)$ cũng là biến ngẫu nhiên, có kỳ vọng tương ứng trong trường hợp rời rạc là

$$E[f(X)] = \sum_{x \in X} f(x)\, p(x) \tag{3.11}$$

và đối với biến ngẫu nhiên liên tục:

$$E[f(X)] = \int f(x)\, p(x)\, dx. \tag{3.12}$$

Kỳ vọng là toán tử tuyến tính, tức là,

$$E\left(\alpha X + \beta Y\right) = \alpha\, EX + \beta\, EY, \tag{3.13}$$

khi α và β không phụ thuộc vào các biến ngẫu nhiên X và Y.

Phương sai của biến ngẫu nhiên cung cấp một thước đo về mật độ các giá trị của biến ngẫu nhiên đó quanh giá trị trung bình của nó:

$$\mathrm{Var}(X) = E\left[(X - EX)^2\right]. \tag{3.14}$$

Khi phương sai càng bé, các giá trị của X tập trung càng gần giá trị kỳ vọng của nó. Căn bậc hai của phương sai được gọi là **độ lệch chuẩn**.

Hiệp phương sai của hai biến ngẫu nhiên cung cấp một khái niệm về mức độ tương quan tuyến tính giữa hai biến, cũng như sự co dãn của các biến này:

$$\mathrm{Cov}(X, Y) = E[(X - EX)(Y - EY)]. \tag{3.15}$$

Giá trị tuyệt đối của hiệp phương sai mà cao có nghĩa là các giá trị của các biến có biên độ thay đổi rất nhiều và đồng thời đều cách xa giá trị trung bình của

chúng. Nếu dấu của hiệp phương sai dương, thì cả hai biến có xu hướng nhận các giá trị cùng tăng hoặc cùng giảm. Nếu dấu của hiệp phương sai âm, thì một biến có xu hướng nhận giá trị tăng trong khi biến kia có xu hướng nhận giá trị giảm. Các thước đo khác như hệ số tương quan dùng để chuẩn hóa đóng góp của mỗi biến để đo mức độ tương quan giữa các biến, không bị ảnh hưởng bởi sự co dãn của từng biến riêng lẻ.

Khái niệm về hiệp phương sai và tính phụ thuộc có liên quan với nhau, nhưng thực tế là hai khái niệm khác biệt. Chúng có liên quan vì hai biến độc lập có hiệp phương sai bằng 0, và hai biến có hiệp phương sai khác 0 là phụ thuộc. Tuy nhiên, tính độc lập là một thuộc tính riêng biệt so với hiệp phương sai. Để hai biến có hiệp phương sai bằng 0, không được có sự phụ thuộc tuyến tính giữa chúng. Tính độc lập là một yêu cầu mạnh hơn so với hiệp phương sai bằng 0, vì tính độc lập cũng loại trừ các mối quan hệ phi tuyến tính. Hai biến có thể phụ thuộc nhưng vẫn có hiệp phương sai bằng 0. Ví dụ, xét hai biến ngẫu nhiên rời rạc X và Y như sau: X là biến ngẫu nhiên có thể nhận các giá trị -1, 0, và 1 với xác suất:

$$P(X = -1) = P(X = 1) = \frac{1}{4}, \quad P(X = 0) = \frac{1}{2}$$

và Y được định nghĩa như sau:

$$Y = X^2.$$

Khi đó, Y sẽ nhận các giá trị 0 và 1 với xác suất:

$$P(Y = 0) = P(X = 0) = \frac{1}{2}, \quad P(Y = 1) = P(X = -1 \text{ hoặc } X = 1) = \frac{1}{2}.$$

Hiệp phương sai giữa X và Y còn có thể được tính bằng công thức:

$$\text{Cov}(X, Y) = E(XY) - EX \cdot EY. \tag{3.16}$$

Ta tính được

$$EX = (-1) \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 0$$
$$EY = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

Ngoài ra, vì $Y = X^2$, nên $XY = X \cdot X^2 = X^3$, có kỳ vọng

$$E(XY) = E\left(X^3\right) = (-1)^3 \cdot \frac{1}{4} + 0^3 \cdot \frac{1}{2} + 1^3 \cdot \frac{1}{4} = -\frac{1}{4} + 0 + \frac{1}{4} = 0$$

Từ đó hiệp phương sai giữa hai biến ngẫu nhiên bằng

$$\text{Cov}(X, Y) = E(XY) - EX \cdot EY = 0 - \left(0 \cdot \frac{1}{2}\right) = 0,$$

trong khi X và Y không độc lập, vì Y phụ thuộc trực tiếp vào X qua công thức $Y = X^2$. Hiệp phương sai của chúng bằng 0 vì X và Y không có mối quan hệ tuyến tính nào. Y chỉ phụ thuộc vào giá trị bình phương của X, nên mối quan hệ giữa chúng là phi tuyến tính.

Ma trận hiệp phương sai của véctơ ngẫu nhiên $\boldsymbol{X} \in \mathbb{R}^n$ là một ma trận cỡ $n \times n$, sao cho

$$\text{Cov}(\boldsymbol{X})_{i,j} = \text{Cov}\left(X_i, X_j\right). \tag{3.17}$$

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

$$\text{Cov}(X_i, X_i) = \text{Var}(X_i). \tag{3.18}$$

3.9 Các phân phối xác suất thường gặp

3.9.1 Phân phối Bernoulli

Phân phối Bernoulli là một phân phối trên một biến ngẫu nhiên nhị phân. Nó được điều khiển bởi một tham số duy nhất $p \in [0, 1]$, biểu thị xác suất để biến ngẫu nhiên bằng 1. Nó có các thuộc tính sau:

$$P(X = 1) = p \tag{3.19}$$

$$P(X = 0) = 1 - p \tag{3.20}$$

$$P(X = x) = p^x (1 - p)^{1-x} \tag{3.21}$$

$$EX = p \tag{3.22}$$

$$\text{Var}(X) = p(1 - p) \tag{3.23}$$

3.9.2 Phân phối Bernoulli bội

Phân phối **Bernoulli bội** hay phân phối **phân loại** là một phân phối trên một biến rời rạc có k trạng thái khác nhau, trong đó k hữu hạn.* Phân phối Bernoulli bội được tham số hóa bởi một véctơ $\boldsymbol{p} \in [0, 1]^{k-1}$, trong đó p_i là xác suất của trạng

*"Multinoulli" – Bernoulli bội – là một thuật ngữ được Gustavo Lacerdo đặt ra gần đây và được phổ biến bởi Murphy (*Machine Learning: A Probabilistic Perspective*, 2012, [9]). Phân phối Bernoulli

thái thứ *i*. Xác suất của trạng thái thứ *k* được cho bởi $1 - \mathbf{1}^T \boldsymbol{p}$. Lưu ý rằng ta phải ràng buộc $\mathbf{1}^T \boldsymbol{p} \leq 1$. Phân phối Bernoulli bội thường được sử dụng để chỉ các phân phối trên các loại đối tượng, vì vậy ta thường giả định rằng trạng thái 1 không phải là giá trị số bằng 1,... Do đó, ta thường không cần tính kỳ vọng hoặc phương sai của các biến ngẫu nhiên phân phối Bernoulli bội.

Phân phối Bernoulli và Bernoulli bội là đủ để mô tả bất kỳ phân phối nào trên miền của chúng. Chúng có thể mô tả bất kỳ phân phối nào trên miền của chúng không phải vì chúng đặc biệt mạnh mẽ, mà là vì miền của chúng đơn giản; chúng mô hình hóa các biến rời rạc mà ta có thể liệt kê được tất cả các trạng thái. Khi làm việc với các biến liên tục, có số trạng thái vô hạn không đếm được, thì bất kỳ phân phối nào được mô tả bởi một số ít tham số đều phải đặt ra các giới hạn nghiêm ngặt lên phân phối đó.

