Danh sách hình vẽ

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGUYỄN VÕ LAN THẢO
VÕ NGỌC TRĂM

CÁC PHƯƠNG PHÁP BAYESIAN TRONG MÁY HỌC

TIỂU LUẬN TỐT NGHIỆP TOÁN HỌC

THÀNH PHỐ HỒ CHÍ MINH - 2019

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

NGUYỄN VÕ LAN THẢO
VÕ NGỌC TRĂM

CÁC PHƯƠNG PHÁP BAYESIAN TRONG MÁY HỌC

TIỂU LUẬN TỐT NGHIỆP TOÁN HỌC

CHUYÊN NGÀNH XÁC SUẤT THỐNG KÊ

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. TRẦN VŨ KHANH

THÀNH PHỐ HỒ CHÍ MINH - 2019

Lời cảm ơn

Lời đầu tiên, chúng em xin bày tỏ lòng kính trọng và biết ơn sâu sắc đến Thầy TS. Trần Vũ Khanh, người đã hết lòng giúp đỡ, hướng dẫn chúng em trong suốt quá trình hoàn thành tiểu luận này.

Kính gửi đến Thầy PGS. TS. Đặng Đức Trọng lòng biết ơn chân thành, Quý Thầy đã tận tình giúp đỡ, tạo điều kiện để chúng em có cơ hội được thực hiện tiểu luận.

Đồng thời, chúng em cũng xin trân trọng cảm ơn Quý Thầy Cô trong và ngoài Khoa Toán - Tin học Trường Đại học Khoa học Tự Nhiên TP. Hồ Chí minh đã truyền đạt kiến thức, nhiệt huyết và kinh nghiệm học thuật cho chúng em trong suốt quá trình học tập tại trường.

Cuối cùng chúng em xin kính chúc Quý Thầy Cô dồi dào sức khỏe và thành công trong sự nghiệp cao quý.
Trân trọng.

TP.HCM, ngày 25 tháng 6 năm 2019

Tác giả

Nguyễn Võ Lan Thảo

Võ Ngọc Trăm

Mục lục

Lời nói đầu

Máy học (machine learning) và trí tuệ nhận tạo (AI - Artifical Intelligence) là các lĩnh vực đang được thế giới quan tâm trong thời gian gần đây. Trong bối cảnh mà thời đại công nghệ lên ngôi, vạn vật kết nối internetthì đòi hỏi phải có một phương thức để giúp chúng ta sử dụng. Đồng thời, từ khối dữ liệu khổng lồ mà con người cần dự đoán, vận hành mà không cần đến quá nhiều bàn tay con người.

Lời phát biểu của Arthur Samuel năm 1959 ghi rằng: "Đây là lĩnh vực nghiên cứu cho phép máy tính có thể học mà không được lập trình một cách rõ ràng".

Đến năm 1997, Tom Mitchell đã đưa ra định nghĩa rõ ràng và mang tính kỹ thuật hơn: "Một chương trình máy tính làm ra để học hỏi kinh nghiệm E liên quan đến các nhiệm vụ T và một số phép đo lường hiệu suất P. Nếu hiệu suất của nó là T, được đo bởi P và cải thiện bởi trải nghiệm E".

Nội dung tiểu luận bao gồm 07 chương:

Chương 1. Kiến thức chuẩn bị.

Nội dung chương này sẽ nhắc lại các kiến thức cơ bản của xác suất và giới thiệu một số khái niệm cơ bản trong máy học.

Chương 2. Maximum likelihood và maximum a posterior.

Nội dung chương này giới thiệu hai phương pháp thường dùng để ước lượng tham số mô hình đó là maximum likelihood estimation và maximum likelihood.

Chương 3. Tiên nghiệm liên hợp

Chương 4. Thuật toán tối đa hóa kỳ vọng

Chương 5. Hồi quy tuyến tính

Chương 6. Bayesian model regression

Chương 7. Hồi quy logistic

Chương 1

KIẾN THỨC CHUẨN BỊ

1.1 Thuật ngữ cơ bản của xác suất

- *Biến ngẫu nhiên* là các biến nhận một giá trị ngẫu nhiên đại diện cho kết quả của phép thử. Mỗi giá trị nhận được của biến ngẫu nhiên X được gọi là một thể hiện của X, đây cũng là kết quả của phép thử hay còn được hiểu là một sự kiện.

- *Phân phối xác suất* là hàm xác định xác suất của các kết quả hoặc các giá trị khác nhau của một biến ngẫu nhiên. Phân phối xác suất liên tục được mô tả bằng các hàm mật độ xác suất trong khi phân phối xác suất rời rạc có thể được biểu diễn bằng các hàm khối xác suất.

- *Hàm khối xác suất (PMF - Probability Mass Function)* là hàm cho biết xác suất tại mỗi giá trị x nào đó trong miền giá trị của nó là bao nhiêu.

- *Xác suất có điều kiện* là xác suất để một biến cố A xảy ra, khi biết biến cố B đã xảy ra, ký hiệu $P(A|B)$, và được cho bởi biểu thức

$$P(A|B) = \frac{P(AB)}{P(B)},$$

với điều kiện $P(B) \neq 0$.

- *Hàm phân phối xác suất đồng thời* hay *hàm phân phối tích luỹ xác suất đồng thời (Joint CDF - Joint Cumulative Probability Distribution Function)* của 2

biến ngẫu nhiên X, Y được định nghĩa như sau:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y), x, y \in \mathbb{R}.$$

1.2 Các khái niệm cơ bản trong máy học

Máy học (machine learning) là một ngành học thuộc khoa học máy tính, giúp máy tính có khả năng tự học mà không phải lập trình một cách rõ ràng.

Thuật toán máy học (machine learning algorithm) được chia làm ba loại chính là học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning). Trong tiểu luận này, chúng ta chỉ tập trung vào học có giám sát.

1.2.1 Học có giám sát

Học có giám sát là thuật toán dự đoán đầu ra của một hoặc nhiều dữ liệu mới dựa trên các cặp (đầu vào, đầu ra) đã biết từ trước. Đây cũng là nhóm phổ biến nhất trong các thuật toán của máy học.
Xét trên phương diện toán học, học có giám sát là khi chúng ta có một tập hợp biến đầu vào $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ và một tập đầu ra tương ứng $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$, trong đó $\mathbf{x}_i, \mathbf{y}_i$ là các vector. Các cặp dữ liệu biết trước $(\mathbf{x}_i, \mathbf{y}_i) \in (\mathbf{X}, \mathbf{Y})$ tạo nên tập huấn luyện. Từ tập huấn luyện này chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập $\mathbf{X}$ sang một phần tử (xấp xỉ) tương ứng của tập $\mathbf{Y}$:

$$\mathbf{y}_i \approx f(\mathbf{x}_i), \forall i = 1, 2, ..., N.$$

Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu $\mathbf{x}$ mới, chúng ta có thể tính được nhãn tương ứng của nó $\mathbf{y} = f(\mathbf{x})$.

1.2.2 Hàm mất mát và tham số của mô hình

Mỗi mô hình machine learning được mô tả bởi các tham số mô hình. Công việc của một thuật toán trong máy học là đi tìm các tham số mô hình phù hợp với mỗi

bài toán. Việc đi tìm các tham số mô hình có liên quan mật thiết đến các phép đánh giá. Mục đích của chúng ta là đi tìm các tham số mô hình sao cho các phép đánh giá cho kết quả tốt nhất. Trong bài toán phân loại, kết quả tốt có thể được hiểu là ít điểm dữ liệu bị phân lớp sai nhất. Trong bài toán hồi quy, kết quả tốt là khi sự sai lệch giữa đầu ra dự đoán và đầu ra thực sự là ít nhất.

Quan hệ giữa một phép đánh giá và các tham số mô hình thường được mô tả thông qua một hàm số được gọi là hàm mất mát (loss function, hay cost function). Hàm mất mát này thường có giá trị nhỏ khi phép đánh giá cho kết quả tốt và ngược lại. Việc đi tìm các tham số mô hình sao cho phép đánh giá trả về kết quả tốt tương đương với việc tối thiểu hàm mất mát. Như vậy, việc xây dựng một mô hình máy học chính là việc đi giải một bài toán tối ưu. Quá trình đó có thể được coi là quá trình học (leanring) của máy (machine).

Tập hợp các tham số mô hình thường được ký hiệu bằng θ , hàm mất mát của mô hình thường được ký hiệu là $\mathcal{L}(\theta)$ hoặc $J(\theta)$, trong tiểu luận này chúng ta kí hiệu là $J(\theta)$. Bài toán tối thiểu hàm mất mát để tìm tham số mô hình thường được viết dưới dạng:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta),$$

ký hiệu $\underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$ được hiểu là giá trị của θ để hàm số $\mathcal{L}(\theta)$ đạt giá trị nhỏ nhất. Khi sử dụng argmin chúng ta phải chỉ rõ nó được thực hiện theo các biến số nào bằng cách ghi các biến số ở dưới min (ở đây là θ). Nếu hàm số chỉ có một biến số, ta có thể bỏ qua biến số đó dưới min. Tuy nhiên, biến số nên được ghi rõ ràng để giảm thiểu sự nhầm lẫn, argmax cũng được sử dụng một cách tương tự khi ta cần tìm giá trị của các biến số để một hàm đạt số đạt giá trị lớn nhất.

Một hàm $\mathcal{L}(\theta)$ bất kỳ có thể có rất nhiều giá trị của θ để nó đạt giá trị nhỏ nhất, hoặc cũng có thể nó không có chặn dưới. Thậm chí, việc tìm giá trị nhỏ nhất của một hàm số đôi khi là không khả thi. Trong machine learning cũng như nhiều bài toán tối ưu thực tế, việc chỉ cần tìm ra một bộ tham số θ làm cho hàm mất mát đạt giá trị nhỏ nhất, hoặc thậm chí đạt một giá trị cực tiểu, thường mang lại các

kết quả khả quan.

1.2.3 Biến tiềm ẩn

Biến tiềm ẩn là một biến ngẫu nhiên không thể được quan sát trực tiếp cả trong giai đoạn đào tạo và giai đoạn thử nghiệm, nhưng có thể được suy ra từ các biến quan sát trực tiếp thông qua mô hình toán học. Cụ thể, các biến như chiều dài, chiều cao, vận tốc, điểm thi,... là các biến có thể đo được trực tiếp. Nhưng một số biến khác như lòng vị tha, niềm hạnh phúc,... chúng ta không có một thang đo định lượng nào có thể đo được và chúng ta gọi đó là các biến tìm ẩn.

Các mô hình toán học nhằm giải thích các biến quan sát theo các biến tiềm ẩn được gọi là các *mô hình biến tiềm ẩn*. Các mô hình biến tìm ẩn thường được sử dụng trong nhiều ngành bao gồm kinh tế học, vật lý, tâm lý học, máy học, quản lý và khoa học xã hội.

Một lợi thế của việc sử dụng các biến tiềm ẩn là chúng có thể phục vụ để giảm số chiều của dữ liệu. Một số lượng lớn các biến quan sát có thể được tổng hợp trong một mô hình để thể hiện một khái niệm cơ bản, giúp dễ hiểu dữ liệu hơn. Theo nghĩa này, biến tiềm ẩn phục vụ một chức năng tương tự như các lý thuyết khoa học. Đồng thời, các biến tìm ẩn liên kết dữ liệu có thể quan sát được trong thế giới thực với dữ liệu tượng trưng trong thế giới được mô hình hóa.

1.3 Mô hình xác suất

Định nghĩa 1.3.1. Mô hình xác suất là một tập hợp các phân phối xác suất $p(x|\theta)$ Chúng ta ký hiệu một họ phân phối là $p(\cdot)$, tham số θ chưa biết.

Ví dụ 1.3.1. *Dữ liệu mô hình tuân theo phân phối Gaussian với hàm mật độ* $p(x|\theta)$*, tham số* $\theta = \{\mu, \sum\}$*. Giả sử dữ liệu của chúng ta độc lập và cùng phân phối (independ identical distribute, viết tắt là iid).*

Khi đó, chúng ta viết $\mathbf{x} \overset{iid}{\sim} p(x|\theta), \forall i = 1, ..., n$.
Hàm mật độ đồng thời được viết dưới dạng

$$p(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

1.4 Phân phối Gaussian

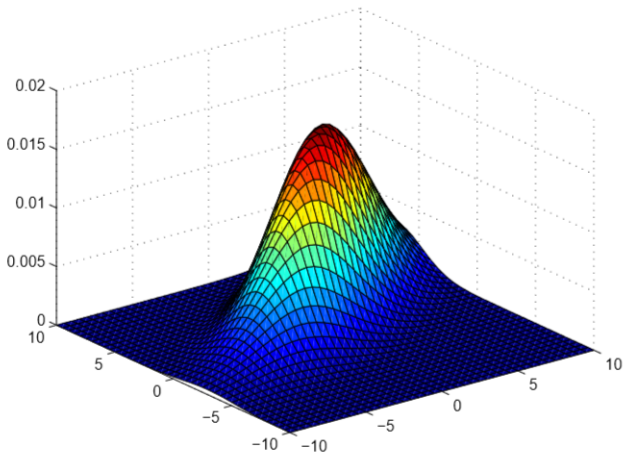
Cho dãy các biến ngẫu nhiên $X_1, .., X_n$ $(X_i \in \mathbf{R}^n)$ có phân phối chuẩn nhiều chiều (hay còn gọi là phân phối Gaussian) với trung bình $\mu \in \mathbf{R}^n$ và hiệp phương sai $\Sigma \in \mathbf{S}^n_{++}$ với $\mathbf{S}^n_{++}$ là không gian của ma trận xác định dương đối xứng được định nghĩa:

$$\mathbf{S}^n_{++} = \left\{A \in \mathbf{R}^{n\times n} : A = A^\top, x^\top Ax > 0, \forall x \in \mathbf{R}^n, x \neq 0\right\}.$$

Khi đó, hàm mật độ của phân phối là:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Chúng ta viết $X \sim N(\mu, \Sigma)$.

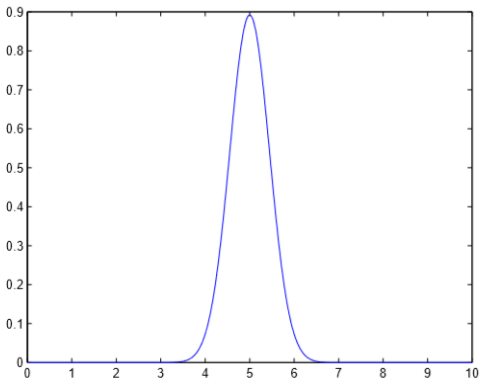


Hình 1.1. Đồ thị cho phân phối Gaussian với 2 biến X_1 và X_2.

1.4.1 Mối quan hệ với Gaussian đơn biến

Hàm mật độ của phân phối Gaussian đơn biến

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$



Hình 1.2. Đồ thị cho phân phối Gaussian đơn biến với $\mu = 0, \sigma = 1$.

Ở đây, đối số của hàm mũ là một hàm bậc hai theo biến x: $-\frac{1}{2\sigma^2}(x-\mu)^2$.
Hơn nữa, vì hệ số của hàm bậc hai là một số âm nên đồ thị của phân phối Gaussian đơn biến có dạng parabol hướng xuống.
Hệ số $\frac{1}{\sqrt{2\pi}\sigma}$ là hằng số không phụ thuộc vào x và là hệ số chuẩn hóa (normalization factor) để

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = 1.$$

Trong trường hợp của hàm mật độ Gaussian nhiều chiều, đối số của hàm mũ là

$$-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu).$$

Đối số này có dạng toàn phương theo vector biến x. Vì Σ xác định dương và nghịch đảo của bất kỳ ma trận xác định dương nào cũng xác định dương nên bất kỳ veơtor z nào khác không đều có $z^\top\Sigma^{-1}z > 0$. Từ đó dẫn đến với bất kỳ vector $x \neq \mu$ thì

$$(x-\mu)^\top\Sigma^{-1}(x-\mu) > 0.$$

Giống như trường hợp của Gaussian đơn biến, chúng ta có thể nghĩ đối số của hàm mũ trong trường hợp nhiều biến như một parabol hướng xuống. Hệ số phía trước $\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}$ tuy phức tạp hơn trường hợp đơn biến nhưng vẫn phụ thuộc vào x và do đó nó được xem là hệ số để đảm bảo rằng

$$\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right) dx_1 dx_2 \ldots dx_n = 1.$$

1.4.2 Ma trận hiệp phương sai

Định nghĩa 1.4.1. Cho X, Y là hai biến số ngẫu nhiên với trung bình $E(X) = \mu_X$, $E(X) = \mu_X$, hiệp phương sai của X, Y ký hiệu $cov(X, Y)$ xác định bởi

$$\mathrm{cov(X, Y)} = \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right].$$

Khi làm việc với mô hình nhiều biến, ma trận hiệp phương sai ký hiệu là Σ, có dạng $n \times n$ và hiệp phương sai tại điểm (i, j) ký hiệu là $Cov[X_i, X_j]$.
Từ đó, chúng ta có cách khác để mô tả ma trận hiệp phương sai của một vector ngẫu nhiên X là

$$\Sigma = \mathrm{E}[(\mathrm{X}-\mu)(\mathrm{X}-\mu)^{\top}] = E[X\mathrm{X}^{\top}] - \mu\mu^{\top}. \tag{1.1}$$

Dấu bằng thứ nhất của (1.1) có thể được chứng minh như sau:

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ Cov[X_2, X_1] & \cdots & Cov[X_n, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_1] \end{bmatrix} \\
&= \begin{bmatrix} E[(X_1 - \mu_1)^2] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & \cdots & E[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & \cdots & E[(X_n - \mu_n)^2] \end{bmatrix} \\
&= E \begin{bmatrix} (X_1 - \mu_1)^2 & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & \cdots & (X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & \cdots & (X_n - \mu_n)^2 \end{bmatrix} \\
&= E \left[\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_n - \mu_n \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 & X_2 - \mu_2 & \cdots & X_n - \mu_n \end{bmatrix} \right] \\
&= E\left[(X - \mu)(X - \mu)^\top\right].
\end{aligned}
$$

Dấu bằng thứ nhất và dấu bằng thứ hai xảy ra do định nghĩa của ma trận hiệp phương sai. Dấu bằng thứ ba xảy ra vì kỳ vọng của một ma trận được tạo ra bằng cách lấy kỳ vọng của từng thành phần trong ma trận. Tiếp đến, dấu bằng thứ tư xảy ra dựa trên thực tế là với mọi vector z bất kỳ, $z \in \mathbf{R}^n$,

$$
zz^\top = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} = \begin{bmatrix} z_1 z_1 & z_1 z_2 & \cdots & z_1 z_n \\ z_2 z_1 & z_2 z_1 & \cdots & z_2 z_n \\ \vdots & \vdots & \ddots & \vdots \\ z_n z_1 & z_n z_2 & \cdots & z_n z_n \end{bmatrix}.
$$

Tiếp theo, chúng ta chứng minh dấu bằng thứ hai của (1.1).

Chúng ta có các biến đổi sau,

$$\begin{aligned}
\Sigma &= E\left[(X-\mu)(X-\mu)^\top\right] \\
&= E\left[(X-\mu)(X^\top-\mu^\top)\right] \\
&= E\left[XX^\top - X\mu^\top - \mu X^\top + \mu\mu^\top\right] \\
&= E[XX^\top] - E[X\mu^\top] - E[X^\top\mu] + \mu\mu^\top \\
&= E[XX^\top] - E[X]\mu^\top - E[X^\top]\mu + \mu\mu^\top \\
&= E[XX^\top] - \mu\mu^\top - \mu\mu^\top + \mu\mu^\top \\
&= E[XX^\top] - \mu\mu^\top. \qquad \square
\end{aligned}$$

1.5 Nhắc lại một số định lý và tính chất dùng trong chứng minh

Định lý 1.5.1. *Nếu* $u = \mathbf{a}^\top\mathbf{x} = \mathbf{x}^\top\mathbf{a}$ *với* $\mathbf{a}^\top = (a_1, a_2, ..., a_p)$ *là vector hằng số, thì:*

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}^\top\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^\top\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}.$$

Định lý 1.5.2. *Nếu* $u = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ *và* $\boldsymbol{A}$ *là ma trận hằng số đối xứng thì :*

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^\top\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

Định lý 1.5.3. *Nếu* $\mathbf{X_1}$ *tuân theo phân phối chuẩn* $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *và* $\mathbf{X_2}$ *tuân theo phân phối chuẩn* $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *thì* $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ *tuân theo phân phối chuẩn*

$$\mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Tính chất 1.5.4. *Nếu* $\mathbf{X}$ *tuân theo phân phối chuẩn* $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *thì*

$$\underset{(n\times d)}{\mathbf{A}} \underset{(d\times 1)}{\mathbf{X}} = \begin{bmatrix} a_{11}\mathbf{X_1} + a_{12}\mathbf{X_2} + ... + a_{1d}\mathbf{X_d} \\ a_{21}\mathbf{X_1} + a_{22}\mathbf{X_2} + ... + a_{2d}\mathbf{X_d} \\ \vdots \\ a_{n1}\mathbf{X_1} + a_{n2}\mathbf{X_2} + ... + a_{nd}\mathbf{X_d} \end{bmatrix}$$

tuân theo phân phối chuẩn $\mathcal{N}_n(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

Tính chất 1.5.5. *Nếu* X *tuân theo phân phối chuẩn* $\mathcal{N}(\mu_1, \sigma_1^2)$ *và* Y *tuân theo phân phối chuẩn* $\mathcal{N}(\mu_2, \sigma_2^2)$ *thì* $X+Y$ *tuân theo phân phối chuẩn* $\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

Chương 2

Maximum Likelihood và Maximum A Posterior

2.1 Giới thiệu

Có rất nhiều mô hình máy học được xây dựng dựa trên các mô hình thống kê (statistical models). Với một mô hình thông kê bất kỳ, ký hiệu θ là tập hợp tất cả các tham số của mô hình đó. Học (learning) chính là quá trình ước lượng (estimate) bộ tham số θ sao cho mô hình tìm được khớp với phân phối của dữ liệu nhất. Quá trình này còn được gọi là ước lượng tham số (parameter estimation).

Trong các mô hình machine learning thống kê, có 2 cách ước lượng tham số thường được dùng. Cách thứ nhất gọi là *maximum likelihood estimation* hay *ML estimation* hoặc gọi tắt là *MLE*: cách này chỉ dựa trên dữ liệu đã biết trong tập huấn luyện. Cách thứ hai gọi là *maximum a posteriori estimation* hay *MAP estimation*: cách này không những dựa trên tập huấn luyện mà còn dựa trên những thông tin đã biết trước của các tham số. Những thông tin này có thể có được bằng *cảm quan* của người xây dựng mô hình. *Cảm quan* càng rõ ràng, càng hợp lý thì khả năng thu được bộ tham số tốt là càng cao. Trong chương này, chúng ta cùng tìm hiểu ý tưởng và cách giải quyết bài toán ước lượng tham số mô hình theo *MLE* hoặc theo *MAP estimation*.

2.2 Maximum likelihood Estimation

Giả sử chúng ta có các điểm dữ liệu $\mathbf{x}_1, ..., \mathbf{x}_n$ và ta biết các điểm dữ liệu này tuân theo một phân phối nào đó được mô tả bởi bộ tham số θ.
Việc đi tìm bộ tham số θ sao cho xác suất sau đây đạt giá trị lớn nhất

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{x}_1, ..., \mathbf{x}_n|\theta). \tag{2.1}$$

được gọi là *maximum likelihood estimation* của θ. Ở đây, $p(\mathbf{x}_1, ..., \mathbf{x}_n)$ chính là xác suất để toàn bộ các sự kiện $\mathbf{x}_1, ..., \mathbf{x}_n$ đồng thời xảy ra, xác suất đồng thời này được gọi là *likelihood*.
Trong mô hình này vì ta đã có trước bộ dữ liệu huấn luyện nên điều chúng ta mong muốn là làm sao để xác suất đồng thời này phải càng cao càng tốt. Từ đó ta đưa đến bài toán tối đa hàm mục tiêu $p(\mathbf{x}_1, ..., \mathbf{x}_n)$.