3.9.3 Phân phối Gauss

Phân phối thường được sử dụng nhất trên các biến ngẫu nhiên nhận giá trị số thực là **phân phối chuẩn**, còn được gọi là **phân phối Gauss**:

$$N\left(x; \mu, \sigma^2\right) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right). \tag{3.24}$$

Xem Hình 3.1 để biết đồ thị của hàm mật độ.

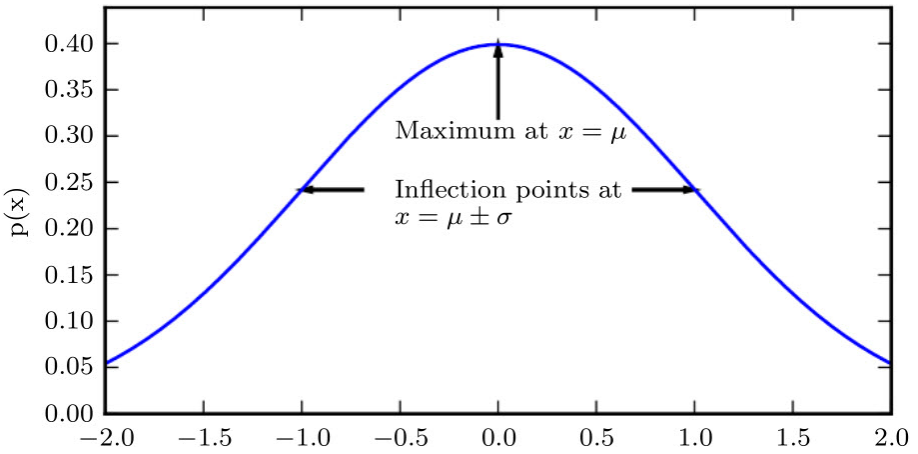
Hai tham số $\mu \in \mathbb{R}$ và $\sigma \in (0, \infty)$ điều khiển phân phối chuẩn. Tham số μ cho tọa độ của đỉnh trung tâm, đây cũng là giá trị trung bình của phân phối: $EX = \mu$. Độ lệch chuẩn của phân phối được cho bởi σ, và phương sai là σ^2.

Khi đánh giá hàm mật độ xác suất, ta cần bình phương và lấy nghịch đảo của σ. Khi cần thường xuyên đánh giá hàm mật độ xác suất với các giá trị tham số khác nhau, một cách tham số hóa hiệu quả hơn cho phân phối là sử dụng tham số $\beta \in (0, \infty)$ để điều khiển **độ chính xác** hoặc nghịch đảo của phương sai của phân phối:

$$N\left(x; \mu, \beta^{-1}\right) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta\left(x-\mu\right)^2\right). \tag{3.25}$$

Phân phối chuẩn là một lựa chọn hợp lý cho nhiều ứng dụng. Khi không có kiến thức trước về dạng của một phân phối trên tập số thực, phân phối chuẩn là một lựa

bội là một trường hợp đặc biệt của phân phối đa thức. Phân phối đa thức là phân phối trên các véctơ trong $\{0, \ldots, n\}^k$, biểu thị số lần mỗi loại trong *k* loại được chọn khi *n* mẫu được rút ra từ một phân phối Bernoulli bội. Nhiều tài liệu sử dụng thuật ngữ "đa thức" để chỉ phân phối Bernoulli bội mà không làm rõ rằng họ chỉ đang đề cập đến trường hợp $n = 1$.



Hình 3.1: **Phân phối chuẩn**: Phân phối chuẩn $N\left(x; \mu, \sigma^2\right)$ có dạng hình "chuông" cổ điển, với tọa độ x của đỉnh trung tâm là μ, và độ rộng của đỉnh được điều khiển bởi σ. Trong ví dụ này, ta biểu diễn **phân phối chuẩn tiêu chuẩn**, với $\mu = 0$ và $\sigma = 1$.

chọn mặc định tốt vì hai lý do chính.

Thứ nhất, nhiều phân phối mà ta muốn mô hình hóa thực sự gần với phân phối chuẩn. **Định lý giới hạn trung tâm** cho thấy tổng của nhiều biến ngẫu nhiên độc lập xấp xỉ phân phối chuẩn. Điều này có nghĩa là trên thực tế, nhiều hệ thống phức tạp có thể được mô hình hóa thành công như nhiễu phân phối chuẩn, ngay cả khi hệ thống đó có thể được phân chia thành các phần có hình thái có cấu trúc hơn.

Thứ hai, trong tất cả các phân phối xác suất có cùng phương sai, phân phối chuẩn mã hóa lượng bất định lớn nhất trên tập số thực. Vì vậy, ta có thể coi phân phối chuẩn là phân phối đưa vào mô hình ít kiến thức tiên nghiệm nhất. Việc phát triển và chứng minh đầy đủ ý tưởng này đòi hỏi nhiều công cụ toán học hơn và sẽ được trình bày trong Mục 10.1.1.

Phân phối chuẩn có thể được tổng quát hóa lên $\mathbb{R}^n$, khi đó nó được gọi là **phân phối chuẩn nhiều chiều**. Nó có thể được tham số hóa bằng một ma trận đối xứng xác định dương Σ:

$$N\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sqrt{\frac{1}{(2\pi)^n \det\left(\boldsymbol{\Sigma}\right)}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right). \quad (3.26)$$

Tham số $\boldsymbol{\mu}$ vẫn cho giá trị trung bình của phân phối, tuy nhiên bây giờ nó có giá trị dạng véctơ. Tham số $\boldsymbol{\Sigma}$ cung cấp ma trận hiệp phương sai của phân phối. Tương tự như trường hợp một biến, khi ta muốn tính hàm mật độ xác suất nhiều lần cho các giá trị khác nhau của tham số, ma trận hiệp phương sai không phải là

cách tham số hóa phân phối hiệu quả về mặt tính toán, vì ta cần phải nghịch đảo $\boldsymbol{\Sigma}$ để tính hàm mật độ xác suất. Thay vào đó, chúng ta có thể sử dụng **ma trận độ chính xác** $\boldsymbol{\beta}$.

$$N\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}\right) = \sqrt{\frac{\det\left(\boldsymbol{\beta}\right)}{(2\pi)^n}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\beta}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right). \tag{3.27}$$

Ta thường cố định ma trận hiệp phương sai là một ma trận đường chéo. Một phiên bản đơn giản hơn nữa là phân phối Gauss **đẳng hướng**, trong đó ma trận hiệp phương sai là một ma trận vô hướng, là ma trận có dạng số vô hướng nhân với ma trận đơn vị.

3.9.4 Phân phối mũ và phân phối Laplace

Trong bối cảnh học sâu, ta thường muốn có một phân phối xác suất có điểm nhọn tại $x = 0$. Để đạt được điều này, ta có thể sử dụng **phân phối mũ**:

$$p\left(x; \lambda\right) = \lambda \mathbf{1}_{x \geq 0} \exp\left(-\lambda x\right). \tag{3.28}$$

Phân phối mũ sử dụng hàm chỉ $\mathbf{1}_{x \geq 0}$ để gán xác suất bằng 0 cho tất cả giá trị x âm của biến ngẫu nhiên.

Một phân phối xác suất có liên quan chặt chẽ, cho phép ta đặt một đỉnh nhọn của trọng số xác suất tại một điểm tùy ý μ, là **phân phối Laplace**

$$\text{Laplace}\left(x; \mu, \gamma\right) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right). \tag{3.29}$$

3.9.5 Phân phối Dirac và phân phối thực nghiệm

Trong một số trường hợp, ta muốn xác định rằng toàn bộ trọng số của một phân phối xác suất tập trung xung quanh một điểm duy nhất. Điều này có thể thực hiện bằng cách định nghĩa một hàm mật độ xác suất sử dụng hàm delta Dirac, $\delta\left(x\right)$:

$$p\left(x\right) = \delta\left(x-\mu\right). \tag{3.30}$$

Hàm delta Dirac được định nghĩa sao cho nó có giá trị bằng 0 ở mọi điểm ngoại trừ tại điểm 0, nhưng tích phân của nó lại bằng 1. Hàm delta Dirac không phải là một hàm thông thường gán mỗi giá trị x với một đầu ra có giá trị thực, mà thay vào đó, nó là một loại đối tượng toán học khác gọi là **hàm suy rộng**, được định nghĩa thông qua các tính chất của nó khi lấy tích phân. Ta có thể nghĩ hàm delta Dirac

như là giới hạn của một dãy các hàm phân phối ngày càng ít trọng số trên tất cả các điểm khác ngoài điểm 0.

Bằng cách định nghĩa $p(x)$ là hàm δ với đối số được tính tiến một lượng $-\mu$, ta thu được một đỉnh có trọng số xác suất vô cùng hẹp và cao tại điểm $x = \mu$. Một ứng dụng phổ biến của phân phối delta Dirac là thành phần của một phân phối thực nghiệm,

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta\left(\boldsymbol{x} - \boldsymbol{x}^{(i)}\right) \tag{3.31}$$

trong đó đặt trọng số xác suất $\frac{1}{m}$ trên mỗi điểm trong tập hợp m điểm $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ tạo nên một tập dữ liệu hoặc một tập hợp các mẫu. Phân phối delta Dirac chỉ cần thiết để định nghĩa phân phối thực nghiệm trên các biến liên tục. Đối với các biến rời rạc, tình huống đơn giản hơn: một phân phối thực nghiệm có thể được khái quát hóa như một phân phối Bernoulli bội, với xác suất gán cho mỗi giá trị đầu vào bằng **tần suất thực nghiệm** của giá trị đó trong tập huấn luyện.

Ta có thể xem phân phối thực nghiệm hình thành từ một tập dữ liệu các ví dụ huấn luyện như là việc xác định phân phối mà từ đó ta lấy mẫu khi huấn luyện một mô hình trên tập dữ liệu này. Một góc nhìn quan trọng khác về phân phối thực nghiệm là nó là hàm mật độ xác suất làm cực đại tính hợp lý của dữ liệu huấn luyện (xem Mục 5.1).

3.9.6 Phân phối hỗn hợp

Một cách rất phổ biến là định nghĩa các phân phối xác suất bằng cách kết hợp các phân phối xác suất đơn giản hơn. Một cách thông dụng để kết hợp các phân phối là xây dựng một **phân phối hỗn hợp**. Phân phối hỗn hợp được tạo thành từ nhiều phân phối thành phần. Trong mỗi lần thử, việc chọn phân phối thành phần nào sẽ tạo ra mẫu được xác định bằng cách lấy mẫu một thành phần từ phân phối Bernoulli bội:

$$P(X = x) = \sum_{i} P(C = i)\ P\left(X = x \mid C = i\right) \tag{3.32}$$

trong đó C có phân phối Bernoulli bội trên các thành phần tố nhận dạng.

Ta đã thấy một ví dụ về phân phối hỗn hợp: phân phối thực nghiệm trên các biến có giá trị thực là một phân phối hỗn hợp với một thành phần Dirac cho mỗi ví dụ 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. Trong Chương 9, 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.

Mô hình hỗn hợp cho ta cái nhìn thoáng qua về một khái niệm sẽ trở nên vô cùng quan trọng sau này – **biến ẩn**. Biến ẩn là một biến ngẫu nhiên mà ta không thể quan sát trực tiếp. Biến nhận dạng thành phần C trong mô hình hỗn hợp là một ví dụ. Các biến ẩn có thể liên quan đến X thông qua phân phối đồng thời, trong trường hợp này, $P(X, C) = P\left(X \mid C\right)\ P(C)$. Phân phối $P(C)$ trên biến ẩn và phân phối $P\left(X \mid C\right)$ liên kết các biến ẩn với các biến quan sát được sẽ xác định hình dạng của phân phối $P(X)$ mặc dù có thể mô tả $P(X)$ mà không cần đề cập đến biến ẩn. Các biến ẩn sẽ được thảo luận chi tiết hơn trong Mục 9.1.

Một loại mô hình hỗn hợp rất mạnh mẽ và phổ biến là mô hình **hỗn hợp Gauss**, trong đó các thành phần $p\left(\boldsymbol{x} \mid C = i\right)$ là các phân phối Gauss. Mỗi thành phần có một giá trị trung bình $\boldsymbol{\mu}^{(i)}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}^{(i)}$ được tham số hóa riêng biệt. Một số hỗn hợp có thể có nhiều ràng buộc hơn. Ví dụ, các ma trận hiệp phương sai có thể được chia sẻ giữa các thành phần thông qua ràng buộc $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma},\ \forall i$. Tương tự như phân phối Gauss đơn lẻ, hỗn hợp Gauss có thể ràng buộc ma trận hiệp phương sai của mỗi thành phần là ma trận đường chéo hoặc đẳng hướng.

Bên cạnh các giá trị trung bình và hiệp phương sai, các tham số của mô hình hỗn hợp Gauss xác định **xác suất tiên nghiệm** $\alpha_i = P(C = i)$ được gán cho mỗi thành phần i. Từ "tiên nghiệm" cho biết nó thể hiện mức tin cậy của mô hình về C trước khi quan sát X. So với đó, $P\left(C \mid X\right)$ là một **xác suất hậu nghiệm**, vì nó được tính sau khi đã quan sát X. Mô hình hỗn hợp Gauss là một **bộ xấp xỉ phổ dụng** của hàm mật độ, theo nghĩa rằng bất kỳ hàm mật độ trơn nào cũng có thể được xấp xỉ bởi một mô hình hỗn hợp Gauss với đủ số lượng thành phần, với một mức sai số khác không bất kỳ.

Hình 3.2 cho thấy các mẫu từ một mô hình hỗn hợp Gauss.

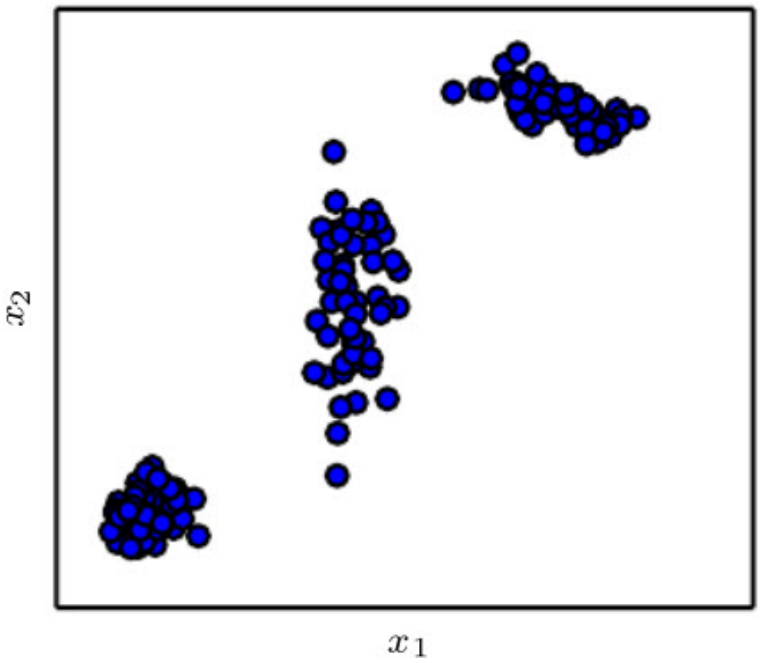
3.10 Các tính chất hữu ích của các hàm thông dụng

Một số hàm thường xuất hiện khi làm việc với các phân phối xác suất, đặc biệt là các phân phối xác suất được sử dụng trong các mô hình học sâu.