Việc đi giải bài toán trong (2.1) thường rất phức tạp. Do đó, cách tiếp cận phổ biến là chúng ta giả sử rằng các điểu dữ liệu $\mathbf{x}_i$ độc lập với nhau. Nhờ dữ kiện độc lập và cùng tuân theo một phân phối mà chúng ta có thể viết:

$$p(\mathbf{x}_1, ..., \mathbf{x}_n) = p(\mathbf{x}_1|\theta)p(\mathbf{x}_2|\theta)...p(\mathbf{x}_n|\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\theta).$$

Lúc này, bài toán ở (2.1) được đưa về bài toán tìm giá trị của tham số θ để phương trình sau đạt giá trị cực đại:

$$\nabla_\theta \prod_{i=1}^{n} p(\mathbf{x}_i|\theta) = 0.$$

Rõ ràng việc tính đạo hàm $\nabla_\theta \prod_{i=1}^{n} p(\mathbf{x}_i|\theta)$ rất phức tạp, do đó chúng ta cần phải tìm giải pháp khác khả quan hơn. Chúng ta biết hàm logarithm là hàm đơn điệu tăng trên R^+ nên thay vì đánh giá hàm $\prod_i f_i$, ta sẽ đánh giá hàm $\ln(\prod_i f_i)$.Và chúng ta có

$$\ln(\prod_i f_i) = \sum_i \ln(f_i).$$

Bên cạnh đó, việc lấy logarithm của một hàm g dương không làm thay đổi vị trí đạt cực đại hay cực tiểu của hàm đó,

$$\underset{y}{\operatorname{argmax}} \ln g(y) = \underset{y}{\operatorname{argmax}}\ g(y).$$

Đối với bài toán tìm ước lượng hợp lý cực đại, chúng ta quan tâm giá trị của tham số θ nhiều hơn là giá trị của hàm tại tham số θ. Do đó, khi $\underset{y}{max} \ln g(y) \neq \max\limits_{y} g(y)$ cũng không làm ảnh hưởng đến mục đích của bài toán. Bởi vì

$$\begin{aligned}
\hat{\theta}_{ML} &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(\mathbf{x}_i|\theta), \\
&= \underset{\theta}{\operatorname{argmax}} \ln \left(\prod_{i} p(\mathbf{x}_i|\theta) \right), \\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i} \ln p(\mathbf{x}_i|\theta).
\end{aligned}$$

Chúng ta tìm $\hat{\theta}_{ML}$ thỏa

$$\nabla_\theta \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\theta) = \sum_{i=1}^{n} \nabla_\theta \ln p(\mathbf{x}_i|\theta) = 0. \tag{2.2}$$

Ví dụ 2.2.1. *Cho mô hình gồm các biến ngẫu nhiên có phân phối Gaussian trên* $\boldsymbol{R}^d$ *với* $\mu \in \boldsymbol{R}^d, \Sigma \in \boldsymbol{S}^d_{++}$, μ, Σ *chưa biết. Giả sử* $x_i \overset{iid}{\sim} p(x|\mu, \Sigma)$. *Tìm MLE của mô hình trên.*

Chúng ta thấy ví dụ này giống như ở (2.2) nên để tìm MLE của mô hình chúng ta cần giải phương trình:

$$\nabla_{(\mu,\Sigma)} \sum_{i=1}^{n} \ln p(x_i|\mu, \Sigma) = 0.$$

Trước tiên, ta tìm $\nabla_\mu \sum\limits_{i=1}^{n} \ln p(x_i|\mu, \Sigma) = 0$.

Chúng ta có các biến đổi tương đương sau:

$$\begin{aligned}
\nabla_\mu \sum_{i=1}^{n} \ln \left[\frac{1}{\sqrt{(2\pi)^d . |\Sigma|}} exp\left(-\frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right) \right] &= 0, \\
\nabla_\mu \sum_{i=1}^{n} \ln \left[\left((2\pi)^d . |\Sigma| \right)^{-\frac{1}{2}} exp\left(-\frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right) \right] &= 0, \\
\nabla_\mu \sum_{i=1}^{n} \left[-\frac{1}{2} \ln \left((2\pi)^d . |\Sigma| \right) - \frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right] &= 0, \\
-\frac{1}{2} \sum_{i=1}^{n} \nabla_\mu \left[\ln \left((2\pi)^d . |\Sigma| \right) + (x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right] &= 0, \\
-\frac{1}{2} \sum_{i=1}^{n} \nabla_\mu \left[{x_i}^\top \Sigma^{-1} x_i - 2\mu^\top \Sigma^{-1} x_i + \mu^\top \Sigma^{-1} \mu \right] &= 0, \\
-\Sigma^{-1} \sum_{i=1}^{n} (x_i - \mu) &= 0.
\end{aligned}$$

Vì Σ xác định dương, do đó

$$\sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Tiếp đó

$$\begin{aligned}
\nabla_\Sigma \sum_{i=1}^{n} \left[-\frac{1}{2} \ln \left((2\pi)^d . |\Sigma| \right) - \frac{1}{2}(x_i-\mu)^\top \Sigma^{-1}(x_i-\mu) \right] &= 0, \\
\frac{-n}{2} \nabla_\Sigma \ln |\Sigma| - \frac{1}{2} \nabla_\Sigma trace \left(\Sigma^{-1} \sum_{i=1}^{n} (x_i-\mu)(x_i-\mu)^\top \right) &= 0, \\
\frac{-n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-2} \sum_{i=1}^{n} (x_i-\mu)(x_i-\mu)^\top &= 0.
\end{aligned}$$

Vậy

$$\begin{aligned}
\mu &= \hat{\mu}_{ML}, \\
\hat{\Sigma}_{ML} &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^\top.
\end{aligned}$$

Ví dụ 2.2.2. *Giả sử, chúng ta có một đồng xu nhưng chúng ta không biết đồng*

xu này có đồng chất hay không. Nói cách khác, chúng ta không biết liệu xác suất nhận mặt ngửa có bằng xác suất nhận mặt sấp hay không? Trong trường hợp này, ta thực hiện tung đồng xu một cách ngẫu nhiên một số lần và xem xét các kết quả mà chúng ta nhận được. Tìm ước lượng cho tham số của mô hình.

Trước tiên, để dễ làm việc chúng ta gọi X là biến ngẫu nhiên đại diện cho kết quả của lần tung (X có giá trị là 1 nếu tung được mặt ngửa và X có giá trị là 0 nếu tung được mặt sấp) và θ là xác suất nhận được mặt ngửa.

Ở đây, vì chúng ta không biết giá trị của θ nên chúng ta sẽ giả sử $\theta = 0.7$ và dùng hàm *np.random.choice ()* để xem kết quả của 10 lần tung đồng xu ngẫu nhiên.

Đoạn code được sử dụng trong phần mềm Python như sau:

```
import numpy as np
n = 10 theta = 0.7
X_arr = np.random.choice([0, 1], p=[1-theta, theta],
size=10)
X_arr
array([1, 1, 0, 1, 0, 1, 1, 1, 1, 1])
```

Kết quả ta nhận được là có 8 mặt sấp và 2 mặt ngửa trong tổng số 10 lần tung ngẫu nhiên. Khi đó, bằng trực giác ta có thể tính được:

$$\hat{\theta} = \frac{8}{10} = 0.8. \tag{2.3}$$

Bây giờ, để kiểm tra lại kết quả trên ta dùng *MLE* để ước lượng tham số cho θ. Vì kết quả của một lần tung đồng xu chỉ có 2 khả năng là ngửa hoặc sấp, do đó ở ví dụ này chúng ta sử dụng phân phối Bernoulli với xác suất như sau:

$$p(x = 1) = \theta,$$

$$p(x = 0) = 1 - \theta.$$

Khi đó, hàm khối xác suất của x có dạng là $p(x|\theta) = \theta^x(1-\theta)^{1-x}$. Do đó, hàm *maximum likelihood* của $\mathbf{x} = (x_1, x_2, ..., x_n)$ lúc này là:

$$\begin{aligned}
\mathcal{L}(\theta; \mathbf{x}) &= \ln \prod_i^n p(x_i|\theta), \\
&= \sum_i^n \ln p(x_i|\theta), \\
&= \sum_i^n \ln \theta^{x_i}(1-\theta)^{1-x_i}, \\
&= \sum_i^n \ln \theta^{x_i} + \sum_i^n \ln(1-\theta)^{1-x_i}, \\
&= \sum_i^n x_i \ln \theta + \sum_i^n (1-x_i)\ln(1-\theta).
\end{aligned}$$

Như ví dụ 2.2.1, chúng ta đi giải phương trình sau

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta; \mathbf{x}) = \frac{\sum\limits_i x_i}{\theta} - \frac{n - \sum\limits_i x_i}{1-\theta} = 0.$$

Chúng ta dễ dàng có các biến đổi tương đương tiếp theo

$$\begin{aligned}
(1-\theta)\sum_i x_i - \theta(n - \sum_i x_i) = 0, \\
\sum_i x_i - n\theta = 0.
\end{aligned}$$

Từ đó ta suy ra

$$\hat{\theta} = \frac{\sum\limits_i x_i}{n}. \tag{2.4}$$

Thay các giá trị mà chúng ta nhận được thừ thí nghiệm, $\mathbf{x} = (1,1,0,1,0,1,1,1,1,1)$ vào (2.4) chúng ta dễ dàng có $\hat{\theta} = \frac{8}{10} = 0.8$.
Chú ý rằng, khi x_i có thể nhận các giá trị 0 hoặc 1 thì ta thấy $\sum_i x_i = n_H$, đây chính là tổng tất cả các lần xảy ra mặt ngửa
từ thí nghiệm. Và trên thực tế, nó phù hợp với kết quả ở (2.3).

2.3 Quy tắc Bayes

2.3.1 Giới thiệu

Quy tắc Bayes cung cấp cho chúng ta cách để cập nhật niềm tin ban dầu (prior belief) dựa trên sự xuất hiện của bằng chứng mới. Ví dụ, chúng ta muốn đưa ra xác suất mà một người bị ung thư, ban đầu chúng ta chỉ có thể đưa ra phần trăm mà một người dân bất kỳ mắc bệnh ung thư. Tuy nhiên, khi có thêm bằng chứng rằng người đó là người hút thuốc lá thì chúng ta có thể cập nhật xác suất của mình, vì xác suất một người bị ung thư sẽ cao hơn nếu biết người này hút thuốc lá.

2.3.2 Ý tưởng

Quy tắc Bayes được xây dựng dựa trên mối quan hệ giữa xác suất có điệu kiện và xác suất đồng thời. Đầu tiên, dựa vào công thức của xác suất có điều kiện chúng ta có:

$$P(A \cap B) = P(A|B)P(B),$$
$$P(B \cap A) = P(B|A)P(A).$$

Từ đây, ta suy ra

$$P(A|B)P(B) = P(B|A)P(A). \tag{2.5}$$

Tiếp theo, ta chia cả hai vế của (2.5) cho $P(B)$, ta nhận được công thức của quy tắc Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.6}$$

Trong đó

- $A \cap \overline{A} = \emptyset, A \cup \overline{A} = U$ với U là toàn bộ không gian .
- $P(A|B)$ là hậu nghiệm (posterior), chỉ xác suất chúng ta muốn tính. Trong ví dụ nêu ở phần giới thiệu thì đây là xác suất một người bị mắc bệnh ung thư

biết rằng người này hút thuốc lá.

- $P(B|A)$ là khả năng (likelihood), chỉ xác suất khi quan sát bằng chứng mới nhận giả thuyết ban đầu của chúng ta. Trong ví dụ trên thì $P(B|A)$ chính là xác suất của một người hút thuốc lá khi biết người đó mắc bệnh ung thư.

- $P(A)$ là tiên nghiệm (prior), chỉ xác suất giả thuyết của chúng ta khi chưa có bất kỳ thông tin nào trước đó. Trong ví dụ trên, $P(A)$ là xác suất một người mắc bệnh ung thư.

- $P(B)$ là bằng chứng (evidence) hay marginal likelihood, được biểu diễn dưới dạng tổng trong trường hợp rời rạc hoặc dưới dạng tích phân trong trường hợp liên tục,

$$P(B) = \begin{cases} \sum_{i=1}^{n} P(B|A_i)P(A_i), \\ \int P(B|A)P(A)dA. \end{cases}$$

  Trong ví dụ trên, P(B) là xác suất một người hút thuốc lá và nó được biểu diễn dưới dạng:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}).$$

Từ các diễn giải này, công thức (2.6) có thể được viết lại dưới dạng

$$Posterior = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}, \tag{2.7}$$

đây cũng là một cách viết khác của quy tắc Bayes.

2.3.3 Ví dụ

Xét ví dụ chẩn đoán bệnh ung thư, chúng ta có thể chỉ ra rằng quy tắc Bayes giúp chúng ta nhận được một ước lượng tốt hơn. Để đánh giá sự khác biệt khi dùng quy tắc Bayes, bây giờ chúng ta sẽ gán những giá trị cụ thể cho ví dụ. Giả sử xác suất mắc bệnh ung thư là 0.05 (nghĩa là có 5% số người bị ung thư), xác suất hút thuốc lá là 0.1 (nghĩa là có 10% số người hút thuốc lá) và 20% người bị ung thư là hút thuốc lá: P(smoker|cancer) = 0.2. Ban đầu, xác suất cho người bị bệnh ung thư của chúng ta chính là niềm tin ban đầu, 0.05. Tuy nhiên, khi chúng ta có thêm bằng chứng mới, chúng ta tính được:

$$P(\text{cancer} \mid \text{smoker}) = \frac{P(\text{smoker} \mid \text{cancer})P(\text{cancer})}{P(\text{smoker})} = \frac{0.2 \times 0.05}{0.1} = 0.1.$$

Rõ ràng, nhờ có thêm thông tin mà chúng ta có thể nhận được một kết quả xác suất tốt hơn. Ban đầu xác suất mắc bệnh ung thư của chúng ta là 0.05, nhưng khi sử dụng thêm bằng chứng "smoker" chúng ta nhận được một xác suất cao hơn, 0.1. Xác suất lúc sau chúng ta tính được gấp đôi xác suất ban đầu. Điều này khá hợp lý vì thực tế cuộc sống chúng ta biết rằng hút thuốc lá gây ung thư. Như vậy, quy tắc Bayes cho phép chúng ta cập nhật niềm tin ban đầu bằng cách sử dụng các thông tin có liên quan.

2.3.4 Maximum a Posterior Estimation

2.3.5 Ý tưởng

Xét lại ví dụ 2.2 về bài toán tung đồng xu. Nếu ta tung đồng xu 5000 lần và nhận được 1000 lần kết quả là mặt head, khi đó ta có thể đánh giá xác suất của mặt head là $\frac{1}{5}$ và việc đánh giá này là đáng tin cậy vì số lượng mẫu lớn.
Ngược lại, nếu chúng ta tung đồng xu 5 lần và chỉ nhận được 1 lần kết quả là mặt head, thì theo như kết quả ở (2.4) của phương pháp MLE, xác suất để có một mặt head được đánh giá là $\frac{1}{5}$. Tuy nhiên, con số $\frac{1}{5}$ với chỉ có 5 kết quả nên ước lượng này không đáng tin, nhiều khả năng việc đánh giá đã bị *overfitting*. Khi tập huấn luyện quá nhỏ, chúng ta cần phải quan tâm tới một giả thiết của các tham số. Trong ví dụ tung đồng xu này, giả thiết của chúng ta là xác suất nhận được mặt head phải gần bằng $\frac{1}{2}$.

Maximum A Posteriori (MAP) ra đời nhằm giải quyết vấn đề này. Trong MAP, chúng ta giới thiệu một giả thiết biết trước, được gọi là prior, của tham số θ. Từ giả thiết này, chúng ta có thể suy ra các khoảng giá trị và phân bố của tham số.
Không giống như MLE, công việc của MAP là tìm θ sao cho tối đa hóa phân phối

posterior của tham số, tức là:

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \underbrace{p(\theta|\mathbf{x}_1, ..., \mathbf{x}_n)}_{posterior}. \tag{2.8}$$

Biểu thức $p(\theta|\mathbf{x}_1, ..., \mathbf{x}_n)$ còn được gọi là xác suất posterior của θ. Chính vì vậy mà việc ước lượng θ theo (2.8) được gọi là *Maximum A Posteriori*.

Thông thường, hàm tối ưu trong (2.8) khó để xác định trực tiếp. Vì vậy, để giải được bài toán MAP, ta thường sử dụng quy tắc Bayes.

$$p(\theta|\mathbf{x}_1, ..., \mathbf{x}_n) = \frac{p(\mathbf{x}_1, ..., \mathbf{x}_n|\theta)p(\theta)}{p(\mathbf{x}_1, ..., \mathbf{x}_n)}. \tag{2.9}$$

Tiếp theo, vì mẫu số của (2.9) không phụ thuộc vào tham số θ nên ta có:

$$p(\theta|\mathbf{x}_1, ..., \mathbf{x}_n) \propto p(\mathbf{x}_1, ..., \mathbf{x}_n|\theta)p(\theta). \tag{2.10}$$

Ở đây, ta dùng kí hiệu $\propto$ để chỉ mối quan hệ tỉ lệ. Bên cạnh đó, nếu chúng ta giả thiết về sự độc lập của các $\mathbf{x}_i$ thì chúng ta nhận được

$$p(\theta|\mathbf{x}_1, ..., \mathbf{x}_n) \propto \prod_{i=1}^{n} p(\mathbf{x}_i|\theta)p(\theta). \tag{2.11}$$

Khi đó, bài toán MAP ở (2.8) trở thành:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{n} p(\mathbf{x}_i|\theta)p(\theta) \right].$$

Như vậy, sự khác biệt lớn nhất giữa hai bài toán tối ưu MLE và MAP là hàm mục tiêu của MAP có thêm $p(\theta)$, tức phân phối của θ. Phân phối này chính là những thông tin ta biết trước về θ và được gọi là *prior*. Ta kết luận rằng *posterior* tỉ lệ thuận với tích của *likelihood* và *prior*.
Vậy việc chọn *prior* như thế nào? Chúng ta sẽ làm rõ vấn đề này ở chương 3.

2.3.6 Ví dụ so sánh hai phương pháp MLE và MAP:

Phương pháp MAP

Cho bộ dữ liệu $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ với các x_i độc lập cùng tuân theo phân phối Poisson(λ). Giả sử chúng ta biết $\lambda \sim \Gamma(x|k, \theta)$ với tham số k = 3 và $\theta = 1$. Bài toán yêu cầu tìm MAP của λ.

Đầu tiên, chúng ta viết hàm mật độ xác suất của phân phối Gamma:

$$\Gamma(x|k, \theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)},$$

với $x > 0, k > 0$ và $\theta > 0$. Hàm $\Gamma(k)$ là hàm Gamma tổng quát hàm giai thừa, nghĩa là khi k là số nguyên thì $\Gamma(k) = (k-1)!$. Ước lượng MAP của tham số cần tìm được viết dưới dạng:

$$\lambda_{MAP} = \arg\max_{\lambda \in (0,\infty)} \{p(\mathcal{D}|\lambda)p(\lambda)\}.$$

Trong đó, hàm likelihood là:

$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n|\lambda), \\ &= \prod_{i=1}^{n} p(x_i|\lambda), \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n}}{\prod_{i=1}^{n} x_i!}. \end{aligned} \tag{2.12}$$

và phân phối tiên nghiệm:

$$p(\lambda) = \frac{\lambda^{k-1}e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Tiếp theo, theo như (2.10) chúng ta có $p(\lambda|\mathcal{D}) \propto p(\mathcal{D}|\lambda)p(\lambda)$.

Lúc này, chúng ta có thể biến đổi hàm logarithm của phân phối hậu nghiệm $p(\lambda|\mathcal{D})$

như sau:

$$\begin{aligned}
\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda), \\
&= \ln(\lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n}) - \sum_{i=1}^{n} \ln x_i! + \ln(\lambda^{k-1} e^{-\frac{\lambda}{\theta}}) - \ln(\theta^k \Gamma(k)), \\
&= \sum_{i=1}^{n} x_i \ln \lambda - \lambda n - \sum_{i=1}^{n} \ln x_i! + (k-1)\ln \lambda - \frac{\lambda}{\theta} - k \ln \theta - \ln \Gamma(k), \\
&= \ln \lambda (k - 1 + \sum_{i=1}^{n} x_i) - \lambda(n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k). \qquad (2.13)
\end{aligned}$$

Để tối ưu (2.13) ta thực hiện lấy đạo hàm của (2.13) theo λ và cho kết quả bằng 0.

$$\begin{aligned}
\frac{\partial}{\partial \lambda} \ln p(\lambda|\mathcal{D}) = (k - 1 + \sum_{i=1}^{n} x_i)\frac{1}{\lambda} - (n + \frac{1}{\theta}) = 0, \\
\lambda_{MAP} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}}. \qquad (2.14)
\end{aligned}$$

Thay các giá trị $k = 3, \theta = 1, n = 6$ và bộ dữ liệu $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ vào (2.14) ta nhận được kết quả $\lambda_{MAP} = 5$. Lưu ý rằng MAP không hoàn toàn là Bayes vì nó chỉ đưa ra ước lượng điểm.

Phương pháp MLE

Cho bộ dữ liệu $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ với các x_i độc lập cùng tuân theo phân phối Poisson với λ chưa biết. Yêu cầu của bài toán tìm MLE cho tham số λ.
Hàm mật độ xác suất của phân phối Poisson được viết dưới dạng: $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, với $\lambda \in \mathbb{R}^+$.
Tham số chúng ta cần ước lượng như sau: $\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\arg\max} \{p(\mathcal{D}|\lambda)\}$, trong đó $p(\mathcal{D}|\lambda)$ như ở (2.12).
Để tìm giá trị λ, chúng ta lấy logarithm của (2.12), ta nhận được

$$\ln p(\mathcal{D}|\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - \lambda n - \sum_{i=1}^{n} (\ln x_i!).$$

Chúng ta tiếp tục lấy đạo hàm kết quả này theo λ,

$$\frac{\partial}{\partial\lambda}\ln p(\mathcal{D}|\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0,$$

$$\lambda_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.15}$$

Thay các giá trị của k và $\mathcal{D}$ vào (2.15) ta nhận được kết quả $\lambda_{ML} = 5.5$.

Nhận xét: Theo kết quả của (2.14) và (2.15) chúng ta thấy khi n tăng thì cả tử số và mẫu số của λ_{MAP} và λ_{ML} cùng tăng một lượng như nhau. Thế nhưng, điều này không còn đảm bảo khi $n \to \infty$ vì có thể xảy ra trường hợp $s_n = \sum_{i=1}^{n} x_i$ không tăng khi n tăng. Do đó, ta tìm hướng giải quyết khác đó là tính kỳ vọng của sự sai lệch giữa λ_{MAP} và λ_{ML}, sau đó chúng ta kiểm tra xem chuyện gì xảy ra khi $n \to \infty$.

Trước tiên, chúng ta cần lưu ý rằng cả hai ước lượng mà chúng ta đang quan tâm có thể được xem như hai biến ngẫu nhiên. Điều này xuất phát từ thực tế là bộ dữ liệu $\mathcal{D}$ của chúng ta được giả sử là độc lập và có cùng phân phối Poisson với tham số λ_t chưa biết.

Đặt $S_n = \sum_{i=1}^{n} X_i$ với $X_i \sim Poisson(\lambda)$ khi đó chúng ta có các ước lượng sau

$$\lambda_{MAP} = \frac{k - 1 + S_n}{n + \frac{1}{\theta}},$$

$$\lambda_{ML} = \frac{S_n}{n}.$$

Bây giờ, chúng ta chứng minh kết quả sau đây hội tụ (theo trung bình):

$$\lim_{n\to\infty} E_{\mathcal{D}}\left|\lambda_{MAP} - \lambda_{ML}\right| = 0,$$

trong đó kỳ vọng được thực hiện trên biến ngẫu nhiên $\mathcal{D}$.

Sự sai lệch giữa hai giá trị ước lượng là:

$$\begin{aligned}
|\lambda_{MAP} - \lambda_{ML}| &= \left| \frac{k-1+S_n}{n+\frac{1}{\theta}} - \frac{S_n}{n} \right|, \\
&= \left| \frac{k-1}{n+\frac{1}{\theta}} - \frac{S_n}{n\left(n+\frac{1}{\theta}\right)} \right|, \\
&\leq \frac{|k-1|}{n+\frac{1}{\theta}} + \frac{S_n}{n\left(n+\frac{1}{\theta}\right)}, \\
&= \varepsilon.
\end{aligned}$$

Tiếp theo, chúng ta biểu diễn kỳ vọng của ε như sau

$$\begin{aligned}
E[\varepsilon] &= \frac{1}{n(n+\frac{1}{\theta})}.E\left[\sum_{i=1}^{n} X_i\right] + \frac{|k-1|}{n+\frac{1}{\theta}}, \\
&= \frac{1}{n+\frac{1}{\theta}}.\lambda + \frac{|k-1|}{n+\frac{1}{\theta}}.
\end{aligned}$$

Khi đó, chúng ta có $\lim_{n\to\infty} E[\varepsilon] = 0$.