Một trong những hàm này là hàm **logistic sigmoid**:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.33}$$

Hàm logistic sigmoid thường được sử dụng để tạo ra tham số p của phân phối Bernoulli vì miền giá trị của nó khoảng $(0, 1)$, nằm trong phạm vi hợp lệ của tham



Hình 3.2: Các mẫu từ một mô hình hỗn hợp Gauss. Trong ví dụ này, có ba thành phần. Từ trái sang phải, thành phần đầu tiên có ma trận hiệp phương sai đẳng hướng, nghĩa là nó có cùng mức phương sai theo mọi hướng. Thành phần thứ hai có ma trận hiệp phương sai là ma trận đường chéo, nghĩa là nó có thể kiểm soát phương sai riêng biệt dọc theo mỗi hướng thẳng hàng với trục tọa độ. Ví dụ này có phương sai hơn dọc theo trục x_2 lớn hơn so với trục x_1. Thành phần thứ ba có ma trận hiệp phương sai có hạng đầy đủ, cho phép nó kiểm soát phương sai riêng biệt theo một cơ sở bất kỳ của các hướng.

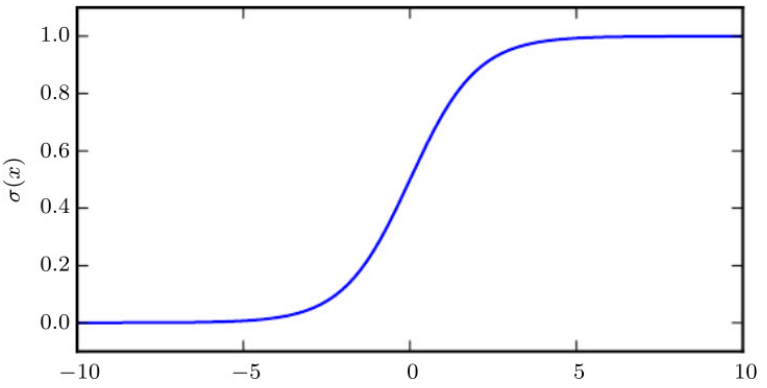
số p. Xem Hình 3.3 để thấy đồ thị của hàm sigmoid. Hàm sigmoid **bão hòa** khi đối số của nó rất lớn (dương) hoặc rất nhỏ (âm), có nghĩa là hàm trở nên rất phẳng và không nhạy với các thay đổi nhỏ trong đầu vào của nó.

Một hàm thường gặp khác là hàm **softplus** (*Incorporating Second-Order Functional Knowledge for Better Option Pricing*, Dugas và các tác giả khác, 2000, [10]):

$$\zeta(x) = \ln\left(1 + e^x\right). \tag{3.34}$$

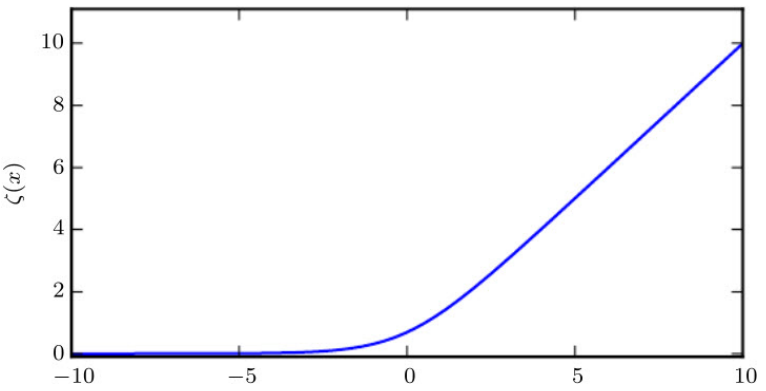
Hàm softplus có thể hữu ích để tạo ra tham số β hoặc σ của phân phối chuẩn vì miền giá trị của nó là $(0, \infty)$. Nó cũng thường xuất hiện khi xử lý các phép tính liên quan đến hàm sigmoid. Tên gọi của hàm softplus xuất phát từ thực tế rằng nó là phiên bản làm mịn hoặc "mềm hóa" của

$$x^+ = \max(0, x) \tag{3.35}$$



Hình 3.3: Hàm logistic sigmoid

Xem Hình 3.4 để thấy đồ thị của hàm softplus.



Hình 3.4: Hàm softplus

Các tính chất sau đây đều đủ hữu ích đến mức bạn có thể muốn ghi nhớ chúng:

$$\sigma(x) = \frac{e^x}{e^x + e^0} \tag{3.36}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{3.37}$$

$$1 - \sigma(x) = \sigma(-x) \tag{3.38}$$

$$\ln \sigma(x) = -\zeta(-x) \tag{3.39}$$

$$\frac{d}{dx}\zeta(x) = \sigma(x) \tag{3.40}$$

$$\forall x \in (0, 1), \quad \sigma^{-1}(x) = \ln \frac{x}{1 - x} \tag{3.41}$$

$$\forall x > 0, \quad \zeta^{-1}(x) = \ln\left(e^x - 1\right) \tag{3.42}$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(t)\, dt \tag{3.43}$$

$$\zeta(x) - \zeta(-x) = x \tag{3.44}$$

Hàm $\sigma^{-1}(x)$ được gọi là logit trong thống kê, nhưng thuật ngữ này ít được sử dụng hơn trong lĩnh vực học máy.

Công thức (3.44) cung cấp thêm lý do để gọi tên là "softplus". Hàm softplus được tạo ra như một phiên bản làm trơn của hàm phần dương, $x^+ = \max\{0, x\}$. Hàm phần dương là đối ngẫu của hàm phần âm, $x^- = \max\{0, -x\}$. Để có một hàm trơn tương tự với phần âm, ta có thể sử dụng $\zeta(-x)$. Giống như có thể khôi phục x từ phần dương và phần âm của nó thông qua đẳng thức $x^+ - x^- = x$, cũng có thể khôi phục x bằng cách sử dụng mối quan hệ tương tự giữa $\zeta(x)$ và $\zeta(-x)$, như được chỉ ra trong công thức (3.44).

3.11 Quy tắc Bayes

Ta thường rơi vào tình huống biết $P\left(Y \mid X\right)$ và cần biết $P\left(X \mid Y\right)$. May mắn là, nếu ta cũng biết $P(X)$, ta có thể tính được giá trị mong muốn bằng cách sử dụng **quy tắc Bayes**:

$$P\left(X \mid Y\right) = \frac{P(X)\ P\left(Y \mid X\right)}{P(Y)}. \tag{3.45}$$

Lưu ý rằng mặc dù $P(Y)$ xuất hiện trong công thức, nhưng thông thường có thể tính được $P(Y) = \sum_x P(x)\ P\left(Y \mid x\right)$, vì vậy ta không cần phải bắt đầu với giả thiết về $P(y)$.

Quy tắc Bayes dễ dàng được suy ra từ định nghĩa xác suất có điều kiện, nhưng việc biết tên của công thức này rất hữu ích vì nhiều tài liệu đề cập đến nó theo tên gọi. Công thức này được đặt tên theo mục sư Thomas Bayes, người đã phát hiện ra một trường hợp đặc biệt của công thức. Phiên bản tổng quát được trình bày ở đây đã được Pierre – Simon Laplace phát hiện độc lập.

3.12 Chi tiết kỹ thuật về các biến liên tục

Để hiểu một cách chính xác và đầy đủ về các biến ngẫu nhiên liên tục và hàm mật độ xác suất, cần phát triển lý thuyết xác suất dựa trên một nhánh của toán học gọi là **lý thuyết độ đo**. Lý thuyết độ đo nằm ngoài phạm vi của cuốn sách này, nhưng ta có thể phác thảo ngắn gọn một số vấn đề mà để giải quyết cần sử dụng lý thuyết độ đo.

Trong Mục 3.3.2, ta đã thấy xác suất của một biến ngẫu nhiên liên tục nhiều chiều X nằm trong một tập hợp S nào đó được xác định bằng tích phân của hàm $p(x)$ trên tập hợp S. Một số lựa chọn tập hợp S có thể dẫn đến nghịch lý. Chẳng hạn, có thể xây dựng hai tập hợp S_1 và S_2 sao cho $P(X \in S_1) + P(X \in S_2) > 1$ nhưng $S_1 \cap S_2 = \varnothing$. Những tập hợp này thường được tạo ra bằng cách tận dụng tối đa độ chính xác vô hạn của các số thực, ví dụ như bằng cách tạo ra các tập hợp có hình dạng fractal hoặc các tập hợp được định nghĩa bằng cách biến đổi tập hợp các số hữu tỉ.* Một trong những đóng góp chính của lý thuyết độ đo là cung cấp một cách phân loại các tập hợp mà ta có thể tính xác suất mà không gặp phải nghịch lý. Trong cuốn sách này, ta chỉ tính tích phân trên các tập hợp có mô tả tương đối đơn giản, do đó khía cạnh này của lý thuyết độ đo sẽ không trở thành mối quan tâm đáng kể.

Đối với mục đích của chúng ta, lý thuyết độ đo hữu ích hơn trong việc mô tả các định lý áp dụng cho hầu hết các điểm trong $\mathbb{R}^n$ nhưng không áp dụng cho một số trường hợp đặc biệt. Lý thuyết độ đo cung cấp một cách mô tả chặt chẽ về việc một tập hợp các điểm có kích thước không đáng kể. Một tập hợp như vậy được gọi là có **độ đo bằng không**. Chúng ta sẽ không định nghĩa chính thức khái niệm này trong cuốn sách này. Đối với mục đích của chúng ta, chỉ cần hiểu trực giác rằng một tập hợp có độ đo bằng không không chiếm thể tích nào trong không gian mà ta đang đo lường. Chẳng hạn, trong $\mathbb{R}^2$, một đường thẳng có độ đo bằng không, trong khi một đa giác có diện tích lấp đầy sẽ có độ đo dương. Tương tự, một điểm đơn lẻ có độ đo bằng không. Bất kỳ hợp của một số đếm được các tập có độ đo bằng không cũng có độ đo bằng không (ví dụ như tập hợp tất cả các số hữu tỉ có độ đo bằng không).

Một thuật ngữ hữu ích khác từ lý thuyết độ đo là **hầu khắp nơi**. Một tính chất được coi là hầu khắp nơi nếu nó đúng trên toàn bộ không gian, ngoại trừ trên một tập hợp có độ đo bằng không. Bởi vì các ngoại lệ này chiếm một lượng không gian không đáng kể, chúng thường có thể bị bỏ qua trong nhiều ứng dụng. Một số kết quả quan trọng trong lý thuyết xác suất áp dụng cho tất cả các giá trị rời rạc nhưng chỉ đúng "hầu khắp nơi" đối với các giá trị liên tục.

Một chi tiết kỹ thuật khác liên quan đến các biến ngẫu nhiên liên tục là cách xử lý các biến ngẫu nhiên liên tục phụ thuộc như là hàm theo các biến ngẫu nhiên khác. Giả sử ta có hai biến ngẫu nhiên, X và Y, sao cho $Y = g(x)$, trong đó g là một phép biến đổi khả vi liên tục và khả nghịch. Có thể bạn sẽ mong đợi rằng

*Định lý Banach – Tarski cung cấp một ví dụ thú vị về các tập hợp như vậy.

$p_Y(y) = p_X\left(g^{-1}(y)\right)$. Tuy nhiên, điều này thực tế không đúng.

Lấy ví dụ đơn giản, giả sử chúng ta có các biến ngẫu nhiên vô hướng X và Y. Giả sử $Y = \frac{X}{2}$ và $X \sim U(0,1)$. Nếu ta áp dụng quy tắc sai lầm $p_Y(y) = p_X(2y)$, thì $p_Y(y)$ sẽ bằng 0 tại mọi điểm ngoại trừ khoảng $\left[0, \frac{1}{2}\right]$, và trong khoảng này nó sẽ có giá trị 1. Điều này dẫn đến

$$\int p_Y(y)\, dy = \frac{1}{2},$$

một kết quả vi phạm định nghĩa của phân phối xác suất. Đây là một sai lầm phổ biến. Vấn đề của cách tiếp cận này là nó không tính đến sự biến dạng của không gian gây ra bởi hàm g. Nhớ rằng xác suất để X nằm trong một vùng rất nhỏ có thể tích $\delta \boldsymbol{x}$ được cho bởi $p_X(\boldsymbol{x})\,\delta \boldsymbol{x}$. Khi hàm g mở rộng hoặc co hẹp không gian, thể tích vô cùng nhỏ xung quanh $\boldsymbol{x}$ trong không gian $\boldsymbol{x}$ có thể có thể tích khác trong không gian $\boldsymbol{y}$.

Để khắc phục vấn đề này, ta sẽ quay lại trường hợp biến ngẫu nhiên vô hướng. Ta cần bảo toàn tính chất của xác suất trong quá trình biến đổi. Cụ thể, xác suất của biến ngẫu nhiên X nằm trong một khoảng vô cùng nhỏ δx cần phải tương đương với xác suất của biến ngẫu nhiên Y nằm trong khoảng vô cùng nhỏ tương ứng δy sau phép biến đổi:

$$\left|p_X(x)\, dx\right| = \left|p_Y(y)\, dy\right|. \tag{3.46}$$

Giải phương trình này, ta được

$$p_Y(y) = p_X\left(g^{-1}(y)\right)\left|\frac{\partial x}{\partial y}\right| \tag{3.47}$$

hoặc tương đương

$$p_X(x) = p_Y(g(x))\left|\frac{\partial g(x)}{\partial x}\right|. \tag{3.48}$$

Trong không gian nhiều chiều, đạo hàm được tổng quát hóa thành định thức của ma trận Jacobi – ma trận có các phần tử $J_{i,j} = \frac{\partial x_i}{\partial y_j}$. Do đó, nếu $\boldsymbol{x}$ và $\boldsymbol{y}$ là các véctơ giá trị thực,

$$p_X(\boldsymbol{x}) = p_Y(\boldsymbol{g}(\boldsymbol{x}))\left|\det\left(\frac{\partial \boldsymbol{g}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|. \tag{3.49}$$

3.13 Lý thuyết thông tin

Lý thuyết thông tin là một nhánh của toán ứng dụng tập trung vào việc định lượng lượng thông tin có trong một tín hiệu. Ban đầu, nó được phát minh để nghiên cứu việc gửi thông điệp từ các bảng chữ cái rời rạc qua một kênh nhiễu, chẳng hạn như truyền thông qua sóng radio. Trong bối cảnh này, lý thuyết thông tin cung cấp cách thiết kế các mã hóa tối ưu và tính toán độ dài kỳ vọng của các thông điệp được lấy mẫu từ các phân phối xác suất cụ thể bằng cách sử dụng các phương pháp mã hóa khác nhau. Trong bối cảnh học máy, ta cũng có thể áp dụng lý thuyết thông tin cho các biến liên tục, nơi mà một số diễn giải về độ dài thông điệp không áp dụng được. Lý thuyết thông tin là nền tảng của nhiều lĩnh vực trong kỹ thuật điện và khoa học máy tính. Trong cuốn sách này, chúng ta chủ yếu sử dụng một số khái niệm quan trọng từ lý thuyết thông tin để mô tả các phân phối xác suất hoặc định lượng sự tương đồng giữa các phân phối xác suất. Để biết thêm chi tiết về lý thuyết thông tin, xem *Elements of Information Theory* (Cover and Thomas, 2006, [11]) hoặc *Information Theory, Inference and Learning Algorithms* (MacKay, 2003, [12]).

Trực giác cơ bản đằng sau lý thuyết thông tin là việc học được rằng một sự kiện hiếm có đã xảy ra sẽ cung cấp nhiều thông tin hơn so với việc học được rằng một sự kiện thường xảy ra đã xảy ra. Một thông điệp như "mặt trời mọc sáng nay" không có nhiều thông tin và gần như không cần thiết phải truyền đạt, nhưng một thông điệp như "đã có nhật thực sáng nay" sẽ rất bổ ích.