Kết quả này cho thấy ước lượng MAP tiếp cận vấn đề của ước lượng MLE cho bộ dữ liệu lớn. Nói cách khác, bộ dữ liệu lớn làm giảm sự quan trọng của prior knowl-edge. Đây là một kết luận rất quan trọng bởi vì nó đơn giản hóa bộ máy toán học cần thiết cho suy luận thực tế.

Chương 3

Tiên nghiệm liên hợp

Có một số hàm hỗ trợ cho sự tồn tại của định lý Bayes. Việc hiểu về các hàm này rất quan trọng.

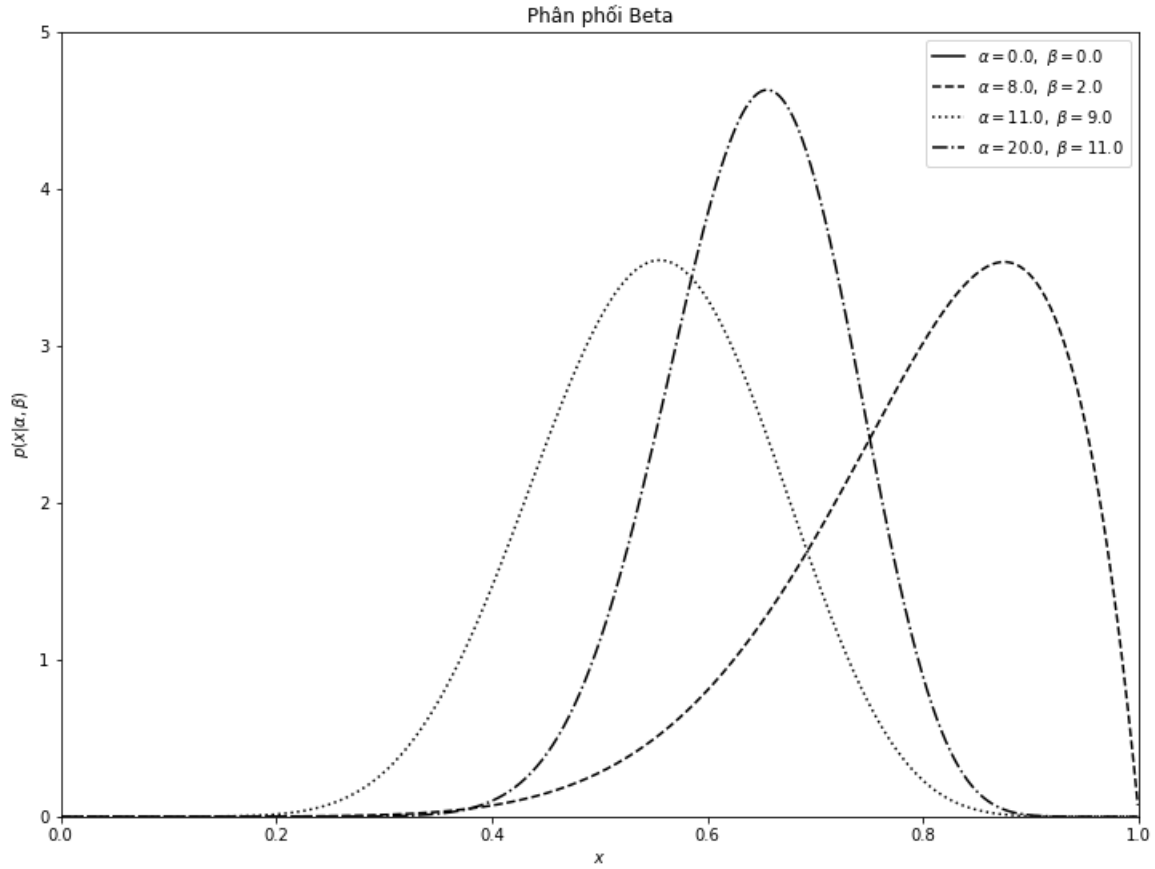
3.1 Phân phối niềm tin ban đầu

Phân phối niềm tin ban đầu được dùng để đại diện cho sức mạnh về sự tin tưởng của chúng ta về tham số dựa trên kinh nghiệm trước đây. Nhưng nếu trong trường hợp chúng ta không có kinh nghiệm trước đó thì sao?

Toán học đã hỗ trợ cho chúng ta phương pháp để khắc phục vấn đề này. Hàm toán học được giới thiệu sử dụng để đại diện cho niềm tin ban đầu (prior beliefs) gọi là phân phối Beta. Phân phối này có các tính chất toán học rất hay cho phép chúng ta mô hình hóa niềm tin ban đầu của mình về phân phối nhị thức.
Hàm mật độ xác suất của phân phối Beta có dạng:

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},$$

với $0 \leq x \leq 1, \alpha > 0, \beta > 0$ và $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$.
Ở đây, α, β là hai tham số quyết định hình dạng của phân phối.
Trong ví dụ tung ngẫu nhiên một đồng xu một số lần thì α là tham số chỉ số lần nhận mặt ngửa trong tổng số các lần thử, β là tham số chỉ số lần nhận mặt sấp.



Hình 3.1. Hàm mật độ của phân phối Beta với các tham số α và β cho trước.

Nhận xét

- Khi chúng ta chưa tung đồng xu, chúng ta tin rằng đồng xu cân bằng, điều này thể hiện ở đường thẳng mà hàm phân phối chúng ta nhận được.

- Khi số lần nhận mặt head nhiều hơn mặt tail (hay $\alpha > \beta$) đồ thị cho thấy đỉnh nhọn dịch chuyển về phía bên phải, điều này cho thấy xác suất xuất hiện mặt head cao hơn và do đó đồng xu không cân bằng.

- Khi nhiều lần tung được thực hiện và số lần nhận mặt head tiếp tục chiếm tỉ lệ lớn hơn thì đỉnh được thu hẹp, điều này giúp gia tăng niềm tin đồng xu này cân bằng.

Chú ý: Ở đây vì để dễ hình dung nên ta gán giá trị trước cho α, β. Trên thực tế, khi biết trước giá trị trung bình (μ) và độ lệch chuẩn (σ) của phân phối, ta có thể

tính α, β dựa vào công thức sau,

$$\mu = \frac{\alpha}{\alpha + \beta},$$
$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}.$$

3.2 Tiên nghiệm liên hợp (Conjugate prior)

Xét phân phối hậu nghiệm $p(\theta|X)$ với tiên nghiệm $p(\theta)$ và hàm likelihood $p(X|\theta)$, được biểu diễn dưới dạng:

$$p(\theta|X) \propto p(X|\theta)p(\theta).$$

Nếu phân phối hậu nghiệm $p(\theta|X)$ có cùng dạng (same family) với phân phối xác suất tiện nghiệm $p(\theta)$ thì tiên nghiệm và hậu nghiệm được gọi là phân phối liên hợp (conjugate distributions) và $p(\theta)$ được gọi là tiên nghiệm liên hợp cho hàm likelihood $p(X|\theta)$. Nếu điều này xảy ra, việc tối ưu bài toán MAP sẽ trở nên tương tự như việc tối ưu bài toán MLE vì nghiệm có cấu trúc giống nhau.

Likelihood	Prior	Posterior
Binomial	Beta	Beta
Negative Binomial	Beta	Beta
Poisson	Gamma	Gamma
Exponential	Gamma	Gamma
Normal (μ chưa biết)	Normal	Normal
Normal (σ chưa biết)	Inverse Gamma	Inverse Gamma
Normal (μ và σ chưa biết)	Normal/Gamma	Normal/Gamma
Multinomial	Dirichlet	Dirichlet

Bảng 3.1 Một số mô hình liên hợp.

3.2.1 Phân phối chuẩn là liên hợp của phân phối chuẩn (μ chưa biết)

Chứng minh Chúng ta có các biển đổi như sau

$$\begin{aligned}
p(\theta|x) &\propto p(x|\theta)p(\theta),\\
p(\theta|x) &\propto \mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1),\\
p(\theta|x) &\propto e^{-\frac{(x-\theta)^2}{2}}e^{-\frac{\theta^2}{2}},\\
p(\theta|x) &\propto e^{-\left(\theta-\frac{x}{2}\right)^2}.
\end{aligned}$$

Vậy $p(\theta|x)$ có phân phối $\mathcal{N}(\theta|\frac{x}{2},\frac{1}{\sqrt{2}})$ $\square$.

3.2.2 Phân phối Beta là liên hợp của phân phối Binomial

Chứng minh Ta có

$$\begin{aligned}
p(\theta|x) &\propto p(x|\theta)p(\theta),\\
&\propto \text{Binomial}(n,\theta)\times\text{Beta}(a,b),\\
&\propto \begin{pmatrix} n,\\ x \end{pmatrix}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},\\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1}.
\end{aligned}$$

Do đó,

$$p(\theta|x)\propto\theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

Vậy nên ta nhận được phân phối hậu nghiệm là phân phối

$$\text{Beta}(x+a,n-x+b). \qquad \square \tag{3.1}$$

Thực tế, phân phối tiên nghiệm của chúng ta đã thêm vào $a-1$ lần thành công và $b-1$ lần thất bại vào bộ dữ liệu.

Ví dụ 3.2.1. *Giả sử ta tung một đồng xu không đồng chất. Gọi xác suất nhận*

được mặt ngửa là $p(\text{head}) = \pi$ *với* $\pi \in [0,1]$ *và xác suất nhận được mặt sấp là* $p(tail) = 1 - \pi$.
Ta thực hiện tung đồng xu nhiều lần và nhận được dãy các giá trị $(x_1, x_2, ..., x_n)$.
Tìm phân phối của $p(\pi|\boldsymbol{x})$.

Trước tiên chúng ta quan sát xu hướng của đồng xu. Giả sử các lần tung đồng xu là độc lập, ta có:

$$p(x_1, ..., x_n|\pi) = \prod_{i=1}^{n} p(x_i|\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}.$$

Tiếp theo, chúng ta chọn $\pi \sim Beta(a, b)$ với $a, b > 0$ và chúng ta có

$$p(\pi) = \text{Beta}(\pi|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}.$$

Khi đó, theo quy tắc Bayes thì

$$p(\pi|x_1, ..., x_n) = \frac{p(x_1, ..., x_n|\pi)p(\pi)}{\int_0^1 p(x_1, ..., x_n|\pi)p(\pi)d\pi}. \tag{3.2}$$

Ta thấy rằng $\int_0^1 p(x_1, ..., x_n|\pi)p(\pi)d\pi$ chỉ là hằng số chuẩn hóa nên không phụ thuộc π.
Do đó, chúng ta có thể viết (3.2) như sau:

$$\begin{aligned} p(\pi|x_1, ..., x_n) &\propto \Big(\prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i}\Big)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}\right), \\ &\propto \pi^{\sum_{i=1}^{n} x_i + a - 1}(1-\pi)^{n - \sum_{i=1}^{n} x_i} + b - 1. \end{aligned}$$

Vậy ta có $p(\pi|x_1, ..., x_n) = \text{Beta}(\sum_{i=1}^{n} x_i + a, n - \sum_{i=1}^{n} x_i + b)$.
Kết quả này hoàn toàn trùng khớp với kết quả ở (3.1).

3.2.3 Hỗn hợp của tiên nghiệm liên hợp

Tiên nghiệm liên hợp là phương pháp tính đơn giản nhưng nó không mạnh và không linh hoạt để mã hóa prior knowl-edge. Tuy nhiên, hóa ra hỗn hợp của tiên

nghiệm liên hợp cũng là một tiên nghiệm và chúng ta có thể xấp xỉ bất cứ loại tiên nghiệm nào. Như vậy, các tiên nghiệm đáp ứng tốt hai yếu tố là thuận tiện tính toán và tính linh hoạt.

Ví dụ, giả sử chúng ta lập mô hình tung đồng xu và chúng ta cho rằng đồng xu này là cân bằng hoặc đồng xu này dễ xảy ra mặt ngửa hơn. Rõ ràng trong trường hợp này phân phối Beta không thể đại diện cho niềm tin của chúng ta. Tuy nhiên, chúng ta có thể mô hình hóa niềm tin đó bằng cách sử dụng hỗn hợp của hai phân phối Beta. Cụ thể, chúng ta có thể cho rằng:

$$p(\theta) = 0.5\,\mathrm{Beta}(\theta|20, 20) + 0.5\,\mathrm{Beta}(\theta|30, 10).$$

Nếu θ đến từ phân phối đầu tiên thì đồng xu cân bằng, ngược lại đồng xu này có khả năng nhận mặt ngửa cao hơn.

Chúng ta gọi z là biến tiềm ẩn, khi $z = k$ nghĩa là tham số θ đến từ thành phần k của hỗn hợp. Lúc này tiên nghiệm được biểu diễn dưới dạng:

$$p(\theta) = \sum_k p(z = k)p(\theta|z = k),$$

trong đó $p(\theta|z = k)$ là tiên nghiệm và $p(z = k)$ được gọi là trọng số hỗn hợp tiên nghiệm.

Khi đó, phân phối hậu nghiệm có thể được viết lại dưới dạng một hỗn hợp của các phân phối liên hợp như sau:

$$p(\theta|X) = \sum_k p(z = k|X)p(\theta|X, z = k),$$

với $p(z = k|X)$ là trọng số hỗn hợp hậu nghiệm được cho bởi:

$$p(z = k|X) = \frac{\sum_k p(z = k)p(X|z = k)}{\sum_k p(z = k'|X)p(\theta|X, z = k')}.$$

Chương 4

Thuật toán tối đa hóa kỳ vọng

4.1 Mô hình hỗn hợp Gaussian

Mô hình hỗn hợp Gaussian cho $\mathbf{x} = (x_1, x_2, .., x_m)$ có dạng như sau:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4.1}$$

với π_k là tỉ lệ hỗn hợp, $\sum\limits_k \pi_k = 1$, $\boldsymbol{\mu}_k$ và $\boldsymbol{\Sigma}_k$ lần lượt là trung bình, phương sai của thành phần Gaussian thứ k.

Cho $\mathbf{z}$ là biến ngẫu nhiên nhị phân K chiều, được biểu diễn dưới dạng vector có kích thước $1 \times K$ trong đó có một thành phần $z_k = 1$ và tất cả các thành phần khác bằng 0. Giá trị của z_k vì thế thỏa mãn $z_k \in \{0, 1\}$ và $\sum\limits_{k=1}^{K} z_k = 1$.

Tiếp theo, chúng ta xác định phân phối đồng thời $p(\mathbf{x}, \mathbf{z})$ có liên hệ với phân phối lề $p(\mathbf{z})$ và phân phối có điều kiện $p(\mathbf{x}|\mathbf{z})$ qua công thức $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$. Phân phối lề trên $\mathbf{z}$ chỉ rõ mối quan hệ với hệ số hỗn hợp π_k thỏa mãn:

$$p(z_k = 1) = \pi_k, \tag{4.2}$$

trong đó tham số $\{\pi_k\}$ thỏa mãn $0 \leq \pi_k \leq 1$ và $\sum_{k=1}^{K} \pi_k = 1$.
Vì $\mathbf{z}$ được biểu diễn đưới dạng $1 \times K$, chúng ta có thể viết phân phối $p(\mathbf{z})$ dưới dạng sau

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{4.3}$$

Tương tự, phân phối có điều kiện của $\mathbf{x}$ khi biết giá trị cụ thể của $\mathbf{z}$ là một Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{4.4}$$

Từ đây chúng ta có

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{4.5}$$

Phân phối đồng thời được xác định bởi $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$, phân phối lề của $\mathbf{x}$ sau đó thu được bằng cách tính tổng phân phối đồng thời trên tất cả các trạng thái phù hợp của $\mathbf{z}$, khi đó kết hợp với (4.4), (4.5) chúng ta nhận được:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{4.6}$$

Vậy nên phân phối lề của $\mathbf{x}$ là một hỗn hợp Gaussian có dạng như ở (4.1). Nếu chúng ta có các quan trắc $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ thì vì phân phối lề được viết dưới dạng $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$ nên với mỗi điểm dữ liệu $\mathbf{x}_n$, tương ứng sẽ có một biến tìm ẩn $\mathbf{z}_n$. Do đó, chúng ta có thể tìm một công thức tương đương của hỗn hợp Gaussian liên quan tới một biến tìm ẩn rõ ràng. Có vẻ như chúng ta đã không đạt được nhiều khi thực hiện bằng cách như vậy.

Tuy nhiên, bây giờ chúng ta có thể làm việc với phân phối đồng thời $p(\mathbf{x}, \mathbf{z})$ thay vì làm việc với phân phối lề $p(\mathbf{x})$. Điều này làm đơn giản đáng kể và đáng chú ý nhất là sự liên hệ với thuật toán tối đa hóa kỳ vọng.
Một đại lượng khác đóng một vai trò quan trọng đó là xác suất có điều kiện của $\mathbf{z}$

khi biết $\mathbf{x}$. Chúng ta sẽ sử dụng $\gamma(z_k)$ để biểu thị cho $p(z_k = 1|\mathbf{x})$ và giá trị này có thể được tìm thấy bằng cách sử dụng định lý Bayes

$$\gamma(z_k) = p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}. \tag{4.7}$$

Từ (4.1) và (4.3) chúng ta nhận được:

$$\gamma(z_k) = p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{4.8}$$

Chúng ta thấy π_k là xác suất tiên nghiệm của $z_k = 1$ và đại lượng $\gamma(z_k)$ được xem như là xác suất hậu nghiệm tương ứng khi chúng ta quan sát $\mathbf{x}$.

4.2 Thuật toán phân nhóm K-means

4.2.1 Giới thiệu

Thuật toán phân nhóm K-means là thuật toán cơ bản nhất trong học không giám sát. Trong thuật toán phân nhóm K-means, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Thuật toán bắt đầu bằng việc chọn ra điểm trung tâm của từng cụm. Chúng ta có thể đơn giản chọn ngẫu nhiên K điểm bất kì trong bộ dữ liệu làm điểm trung tâm, hoặc dùng các cách tiếp cận khác, nhưng nhìn chung chọn ngẫu nhiên vẫn là cách tốt nhất. Sau đó chúng ta luân phiên lặp lại 2 giai đoạn dưới đây.

1. Giai đoạn gán: gán từng phần tử trong bộ dữ liệu của chúng ta vào các cụm. Cách thức tiến hành đó là: với mỗi điểm, hãy tính khoảng cách từ điểm đó tới vị trí các điểm trung tâm, sau cùng: điểm trung tâm nào gần nhất thì gán vào cụm ứng với điểm trung tâm đó.

2. Giai đoạn cập nhật: duyệt từng cụm, cập nhật lại tọa độ của điểm trung

tâm: như đã biết, sau giai đoạn gán, chúng ta đã thu được K cụm ứng với dãy các điểm được gán cho từng cụm. Tọa độ điểm trung tâm mới của cụm sẽ bằng trung bình cộng tọa độ các điểm trong cụm.

Sau càng nhiều vòng lặp, các điểm trung tâm càng di chuyển chậm dần và tổng khoảng cách từ mỗi điểm trong cụm tới cụm lại càng nhỏ đi. Quá trình sẽ kết thúc cho tới khi hàm tổng khoảng cách hội tụ. Lúc này tọa độ điểm trung tâm sẽ vẫn bằng trung bình cộng các điểm dữ liệu hiện tại trong cụm, hay nói cách khác điểm trung tâm sẽ không còn di chuyển nữa.

Chú ý: thuật toán K-means chỉ đảm bảo được quá trình này sẽ đưa hàm tổng khoảng cách hội tụ tới điểm cực tiểu địa phương, chứ không chắc chắn đó sẽ là giá trị nhỏ nhất của toàn bộ hàm số.

4.2.2 Lý thuyết toán học cho thuật toán K-means

Giả sử có N điểm dữ liệu là $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ và $K < N$ là số nhóm chúng ta muốn phân chia. Chúng ta cần tìm các điểm trung tâm $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K \in \mathbb{R}^{d \times 1}$ và nhãn của mỗi điểm dữ liệu.

Với mỗi điểm dữ liệu $\mathbf{x}_i$, ta cần tìm nhãn $y_i = k$ của nó, với $k \in \{1, 2, ..., K\}$. Ngoài ra, để dễ làm việc thông thường mỗi nhãn k được thay thế bằng một vector hàng có dạng $\mathbf{y}_i = [y_{i1}, y_{i2}, ..., y_{iK}]$ được gọi là *label vector*, trong đó nếu $\mathbf{x}_i$ được phân vào lớp k thì các phần tử của $\mathbf{y}_i$ bằng 0, ngoại trừ phần tử ở vị trí thứ k bằng 1. Cụ thể $y_{\text{ij}} = 0, \forall j \neq k, y_{ik} = 1$.

Khi chồng các vector $\mathbf{y}_i$ lên nhau ta được một ma trận nhãn $\mathbf{Y} \in \mathbb{R}^{n \times K}$. Nhắc lại rằng y_{ij} là phần tử hàng thứ i, cột thứ j của ma trận $\mathbf{Y} \in \mathbb{R}^{N \times K}$ và nó cũng chính là phần tử thứ j của vector $\mathbf{y}_i$. Khi một điểm dữ liệu có label vector là $[1, 0, ..., 0]$ thì điểm dữ liệu đó thuộc vào nhóm thứ nhất, là $[0, 1, ..., 0]$ thì nó thuộc vào nhóm thứ hai.

Các ràng buộc của $\mathbf{y}_i$ trong toán học như sau

$$y_{ij} \in \{0, 1\}, \forall i, j \text{ và } \sum_{k=1}^{K} y_{ik} = 1 \forall i. \tag{4.9}$$

4.2.3 Hàm mất mát và bài toán tối ưu

Nếu gọi $\mathbf{m}_k$ là điểm trung tâm của mỗi nhóm và thay thế tất cả các điểm được phân vào nhóm này bởi $\mathbf{m}_k$, thì một điểm dữ liệu $\mathbf{x}_i$ khi được phân vào nhóm k sẽ bị sai số là $\mathbf{x}_i - \mathbf{m}_k$.

Chúng ta mong muốn vector sai số này gần với vector không, tức $\mathbf{x}_i$ gần với $\mathbf{m}_k$. Và chúng ta nhớ đến một đại lượng giúp đo khoảng cách giữa hai điểm là (bình phương) khoảng cách Euclid $\|\mathbf{x}_i - \mathbf{m}_k\|_2^2$.

Hơn nữa, vì $\mathbf{x}_i$ được phân vào nhóm k nên $y_{ik} = 1$, $y_{ij} = 0, \forall j \neq k$.

Khi đó, biểu thức khoảng cách Euclid có thể được viết lại thành:

$$\|\mathbf{x}_i - \mathbf{m}_k\|_2^2 = y_{ik} \|\mathbf{x}_i - \mathbf{m}_k\|_2^2 = \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 . \tag{4.10}$$

Sai số trung bình cho toàn bộ dữ liệu sẽ là:

$$\mathcal{L}(\mathbf{Y}, \mathbf{M}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2. \tag{4.11}$$

Trong đó $\mathbf{Y} = [y_1, y_2, ..., y_N]$ là ma trận tạo bởi nhãn của mỗi điểm dữ liệu,
$\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K] \in \mathbb{R}^{d \times K}$ là ma trận tạo bởi K điểm trung tâm. Hàm mất mát trong bài toán phân nhóm K-means là hàm $\mathcal{L}(\mathbf{Y}, \mathbf{M})$ với ràng buộc như được nêu ở (4.9).

Tóm lại, chúng ta cần tối ưu bài toán sau:

$$\mathbf{Y}, \mathbf{M} = \arg\min_{\mathbf{Y}, \mathbf{M}} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2, \tag{4.12}$$

thỏa mãn điều kiện $y_{ij} \in \{0, 1\}, \forall i, j$ và $\sum_{j=1}^{K} y_{ij} = 1, \forall i$.

4.2.4 Thuật toán tối ưu hàm mất mát

Bài toán (4.12) là bài toán khó tìm điểm tối ưu vì nó có thêm nhiều điều kiện ràng buộc. Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương

pháp để tìm được nghiệm gần đúng. Một kỹ thuật đơn giản và phổ biến để giải bài toán (4.12) là xen kẽ giải $\mathbf{Y}$ và $\mathbf{M}$ khi biến còn lại được cố định tới khi hàm mất mát hội tụ. Chúng ta sẽ lần lượt giải quyết hai bài toán dưới đây.