Ta muốn định lượng thông tin theo cách chính thức hóa trực giác này. Cụ thể:

- Những sự kiện có xác suất cao nên có lượng thông tin thấp, và trong trường hợp cực đoan, các sự kiện được đảm bảo sẽ xảy ra (xác suất bằng 1) sẽ không có lượng thông tin nào cả.
- Những sự kiện có xác suất thấp nên có lượng thông tin cao hơn.
- Các sự kiện độc lập nên có lượng thông tin cộng dồn. Ví dụ, biết rằng một đồng xu được tung hai lần và cả hai lần đều ra mặt sấp nên cung cấp gấp đôi lượng thông tin so với việc biết rằng đồng xu chỉ được tung một lần và ra mặt sấp.

Để thỏa mãn cả ba thuộc tính này, chúng ta định nghĩa **lượng thông tin tự thân** của một sự kiện $X = x$ là:

$$I(x) = -\log p(x). \tag{3.50}$$

Trong cuốn sách này, chúng ta luôn sử dụng hàm logarit với cơ số tự nhiên e. Do đó, định nghĩa của $I(x)$ được viết theo đơn vị **nats**. Một nat là lượng thông tin thu được khi quan sát một sự kiện có xác suất $\frac{1}{2}$. Một số tài liệu khác sử dụng logarit cơ số 2 và đơn vị gọi là **bits** hoặc **shannons**; thông tin đo bằng bits chỉ là một sự thay đổi tỷ lệ từ thông tin đo bằng nats.

Khi X là biến liên tục, ta vẫn sử dụng định nghĩa tương tự cho lượng thông tin, nhưng một số thuộc tính từ trường hợp rời rạc sẽ bị mất. Ví dụ, trong trường hợp biến liên tục, một sự kiện với mật độ đơn vị vẫn có lượng thông tin bằng 0, mặc dù đây không phải là một sự kiện được đảm bảo xảy ra (xác suất bằng 1).

Lượng thông tin tự thân chỉ xem xét một kết quả duy nhất. Ta có thể định lượng mức độ bất định của toàn bộ một phân phối xác suất bằng **entropy Shannon**:

$$H(X) = E[I(X)] = -E\left[\log p(X)\right], \tag{3.51}$$

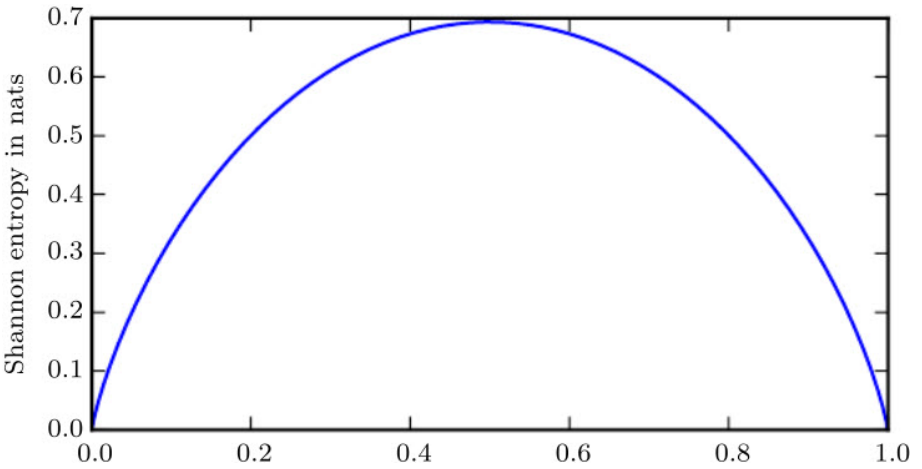
còn được ký hiệu là $H(p)$. Nói cách khác, entropy Shannon của một phân phối là lượng thông tin kỳ vọng từ một sự kiện được lấy mẫu từ phân phối đó. Nó cung cấp một giới hạn dưới cho số lượng bit trung bình (nếu sử dụng logarit cơ số 2, nếu không đơn vị sẽ khác) cần thiết để mã hóa các ký hiệu được lấy từ một phân phối p. Các phân phối gần như xác định (khi kết quả gần như chắc chắn) có entropy thấp. Các phân phối càng gần phân phối đều (mỗi kết quả có xác suất tương tự nhau) có entropy cao hơn. Xem minh họa trong Hình 3.5. Khi X là biến liên tục, entropy Shannon được gọi là **entropy vi phân**.

Nếu X có phân phối xác suất $p(x)$, và $q(x)$ là một phân phối xác suất nào đó, ta có thể đo lường sự khác biệt giữa hai phân phối này bằng **lượng phân kỳ Kullback – Leibler (KL)**:

$$D_{\mathrm{KL}}\left(p \parallel q\right) = E\left[\log \frac{p(X)}{q(X)}\right] = E\left[\log p(X) - \log q(X)\right]. \tag{3.52}$$

Trong trường hợp biến rời rạc, lượng phân kỳ KL đo lượng thông tin dư thừa (tính bằng bits nếu sử dụng logarit cơ số 2, nhưng trong học máy ta thường sử dụng nats và logarit tự nhiên) cần để gửi một thông điệp chứa các ký hiệu được lấy từ phân phối xác suất p, khi ta sử dụng một mã được thiết kế để tối thiểu hóa độ dài của các thông điệp được lấy từ phân phối q.

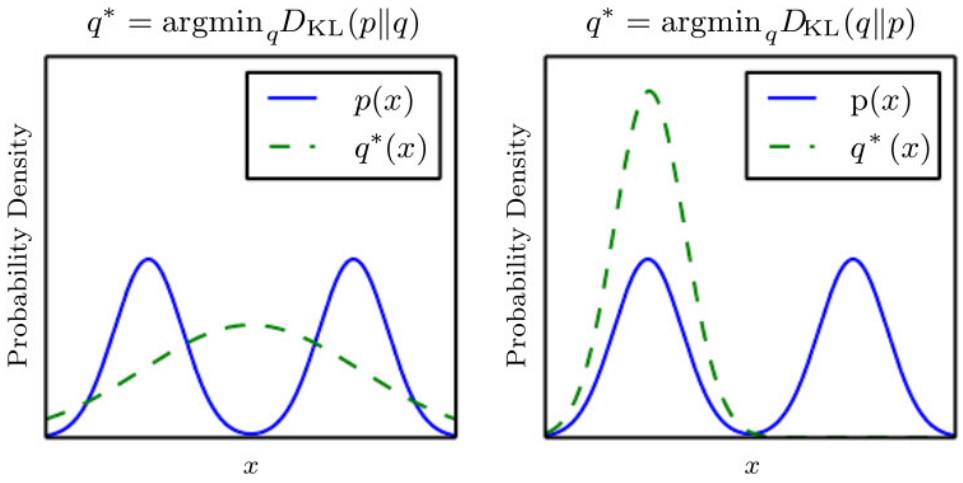
Lượng phân kỳ KL có nhiều tính chất hữu ích, đáng chú ý nhất là tính không âm. Lượng phân kỳ KL bằng 0 chỉ khi p và q là cùng một phân phối trong trường hợp biến rời rạc, hoặc bằng nhau "hầu khắp nơi" trong trường hợp biến liên tục. Vì lượng phân kỳ KL không âm và đo lường sự khác biệt giữa hai phân phối, nó