Cố định M, tìm Y

Giả sử chúng ta đã tìm được các điểm trung tâm, việc tiếp theo là tìm các *label vector* để hàm mất mát đạt giá trị nhỏ nhất. Điều này tương đương với việc tìm nhóm cho mỗi điểm dữ liệu.
Khi các điểm trung tâm là cố định, bài toán tìm *label vector* cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm *label vector* cho từng điểm dữ liệu $\mathbf{x}_i$ như sau:

$$\mathbf{y}_i = \underset{\mathbf{y}_i}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{K} y_{ij} \left\| \mathbf{x}_i - \mathbf{m}_j \right\|_2^2, \tag{4.13}$$

thỏa mãn $y_{ij} \in \{0, 1\}, \forall i, j$ và $\sum_{j=1}^{K} y_{ij} = 1, \forall i$.
Vì chỉ có một phần tử của *label vector* $\mathbf{y}_i$ bằng 1 nên bài toán (4.13) chính là bài toán đi tìm điểm trung tâm gần điểm $\mathbf{x}_i$ nhất: $j = \underset{j}{\operatorname{argmin}} \left\| \mathbf{x}_i - \mathbf{m}_j \right\|_2^2$.
Vì $\left\| \mathbf{x}_i - \mathbf{m}_j \right\|_2^2$ chính là bình phương khoảng cách Euclid tính từ điểm $\mathbf{x}_i$ tới điểm trung tâm $\mathbf{m}_i$, ta có thể kết luận rằng mỗi điểm $\mathbf{x}_i$ thuộc vào nhóm có điểm trung tâm gần $\mathbf{x}_i$ nhất. Từ đó ta dễ dàng suy ra *label vector* của từng điểm dữ liệu.

Cố định Y, tìm M

Giả sử chúng ta đã tìm được nhóm cho từng điểm, việc tiếp theo chúng ta tìm điểm trung tâm mới cho mỗi nhóm để hàm mất mát đạt giá trị nhỏ nhất
Một khi chúng ta đã xác định được *label vector* cho từng điểm dữ liệu, bài toán tìm điểm trung tâm cho mỗi nhóm được rút gọn thành:

$$\mathbf{m}_j = arg \min_{\mathbf{m}_j} \frac{1}{N} \sum_{i=1}^{N} y_{ij} \left\| \mathbf{x}_i - \mathbf{m}_j \right\|_2^2. \tag{4.14}$$

Vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm $\mathbf{m}_j$ nên để tìm nghiệm của (4.14) chúng ta có thể dùng phương pháp giải phương trình đạo hàm bằng không. Đặt $l(\mathbf{m}_j)$ là hàm bên trong dấu argmin ở (4.14), lấy đạo hàm của (4.14) theo $\mathbf{m}_j$ chúng ta có

$$\nabla_{\mathbf{m_j}} l(\mathbf{m_j}) = \frac{2}{N} \sum_{i=1}^{N} y_{ij}(\mathbf{m_j} - \mathbf{x_i}).$$

Chúng ta cần giải phương trình sau:

$$\begin{aligned} \nabla_{\mathbf{m_j}} l(\mathbf{m_j}) = \frac{2}{N} \sum_{i=1}^{N} y_{ij}(\mathbf{m_j} - \mathbf{x_i}) = 0, \\ \Leftrightarrow \mathbf{m_j} \sum_{i=1}^{N} y_{ij} = \sum_{i=1}^{N} y_{ij}\mathbf{x_i}, \\ \Leftrightarrow \mathbf{m_j} = \frac{\sum_{i=1}^{N} y_{ij}\mathbf{x_i}}{\sum_{i=1}^{N} y_{ij}}. \end{aligned} \tag{4.15}$$

Công thức (4.15) có mẫu số là số lượng các điểm dữ liệu trong nhóm j và tử số chính là tổng các điểm dữ liệu trong nhóm j. Nghĩa là $\mathbf{m}_j$ là trung bình cộng của các điểm trong nhóm j.

Trường hợp nếu tồn tại một nhóm không chứa điểm nào, mẫu số của (4.15) sẽ bằng không và phép chia sẽ không thực hiện được. Vì vậy, K điểm bất kỳ trong tập huấn luyện sẽ được chọn để làm điểm trung tâm ban đầu, điều này để đảm bảo mỗi nhóm có ít nhất một điểm.

Trong quá trình huấn luyện, nếu tồn tại một nhóm không chứa điểm nào thì có hai cách giải quyết. Cách thứ nhất là bỏ đi nhóm đó và giảm K đi 1 đơn vị. Cách thứ hai là thay điểm trung tâm của nhóm đó bằng một điểm bất kỳ trong tập huấn luyện ví dụ điểm gần điểm trung tâm hiện tại của nó nhất.

4.3 Ví dụ thuật toán K-means

Giả sử có 4 loại thuốc A, B, C, D. Mỗi loại thuốc được biểu diễn bởi 2 đặc trưng X và Y như trong bảng cho dưới đây. Mục đích của chúng ta là nhóm các loại thuốc

đã cho vào 2 nhóm ($K = 2$) dựa vào các đặc trưng cho dưới đây.

Đối tượng	Chỉ số cân nặng (X)	Độ pH (Y)
Thuốc A	1	1
Thuốc B	2	1
Thuốc C	4	3
Thuốc D	5	4

Bước 1. Gán điểm trung tâm cho 2 nhóm.

Giả sử ta chọn A là tâm của nhóm thứ nhất và B là tâm của nhóm thứ hai. Khi đó tọa độ tâm của nhóm thứ nhất là $c_1 = (1, 1)$ và tọa độ tâm của nhóm thứ hai là $c_2 = (2, 1)$.

Bước 2. Tính khoảng cách từ các đối tượng đến tâm của các nhóm

Trước tiên, chúng ta có thể biểu diễn số liệu đã cho dưới dạng ma trận sau

$$\begin{matrix} A & B & C & D \end{matrix}$$
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix} .$$

Tiếp theo, chúng ta tiến hành tính khoảng cách từ các loại thuốc A, B, C, D đến tâm $c_1 = (1, 1)$. Ví dụ, khoảng cách từ loại thuốc C (có tọa độ $(4, 3)$) đến tâm $c_1 = (1, 1)$ là $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$. Thực hiện tương tự để tính khoảng cách từ các loại thuốc A, B, C, D đến tâm $c_2 = (2, 1)$. Sau khi tính chúng ta thu được ma trận khoảng cách như sau

$$D^0 = \begin{matrix} \\ \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \\ A \quad B \quad C \quad D \end{matrix} .$$

Mỗi cột trong ma trận khoảng cách D^0 là một đối tượng (cột thứ nhất tương ứng với đối tượng A, cột thứ hai tương ứng với đối tượng B,...). Hàng thứ nhất trong ma trận khoảng cách biểu diễn khoảng cách giữa các đối tượng đến tâm của nhóm thứ nhất (c_1) và hàng thứ hai trong ma trận khoảng cách biểu diễn khoảng cách của các đối tượng đến tâm của nhóm thứ hai (c_2).

Bước 3. Nhóm các đối tượng vào nhóm gần nhất

Quan sát ma trận khoảng cách chúng ta thấy rằng khoảng cách từ đối tượng A đến tâm c_1 là bé nhất, do đó chúng ta xếp đối tượng A vào nhóm 1 và các đối tượng còn lại là B, C, D vào nhóm 2. Ta thu được ma trận

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \text{nhóm 1} \\ \text{nhóm 2} \end{matrix}.$$

Bước 4. Tính lại tọa độ các tâm cho các nhóm mới dựa vào tọa độ của các đối tượng trong nhóm.

Vì nhóm 1 chỉ có một đối tượng A nên tâm nhóm 1 vẫn là $c_1 = (1, 1)$. Bên cạnh đó, tâm nhóm 2 được tính dựa vào số liệu của ba đối tượng B, C, D như sau

$$c_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\frac{11}{3}, \frac{8}{3}\right).$$

Bước 5. Tính lại khoảng cách từ các đối tượng đến tâm mới

Tâm mới của chúng ta là $c_1 = (1, 1)$ và $c_2 = (\frac{11}{3}, \frac{8}{3})$.

Tính toán tương tự như ở bước 2, chúng ta thu được ma trận khoảng cách mới là

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

Bước 6. Nhóm các đối tượng vào nhóm

Vì khoảng cách từ đối tượng B đến tâm c_1 có giá trị nhỏ nhất do đó chúng ta thêm đối tượng B vào nhóm 1. Mặt khác, khoảng cách từ đối tượng C, D đến tâm c_2 có giá trị nhỏ nên nhóm 2 của chúng ta chứa các đối tượng là C và D. Như vậy, ma trận thu được là

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{nhóm 1} \\ \text{nhóm 2} \end{matrix}.$$

Bước 7. Tính lại tâm cho nhóm mới

Dựa vào kết quả của ma trận G^1 chúng ta tính được tâm mới cho các nhóm như sau

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(1\frac{1}{2}, 1\right).$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = \left(4\frac{1}{2}, 3\frac{1}{2}\right).$$

Bước 8. Tính lại khoảng cách từ các đối tượng đến tâm mới

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.2 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

Bước 9. Nhóm các đối tượng vào nhóm

Chúng ta thấy rằng khoảng cách từ các đối tượng A, B đến tâm c_1 là nhỏ nhất nên nhóm 1 vẫn gồm hai đối tượng là A, B. Đồng thời, khoảng cách từ các đối tượng C, D đến tâm c_2 là nhỏ nhất nên nhóm 2 vẫn gồm hai đối tượng là C và D. Ma trận nhận được là

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{nhóm 1} \\ \text{nhóm 2} \end{matrix}$$

Ta thấy $G^2 = G^1$ (không có sự thay đổi nhóm nào của các đối tượng) nên thuật toán dừng lại và kết quả phân nhóm là nhóm 1 gồm có đối tượng A, B và nhóm 2 gồm có đối tượng C, D.

Nhận xét

Thuật toán K-means có ưu điểm là đơn giản và dễ hiểu. Tuy nhiên, hạn chế của thuật toán K-means là độ tốt của thuật toán phụ thuộc vào việc chọn số nhóm K (phải xác định trước) và chi phí cao để thực hiện vòng lặp tính khoảng cách khi số cụm K và dữ liệu phân cụm lớn.

4.4 Thuật toán tối đa hóa kỳ vọng

Trong máy học và thống kê, có rất nhiều mô hình chúng ta có thể dễ dàng ước lượng tham số ML hay MAP nếu chúng ta có đầy đủ các giá trị của tất cả các biến trong bộ dữ liệu. Tuy nhiên, trong trường hợp bộ dữ liệu của chúng ta bị thiếu giá trị thì việc ước lượng lúc này sẽ trở nên rất khó khăn. Thuật toán tối đa hóa kỳ vọng (expectation maximization) sẽ giúp chúng ta giải quyết vấn đề này.

4.4.1 Bất đẳng thức Jensen

Trước khi trình bày cách mà thuật toán EM giúp chúng ta giải quyết những khó khăn thì việc đầu tiên chúng ta cần tìm hiểu về bất đẳng thức Jensen.

Cho f là một hàm có miền xác định trên $\mathbb{R}$. Nhắc lại rằng f là hàm lồi nếu $f''(x) \geq 0$ với mọi $x \in \mathbb{R}$. Trong trường hợp f là một vector thì cần có điều kiện ma trận hessian H là ma trận nửa xác định dương ($H \geq 0$). Nếu $f''(x) > 0$ với mọi x thì ta nói f là hàm lồi nghiêm ngặt. Bất đẳng thức được phát biểu như dưới đây.

Định nghĩa 4.4.1. Cho f là một hàm lồi và X là biến ngẫu nhiên. Khi đó theo bất đẳng thức Jensen, ta có

$$f(E[X]) \leq E[f(X)].$$

Hơn thế nữa, nếu f là hàm lồi ngặt thì khi đó $f(E[X]) = E[f(X)]$ nếu và chỉ nếu $X = E[X]$. Nói cách khác, f là hàm lõm nếu $f'' < 0$ và ta có $f(E[X]) \geq E[f(X)]$.

Ví dụ 4.4.1. *Cho X là biến ngẫu nhiên. Chứng minh rằng $lnE[X] \geq E[\ln X]$, $\forall X > 0$.*

Chứng minh Trước tiên, chúng ta đặt $f(X) = \ln(X)$. Khi đó ta có

$$f''(X) = \left(\frac{1}{X}\right)' = -\frac{1}{X^2} < 0, \forall X > 0.$$

Do đó, f là hàm lõm với mọi X $>$ 0. Vậy nên theo bất đẳng thức Jensen ta có

$$lnE[X] \geq E[\ln X], X > 0. \qquad \square \tag{4.16}$$

4.4.2 Ý tưởng cơ bản

Giả sử chúng ta có bộ dữ liệu được cho bởi $\mathbf{x} = (x_1, x_2, ..., x_m)$. và chúng ta không biết nhãn của mỗi điểm dữ liệu trong thực tế. Do đó, ta quan tâm đến biến tiềm ẩn $\mathbf{z} = (z_1, z_2, ..., z_m)$, trong đó mỗi z_i cho biết phân phối của một x_i thuộc về. Trong trường hợp này, chúng ta có mô hình $p(\mathbf{x},\mathbf{z}|\Theta)$, với $\mathbf{x}$ là biến có thể quan sát được.

Mục đích của chúng ta là tối đa hóa $p(\mathbf{x}|\Theta) = \prod_i p(x_i|\Theta)$. Tuy nhiên, thay vì tối đa hóa hàm trên ta sẽ đi tối đa hóa hàm $\ln p(\mathbf{x}|\Theta)$.

Chúng ta có các biến đổi sau

$$\begin{aligned}
\ln p(\mathbf{x}|\Theta) &= \ln \prod_i p(x_i|\Theta), \\
&= \sum_i \ln p(x_i|\Theta), && (4.17)\\
&= \sum_i \ln \sum_{z_i} p(x_i, z_i|\Theta), && (4.18)\\
&= \sum_i \ln \sum_{z_i} q_i(z_i) \frac{p(x_i, z_i|\Theta)}{q_i(z_i)}, && (4.19)\\
&= \sum_i \ln E\left[\frac{p(x_i, z_i|\Theta)}{q_i(z_i)}\right]. && (4.20)
\end{aligned}$$

trong đó $q_i(z_i)$ là phân phối xác suất tùy ý cho z_i và do đó $q_i(z_i) \geq 0$ và $\sum_{z_i} q_i(z_i) = 1$. Dấu bằng ở (4.20) xảy ra theo công thức $E(f(x)) = \sum_i p(x_i) f(x_i)$ với $p(x_i)$ là hàm khối xác suất cho x_i, do đó $\sum_{z_i} q_i(z_i) \frac{p(x_i, z_i|\Theta)}{q_i(z_i)}$ chính là kỳ vọng của $\frac{p(x_i, z_i|\Theta)}{q_i(z_i)}$ được cho bởi x_i.

Tiếp tục chúng ta có các biến đổi tiếp theo

$$\begin{aligned}
\ln p(\mathbf{x}|\Theta) &= \sum_i \ln E_{z_i \sim q_i(z_i)}\Big(\frac{p(x_i, z_i|\Theta)}{q_i(z_i)}\Big), \\
&\geq \sum_i E\Big(\ln\big(\frac{p(x_i, z_i|\Theta)}{q_i(z_i)}\big)\Big) \text{ (theo (4.16))}, \\
&= \sum_i \sum_{z_i} q_i(z_i) \ln \frac{p(x_i, z_i|\Theta)}{q_i(z_i)}. && (4.21)
\end{aligned}$$

Như vậy, chúng ta đã xây dựng được giới hạn dưới cho $\ln p(\mathbf{x}|\Theta)$. Để dấu bằng của bất đẳng thức Jensen xảy ra thì chúng ta cần

$$\frac{p(x_i, z_i|\Theta)}{q_i(z_i)} = c \text{ , với c là hằng số.}$$

Lúc này chúng ta suy ra

$$q_i(z_i) \propto p(x_i, z_i|\Theta) \text{ với} \sum_{z_i} q_i(z_i) = 1.$$

Từ công thức này chúng ta có thể chọn q

$$q_i(z_i) = \frac{p(x_i, z_i|\Theta)}{\sum_z p(x_i, z|\Theta)}, \tag{4.22}$$

$$= \frac{p(x_i, z_i|\Theta)}{p(x_i|\Theta)}, \tag{4.23}$$

$$= p(z_i|x_i, \Theta). \tag{4.24}$$

Biểu thức (4.24) chính là phân phối hậu nghiệm của z_i được cho bởi x_i và tham số Θ.

4.4.3 Hoạt động của thuật toán

Thuật toán EM được xây dựng bằng cách thực hiện các bước lặp, bao gồm 2 bước:

- Bước E. cho $q_i(z_i) = p(z_i|x_i, \Theta)$ và đưa ra một giới hạn dưới cho $\ln p(\mathbf{x}|\Theta)$.

- Bước M. Thiết lập lại tham số để tối đa hóa hàm giới hạn dưới

$$\Theta = \underset{\Theta}{\operatorname{argmax}} \sum_i \sum_{z_i} q_i(z_i) \ln \frac{p(x_i, z_i|\Theta)}{q_i(z_i)}.$$

4.4.4 Sự hội tụ

Để chứng minh sự hội tụ của thuật toán EM, chúng ta có thể chứng minh thực tế rằng

$$\ln p(\mathbf{x}|\Theta^{(t+1)}) > \ln p(\mathbf{x}|\Theta^{(t)}) \text{ với mọi t và } \Theta^{(t)} \text{ là tham số tại bước lặp t.}$$

Nhớ lại rằng chúng ta đã chọn $q_i(z_i) = p(z_i|x_i, \Theta)$, do đó khi dấu bằng của bất đẳng thức Jensen xảy ra chúng ta có

$$\ln p(\mathbf{x}|\Theta^{(t)}) = \sum_i \sum_{z_i} q_i^{(t)}(z_i) \ln(\frac{p(x_i, z_i|\Theta^{(t)})}{q_i^{(t)}(z_i)}). \tag{4.25}$$

Trong bước M, chúng ta cập nhật lại tham số $\Theta^{(t)}$ thành $\Theta^{(t+1)}$ để tối đa hóa vế phải của (4.25).

$$\begin{aligned}
\ln p(\mathbf{x}|\Theta^{(t+1)}) &= \sum_i \sum_{z_i} q_i^{(t+1)}(z_i) \ln \frac{p(x_i, z_i|\Theta^{(t+1)})}{q_i^{(t+1)}(z_i)} \\
&\geq \sum_i \sum_{z_i} q_i^{(t)}(z_i) \ln \frac{p(x_i, z_i|\Theta^{(t+1)})}{q_i^{(t)}(z_i)} \qquad (4.26) \\
&\geq \sum_i \sum_{z_i} q_i^{(t)}(z_i) \ln \frac{p(x_i, z_i|\Theta^{(t)})}{q_i^{(t)}(z_i)} \qquad (4.27) \\
&= \ln p(\mathbf{x}|\Theta^{(t)}). \qquad (4.28)
\end{aligned}$$

Như vậy chúng ta chứng minh được

$$\ln p(\mathbf{x}|\Theta^{(t+1)}) \geq \ln p(\mathbf{x}|\Theta^{(t)}).$$

Rõ ràng thuật toán tối đa hóa kỳ vọng làm hàm log likelihood thay đổi đơn điệu. Trong thực tế, khi xét sự hội tụ của thuật toán tối đa hóa kỳ vọng đồng nghĩa với việc chứng minh $\left|\ln p(\mathbf{x}|\Theta^{(t+1)}) - \ln p(\mathbf{x}|\Theta^{(t)})\right| < \varepsilon$ thỏa mãn ε là giá trị cho trước.

4.4.5 Hiểu hơn về thuật toán tối đa hóa kỳ vọng

Phần trên tiểu luận đã giới thiệu một cách chi tiết về thuật toán EM. Phần này chúng ta sẽ đi sâu vào phân tích khai triển của $\ln p(\mathbf{x}|\Theta)$. Nhưng trước tiên, chúng ta sẽ làm quen với phân kì Kullback - Leibler và giới hạn dưới.

Định nghĩa 4.4.2. Cho hai phân phối xác suất $q(z)$ và $p(z)$ của một biến ngẫu

nhiên Z, khi đó phân kì Kullback-Leibler $\mathcal{KL}$ được biểu diễn dưới dạng

$$\mathcal{KL}(q||p) = \sum_z q(z) \ln \frac{q(z)}{p(z|\mathbf{x}, \Theta)}.$$

Phân kỳ $\mathcal{KL}$ dùng để đo lường sự khác biệt giữa hai phân phối xác suất trên cùng một biến ngẫu nhiên.

Tính chất 4.4.3. *Phân kỳ $\mathcal{KL}$ có các tính chất sau.*

- $\mathcal{KL}(q(z)||p(z)) \geq 0$.

- $\mathcal{KL} = 0$ *nếu và chỉ nếu* $q(z) = p(z)$ *với mọi* $z \in Z$.

- $\mathcal{KL}(q(z)||p(z)) \neq \mathcal{KL}(p(z)||q(z)$.

Giới hạn dưới

Giới hạn dưới của $\ln p(\mathbf{x}|\Theta)$ được cho bởi

$$\mathcal{L}(\mathbf{x}|\Theta) = \sum_z q(z) \ln \frac{p(\mathbf{x}, z|\Theta)}{q(z)}. \tag{4.29}$$

Trong thực tế, chúng ta có thể biểu diễn $\ln p(\mathbf{x}|\Theta)$ dưới dạng tổng của giới hạn dưới và phân kỳ $\mathcal{KL}$,

$$\ln p(\mathbf{x}|\Theta) = \mathcal{L}(\mathbf{x}|\Theta) + \mathcal{KL}(q||p). \tag{4.30}$$

Để làm sáng tỏ vấn đề ở (4.30) chúng ta sẽ tiến hành khai triển $\ln p(\mathbf{x}|\Theta)$. Chúng ta có một loạt các biến đổi dưới đây,

$$\begin{aligned}
\ln p(\mathbf{x}|\Theta) &= \sum_z q(z) \ln p(\mathbf{x}|\Theta) \text{ do } \sum_z q(z) = 1, \\
&= \sum_z q(z) \ln \frac{p(\mathbf{x}, z|\Theta)}{p(z|\mathbf{x}, \Theta)} \text{ (theo công thức xác suất có điều kiện)}, \\
&= \sum_z q(z) \ln \frac{p(\mathbf{x}, z|\Theta) q(z)}{p(z|\mathbf{x}, \Theta) q(z)}, \\
&= \sum_z q(z) \ln \frac{p(\mathbf{x}, z|\Theta)}{q(z)} + \sum_z q(z) \ln \frac{q(z)}{p(z|\mathbf{x}, \Theta)} \text{ (dùng tính chất logarithm của một tích)}, \\
&= \mathcal{L}(\mathbf{x}|\Theta) + \mathcal{KL}(q||p). \qquad \square
\end{aligned}$$

Nhận xét Biểu thức (4.30) là khoảng cách giữa hàm log likelihood cận biên và giới hạn dưới chính bằng với tổng phân kỳ $\mathcal{KL}$. Vì vậy, mục đích trong tối đa hóa kỳ vọng là tối đa hóa giới hạn dưới này với Θ hay nói cách khác là đẩy giới hạn dưới càng gần với marginal likelihood càng tốt. Mặt khác, như chúng ta đã biết, phân kỳ $\mathcal{KL}$ là dùng để đo sự khác biệt của hai phân phối nên vấn đề tối đa hóa giới hạn dưới sẽ tương đương với việc tối thiểu phân kỳ $\mathcal{KL}$.

4.5 Áp dụng thuật toán EM vào mô hình hỗn hợp Gaussian

Trong phần này chúng ta sẽ dùng hỗn hợp Gaussian để giải thích ứng dụng của thuật toán EM.

Xét mô hình hỗn hợp Gaussian được giới thiệu như ở (4.1), ϕ_k là tham số đa thức của một điểm dữ liệu cụ thể của thành phần thứ k và z_i là biến tiềm ẩn của x_i. Mục đích của chúng ta là tìm $\max_{\mu,\Sigma,\phi} \ln p(\mathbf{x}|\mu, \Sigma, \phi)$.
Trước tiên, chúng ta giả sử số chiều của mỗi x_i là n chiều. Theo hoạt động của thuật toán EM, chúng ta lần lượt thực hiện 2 bước.
Bước E. Thiết lập $w_j^{(i)} = q_i(z_i = j) = p(z_i = j|x_i, \mu, \Sigma, \phi)$.

Bước M. Sử dụng (4.29) chúng ta có giới hạn dưới của $\ln p(\mathbf{x}|\mu,\Sigma,\phi)$ là

$$\begin{aligned}\mathcal{L}(\mathbf{x}|\Theta) &= \sum_i^m \sum_j^K q_i(z_i=j)\ln\frac{p(x_i,z_i=j|\mu,\Sigma,\phi)}{q_i(z_i=j)},\\ &= \sum_i^m \sum_j^K q_i(z_i=j)\ln\frac{p(x_i|z_i=j;\mu,\Sigma,\phi)p(z_i=j|\phi)}{q_i(z_i=j)}. \qquad (4.31)\end{aligned}$$

Chú ý rằng $x_i|z_i=j,\mu,\Sigma\sim\mathcal{N}(\mu_j,\Sigma_j)$ và $z_i=j|\phi\sim Multi(\phi)$. Như vậy chúng ta có thể tận dụng hàm mật độ xác suất của các phân phối này để tiếp tục mục đích của bài toán. Khi đó $\mathcal{L}(\mathbf{x}|\Theta)$ tiếp tục được biến đổi thành

$$\mathcal{L}(\mathbf{x}|\Theta)=\sum_i^m\sum_j^K w_j^{(i)} ln\ \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}}\ exp\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big)\ \phi_j}{w_j^{(i)}}. \qquad (4.32)$$

Tiếp theo, chúng ta cần tối đa giới hạn dưới $\mathcal{L}(\mathbf{x}|\Theta)$ cho các tham số μ,Σ,ϕ. Để làm được điều này chúng ta lần lượt lấy đạo hàm của $\mathcal{L}(\mathbf{x}|\Theta)$ theo các tham số μ,Σ,ϕ và cho kết quả đạo hàm đó bằng không.

Thực hiện lấy đạo hàm của $\mathcal{L}(\mathbf{x}|\Theta)$ theo μ_j ta được,

$$\begin{aligned}\nabla_{\mu_j}\mathcal{L}(\mathbf{x}|\Theta) &= \nabla_{\mu_j}\sum_i^m w_j^{(i)}\ln\ \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}}\ exp\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big)\ \phi_j}{w_j^{(i)}},\\ &= \nabla_{\mu_j}\sum_i^m w_j^{(i)}\left[\ln\ \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}}\phi_j}{w_j^{(i)}}+\ln exp\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big)\right],\\ &= \nabla_{\mu_j}\sum_i^m w_j^{(i)}\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big),\\ &= -\frac{1}{2}\sum_i^m w_j^{(i)}\nabla_{\mu_j}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j),\\ &= -\frac{1}{2}\sum_i^m w_j^{(i)}(-2)\Sigma_j^{-1}(x_i-\mu_j),\\ &\text{(vì } f(x,s)=(x-s)^\top A(x-s) \text{ thì } \nabla_s f(x,s)=-2A(x-s).),\\ &= \sum_i^m w_j^{(i)}\Sigma_j^{-1}(x_i-\mu_j).\end{aligned}$$

Tiếp đến, cho đạo hàm bằng không chúng ta có các biến đổi tương đương sau.

$$\begin{aligned}
\nabla_{\mu_j}\mathcal{L}(\mathbf{x}|\Theta) &= 0,\\
\sum_i^m w_j^{(i)}\Sigma_j^{-1}(x_i-\mu_j) &= 0,\\
\Sigma_j^{-1}\sum_i^m w_j^{(i)}(x_i-\mu_j) &= 0,\\
\sum_i^m w_j^{(i)}(x_i-\mu_j) &= 0,\\
\sum_i^m w_j^{(i)}x_i - \mu_j\sum_i^m w_j^{(i)} &= 0,\\
\mu_j\sum_i^m w_j^{(i)} &= \sum_i^m w_j^{(i)}x_i,\\
\mu_j &= \frac{\sum_i^m w_j^{(i)}x_i}{\sum_i^m w_j^{(i)}}.
\end{aligned}$$

Tiếp tục lấy đạo hàm của $\mathcal{L}(\mathbf{x}|\Theta)$ theo Σ_j ta được,

$$\begin{aligned}
\nabla_{\Sigma_j}\mathcal{L}(\mathbf{x}|\Theta) &= \nabla_{\Sigma_j}\sum_i^m w_j^{(i)} ln\ \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}}\ exp\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big)\ \phi_j}{w_j^{(i)}},\\
&= \nabla_{\Sigma_j}\sum_i^m w_j^{(i)}\Big[\ln\frac{1}{\sqrt{|\Sigma_j|}} + \ln exp\Big(-\frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big) + \ln\frac{1}{\sqrt{(2\pi)^n w_j^{(i)}}}\phi_j\Big],\\
&= \sum_i^m w_j^{(i)}\nabla_{\Sigma_j}\Big[\ln\frac{1}{\sqrt{|\Sigma_j|}} - \frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big],\\
&= \sum_i^m w_j^{(i)}\nabla_{\Sigma_j}\Big[\ln|\Sigma_j|^{-\frac{1}{2}} - \frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big],\\
&= \sum_i^m w_j^{(i)}\nabla_{\Sigma_j}\Big[-\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big],\\
&= -\frac{1}{2}\sum_i^m w_j^{(i)}\nabla_{\Sigma_j}\Big[\ln|\Sigma_j| + (x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big],\\
&= -\frac{1}{2}\sum_i^m w_j^{(i)}\Big[\frac{\partial}{\partial\Sigma_j}\ln|\Sigma_j| + \frac{\partial}{\partial\Sigma_j}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j)\Big].
\end{aligned}$$

Tới đây, chúng ta quan sát đạo hàm đầu tiên trong ngoặc vuông và có các biến đổi

dưới đây,

$$\begin{aligned}\frac{\partial \ln|\Sigma_j|}{\partial \Sigma_j} &= \frac{1}{|\Sigma_j|}\frac{\partial|\Sigma_j|}{\partial \Sigma_j} \ \Big(\text{vì } \frac{\partial \det(X)}{\partial X} = \det(X)(X^{-1})^\top\Big),\\ &= \frac{1}{|\Sigma_j|}|\Sigma_j|(\Sigma_j^{-1})^\top,\\ &= (\Sigma_j^{-1})^\top,\\ &= \Sigma_j^{-1} \text{ (vì } \Sigma \text{ là ma trận xác định dương).}\end{aligned} \tag{4.33}$$

Tiếp tục, chúng ta xét đạo hàm thứ 2 trong ngoặc vuông.

$$\frac{\partial}{\partial \Sigma_j}(x_i-\mu_j)^\top\Sigma_j^{-1}(x_i-\mu_j) = \Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Sigma_j^{-1} \ (\text{vì } \frac{\partial a^\top X^{-1}b}{\partial X} = -X^{-\top}ab^\top X^{-\top}). \tag{4.34}$$

Kết hợp các kết quả này lại, chúng ta nhận được:

$$\begin{aligned}\nabla_{\Sigma_j}\mathcal{L}(\mathbf{x}|\Theta) &= -\frac{1}{2}\sum_i^m w_j^{(i)}\Big[\Sigma_j^{-1} - \Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Sigma_j^{-1}\Big],\\ &= -\frac{1}{2}\sum_i^m w_j^{(i)}\Big[I - \Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Big]\Sigma_j^{-1}.\end{aligned}$$

Khi đó, cho kết quả đạo hàm của $\mathcal{L}(\mathbf{x}|\Theta)$ theo Σ_j bằng không chúng ta có các biến đổi tương đương sau.

$$
\begin{aligned}
-\frac{1}{2}\sum_{i}^{m} w_j^{(i)}\Big[I-\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Big]\Sigma_j^{-1}&=0,\\
\sum_{i}^{m} w_j^{(i)}\Big[I-\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Big]&=0,\\
\sum_{i}^{m} w_j^{(i)}\Big[\Sigma_j^{-1}\Sigma_j-\Sigma_j^{-1}(x_i-\mu_j)(x_i-\mu_j)^\top\Big]&=0,\\
\Sigma_j^{-1}\sum_{i}^{m} w_j^{(i)}\Big[\Sigma_j-(x_i-\mu_j)(x_i-\mu_j)^\top\Big]&=0,\\
\sum_{i}^{m} w_j^{(i)}\Big[\Sigma_j-(x_i-\mu_j)(x_i-\mu_j)^\top\Big]&=0,\\
\sum_{i}^{m} w_j^{(i)}\Sigma_j-\sum_{i}^{m} w_j^{(i)}(x_i-\mu_j)(x_i-\mu_j)^\top&=0,
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i}^{m} w_j^{(i)}\Sigma_j&=\sum_{i}^{m} w_j^{(i)}(x_i-\mu_j)(x_i-\mu_j)^\top,\\
\Sigma_j&=\frac{\sum_i^m w_j^{(i)}(x_i-\mu_j)(x_i-\mu_j)^\top}{\sum_i^m w_j^{(i)}}.
\end{aligned}
$$

Tiếp theo, chúng ta tìm ϕ_j. Quan sát thấy rằng $\sum_j \phi_j = 1$ do đó thay vì trực tiếp lấy đạo hàm của $\mathcal{L}(\mathbf{x}|\Theta)$ theo ϕ_j, chúng ta sẽ áp dụng nhân tử Lagrange để có một số bước biến đổi nhỏ sau.

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}|\Theta)&=\sum_{i}^{m}\sum_{l}^{k} w_l^{(i)} ln\ \frac{\frac{1}{\sqrt{(2\pi)^n|\Sigma_l|}}\ exp\Big(-\frac{1}{2}(x_i-\mu_l)^\top\Sigma_l^{-1}(x_i-\mu_l)\Big)\ \phi_l}{w_l^{(i)}},\\
&=\sum_{i}^{m}\sum_{l}^{k} w_l^{(i)} ln\ \phi_l.
\end{aligned}
$$

Đến đây, chúng ta cần xây dựng hàm Lagrange với nhân tử Lagrange là λ,

$$\mathcal{L}(\phi)=\mathcal{L}(\mathbf{x}|\Theta)+\lambda(\sum_{l}^{k}\phi_l-1).$$

Bây giờ, ta thực hiện lấy đạo hàm của $\mathcal{L}(\phi)$ theo ϕ_j và cho kết quả bằng không.

$$\begin{aligned}\frac{\partial \mathcal{L}(\phi)}{\partial \phi_j} &= \frac{\partial}{\partial \phi_j}\Big[\mathcal{L}(\mathbf{x}|\Theta) + \lambda(\sum_{l}^{k}\phi_l - 1)\Big],\\ &= \sum_i w_j^{(i)}\frac{1}{\phi_j} + \lambda = 0.\end{aligned}$$

Chúng ta tiếp tục thực hiện các biến đổi đại số,

$$\begin{aligned}\frac{1}{\phi_j}\sum_i w_j^{(i)} &= -\lambda,\\ \frac{1}{\phi_j} &= \frac{-\lambda}{\sum_i w_j^{(i)}},\\ \phi_j &= -\frac{\sum_i w_j^{(i)}}{\lambda}.\end{aligned}$$

Với $\phi_j = -\frac{\sum_i w_j^{(i)}}{\lambda}$ và $\sum_j \phi_j = 1$ ta có:

$$\begin{aligned}\sum_j \phi_j = \sum_j -\frac{\sum_i w_j^{(i)}}{\lambda} &= 1,\\ \lambda = -\sum_j\sum_i w_j^{(i)}&,\\ = -\sum_j\sum_i p(z^{(i)} = j|x^{(i)})&\\ = -\sum_i 1 = -m.&\end{aligned}$$

Từ đó, ta được biểu thức sau:

$$\phi_j = \frac{\sum_i w_j^{(i)}}{m}.$$

Cuối cùng, chúng ta thực hiện tính $\ln p(\mathbf{x}|\Theta)$ và kiểm tra sự hội tụ của các tham số. Nếu điều kiện hội tụ không được thỏa mãn thì quay lại bước E.

Chương 5

Hồi quy tuyến tính

5.1 Hồi quy

Hồi quy giống như bài toán phân loại, tuy nhiên biến phản hồi là biến liên tục. Một số vấn đề cơ bản có thể phát sinh trong hồi quy đó là giá trị đầu vào nhiều chiều hay bộ dữ liệu có giá trị ngoại lai (outliers).
Các bài toán hồi quy thường gặp trong thực tế như sau:

- Dự đoán giá cổ phiếu ngày mai khi biết giá cổ phiếu hôm nay và một số thông tin khác.
- Dự đoán tuổi của một người đang xem một video nhất định trên Youtube.
- Dự đoán nhiệt độ tại bất kỳ vị trí nào trong tòa nhà bằng dữ liệu thời tiết, thời gian, cửa cảm biến,..
- Dự đoán số lượng kháng nguyên đặc hiệu tuyến tiền liệt (PSA) trong cơ thể như là một chức năng của một số phép đo lâm sàng khác nhau.

5.2 Hồi quy tuyến tính

Trong máy học, vấn đề đáng quan tâm đầu tiên đó là hồi quy tuyến tính.
Giả sử chúng ta quan tâm đến giá trị của hàm số

$$y(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R},$$

với $\mathbf{x}$ là vector giá trị đầu vào d chiều và $\mathcal{D} = \{(\mathbf{x}_i, y_i)_{i=1}^N$ là các quan trắc của ánh xạ, ta dùng $\mathcal{D}$ để phục vụ cho dữ liệu huấn luyện.

Mục tiêu của bài toán hồi quy là có thể *dự đoán giá trị đầu ra y* dựa trên vector đặc trưng đầu vào mới $\mathbf{x}_*$. Nói cách khác, chúng ta muốn học từ dữ liệu huấn luyện $\mathcal{D}$ để tìm được một hàm dự đoán có tính chính xác cao, hàm này ký hiệu là $\hat{y}(\mathbf{x}_*)$. Nhìn chung, chúng ta có rất nhiều hàm $\hat{y}$. Vậy chúng ta làm thế nào để chọn được một hàm dự đoán $\hat{y}$ tốt nhất? Bài toán đưa về việc chúng ta làm thế nào để sử dụng tập các quan trắc huấn luyện $\mathcal{D}$ một cách hiệu quả?

Không gian của tất cả các hàm hồi quy $\hat{y}$ là rất rộng, để giới hạn không gian này và làm cho quá trình học dễ kiểm soát hơn thì bài toán hồi quy tuyến tính giả sử mối quan hệ giữa $\mathbf{x}$ và y là tuyến tính:

$$y(x) = \mathbf{x}^\top \mathbf{w} + \varepsilon(\mathbf{x}), \tag{5.1}$$

trong đó $\varepsilon(\mathbf{x})$ là sai số hay phần dư (residual) của mô hình, $\mathbf{x}_i = [x_{i1}, x_{i2,}, ..., x_{id}]$ với $\mathrm{i} = \overline{1, n}$ và $\mathbf{w} = [\mathrm{w}_0, \mathrm{w}_1, ..., \mathrm{w}_{\mathrm{d}-1}]$ là vector hệ số chúng ta cần đi tìm. Đây cũng chính là *tham số mô hình* của bài toán. Chúng ta thường giả sử ϵ có phân phối chuẩn $\varepsilon \sim N(\mu, \sigma^2)$.

Một vấn đề khác nữa đó là khi chúng ta có bộ dữ liệu $\mathcal{D} = \{(\mathbf{x}_i, y_i)_{i=1}^N$, thực hiện xếp các vector $\mathbf{x_i}$ chồng lên nhau chúng ta sẽ nhận được ma trận giá trị đầu vào là $\mathbf{X}^{N\times d}$. Tương tự như vậy chúng ta cũng nhận được vector giá trị đầu ra là ${\mathbf{y}_i}^{N\times 1}$ và vector sai số là $\boldsymbol{\epsilon}$. Lúc này, dựa vào mối quan hệ tuyến tính được giả sử ở (5.1) chúng ta có thể biểu diễn bộ dữ liệu huấn luyện dưới dạng

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}. \tag{5.2}$$

Bên cạnh đó, như đã nói y là giá trị đầu ra chính xác, trong khi $\hat{y}$ là giá trị đầu ra dự đoán của mô hình hồi quy tuyến tính ở (5.1). Hai giá trị y và $\hat{y}$ là khác nhau do đó dĩ nhiên chúng ta sẽ có sai số mô hình. Và khi đường thẳng hồi quy có xu hướng trùng với đường nối của các điểm dữ liệu thì sự sai số này là rất nhỏ.

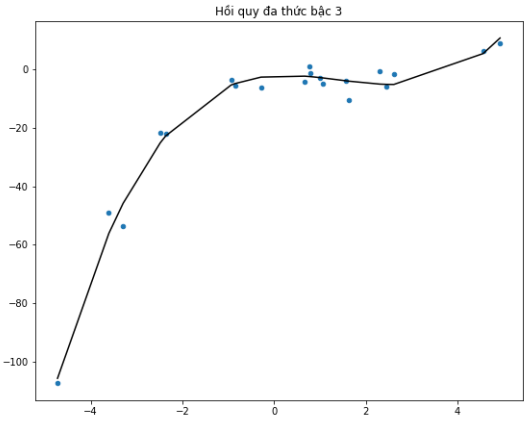
Chúng ta cần chú ý rằng mô hình được giới thiệu ở trên không chứa hệ số chặn. Trong trường hợp chúng ta thêm một thành phần bằng 1 vào vector giá trị đầu vào $\mathbf{x}' = [1, \mathbf{x}]^\top$ thì lúc này giá trị w_0 chính là hệ số chặn,

$$y = w_0 + \mathbf{x}_1 w_1 + \mathbf{x}_2 w_2 + ... + \mathbf{x}_{d-1} w_{d-1} + \varepsilon. \tag{5.3}$$

Tổng quát hơn, chúng ta có thể chọn bất kỳ hàm nào với $\mathbf{x}$ đóng vai trò là giá trị đầu vào cho mô hình hồi quy tuyến tính. Ví dụ, để thực hiện hồi quy đa thức bậc k có dạng

$$y = w_0 + w_1 \mathbf{x} + w_2 \mathbf{x}^2 + ... + w_k \mathbf{x}^k + \varepsilon, \tag{5.4}$$

chúng ta lấy $\mathbf{x}' = [1, \mathbf{x}, \mathbf{x}^2, ..., \mathbf{x}^k]^\top$ làm giá trị đầu vào. Quay lại vấn đề của bài



Hình 5.1. Đồ thị cho hàm số hồi quy theo đa thức bậc 3.

toán hồi quy đó là ước lượng tham số $\mathbf{w}$ của mô hình. Tiểu luận xin giới thiệu hai phương pháp ước lượng phổ biến nhất đó là phương pháp bình phương bé nhất và phương pháp hồi quy ridge.

5.3 Phương pháp bình phương bé nhất

Như đã đề cập trước đó, một ánh xạ tuyến tính $\mathbf{x}^\top \mathbf{w}$ tốt sẽ cho phần dư rất nhỏ. Vậy nên chúng ta mong muốn giá trị sau đây càng nhỏ càng tốt:

$$\hat{\varepsilon}^2 = (y - \hat{y})^2.$$

Ở đây chúng ta lấy bình phương vì $\hat{\varepsilon} = y - \hat{y}$ có thể là một số âm. Việc sai số là nhỏ nhất có thể được mô tả bằng cách lấy giá trị tuyệt đối $\hat{\varepsilon} = |y - \hat{y}|$. Tuy nhiên, cách làm này ít được sử dụng vì hàm trị tuyệt đối không khả vi tại mọi điểm nên không thuận tiện cho việc tối ưu sau này.
Bài toán dẫn đến việc tìm giá trị nhỏ nhất của hàm số:

$$\sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2,$$

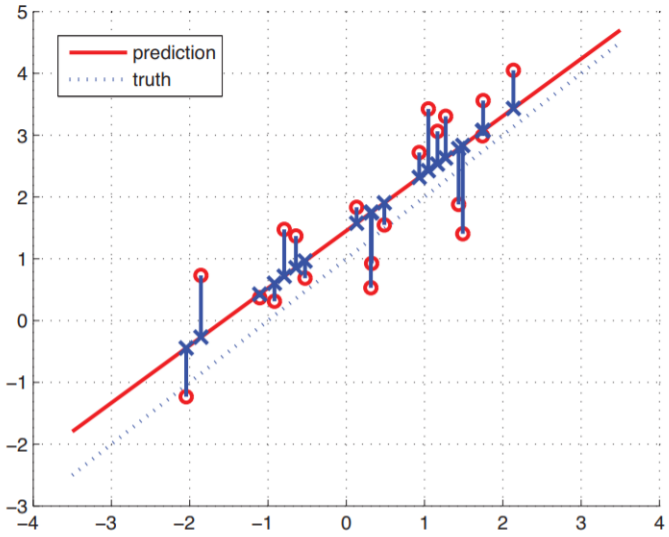
hay nói cách khác, chúng ta đi tìm giá trị $\mathbf{w}$ sao cho:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\mathbf{x}_i^\top \mathbf{w} - y_i\right)^2. \tag{5.5}$$

Cách tiếp cận như thế được gọi là phương pháp bình phương bé nhất. Nhờ có giả định độ lớn của tất cả các phần dư đều độc lập với nhau, do đó chúng ta có thể lấy tổng bình phương của phần dư và tìm cực tiểu của chúng.

Trong phương pháp bình phương bé nhất của mô hình tuyến tính, chúng ta cố gắng làm cực tiểu khoảng cách từ mỗi điểm trong tập huấn luyện (được biểu thị bởi một vòng tròn màu đỏ) đến điểm chính xác của nó (được biểu thị bởi một dấu thập màu xanh). Nghĩa là, chúng ta phải giảm thiểu tổng chiều dài của các đường dọc màu xanh. Đường thẳng màu đỏ đại điện cho phương trình hồi quy $\hat{y}(x) = w_0 + w_1 x$, đây chính là đường thẳng hồi quy.

Để viết được phương trình hồi quy thì trước tiên chúng ta cần phải tìm ước lượng



Hình 5.2. Phân bố của các điểm dữ liệu (màu đỏ) và đường thẳng xấp xỉ tìm được nhờ phương pháp hồi quy tuyến tính.

tham số $\mathbf{w}$. Hàm mất mát của bài toán hồi quy tuyến tính có dạng

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}), \tag{5.6}$$

với $\mathbf{y} = [y_1, y_2, ..., y_N]^\top$ và $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$.
Khai triển hàm mất mát chúng ta có,

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= (\mathbf{y}^\top - \mathbf{w}^\top \mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\mathbf{w}), \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}. \end{aligned} \tag{5.7}$$

Mặt khác, vì $(\mathbf{y}^\top \mathbf{X}\mathbf{w})^\top = \mathbf{w}^\top \mathbf{X}^\top \mathbf{y}$ và $(\mathbf{w}^\top \mathbf{X}^\top \mathbf{y})^\top = \mathbf{y}^\top \mathbf{X}\mathbf{w}$ nên $\mathbf{y}^\top \mathbf{X}\mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{y}$.
Do đó, phương trình (5.7) trở thành

$$\mathcal{L}(\mathbf{w}) = \mathbf{y}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{y}^\top \mathbf{X}\mathbf{w}.$$

Tiếp theo, chúng ta thực hiện lấy đạo hàm của $\mathcal{L}(\mathbf{w})$ theo $\mathbf{w}$ và cho kết quả bằng không ta nhận được,

$$\frac{\nabla \mathcal{L}(\mathbf{w})}{\nabla \mathbf{w}} = \nabla_{\mathbf{w}}(\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}) - 2\nabla_{\mathbf{w}}(\mathbf{y}^\top \mathbf{X}\mathbf{w}) = 0. \tag{5.8}$$

Áp dụng định lý 2 cho đạo hàm thứ nhất với $\mathbf{X}^\top\mathbf{X}$ là ma trận đối xứng, chúng ta có

$$\nabla_{\mathbf{w}}(\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}) = 2\mathbf{X}^\top\mathbf{X}\mathbf{w}. \tag{5.9}$$

Với đạo hàm thứ hai, vì $(\mathbf{X}^\top\mathbf{y})^\top\mathbf{w} = \mathbf{w}^\top(\mathbf{X}^\top\mathbf{y})$ nên theo định lý 1 thì

$$\nabla_{\mathbf{w}}(\mathbf{y}^\top\mathbf{X}\mathbf{w}) = \nabla_{\mathbf{w}}[(\mathbf{X}^\top\mathbf{y})^\top\mathbf{w}] = \mathbf{X}^\top\mathbf{y}. \tag{5.10}$$

Lúc này, phương trình (5.8) trở thành

$$\begin{aligned} 2\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{X}^\top\mathbf{y} &= 0, \\ \mathbf{X}^\top\mathbf{X}\mathbf{w} &= \mathbf{X}^\top\mathbf{y}. \end{aligned}$$

Giả sử $\mathbf{X}^\top\mathbf{X}$ là ma trận khả nghịch thì khi đó,

$$(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Đến đây, gọi $\widehat{\mathbf{w}}$ là ước lượng của $\mathbf{w}$ chúng ta nhận được

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{5.11}$$

$$\hat{y}(\mathbf{x}_*) = \mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{5.12}$$

Như vậy chúng ta đã tìm được hàm dự đoán giá trị đầu ra y cho vector giá trị đầu vào mới $\mathbf{x}_*$. Phương pháp bình phương bé nhất là phương pháp thường dùng trong hồi quy tuyến tính.