Hình 3.5: Đồ thị này minh họa cách mà các phân phối gần với phân phối xác định có entropy Shannon thấp, trong khi các phân phối gần như đồng nhất có entropy Shannon cao. Trục hoành biểu diễn giá trị p, là xác suất để biến ngẫu nhiên nhị phân bằng 1. Entropy được cho bởi công thức: $-(1-p)\log(1-p) - p\log p$. Khi p gần bằng 0, phân phối gần như xác định vì biến ngẫu nhiên gần như luôn bằng 0, và do đó entropy thấp. Tương tự, khi p gần bằng 1, phân phối cũng gần như xác định vì biến ngẫu nhiên gần như luôn bằng 1, và entropy cũng thấp. Ngược lại, khi $p = 0.5$, entropy đạt cực đại, vì phân phối là đồng nhất giữa hai kết quả (cả hai kết quả có xác suất xảy ra bằng nhau), cho thấy sự bất định lớn nhất trong việc dự đoán kết quả. Entropy đạt cực đại khi có sự bất định lớn nhất, tức là khi các kết quả đều có xác suất như nhau, và nó giảm dần khi một trong hai kết quả trở nên chắc chắn hơn.

thường được xem như là một phép đo "khoảng cách" giữa các phân phối này. Tuy nhiên, lượng phân kỳ KL không phải là một phép đo khoảng cách thực sự vì nó không đối xứng: $D_{\mathrm{KL}}\left(p \parallel q\right) \neq D_{\mathrm{KL}}\left(q \parallel p\right)$ đối với một số phân phối p và q. Sự bất đối xứng này dẫn đến những hệ quả quan trọng khi chọn sử dụng $D_{\mathrm{KL}}\left(p \parallel q\right)$ hay $D_{\mathrm{KL}}\left(q \parallel p\right)$. Tùy thuộc vào thứ tự sử dụng các phân phối trong lượng phân kỳ KL, kết quả có thể khác biệt lớn, phản ánh hướng mà ta đang đo lường sự khác nhau giữa các phân phối. Ví dụ: $D_{\mathrm{KL}}\left(p \parallel q\right)$ thường được diễn giải như là lượng thông tin mất đi khi sử dụng phân phối q để xấp xỉ phân phối thực tế p, còn $D_{\mathrm{KL}}\left(q \parallel p\right)$ lại đo lường điều ngược lại: lượng thông tin mất đi khi dùng p để mô hình hóa phân phối q. Hình 3.6 có thể minh họa rõ hơn sự khác biệt này.

Một đại lượng liên quan chặt chẽ với lượng phân kỳ KL là **cross-entropy** $H(p,q) = H(p) + D_{\mathrm{KL}}\left(p \parallel q\right)$, mà về cơ bản tương tự lượng phân kỳ KL nhưng



Hình 3.6: Lượng phân kỳ KL là bất đối xứng. Giả sử ta có một phân phối $p(x)$ và muốn xấp xỉ nó bằng một phân phối khác $q(x)$. Ta có thể chọn giữa việc cực tiểu hóa $D_{\text{KL}}\left(p \parallel q\right)$ hoặc $D_{\text{KL}}\left(q \parallel p\right)$. Hiệu ứng của sự lựa chọn này được minh họa bằng cách sử dụng phân phối hỗn hợp của hai phân phối Gauss cho p, và một phân phối Gauss đơn cho q. Hướng lựa chọn của lượng phân kỳ KL phụ thuộc vào vấn đề cụ thể. Một số ứng dụng yêu cầu một xấp xỉ thường đặt xác suất cao ở bất kỳ đâu mà phân phối thực p đặt xác suất cao, trong khi một số ứng dụng khác lại yêu cầu xấp xỉ mà phân phối q hiếm khi đặt xác suất cao ở những nơi mà phân phối thực p có xác suất thấp. Sự lựa chọn hướng của KL divergence phản ánh yếu tố nào được ưu tiên trong mỗi ứng dụng. (*Bên trái*) Hiệu ứng của việc cực tiểu hóa $D_{\text{KL}}\left(p \parallel q\right)$. Trong trường hợp này, ta chọn một q có xác suất cao ở những nơi mà p có xác suất cao. Khi p có nhiều đỉnh, q chọn cách làm mờ các đỉnh lại với nhau để đặt trọng số xác suất lớn lên tất cả các đỉnh. Điều này có nghĩa là q sẽ bao quát toàn bộ phân phối p, ngay cả khi nó phải tạo ra một phân phối mờ hơn. (*Bên phải*) Hiệu ứng của việc cực tiểu hóa $D_{\text{KL}}\left(q \parallel p\right)$. Trong trường hợp này, ta chọn một q có xác suất thấp ở những nơi mà p có xác suất thấp. Khi p có nhiều đỉnh tách biệt nhau, lượng phân kỳ KL được tối thiểu hóa bằng cách chọn một đỉnh duy nhất, để tránh đặt trọng số xác suất vào các vùng có xác suất thấp giữa các đỉnh của p. Trong ví dụ này, q được chọn để nhấn mạnh đỉnh bên trái của p. Ta cũng có thể đạt được giá trị lượng phân kỳ KL tương tự bằng cách chọn đỉnh bên phải. Nếu các đỉnh không bị ngăn cách bởi một vùng có xác suất thấp đủ mạnh, thì hướng này của lượng phân kỳ KL cũng có thể chọn cách làm mờ các đỉnh, tương tự như trường hợp bên trái.

không có hạng tử bên trái:

$$H(p, q) = -E\left[\log q(X)\right]. \tag{3.53}$$

Tối thiểu hóa cross−entropy theo q tương đương với việc cực tiểu hóa lượng phân kỳ KL, vì q không xuất hiện trong hạng tử bị bỏ qua $H(p)$. Điều này có nghĩa là khi tối ưu hóa q, ta có thể xem việc cực tiểu hóa cross−entropy là tương đương với việc xấp xỉ tốt nhất phân phối p bằng phân phối q.

Khi tính toán các đại lượng này, ta thường gặp những biểu thức có dạng $0 \log 0$. Theo quy ước trong lý thuyết thông tin, những biểu thức này được xử lý như giới hạn $\lim_{x \to 0} x \log x = 0$. Quy ước này giúp tránh các giá trị không xác định khi xác suất bằng 0 trong các phân phối xác suất, đảm bảo các phép tính về entropy và lượng phân kỳ được thực hiện một cách ổn định.

3.14 Mô hình xác suất có cấu trúc

Các thuật toán học máy thường liên quan đến các phân phối xác suất trên một số lượng lớn các biến ngẫu nhiên. Thông thường, các phân phối xác suất này bao gồm tương tác trực tiếp giữa một số ít biến. Sử dụng một hàm duy nhất để mô tả toàn bộ phân phối xác suất đồng thời có thể rất kém hiệu quả (cả về mặt tính toán lẫn thống kê).

Thay vì sử dụng một hàm duy nhất để biểu diễn phân phối xác suất, ta có thể tách phân phối xác suất thành nhiều thành phần và nhân chúng với nhau. Ví dụ, 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 độc lập với nhau khi biết b. Từ công thức nhân xác suất (3.6), ta có thể biểu diễn phân phối xác suất trên cả ba biến này dưới dạng tích của các phân phối xác suất trên hai biến:

$$\begin{aligned} P(a, b, c) &= P(a) \cdot P\left(b \mid a\right) \cdot P\left(c \mid a, b\right) \\ &= P(a) \cdot P\left(b \mid a\right) \cdot P\left(c \mid b\right) \end{aligned} \tag{3.54}$$

Những phép phân tích này có thể giảm đáng kể số lượng tham số cần thiết để mô tả phân phối. Mỗi thành phần trong phân tích sử dụng số lượng tham số theo cấp số nhân với số biến trong thành phần đó. Điều này có nghĩa là ta có thể giảm mạnh chi phí biểu diễn một phân phối nếu có thể tìm ra cách phân tích thành các phân phối trên ít biến hơn.