5.4 Phương pháp hồi quy ride

Trong trường hợp ma trận $\mathbf{X}^\top\mathbf{X}$ không khả nghịch, có một kỹ thuật nhỏ để tránh vấn đề này là biến đổi $\mathbf{X}^\top\mathbf{X}$ một chút để biến nó trở thành $\mathbf{A} = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ với λ là một số dương rất nhỏ và $\mathbf{I}$ là ma trận đơn vị với bậc phù hợp.

Ma trận $\mathbf{A}$ là khả nghịch vì nó là một ma trận xác định dương. Thật vậy, với mọi

$\mathbf{w} \neq 0$,

$$\begin{aligned}\mathbf{w}^\top \mathbf{A}\mathbf{w} &= \mathbf{w}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\mathbf{w} \\ &= \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} + \lambda \mathbf{w}^\top \mathbf{w} \\ &= \left\|\mathbf{X}^\top \mathbf{w}\right\|_2^2 + \lambda \left\|\mathbf{w}\right\|_2^2 > 0. \end{aligned} \tag{5.13}$$

Hàm mất mát của bài toán lúc này là:

$$\mathcal{L}_2(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \mathbf{w}\right)^2 + \lambda \left\|\mathbf{w}\right\|_2^2 \text{ với } \lambda \geq 0. \tag{5.14}$$

Tiếp tục khai triển,

$$\mathcal{L}_2(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}. \tag{5.15}$$

Lấy đạo hàm của $\mathcal{L}_2(\mathbf{w})$ theo $\mathbf{w}$ và cho kết quả bằng không,

$$\frac{\nabla \mathcal{L}_2(\mathbf{w})}{\nabla \mathbf{w}} = 0.$$

Thực hiện như ở phương pháp bình phương bé nhất chúng ta nhận được:

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{5.16}$$

Vậy $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ chính là nghiệm của bài toán tối thiểu $\mathcal{L}_2(\mathbf{w})$ ở (5.14). Mô hình máy học với hàm mất mát như ở (5.14) được gọi là hồi quy ridge.

Phương pháp hồi quy ridge rất hữu ích vì đôi khi thực hiện theo phương pháp bình phương bé nhất phát sinh các vấn đề không mong muốn. Cụ thể, khi dữ liệu huấn luyện rất gần với dữ liệu chính xác hoặc khi số lượng quan sát ít hơn số đặc trưng. Lúc này, độ lớn của tham số chúng ta cần ước lượng $\hat{\mathbf{w}}$ có thể rất xấu và cực lớn. Và đặc biệt hơn còn có thể xuất hiện hiện tượng *overfitting*.

Nhận xét.

Vì tham số $\lambda \geq 0$ cũng là một tham số của mô hình mà chúng ta có thể điều chỉnh

theo ý muốn nên khi $\lambda \to 0$ thì $\hat{\mathbf{w}}_{ridge} = \hat{\mathbf{w}}_{OLS}$.
Ngoài ra, khi $\lambda \to \infty$ thì $\hat{\mathbf{w}}_{ridge} \to 0$.

5.5 Hồi quy tuyến tính Bayesian

Xét một cách tiếp cận Bayesian trong hồi quy tuyến tính. Giả sử chúng ta có mối quan hệ tuyến tính:

$$y(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + \varepsilon(\mathbf{x})$$

như được giới thiệu ở (5.1). Ở đây, vector $\mathbf{w}$ xem như tham số chưa biết của mô hình, tham số này dùng để thảo luận cách sử dụng của phương pháp Bayesian.
Để thuận tiện, chúng ta chọn một phân phối tiên nghiệm Gaussian trên $\mathbf{w}$ có dạng

$$p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{5.17}$$

Như đã nói trước đó, chúng ta mong đợi một tiên nghiệm làm cho entries của $\mathbf{w}$ nhỏ. Thêm vào đó, nếu không có lập luận nào khác, chúng ta có thể giả định rằng tất cả các hướng tiềm năng của $\mathbf{w}$ là gần như bằng nhau. Hai giả định đó có thể được mã hóa bằng cách chọn phân phối Gaussian cho $\mathbf{w}$ với trung bình bằng 0 và hiệp phương sai là đường đẳng hướng $s^2\mathbf{I}$(isotropic diagonal covariance).

$$p(\mathbf{w}|s^2) = \mathcal{N}(\mathbf{w}|\mathbf{0}, s^2\mathbf{I}). \tag{5.18}$$

Cho dữ liệu $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, việc cần làm là tìm được công thức của phân phối hậu nghiệm $p(\mathbf{w}|\mathcal{D})$. Chúng ta thực hiện bằng cách lấy phân phối đồng thời giữa vector trọng số $\mathbf{w}$ và vector quan trắc của giá trị $\mathbf{y}$ nhận giá trị đầu vào $\mathbf{X}$. Chúng ta nhắc lại giả định mối quan hệ tuyến tính trên dữ liệu huấn luyện,

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}.$$

Ở đây $\mathbf{X}$ và $\mathbf{w}$ là tham số đã biết của mô hình và yếu tố không đảm bảo ở đây là phần dư ε. Chúng ta giả sử phần dư ε độc lập, không chệch và có xu hướng "nhỏ".

Muốn được như vậy, chúng ta lập mô hình cho ε như sau:

$$p(\varepsilon|\sigma^2) = \mathcal{N}(\varepsilon|\mathbf{0}, \sigma^2\mathbf{I}). \tag{5.19}$$

Đặt $\mathbf{f} = \mathbf{X}\mathbf{w}$, khi đó $\mathbf{f}$ là một phép biến đổi tuyến tính của phân phối Gaussian. Áp dụng **tính chất 1**, ta có:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{f}; \boldsymbol{X}\boldsymbol{\mu}, \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top). \tag{5.20}$$

Bên cạnh đó, khi đặt $\mathbf{f} = \mathbf{X}\mathbf{w}$ thì lúc này $\mathbf{y} = \mathbf{X}\mathrm{w} + \varepsilon$, vector $\mathbf{y}$ là tổng của hai vector độc lập $\mathbf{f}$ và ε có phân phối Gaussian đa biến. Dùng **tính chất 2**, ta có:

$$\left.\begin{array}{l} \mathbf{X}\mathbf{w} = \mathbf{f} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\mu}, \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top) \\ \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \end{array}\right\} \Rightarrow \mathbf{y} = \mathbf{f} + \varepsilon \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\mu}, \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top + \sigma^2\mathbf{I})$$

Do đó ta viết

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2) = \mathcal{N}(\mathbf{y}; \boldsymbol{X}\boldsymbol{\mu}, \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top + \sigma^2\mathbf{I}). \tag{5.21}$$

Cuối cùng, chúng ta ước lượng hiệp phương sai giữa $\mathbf{y}$ và $\mathbf{w}$:

$\mathrm{cov}[\mathbf{y}, \mathbf{w}] = \mathrm{cov}[\mathbf{X}\mathbf{w} + \varepsilon, \mathbf{w}]$,
$= \mathbf{X}\mathrm{cov}[\mathbf{w}, \mathbf{w}] + \mathrm{cov}[\varepsilon, \mathbf{w}]$ (áp dụng $\mathrm{cov}[aX_1 + bX_2, Y] = a\mathrm{cov}[X_1, Y] + b\mathrm{cov}[X_2, Y]$),
$= \mathbf{X}\mathrm{cov}[\mathbf{w}, \mathbf{w}]$ (vì ε độc lập với $\mathbf{w}$ nên $\mathrm{cov}[\varepsilon, \mathbf{w}] = 0$).
Do đó

$$\mathrm{cov}[\mathbf{y}, \mathbf{w}] = \mathbf{X}\mathrm{cov}[\mathbf{w}, \mathbf{w}] = \mathbf{X}\boldsymbol{\Sigma}. \tag{5.22}$$

Tiếp theo, ta dùng **định lý 3:**
Khi đó, phân phối đồng thời của $\mathbf{w}$ và $\mathbf{y}$ là một phân phối Gaussian đa biến:

$$p\left(\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix} |\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma^2\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{X}\boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Sigma} & \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top + \sigma^2\mathbf{I} \end{bmatrix}\right). \tag{5.23}$$

Chúng ta tiếp tục sử dụng công thức Gaussian đa biến có điều kiện trên vector để nhận được phân phối hậu nghiệm của $\mathbf{w}$ cho bởi dữ liệu $\mathcal{D}$ như sau:

$$p(\mathbf{w}|\mathcal{D}, \mu, \mathbf{\Sigma}, \sigma^2) = \mathcal{N}(\mathbf{w}; \mu_{\mathbf{w}|\mathcal{D}}, \mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}}) \tag{5.24}$$

với

Chương 6

Bayesian model regression

6.1 Mô hình chọn

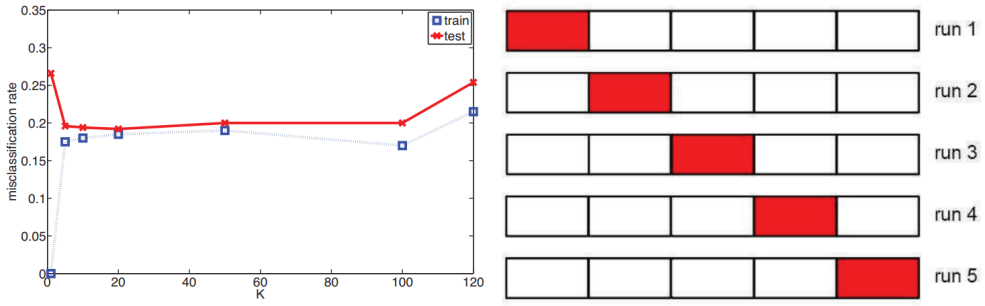
Khi chúng ta có nhiều mô hình có độ phức tạp khác nhau (ví dụ: mô hình hồi quy tuyến tính hoặc hồi quy logistic với bậc đa thức khác nhau hoặc phân loại KNN với các giá trị khác nhau của K), chúng ta nên chọn mô hình như thế nào cho đúng? Một cách tiếp cận tự nhiên là tính toán tỷ lệ phân loại sai trên tập huấn luyện cho từng phương pháp. Điều này được định nghĩa như sau:

$$err(f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) \neq y_i),$$

với $f(x)$ là phân loại của bài toán. Trong hình (6.1a) tỉ lệ lỗi này được vẽ so với K cho phân loại KNN (đường màu xanh chấm). Ta thấy rằng tăng K sẽ làm tăng tỉ lệ lỗi trên tập huấn luyện, bởi vì ta đã làm trơn quá mức (over-smoothing). Như đã nói ở trên, ta có thể nhận được lỗi tối thiểu trong tập huấn luyện bằng cách sử dụng K = 1, vì mô hình này chỉ ghi nhớ dữ liệu.

Tuy nhiên, điều quan tâm là tổng quát hóa lỗi (generalization error) là giá trị dự kiến của tỷ lệ phân loại sai (misclassification) khi tính trung bình trên dữ liệu trong tương lai (future data). Điều này có thể được xấp xỉ bằng cách tính tỷ lệ phân loại sai trên bộ thử nghiệm (test set) độc lập lớn, không được sử dụng trong quá trình đào tạo mô hình. Ta vẽ lỗi kiểm tra (test error) với K trong hình (6.1a) bằng màu

đỏ đặc (đường cong trên). Bây giờ thấy có một đường cong hình chữ U(U-shaped curve): cho các mô hình phức tạp (K nhỏ),phương pháp *overfits*, và cho các mô hình đơn giản hơn (K lớn), phương pháp *underfits*. Do đó, một cách hiển nhiên để chọn K là chọn giá trị với sai số tối thiểu trên bộ thử nghiệm.



(a) **(b)**

Hình 5.1(a) Tỷ lệ phân loại sai với K trong phân loại *K-nearest neighbor*. Ở bên trái, với K nhỏ, mô hình phức tạp và do đó chúng trở nên *overfitting*. Ở bên phải, với K lớn, mô hình đơn giản và chúng trở nên *underfitting*. Đường chấm màu xanh: tập huấn luyện (cỡ 200). Đường liền màu đỏ: bộ kiểm tra (kích thước 500). (b) Sơ đồ của 5-fold cross validation.

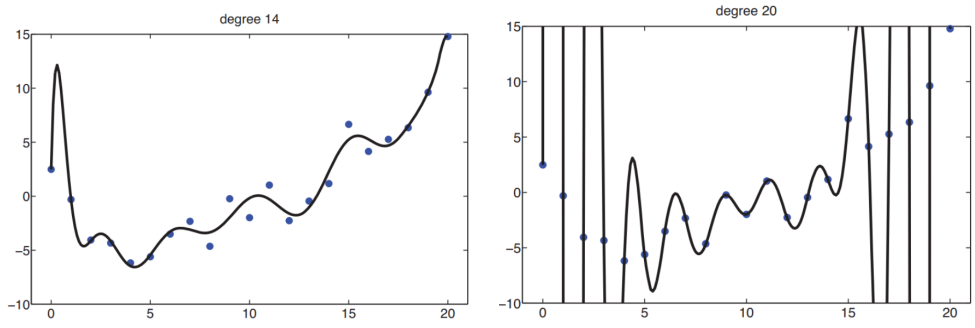
Thật không may, khi đào tạo mô hình, ta không có quyền truy cập vào bộ thử nghiệm (theo giả định), vì vậy không thể sử dụng bộ thử nghiệm để chọn mô hình có độ phức tạp phù hợp. Tức là, trong các môi trường học thuật, ta thường có quyền truy cập vào bộ thử nghiệm (test set), nhưng không nên sử dụng nó để điều chỉnh mô hình hoặc lựa chọn mô hình, nếu không chúng ta sẽ nhận được hiệu suất ước lượng "unrealistic optimism" trong phương pháp. Đây là một trong những "quy tắc vàng" về nghiên cứu máy học. Tuy nhiên, có thể tạo bộ thử nghiệm bằng cách phân vùng tập huấn luyện thành hai: phần được sử dụng để huấn luyện mô hình và phần thứ hai, được gọi là tập xác thực(validation set), được sử dụng để chọn độ phức tạp của mô hình. Sau đó, chúng ta phù hợp với tất cả các mô hình trên tập huấn luyện và đánh giá hiệu suất của chúng trên tập xác thực và chọn ra mô hình tốt nhất. Khi đã chọn được mô hình tốt nhất, chúng ta có thể chỉnh lại nó cho tất cả các dữ liệu có sẵn. Nếu có một bộ kiểm tra riêng, ta có thể đánh giá hiệu suất trên tập xác thực , để ước lượng độ chính xác của phương pháp.

Thông thường sử dụng khoảng 80% dữ liệu cho tập huấn luyện và 20% cho tập

xác thực. Nhưng nếu số lượng các trường hợp đào tạo ít, thì kỹ thuật này gặp vài vấn đề, bởi vì mô hình đã không có đủ dữ liệu để đào tạo và đã không có đủ dữ liệu để ước lượng hiệu suất (performance) trong tương lai.
Một giải pháp đơn giản nhưng phổ biến cho vấn đề này là sử dụng xác thực chéo (cross validation (CV)). Ý tưởng rất đơn giản: chúng ta chia dữ liệu đào tạo thành K nếp gấp (fold); sau đó, với mỗi nếp gấp $k \in \{1, ..., K\}$, chúng ta huấn luyện trên tất cả các nếp gấp ngoại trừ nếp thứ k và kiểm tra trên nếp thứ k, và thực hiện theo kiểu vòng tròn (we train on all the folds but the k'th, and test on the k'th, in a round-robin fashion), được phát họa trong hình (6.1b). Sau đó, tính toán sai số trung bình trên tất cả các nếp gấp và sử dụng điều này như một *proxy* cho lỗi kiểm tra. (Lưu ý rằng mỗi điểm chỉ được dự đoán một lần, mặc dù nó sẽ được sử dụng để huấn luyện $K - 1$ lần). Thông thường sử dụng $K = 5$; đây được gọi là *5-fold CV*. Nếu đặt $K = N$, thì sẽ nhận được một phương thức gọi là *leave-one out cross validation* hoặc LOOCV, vì trong lần i, đã huấn luyện tất cả các trường hợp dữ liệu trừ i, và sau đó kiểm tra trên i.

Chọn K cho trình phân lớp KNN là trường hợp đặc biệt của một vấn đề chung hơn được gọi là lựa chọn mô hình, trong đó chúng ta phải chọn giữa các mô hình với mức độ linh hoạt khác nhau.

6.2 Bayesian model regression



(a) **(b)**

Hình 6.2. Đa thức bậc 14 và 20 phù hợp bằng bình phương tối thiểu đến 21 điểm dữ liệu.

Trong hình (6.2) việc sử dụng đa thức bậc quá cao sẽ cho kết quả *overfitting* và

sử dụng bậc đa thức quá thấp sẽ cho kết quả *underfitting*. Tương tự trong hình (6.6a) việc sử dụng tham số chính quy hóa quá nhỏ sẽ dẫn đến *overfitig* và giá trị quá lớn dẫn đến *underfiting*. Nói chung, khi gặp một tập hợp các mô hình (nghĩa là các họ của những phân ơhoois tham số (parametric distributions)) có độ phức tạp khác nhau, chúng ta nên chọn mô hình tốt nhất bằng cách nào? Đây được gọi là vấn đề lựa chọn mô hình.

Một cách tiếp cận là sử dụng xác thực chéo (cross-validation) để ước lượng tổng quát hóa lỗi trên các mô hình ứng cử (candidate models) và sau đó chọn mô hình tốt nhất. Tuy nhiên, điều này đòi hỏi phải phù hợp với từng mô hình K lần, trong đó K là số lần gấp của CV. Một cách tiếp cận hiệu quả hơn là tính toán hậu nghiệm trên các mô hình,

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m,\mathcal{D})}.$$

Từ điều này, chúng ta có thể dễ dàng tính mô hình MAP, $\hat{m} = \operatorname{argmax} p(m|\mathcal{D})$. Đây chính là **Bayesian model selection**.
Nếu chúng ta dùng uniform prior trên các mô hình, $p(m) \propto 1$, để chọn ra mô hình tối đa hóa.

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}, \tag{6.1}$$

Đại lượng này được gọi là marginal likelihood cho mô hình m. Các chi tiết về cách thực hiện tích phân này sẽ được thảo luận trong phần (6.2.2) . Nhưng trước tiên, chúng ta đưa ra một giải thích trực quan về ý nghĩa của số lượng này.

6.2.1 Bayesian Occam's razor

Chúng ta luôn nghĩ rằng dùng $p(\mathcal{D}|m)$ để chọn các mô hình sẽ luôn ưu tiên mô hình có nhiều tham số nhất. Điều này là đúng nếu dùng $p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m)$ để chọn các mô hình, với $\hat{\boldsymbol{\theta}}_m$ là ước lượng MLE hay MAP của các tham số cho mô hình m, bởi vì các mô hình với nhiều các tham số hơn sẽ phù hợp với bộ dữ liệu hơn và do đó đạt được *likelihood* cao hơn. Tuy nhiên, nếu chúng ta tích hợp các tham số, thay vì tối đa hóa chúng, chúng được bảo vệ một cách tự động nhằm tránh hiện tượng

overfiting: tức là các mô hình có nhiều tham số không nhất thiết phải có *marginal likelihood* cao. Diều này được gọi là hiệu quả *Bayesian Occam's razor*, sau đó được biết đến với tên **Occam's razor**, trong đó nói rằng ta nên chọn mô hình đơn giản nhất giải thích đầy đủ dữ liệu.

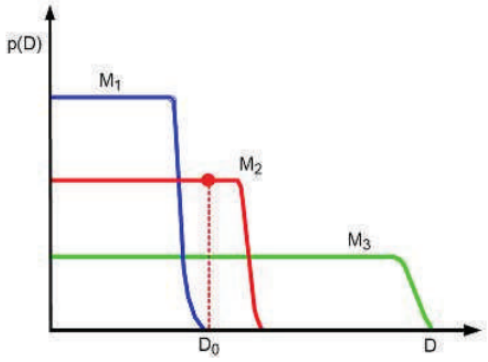
Một cách để hiểu về **Bayesian Occam's razor** là chú ý rằng *marginal likelihood* được viết dưới dạng sau, dựa trên quy tắc chuỗi của xác suất:

$$p(\mathcal{D}) = p(y_1)p(y_2|y_1)p(y_3|y_2, y_1)...p(y_N|y_{1:N-1}),$$

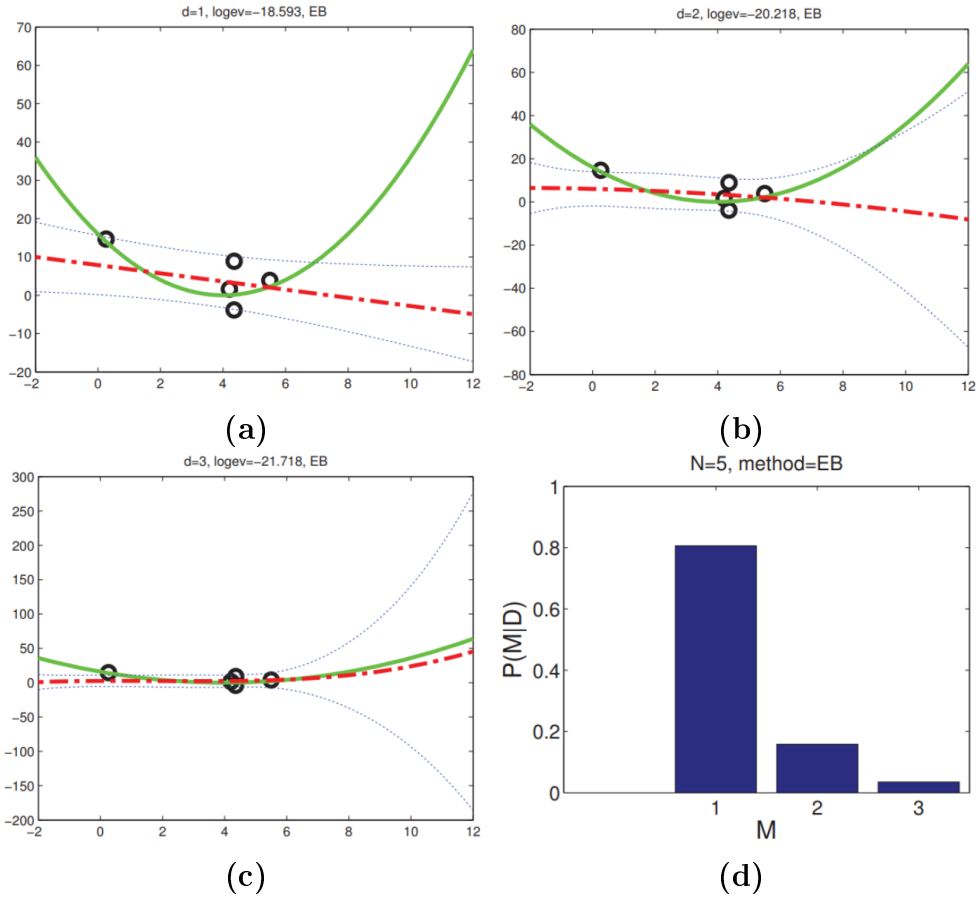
công thức trên đã bỏ qua điều kiện trên $\mathbf{x}$ cho ngắn gọn. Diều này tương tự với ước lượng *eave-one-out cross-validation* của likelihood, vì chúng ta dự đoán từng điểm trong tương lai được đưa ra từ các điểm trước đó.(Dĩ nhiên, bậc của dữ liệu không có vấn đề gì trong biểu thức trên.) Nếu mô hình quá phức tạp , it will overfit the "early" examples and will then predict the remaining ones poorly.

Một cách khác để hiểu về hiệu ứng của Bayesian Occam's razor là tổng các xác suất phải bằng 1, tức là $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$ với tổng này là trên tất cả các tập dữ liệu có thể. Các mô hình phức tạp có thể dự đoán được nhiều thứ, phải phân tán hàm khối xác suất của chúng một cách mỏng manh và do đó sẽ không thu được xác suất lớn cho bất kỳ tập dữ liệu đã cho như các mô hình đơn giản hơn. Diều này đôi khi được gọi là bảo tồn nguyên lý khối lượng xác suất, và được minh họa ở hình (6.3). Trên trục hoành chúng ta vẽ tất cả các tập dữ liệu có thể theo thứ tự độ phức tạp tăng dần (được đo bằng một số ý nghĩa trừu tượng). Trên trục tung, là đồ thị dự đoán của ba mô hình có thể: mô hình đơn giản là mô hình M_1, mô hình trung bình M_2 và mô hình phức tạm M_3. Dữ liệu được quan sát $\mathcal{D}_\prime$ được thể hiện trên trục tung. Mô hình 1 quá đơn giản và gán xác suất thấp cho $\mathcal{D}_\prime$. Mô hình 3 cũng gán cho $\mathcal{D}_\prime$ xác suất tương đối thấp, bởi vì có nó có thể dự đoán nhiều tập dữ liệu và do đó nó lan truyền xác suất của nó khá rộng rãi và mỏng. Mô hình 2 là phù hợp: nó dự đoán dự liệu được quan sát với mức độ tin cậy hợp lý, nhưng không dự đoán quá nhiều thứ khác. Vì thể, mô hình 2 là mô hình có thể xảy ra nhất.

Một ví dụ cụ thể về Bayesian Occam's razor, bây giờ ta hãy xem xét ví dụ trong hình (6.4). Chúng ta dùng các đa thức bậc 1, 2 và 3 để phù hợp với $N = 5$ bộ



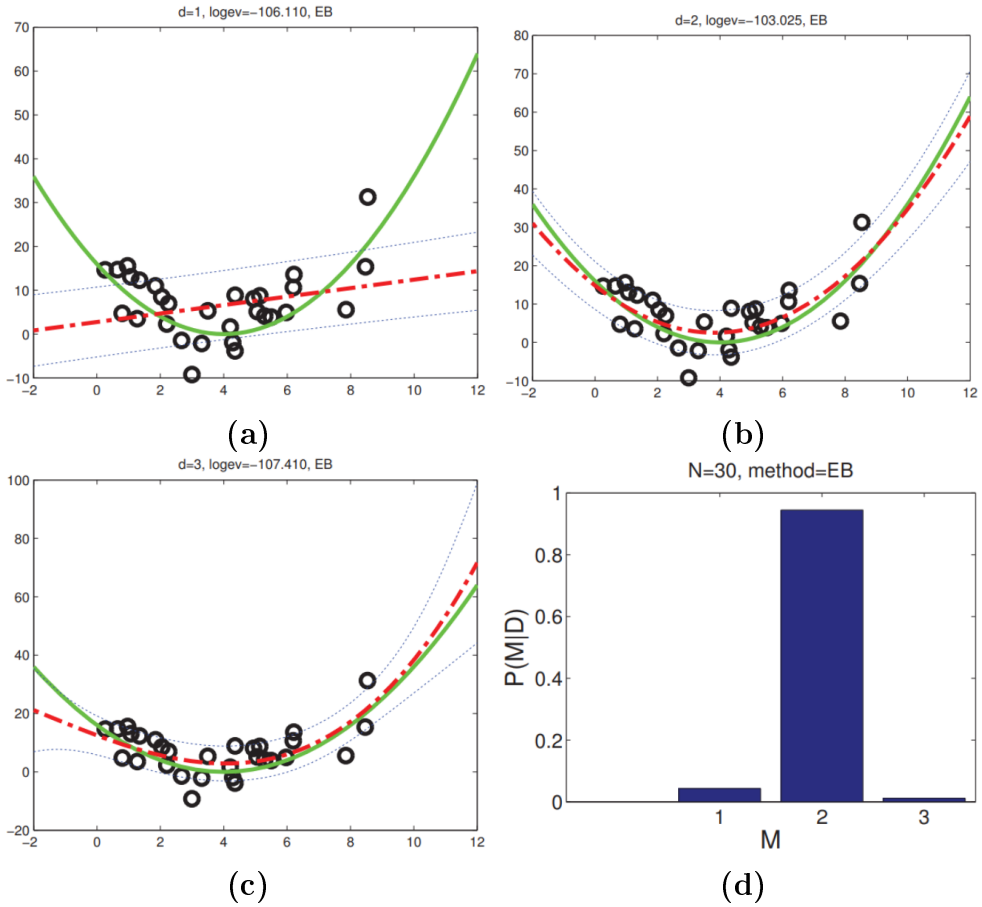
Hình 6.3. Một minh họa sơ đồ của Bayesian Occam's razor. Đường cong rộng (màu xanh lá cây) tương ứng với một mô hình phức tạp, đường cong hẹp (màu xanh) với một mô hình đơn giản và đường cong giữa (màu đỏ) là vừa phải.



Hình 6.4. (a-c) Đa thức bậc 1, 2 và 3 phù hợp với $N = 5$ điểm dữ liệu bằng Bayes theo kinh nghiệm (empirical Bayes). Đường cong màu xanh lá cây đặc là chức năng thực sự, đường cong màu đỏ nét đứt là dự đoán (đường màu xanh chấm chấm biểu thị $\pm\sigma$ xung quanh giá trị trung bình). (d) Chúng tôi vẽ sơ đồ sau cho các mô hình, $p(d|\mathcal{D})$, giả sử đồng phục trước $p(d) \propto 1$.

dữ liệu. Nó cũng cho thấy hậu nghiệm sp với các mô hình, ở đây ta sử dụng tiên nghiệm Gaussian. Ta không có đủ dữ liệu để chứng minh một mô hình phức tạp ,

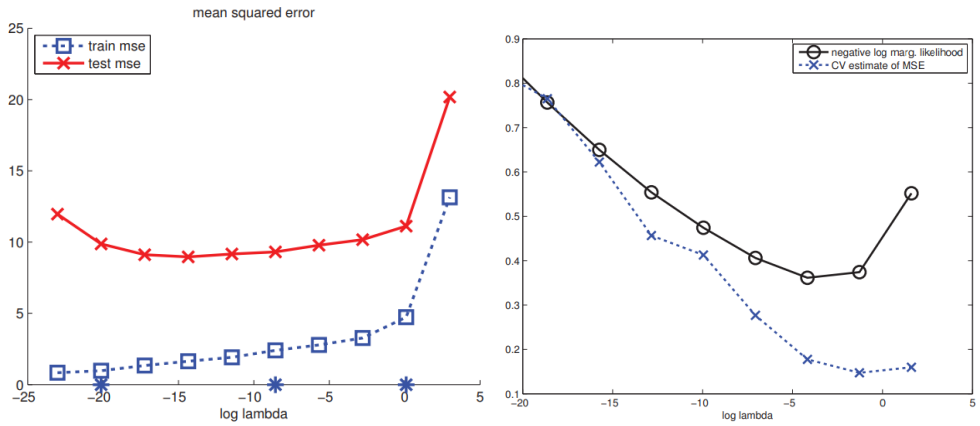
vì vậy mô hình MAP với $d = 1$. Trong hình (6.5) với $N = 30$ cho thấy $d = 2$ là mô hình phù hợp.(thực tế dữ liệu được tạo ra từ một bậc 2). Một ví dụ khác, Hình



Hình 6.5. Tương tự như hình (6.4a) nhưng với $N = 30$.

(6.6a) vẽ sơ đồ $p(\mathcal{D}|\lambda)$ với $\ln \lambda$, cho mô hình hồi quy ridge đa thức, trong đó λ nằm trên cùng một bộ giá trị được sử dụng trong thử nghiệm CV. Chúng ta thấy rằng tối đa hóa bằng chứng xảy ra ở cùng một điểm với mức tối thiểu của thử nghiệm MES, cũng tương ứng với điểm được chọn bởi CV.

Khi sử dụng xấp xỉ Bayesian, chúng ta không bị giới hạn việc đánh giá bằng chứng tại một lưới các giá trị hữu hạn. Thay vào đó, chúng ta có thể sử dụng tối ưu hóa để tìm $\lambda^* = \text{argmax}_\lambda p(\mathcal{D}|\lambda)$. Phương pháp này được gọi là *empirical Bayes* hoặc *loại II maximum likelihood*. Một ví dụ được hiển thị trong hình (6.6b): chúng ta thấy rằng đường cong có hình dạng như ước lượng CV, nhưng nó có thể được tính toán một cách hiệu quả hơn.



(a) (b)

Hình 6.6. Lỗi đào tạo (màu xanh chấm) và lỗi kiểm tra (màu đỏ đặc) cho phù hợp đa thức bậc 14 bằng *hồi quy rigde*, được vẽ so với *ln(lambda)*. Dữ liệu được tạo từ nhiễu với phương sai $sigma^2 = 4$ (tập huấn luyện có kích thước $N = 21$). Lưu ý: Các mô hình được sắp xếp từ phức tạp (small regularizer) ở bên trái đến đơn giản (large Regularizer) ở bên phải. Ước ;lượng hiệu suất sử dụng tập huấn luyện. Chấm màu xanh: Ước lượng xác thực chéo 5-fold của MSE trong tương lai. Màu đen đặc: khả năng cận biên bản ghi, $-\ln p(\mathcal{D}|\lambda)$. Cả hai đường cong đã được thay đổi kích thước theo chiều dọc thành [0,1] để làm cho chúng có thể so sánh được.

6.2.2 Tính toán Likelihood cận biên (bằng chứng)

Khi thảo luận về suy luận tham số cho một mô hình cố định,ta thường viết

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m),$$

đã bỏ qua hằng số chuẩn hóa $p(\mathcal{D}|m)$. Điều này là hợp lệ vì $p(\mathcal{D}|m)$ là hằng số theo tham số $\boldsymbol{\theta}$. Tuy nhiên, khi so sánh các mô hình, cần biết làm thế nào để tính toán *Likelihood cận biên*, $p(\mathcal{D}|m)$. Thông thường, điều này có thể khá khó khăn, vì phải tích hợp tất cả các giá trị tham số có thể, nhưng khi có liên hợp tiên nghiệm, thì rất dễ dàng để tính toán.

Lấy $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z_0$ là tiên nghiệm của chúng ta, với $q(\boldsymbol{\theta})$ là phân phối chưa được chuẩn hóa và Z_0 là hằng số chuẩn hóa của tiên nghiệm. Lấy $p(\mathcal{D}|\boldsymbol{\theta}) = q(\mathcal{D}|\boldsymbol{\theta})/Z_N$ là hậu nghiệm của chúng ta, với $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ là hậu nghiệm chưa chuẩn hóa

và Z_N là hằng số chuẩn hóa của hậu nghiệm. Khi đó,

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{6.2}$$

$$\frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_l Z_0 p(\mathcal{D})} \tag{6.3}$$

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l} \tag{6.4}$$

Vì vậy, giả sử các hằng số chuẩn hóa có liên quan là dễ xử lý (tractable), chúng ta có một cách dễ dàng tính toán *Likelihood cận biên*. Xem ví dụ dưới đây.

Beta-binomial model

Bây giờ, ta sẽ áp dụng kết quả trên cho mô hình Beta-binomial. Theo giả thuyết $p(\theta|a', b')$, với $a' = a + N_1$ và $b' = b + N_0$, với hằng số chuẩn hóa của hậu nghiệm là $B(a', b')$. Do đó,

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \frac{1}{p(\mathcal{D})}\Big[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\Big]\Big[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\Big] \\ &= \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}\big[\theta^{a+N_1-1}(1-\theta)^{b+N_0-1}\big] \end{aligned}$$

Vì vậy

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(s+N_1, b+N_0)}{B(a,b)}$$

Tương tự, *Likelihood cận biên* của mô hình *Beta-Bernoulli* tương tự như trên, ngoại trừ việc thiếu thuật ngữ $\binom{N}{N_1}$.

6.2.3 Xấp xỉ BIC cho logarit Likelihood cận biên

Nhìn chung, việc tính toán tích phân trong phương trình (6.1) có thể khá khó khăn. Một xấp xỉ đơn giản nhưng phổ biến được gọi là *Bayesian information criterion* hoặc BIC, có dạng sau:

$$BIC = \ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{dof(\hat{\boldsymbol{\theta}})}{2} \ln N \approx -2 \ln p(\mathcal{D}),$$

với dof($\hat{\boldsymbol{\theta}}$) là số bậc tự do trong mô hình , và $\hat{\boldsymbol{\theta}}$ là ước lượng MLE cho mô hình. Theo truyền thống, điểm BIC (BIC score) được xác định bằng cách sử dụng ước lượng MLE cho $\hat{\boldsymbol{\theta}}$, vì vậy nó độc lập với tiên nghiệm. Tuy nhiên, đối với các mô hình như mô hình hỗn hợp Gaussian, ước tính MLE có thể hoạt động kém, vì vậy tốt hơn là đánh giá điểm BIC bằng cách sử dụng ước lượng MAP. Chúng ta thấy rằng điều này có dạng logarit likelihood bị phạt (penalized), trong đó thuật ngữ phạt phụ thuộc vào độ phức tạp của mô hình.
Ví dụ, xét mô hình hồi quy tuyến tính. Ta thấy rằng, MLE được cho bởi $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ và $\hat{\sigma}^2 = RSS/N$, với $RSS = \sum_{i=1}^{N} (y_i - \hat{\mathbf{w}}_{mle}^\top \mathbf{x}_i)^2$. Ở mức tối đa có một dạng đặc biệt đơn giản cho logarit likelihood đã được các tài liệu chứng minh có dạng như sau: (xem thêm trên Wiki)

$$\ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}) = -\frac{N}{2} \ln(2\pi\hat{\sigma}^2) - \frac{N}{2}.$$

Do đó, điểm BIC như sau (bỏ qua các thuật ngữ về hằng số)

$$BIC = -\frac{N}{2} \ln(\hat{\sigma}^2) - \frac{d}{2} \ln(N),$$

với d là số biến ngẫu nhiên trong mô hình. Trong tài liệu thống kê, người ta thường sử dụng *định nghĩa* thay thế của BIC, mà chúng ta gọi là chi phí BIC (BIC cost) (vì chúng ta muốn giảm thiểu nó):

$$BIC - cost = -2 \ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}) + dof(\hat{\boldsymbol{\theta}}) \ln N \approx p(\mathcal{D}).$$

Trong trường hợp hồi quy tuyến tính, điều này trở thành

$$\begin{aligned} BIC - cost &= -2(-\frac{N}{2}\ln(2\pi\hat{\sigma}^2) - \frac{N}{2}) + D\ln N, \\ &= N\ln(2\pi\hat{\sigma}^2) - N + D\ln N, \\ &= N\ln(\hat{\sigma}^2) + D\ln(N). \end{aligned}$$

6.3 Bayes factors

Giả sử, tiên nghiệm trên các mô hình là **uniform**, $p(m) \propto 1$. Khi đó, lựa chọn mô hình tương đương với việc chọn mô hình có marginal likelihood cao nhất. Bây giờ, giả sử ta có 2 mô hình mà chúng ta đang xem xét, được gọi là giả thuyết không-$\mathcal{M}_0$ và đối thuyết $\mathcal{M}_1$.

Định nghĩa nhân tố Bayes chính là tỷ lệ của likelihood marginal:

$$BF_{1,2} = \frac{Pr(\mathcal{M}_1|\mathcal{D})}{Pr(\mathcal{M}_2|\mathcal{D})} = \frac{Pr(\mathcal{M}_1)p(\mathcal{D}|\mathcal{M}_1)}{Pr(\mathcal{M}_2)p(\mathcal{D}|\mathcal{M}_2)} = \frac{Pr(\mathcal{M}_1)\int p(\mathcal{D}|\theta_1,\mathcal{M}_1)p(\theta_1|\mathcal{M}_1)d\theta_1}{Pr(\mathcal{M}_2)\int p(\mathcal{D}|\theta_2,\mathcal{M}_2)p(\theta_2|\mathcal{M}_2)d\theta_2},$$

(Đây giống như tỷ lệ likehood, ngoại trừ chúng ta tích hợp các tham số, cho phép so sánh các mô hình có độ phức tạp khác nhau.) Nếu $BF_{1,0} > 1$ thì chúng ta thích mô hình 1, còn nếu không thì mô hình 2 sẽ đưa ưa thích hơn.
Tất nhiên, có thể là $BF_{1,0}$ chỉ lớn hơn một chút so với 1. Trong trường hợp đó, chúng ta sẽ không tự tin rằng mô hình 1 sẽ tốt hơn. Jeffreys (1961) đã đề xuất thang đo bằng chứng để diễn giải mức độ của yếu tố Bayes, được thể hiện trong Bảng (6.3). Đây là một thay thế Bayes cho khái niệm thường xuyên (frequentist concept) về p-giá trị. Ngoài ra, chúng ta chỉ có thể chuyển đổi nhân tố Bayes thành một hậu thế trên các mô hình. Nếu $p(\mathcal{M}_1 = P(\mathcal{M}_2) = 0.5$ chúng ta có:

$$p(\mathcal{M}_2|\mathcal{D}) = \frac{BF_{2,1}}{1 + BF_{2,1}} = \frac{1}{BF_{2,1} + 1},$$

Bayes factor BF(1, 2)	Giải thích
$BF < \frac{1}{100}$	Bằng chứng quyết định cho $\mathcal{M}_2$
$BF < \frac{1}{10}$	Bằng chứng mạnh mẽ cho $\mathcal{M}_2$
$\frac{1}{10} < BF < \frac{1}{3}$	Bằng chứng vừa phải cho $\mathcal{M}_2$
$\frac{1}{3} < BF < 1$	Bằng chứng yếu $\mathcal{M}_2$
$1 < BF < 3$	Bằng chứng yếu cho $\mathcal{M}_1$
$3 < BF < 10$	Bằng chứng vừa phải cho $\mathcal{M}_1$
$BF > 10$	Bằng chứng mạnh mẽ cho $\mathcal{M}_1$
$BF > 100$	Bằng chứng quyết định cho $\mathcal{M}_1$

6.3.1 Ví dụ: Testing if a coin is fair

Giả sử có một đồng xu và ta muốn so sánh hai mô hình nhằm giải thích xu hướng nhận mặt sấp hay ngửa của đồng xu. Trong mô hình đầu tiên $\mathcal{M}_1$, ta giả sử rằng xác suất nhận mặt ngửa là cố định và bằng $\frac{1}{2}$. Chú ý rằng, mô hình này không có bất kì tham số nào. Mô hình thứ hai $\mathcal{M}_2$, giả sử rằng xác suất mặt ngửa là cố định và có giá trị là $\theta \in (0,1)$, với tiên nghiệm đều trên θ: $p(\theta|\mathcal{M}_2) = 1$ (điều này tương đương với tiên nghiệm beta trên θ với $\alpha = \beta = 1$). Để đơn giản, ta chọn mô hình tiên nghiệm đều: $Pr(\mathcal{M}_1) = Pr(\mathcal{M}_2) = \frac{1}{2}$.

Gọi x là số lần xuất hiện mặt ngửa. Bây giờ ta thực hiện tung đồng xu 200 lần và quan sát thấy số lần xảy ra mặt ngửa là $x = 115$ và số lần xảy ra mặt sấp là $n-x = 85$ lần. Câu hỏi đặt ra bây giờ là: Chúng ta nên chọn mô hình nào để có thể phù hợp với dữ liệu này? Bây giờ, chúng ta tính model evidence $p(x|n, \mathcal{M}_1), p(x|n, \mathcal{M}_2)$ cho các mô hình. Model evidence cho mô hình $\mathcal{M}_1$ khá là đơn giản, vì nó không có tham số:

$$Pr(x|n, \mathcal{M}_1) = Binomial(n, x, \frac{1}{2}) = \binom{200}{115} \frac{1}{2^{115}} \frac{1}{2^{85}} = \binom{200}{115} \frac{1}{2^{200}} \approx 0.005956.$$

Model evidence trên mô hình $\mathcal{M}_2$ với tham số θ được tính như sau:

$$\begin{aligned} Pr(x|n,\mathcal{M}_2) &= \int Pr(x|n,\theta,\mathcal{M}_2)p(\theta|\mathcal{M}_2)d\theta \\ &= \int_0^1 \binom{200}{115} \theta^{115}(1-\theta)^{200-115}d\theta \qquad \theta \in (0,1) \\ &= \frac{1}{201} \\ &\approx 0.004975. \end{aligned}$$

Nhân tố Bayes in favor of $\mathcal{M}_1$ xấp xỉ bằng $BF_{0,1} = \frac{0.005956}{0.004975} = 1.197$, vì vậy dữ liệu đưa ra bằng chứng rất yếu ủng hộ (in favor of) mô hình đơn giản hơn là mô hình $\mathcal{M}_1$. Hay nói cách khác, tỷ lệ này là 1.197 ..., "hầu như không đáng nhắc đến" ngay cả khi nó chỉ hơi hướng về phía $\mathcal{M}_1$.
Quay lại kiểm định giả thuyế theo phương pháp truyền thống - phương pháp tần số sẽ tạo ra một kết quả khác. Cụ thể ta sẽ tiến hành kiểm định tỷ lệ, với các giả thuyết sau:

Kiểm định trên nói rằng, giả thuyết không: $\theta = \frac{1}{2}$ sẽ bị bác bỏ với mức ý nghĩa $\alpha = 5\%$. Xác suất tạo ra ít nhất 115 mặt ngửa trên mô hình $\mathcal{M}_1$ là khoảng 0.02 (một cách tương tự, xác suất phát sinh ít nhất 115 lần xuất hiện mặt sấp là 0.02). Vì vậy, kiểm định hai phía trên sẽ cho p giá trị xấp xỉ 4%.

6.4 Uninformative priors

Như đã biết, ta có Binomial Likelihood $\propto p^x(1-p)^x$ với x là số lần thành công trong n lần thử.
Phân phối Beta prior với α, β là tham số hình dạng:

$$\text{Beta prior} \propto p^{\alpha-1}(1-p)^{\beta-1},$$

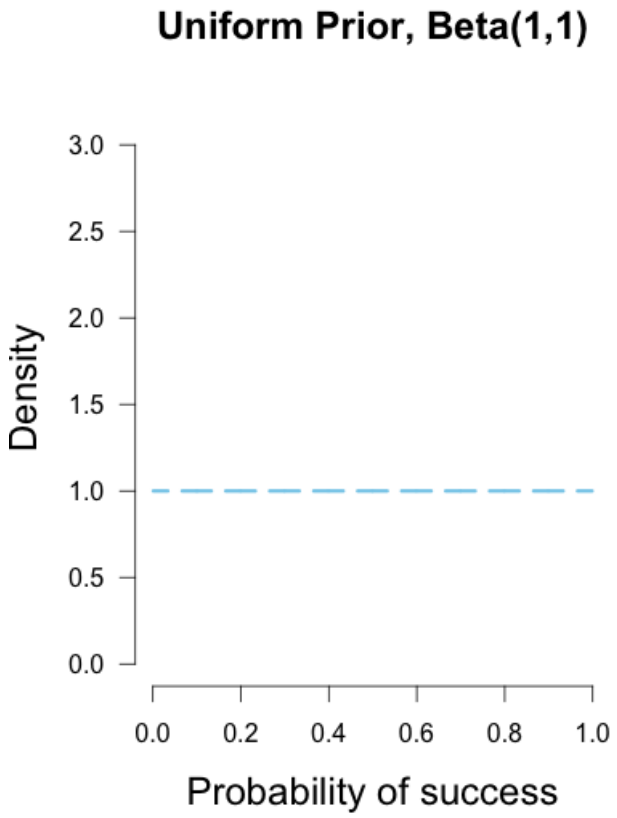
Khi đó phân phối hậu nghiệm có dạng:

$$\begin{aligned}\text{Beta posterior} &\propto p^x(1-p)^{n-x}p^{\alpha-1}(1-p)^{\alpha-1}\\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}.\end{aligned}$$

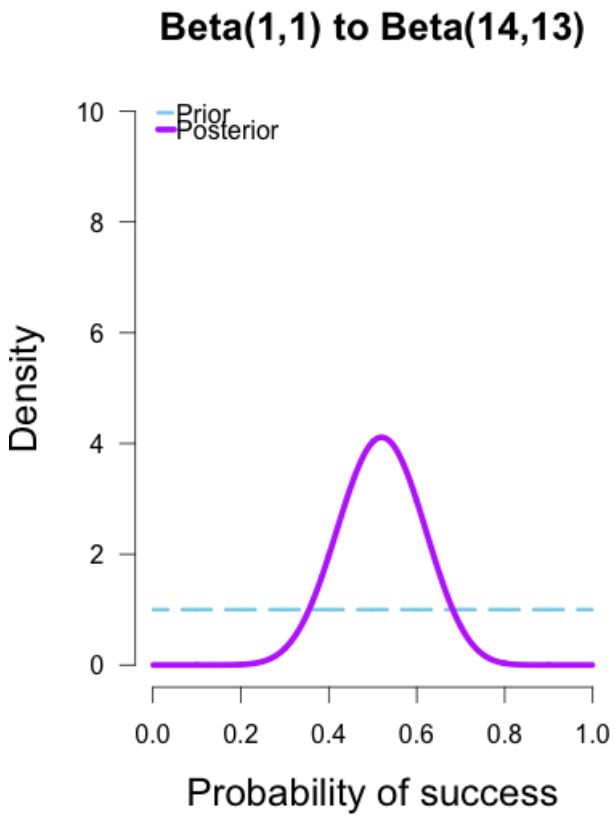
Ta nhận được tiên nghiệm, thêm vào số lần thành công và thất bại với số mũ khác nhau. Nói cách khác, ta nhận được tiên nghiệm Beta(α, β) với số lần thành công là $x+\alpha$ và thất bại là $n-x+\beta$. Khi đó, hậu nghiệm nhận được là $B(\alpha+x, \beta+n-x)$. Nếu $\alpha = \beta = 1$ thì prior the posterior sẽ có hình dạng chính xác của Binomial likelihood, hậu nghiệm lúc này được viết là $Beta(1+x, 1+n-x)$.