Ta có thể mô tả những loại phân tích thành phần này bằng cách sử dụng đồ thị. Ở đây, từ "đồ thị" được hiểu theo lý thuyết đồ thị: một tập hợp các đỉnh có thể được

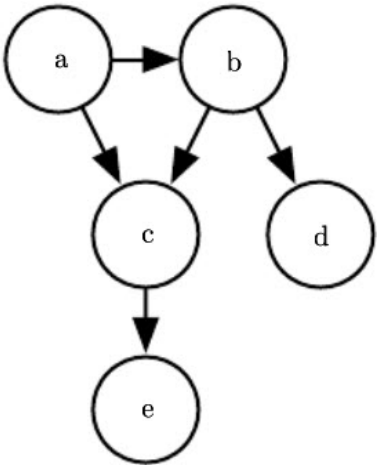
kết nối với nhau bằng các cạnh. Khi ta biểu diễn phép phân tích một phân phối xác suất bằng đồ thị, ta gọi nó là một **mô hình xác suất có cấu trúc** hoặc **mô hình đồ thị**.

Có hai loại chính của mô hình xác suất có cấu trúc: mô hình có hướng và mô hình vô hướng. Cả hai loại mô hình đồ thị này đều sử dụng một đồ thị G trong đó mỗi đỉnh trong đồ thị tương ứng với một biến ngẫu nhiên, và một cạnh kết nối hai biến ngẫu nhiên có nghĩa là phân phối xác suất có thể biểu diễn tương tác trực tiếp giữa hai biến ngẫu nhiên đó.

Các mô hình có hướng sử dụng đồ thị với các cạnh có hướng và biểu diễn sự phân tích thành các phân phối xác suất có điều kiện, như trong ví dụ ở trên. Cụ thể, một mô hình có hướng chứa một thành phần cho mỗi biến ngẫu nhiên X_i trong phân phối, và thành phần đó bao gồm phân phối có điều kiện của X_i dựa trên các đỉnh cha "cha" của X_i, được ký hiệu là $\text{Pa}_G(x_i)$:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \text{Pa}_G(X_i)\right). \tag{3.55}$$

Xem Hình 3.7 để có ví dụ về một đồ thị có hướng và cách phân tích các phân phối xác suất mà nó biểu diễn.



Hình 3.7: Mô hình đồ thị có hướng trên các biến ngẫu nhiên a, b, c, d, và e. Đồ thị này tương ứng với các phân phối xác suất có thể được phân tích là $P(a, b, c, d, e) = P(a) \cdot P\left(b \mid a\right) \cdot P\left(c \mid a, b\right) \cdot P\left(d \mid b\right) \cdot P\left(e \mid c\right)$. Đồ thị này cho phép ta nhanh chóng thấy một số tính chất của phân phối. Ví dụ, a và c tương tác trực tiếp, nhưng a và e chỉ tương tác gián tiếp qua c.

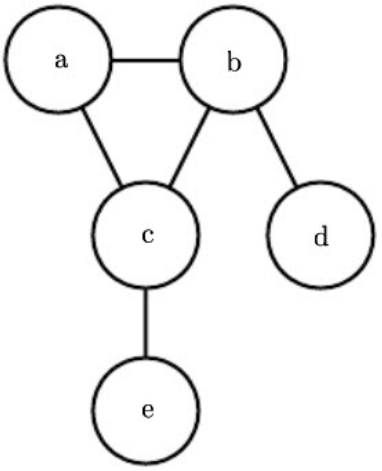
Các mô hình vô hướng sử dụng đồ thị với các cạnh vô hướng và biểu diễn phép phân tích thành một tập các hàm; không giống như trường hợp có hướng, các hàm

này thường không phải là phân phối xác suất. Bất kỳ đồ thị con đầy đủ của G (tập con các đỉnh mà mọi cặp đỉnh đều có cạnh) được gọi là một "clique". Mỗi clique $C^{(i)}$ trong một mô hình vô hướng được liên kết với một nhân tử $\phi^{(i)}\left(C^{(i)}\right)$. Những nhân tử này chỉ là các hàm, không phải là phân phối xác suất. Mỗi nhân tử phải không âm, nhưng không có ràng buộc rằng chúng phải có tổng hoặc tích phân bằng 1 như phân phối xác suất.

Xác suất của một cấu hình của các biến ngẫu nhiên tỷ lệ thuận với tích của tất cả các nhân tử này – những phép gán nào dẫn đến giá trị nhân tử lớn hơn sẽ có khả năng xảy ra cao hơn. Tất nhiên, không có gì đảm bảo rằng tích này sẽ có tổng bằng 1. Vì vậy, ta chia cho một hằng số chuẩn hóa Z, được định nghĩa là tổng hoặc tích phân trên tất cả các trạng thái của tích các hàm ϕ, để có được phân phối xác suất được chuẩn hóa:

$$p\left(X_1, X_2, \ldots, X_n\right) = \frac{1}{Z} \prod_i \phi^{(i)}\left(C^{(i)}\right). \tag{3.56}$$

Xem Hình 3.8 để có ví dụ về một đồ thị có hướng và cách phân tích các phân phối xác suất mà nó biểu diễn.



Hình 3.8: Một mô hình đồ thị vô hướng trên các biến ngẫu nhiên a, b, c, d, và e. Đồ thị này tương ứng với các phân phối xác suất có thể được phân tích bởi $p\left(a, b, c, d, e\right) = \frac{1}{Z}\phi^{(1)}\left(a, b, c\right) \cdot \phi^{(2)}\left(b, d\right) \cdot \phi^{(3)}\left(c, e\right)$. Đồ thị này cho phép ta nhanh chóng thấy một số tính chất của phân phối. Ví dụ, a và c tương tác trực tiếp, nhưng a và e chỉ tương tác gián tiếp thông qua c.

Hãy nhớ rằng những biểu diễn đồ thị này của các phép phân tích là một ngôn ngữ để mô tả các phân phối xác suất. Chúng không phải là các họ phân phối xác

suất loại trừ lẫn nhau. Có hướng hoặc vô hướng không phải là thuộc tính của một phân phối xác suất; đó là thuộc tính của một cách mô tả cụ thể về phân phối xác suất, nhưng bất kỳ phân phối xác suất nào cũng có thể được mô tả theo cả hai cách.

Trong các Phần I và Phần II của cuốn sách này, chúng ta sẽ chỉ sử dụng các mô hình xác suất có cấu trúc như một ngôn ngữ để mô tả các mối quan hệ xác suất trực tiếp mà các thuật toán học máy khác nhau lựa chọn để biểu diễn. Không cần phải hiểu thêm về các mô hình xác suất có cấu trúc cho đến Phần III, khi chúng ta sẽ khám phá các mô hình xác suất có cấu trúc chi tiết hơn trong các chủ đề nghiên cứu.

Chương này đã tổng quan các khái niệm cơ bản của lý thuyết xác suất có liên quan nhất đến học sâu. Còn một tập hợp các công cụ toán học cơ bản khác: các phương pháp tính.

Tài liệu tham khảo

1. Kaare Brandt Petersen, M. S. P. *The Matrix Cookbook* 72 trang (2012).
2. Shilov, G. E. *Linear Algebra* 864 trang (Dover Publications, 1977).
3. Bonaccorso, G. *Machine Learning Algorithms* In lần thứ 2. 514 trang (Packt, 2018).
4. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* In lần thứ 3. 864 trang (O'Reilly, 2022).
5. Goodfellow, I. *Deep Learning* 801 trang (2016).
6. Jaynes, E. T. *Probability Theory: The Logic of Science* In lần thứ 22. 759 trang (Cambridge University Press, 2003).
7. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* 573 trang (Morgan Kaufmann, 1988).
8. Ramsey, F. P. *The Foundations of Mathematics and Other Logical Essays* 310 trang (Routledge, 1926).
9. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* 1098 trang (MIT Press, 2012).
10. Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., và Garcia, R. *Incorporating Second-Order Functional Knowledge for Better Option Pricing* **in** *Advances in Neural Information Processing Systems* (**editors** Leen, T., Dietterich, T., và Tresp, V.) Tập 13 (MIT Press, 2000).
11. Cover, T. M., và Thomas, J. A. *Elements of Information Theory* In lần thứ 2. 774 trang (Wiley-Interscience, 2006).
12. D.J.C., M. *Information Theory, Inference and Learning Algorithms* 642 trang (Cambridge University Press, 2003).