Đồ thị

Nếu chúng ta bắt đầu với Uniform prior, Beta(1,1), nó sẽ trong như thế này:



Nếu bạn nhận được 13 thành công trong 25 lần thử thì hậu nghiệm mới là $Beta(1+13, 1+12)$ hoặc $Beta(14, 13)$, đồ thị bên dưới:



6.5 Mô hình chọn cho hồi quy tuyến tính Bayes

Một ứng dụng phổ biến cho mô hình chọn là chọn giữa những hàm đặc trưng mở rộng $\phi(\mathbf{x})$ trong hồi quy tuyến tính Bayes. Ở đây mô hình $\mathcal{M}_i$ có thể ví dụ tương ứng với hồi quy đa thức bậc i với

$$\phi_i(\mathbf{x}) = [1, \mathbf{x}, \mathbf{x}^2, ..., \mathbf{x}^i]^\top.$$

Sau khi chọn một tập hợp các mô hình để so sánh, cũng như xác suất tiền nghiệm cho mỗi mô hình, nhiệm vụ duy nhất còn lại là tính toán bằng chứng cho từng mô hình trong dữ liệu được quan sát $(\mathbf{X}, \mathbf{y})$. Trong thảo luận của chúng ta về hồi quy tuyến tính Bayesian, chúng ta đã thực sự tính toán được lượng mong muốn:

$$p(\mathbf{y}|\mathbf{X}, \sigma^2, \mathcal{M}_i) = \mathcal{N}(\mathbf{y}; \phi_i(\mathbf{X})\mu, \phi_\mathbf{i}(\mathbf{X})\mathbf{\Sigma}\phi_\mathbf{i}(\mathbf{X})^\top + \sigma^\mathbf{2}\mathbf{I}),$$

với ϕ_i là một mở rộng cơ bản.
Chú ý rằng mô hình ϕ_i cũng có thể giải thích một cách rõ ràng tất cả các bộ dữ liệu được giải thích tốt (well-explained)d bởi những mô hình ϕ_j với $j < i$, bằng một thiết lập đơn giản là đặt trọng số trên bậc cao hơn và có giới hạn về 0. Tuy nhiên, ta nhớ lại rằng một mô hình đơn giản hơn sẽ được ưu tiên bởi Occam's razor được mô tả ở trên.

6.6 Bayesian Model Averaging

Chú ý rằng một cách xử lý của "full Bayesian" về một vấn đề sẽ tránh hoàn toàn việc lựa chọn mô hình. Thay vào đó, khi đưa ra dự đoán, về mặt lý thuyết, chúng ta nên sử dụng luật tổng để marginalize mô hình chưa biết, ví dụ:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \sum_i p(y_*|\mathbf{x}_*, \mathcal{D}, \mathcal{M}_i) Pr(\mathcal{M}_i|\mathcal{D}).$$

Cách tiếp cận nhu vậy được gọi là *mô hình trung bình Bayesian*. Mặc dù, điều nay đôi khi đã được nhìn thấy, việc lựa chọn mô hình vẫn được dùng rộng rãi trong thực tế. Lý do là tổng chi phí tính toán của việc sử dụng mô hình đơn giản là thấp hơn nhiều so vơi việc phải liên tục đào tạo lại nhiều mô hình, và mô hình trung bình Bayesian sử dụng phân phối hỗn hợp để thự hiện các dự đoán, cái mà có thể có các đặc tính gây nhiễu (chẳng hạn, phân phối dữ đoán có thể là nhiều mốt (multimodal)).

Chương 7

Hồi quy logistic

Cho đến bây giờ ta vẫn làm việc với sự liên hợp của likelihood và phân phối tiên nghiệm, chúng cho phép tính toán phân phối hậu nghiệm ở dạng đóng. Thật không may, sẽ có nhiều tính huống không xảy ra, buộc chúng ra phải xấp xỉ phân phối hậu nghiệm và số lượng liên quan (chẳng hạn như *mô hình bằng chứng* hoặc kỳ vọng theo phân phối sau). Hồi quy logistic là phương pháp hồi quy thông thường cho phân lớp nhị phân, và việc cố gắng dùng xấp xỉ Bayesian một cách trực tiếp sẽ không thể thực hiện được.

7.1 Khai triển Taylor cho hàm nhiều biến

7.1.1 Ma trận đạo hàm riêng

Xét hàm có giá trị vô hướng, chẳng hạn như $f(x, y)$ hay $f(x, y, z)$... Khi đó, chúng ta có thể xem đạo hàm riêng như the rates of increase của hàm trong hệ trục tọa độ. Nếu hàm là khả vi thì đạo hàm riêng là một ma trận hàng thu được từ tất cả những đạo hàm riêng theo từng biến. Do đó, ma trận này được gọi là ma trận đạo hàm riêng (hay ma trận Jacobian).
Cho

$$f(\mathbf{x}) : \mathbb{R}^{\mathbf{n}} \to \mathbb{R},$$

Khi đó, ma trận đạo hàm riêng tại $\mathbf{x} = \mathbf{a}$ là:

$$\nabla f(\boldsymbol{a}) = \left[\frac{\partial f}{\partial x_1}(\boldsymbol{a})\frac{\partial f}{\partial x_2}(\boldsymbol{a})\dots\frac{\partial f}{\partial x_n}(\boldsymbol{a})\right].$$

Xét hàm tổng quát có giá trị là véc tơ $\mathbf{f}:\mathbb{R}^n \to \mathbb{R}^m$. Ở đây $\mathbf{f}(\mathbf{x})$ là hàm của véc tơ $\mathbf{x} = (x_1, x_2, ..., x_n)$ với đầu ra là một véc tơ m thành phần. Chúng ta có thể viết f như sau:

$$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_m(\mathbf{x})) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}$$

Để tìm ma trận đạo hàm riêng, chúng ta xem $\mathbf{f}(\mathbf{x})$ như một ma trận cột, trong đó mỗi thành phần có giá trị vô hướng. Ma trận đạo hàm riêng của từng thành phần $f_i(\mathbf{x})$ sẽ là một ma trận $1 \times n$. Chúng ta chỉ cần xếp các ma trận này lên nhau để tạo thành một ma trận lớn hơn. Khi đó, ma trận đạo hàm riêng tại $\mathbf{x} = \mathbf{a}$ là ma trận $m \times n$:

$$\nabla\mathbf{f}(\mathbf{a}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \frac{\partial f_1}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{a}) & \frac{\partial f_2}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \frac{\partial f_m}{\partial x_2}(\mathbf{a}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{bmatrix}.$$

7.1.2 Khai triển Taylor cho hàm nhiều biến

Nhớ lại, ta có khai triển Taylor của hàm f xung quanh điểm $x = a$ trong trường hợp một chiều có dạng:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \frac{1}{6}f'''(a)(x-a)^3 + \cdots.$$

Trong trường hợp nhiều chiều, ta sẽ tổng quát hóa khai triển đa thức Taylor cho hàm nhiều biến:

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$$

Như chúng ta đã biết, xấp xỉ tuyến tính tốt nhất của f có dạng như sau:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

với $\nabla f(\mathbf{a})$ là ma trận đạo hàm riêng. Xấp xỉ tuyến tính là đa thức Taylor bậc 1. Khi $f(\mathbf{x})$ là một vô hướng, đạo hàm bậc một là $\nabla f(\mathbf{x})$, là một ma trận $1 \times n$. Khi đó, đạo hàm bậc 2 của $f(\mathbf{x})$ là một ma trận của ma trận đạo hàm riêng của hàm $\nabla f(\mathbf{x})$ và có thể viết là $\nabla\nabla f(\mathbf{x})$. Ma trận phát sinh này là ma trận $n \times n$ và được gọi là ma trận Hessian của f. Kí hiệu là:

$$\mathbf{H} = -\nabla\nabla f(\mathbf{x}) \tag{7.1}$$

Khi f là hàm nhiều biến, thuật ngữ phá sinh thứ 2 trong chuỗi Taylor sẽ sử dụng Hessian $\mathbf{H}f(\mathbf{a})$. Khi đó biểu thức bậc 2 sẽ là:

$$-\frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H}(\mathbf{x} - \mathbf{a})$$

Vậy đa thức Taylor bậc 2 cho trường hợp nhiều biến xung quanh điểm $\mathbf{x} = \mathbf{a}$ là:

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a}) - \frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H}(\mathbf{x} - \mathbf{a})$$

Đa thức Taylor bậc 2 của $f(\boldsymbol{x})$ xung quanh điểm $\mathbf{x} = \mathbf{a}$ là xấp xỉ tốt hơn xấp xỉ tuyến tính (hay xấp xỉ bậc 1). Chúng ta có thể dùng nó cho các vấn đề tìm cực tiểu hay cực đại địa phương của hàm bất kì.

7.2 Hồi quy Logistic

Trong hồi quy tuyến tính, ta quan tâm đến các hàm mang giá trị thực.

$$y(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R},$$

với $\mathbf{x}$ là véc tơ chứa giá trị đầu vào d-chiều. Ở đây, ta sẽ xét thiết lập tương tự, nhưng với một chút thay đổi: chúng ta hạn chế giá trị đầu ra của hàm y, tức là chỉ nhận không gian giá trị $y \in \{0, 1\}$. Trong máy học, các vấn đề của dạng này thuộc

loại phân lớp nhị phân: cho một giá trị đầu vào $\mathbf{x}$ và chúng ta muốn phân lớp nó thành một trong hai loại, trong trường hợp này là lớp 0 hoặc 1.

Ở đây lại giả sử rằng ta nhận được một vài quan sát của ánh xạ, $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, để phục vụ cho quá trình huấn luyện dữ liệu. Cho ví dụ sau, mục đích của phân lớp nhị phân là có thể dự đoán nhãn tại một vị trí đầu vào mới là $\mathbf{x}_*$.
Cũng như hồi quy tuyến tính, vấn đề này vẫn chưa được đặt ra đúng đắn mà không có một số giới hạn trên y. Trong hồi quy tuyến tính ta đã giả sử rằng mối quan hệ giữa $\mathbf{x}$ và y là như sau:

$$y(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + \epsilon(\mathbf{x}),$$

với $\mathbf{w} \in \mathbb{R}^d$ là bộ các véc tơ tham số, và $\epsilon(\mathbf{x})$ là sai số. Giả định này không được mong muốn trong trường hợp phân lớp, nơi mà các giá trị đầu ra bị hạn chế trong $\{0, 1\}$ (chú ý rằng, ví dụ, $\mathbf{x}^\top \mathbf{w}$ không bị ràng buộc khi chuẩn của $\mathbf{x}$ tăng lên, bắt buộc phần dư phải tăng lớn hơn.)

Trong các phương pháp phân lớp tuyến tính, thay vì chúng ta giả sử rằng xác suất có điều kiện của phân lớp thuộc về lớp "1" được cho bởi biến đổi phi tuyến của hàm tuyến tính cơ bản theo $\mathbf{x}$:

$$Pr(y = 1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{x}^\top \mathbf{w}),$$

với σ được gọi là hàm "Sigmoid" (có hình dạng giống hình chữ "S") ánh xạ hàm tăng theo xác suất trong khoảng (0,1).

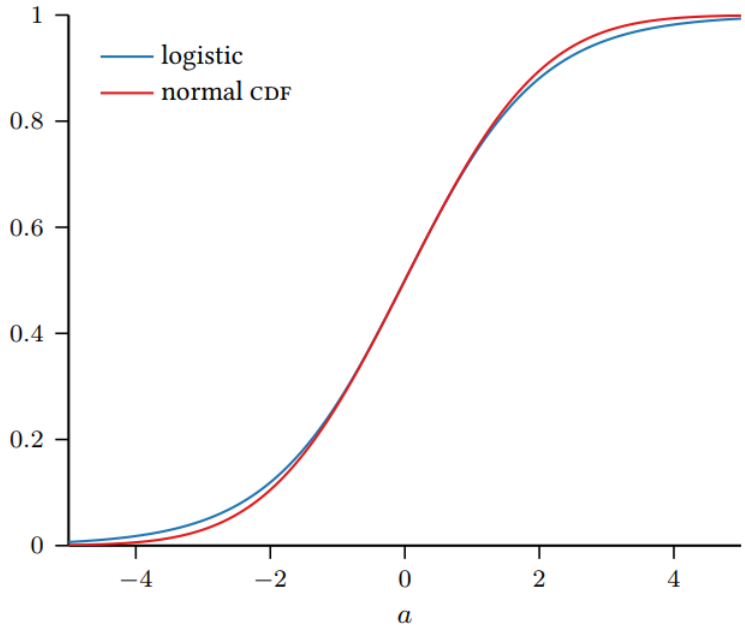
Chúng ta thường thêm một tham số chệch b vào mô hình, để tạo ra xác suất $\sigma(\mathbf{x}^T \mathbf{w} + b)$. Mặc dù, tham số chệch này thường được loại bỏ trong trình bày, để giảm sự hỗn loạn. Chúng ta luôn có thể tìm ra cách để thêm lại một tham số chệch vào mô hình bằng một hằng số vào đặc trưng đầu vào $\mathbf{x}$.

Các hàm σ được sử dụng phổ biến nhất là hàm logistic:

$$\sigma(a) = \frac{exp(a)}{1 + exp(a)},$$

hoặc hàm phân phối tích lũy của phân phối chuẩn hóa:

$$\sigma(a) = \Phi(a) = \int_{-\infty}^{a} \mathcal{N}(0, 1^2)dx,$$



Hình 7.1. Một so sánh của hai chức năng sigmoid được mô tả trong hình. Đường cong CDF của phân phối chuẩn trong ví dụ này sử dụng phép biến đổi $\Psi(\sqrt{\frac{\pi}{8}}a)$, đảm bảo độ dốc của hai đường cong bằng nhau tại điểm gốc. Ở đây $\lambda = \frac{\pi}{8}$, được chọn sao cho đạo hàm của hai đường cong khớp với nhau tại $x = 0$.

Có hai sự lựa chọn được so sánh trong hình (7.1). Sự khác biết định tính chính là hàm logistic có đuôi nặng hơn so với hàm CDF của phân phối chuẩn. Phân lớp logistic dùng trong hàm logistic được gọi là hồi quy logistic; phân lớp tuyến tính dùng trong CDF của phân phối chuẩn được gọi là hồi quy probit. Hồi quy logistic bắt gặp trong thực tế. Chú ý rằng, giả định tuyến tính trên được kết hợp với hàm

ẩn logistic sigmoid để chỉ logg odds là hàm tuyến tính của đầu vào $\mathbf{x}$.

$$\begin{aligned}
\ln \frac{Pr(y=1|\mathbf{x},\mathbf{w})}{Pr(y=0|\mathbf{x},\mathbf{w})} &= \ln \frac{\sigma(\mathbf{x}^\top \mathbf{w})}{1-\sigma(\mathbf{x}^\top \mathbf{w})}, \\
&= \ln \sigma(\mathbf{x}^\top \mathbf{w}) - \ln(1-\sigma(\mathbf{x}^\top \mathbf{w})), \\
&= \ln \frac{\exp(\mathbf{x}^\top \mathbf{w})}{1+\exp(\mathbf{x}^\top \mathbf{w})} - \ln\left(1 - \frac{\exp(\mathbf{x}^\top \mathbf{w})}{1+\exp(\mathbf{x}^\top \mathbf{w})}\right), \\
&= \ln \frac{\exp(\mathbf{x}^\top \mathbf{w})}{1+\exp(\mathbf{x}^\top \mathbf{w})} - \ln\left(\frac{1}{1+\exp(\mathbf{x}^\top \mathbf{w})}\right), \\
&= \ln \exp(\mathbf{x}^\top \mathbf{w}), \\
&= \mathbf{x}^\top \mathbf{w}.
\end{aligned}$$

Đối với dữ liệu $\mathbf{x}$ chưa từng được thấy trước đó và tập hợp các hệ số $\mathbf{w}^*$ được tìm thấy từ phương trình ước lượng hồi quy truyến tính. Khi đó, ta dễ dàng tìm được xác suất hậu nghiệm:

$$p(y=1|\mathbf{x},\mathbf{w}^*) = \frac{\exp(\mathbf{x}^\top \mathbf{w}^*)}{1+\exp(\mathbf{x}^\top \mathbf{w}^*)}$$

Nếu $p(y=1|\mathbf{x},\mathbf{w}^*) \geq 0.5$ thì chúng ta kết luận rằng dữ liệu tại điểm $\mathbf{x}$ nên được gán nhãn dương ($\hat{y}=1$). Ngược lại, gán nhãn âm ($\hat{y}=0$). Do đó, giá trị dự đoán cho bộ vector $\mathbf{x} = (x_0 = 1, x_1, ..., x_d)$ là 0 hoặc 1. Chú ý rằng, $p(y=1|\mathbf{x},\mathbf{w}^*) \geq 0.5$ khi $\mathbf{x}^\top \mathbf{w} \geq \mathbf{0}$. Biểu thức, $\mathbf{x}^\top \mathbf{w} = \mathbf{0}$ đại diện cho một siêu phẳng phân tách các lớp, chẳng hạn, lớp âm và dương. Do đó, mô hình hồi quy Logistic là một mô hình phân lớp tuyến tính.

Quay lại với ước lượng tham số, chúng ta đã giả sử tập dữ liệu đang xét có dạng $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ là $i.i.d$ và mẫu được lấy từ một phân phối cố định nhưng không xác định $p(\mathbf{x}, y)$. Cụ thể hơn nữa, chúng tôi sẽ giả định rằng quá trình tạo dữ liệu sẽ rút ngẫu nhiên một điểm dữ liệu, nhận được vectơ ngẫu nhiên $(x_0 = 1, x_1, ..., x_{d-1})$,

phụ thuộc vào $p(\mathbf{x})$ và sau đó, đặt nhãn phân lớp y theo phân phối Bernoulli:

$$p(y|x) = \begin{cases} \left(\frac{\exp(\mathbf{x}^\top \mathbf{w})}{1+\exp(\mathbf{x}^\top \mathbf{w})}\right)^y, & \text{khi } y = 1 \\ \left(1 - \frac{\exp(\mathbf{x}^T \mathbf{w})}{1+\exp(\mathbf{x}^\top \mathbf{w})}\right)^{1-y}, & \text{khi } y = 0 \end{cases}$$
$$= \sigma(\mathbf{x}^\top \mathbf{w})^y (1 - \sigma(\mathbf{x}^T \mathbf{w}))^{1-y}$$

với $\mathbf{w} = (w_0, w_1, ..., w_{d-1})$ là tập các hệ số chưa biết mà ta muốn tìm thông qua bộ dữ liệu đã quan quan sát $\mathcal{D}$. Thông qua những nguyên tắc ước lượng, chúng ta có thể ước lượng $\mathbf{w}$ bằng cách tối đa hóa hàm điều kiện Likelihood của các nhãn được quan sát $\mathbf{y} = (y_1, y_2, ..., y_n)$ được cho bởi dữ liệu đầu vào $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_n^\top)$.
Cũng như ước lượng hợp lý cực đại (MLE) đã được đề cập trước đó, đầu tiên chúng ta sẽ viết hàm điều kiện cho *Likelihood* $p(\mathbf{y}|\mathbf{X},\mathbf{w})$:

$$Pr(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w})$$
$$= \prod_{i=1}^{N} \sigma(\mathbf{x}^\top \mathbf{w})^{y_i} (1 - \sigma(\mathbf{x}^\top \mathbf{w}))^{1-y_i}$$

Để kiểm tra phương trình này, chú ý rằng mỗi y_i sẽ nhận giá trị 0 hoặc 1, vì vậy chính xác một trong những y_i hay $1 - y_i$ sẽ khác không, sẽ được chọn ra chính xác cho likelihod.
Cách tiếp cận truyền thống của hồi quy logistic là tối đa hóa hàm likelihood của dữ liệu huấn luyện như một hàm theo tham số $\mathbf{w}$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{X}, \mathbf{w});$$

với $\hat{\mathbf{w}}$ là một ước lượng hợp lý cực đại (MLE). Không giống như hồi quy tuyến tính, ở đây có một biểu thức dạng đóng cho ước lượng hợp lý cực đại, không có giải pháp nào cho hồi quy logistic. Mặc dù vậy, mọi thứ lại quá tệ vì hóa ra đối với hồi quy logistic, *logarit likelihood* âm là lồi và xác định dương, có nghĩa là một mức tối thiểu toàn cục duy nhất (và do đó MLE là duy nhất).

7.3 Tính chất của hồi quy Logistic

7.3.1 Hồi quy Logistic thực ra được sử dụng nhiều trong các bài toán phân loại

.

Mặc dù có tên là "*Hồi quy*", tức một mô hình cho *fitting*, Hồi quy Logistic lại được sử dụng nhiều trong các bài toán phân loại. Sau khi tìm được mô hình, việc xác định lớp y cho một điểm dữ liệu $\mathbf{x}$ được xác định bằng việc so sánh hai biểu thức xác suất:

$$p(y = 1|\mathbf{x}; \mathbf{w}); \qquad p(y = 0|\mathbf{x}; \mathbf{w})$$

Nếu biểu thức thứ nhất lớn hơn thì ta kết luận điểm dữ liệu thuộc phân lớp 1, ngược lại thì nó thuộc phân lớp 0. Vì tổng hai biểu thức này luôn bằng 1 nên một cách gọn hơn, ta chỉ cần xác định xem $p(y = 1|\mathbf{x}; \mathbf{w})$ lớn hơn 0.5 hay không. Nếu có, phân lớp 1. Nếu không, là phân lớp 0.

7.3.2 Biên tạo bởi hồi quy logistic có dạng tuyến tính

Thật vậy, chúng ta cần kiểm tra:

$$\begin{aligned} P(y = 1|\mathbf{x}; \mathbf{w}) &> 0.5 && (7.2)\\ \Leftrightarrow \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} &> 0.5 && (7.3)\\ \Leftrightarrow e^{-\mathbf{w}^\top \mathbf{x}} &< 1 && (7.4)\\ \Leftrightarrow \mathbf{w}^\top \mathbf{x} &> 0 && (7.5) \end{aligned}$$

Nói cách khác, biên giữa hai lớp là đường có phương trình $\mathbf{w}^\top \mathbf{x}$. Đây chính là phương trình của một siêu mặt phẳng. Vậy hồi quy logistic tạo ra biên có dạng tuyến tính.

7.4 Bayesian logistic regression

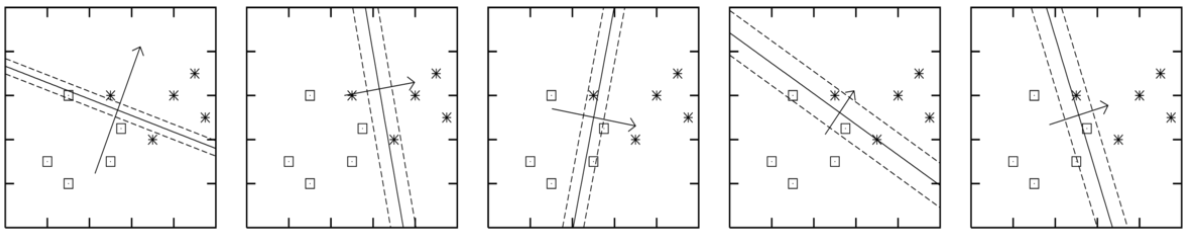
Xấp xỉ Bayesian trong hồi quy logistic yêu cầu chúng ta chọn phân phối tiên nghiệm cho các tham số $\mathbf{w}$ và nhận phân phối hậu nghiệm theo luật Bayes có dạng như sau:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$$

Với $p(\mathcal{D})$ là hằng số chuẩn hóa và được tính bằng tích phân để làm cho tích phân của phân phối hậu nghiệm bằng 1:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathbf{dw}.$$

Hình bên dưới đại diện cho năm bộ các tham số hợp lí khác nhau, được lấy mẫu từ phân phối hậu nghiệm $p(\mathbf{w}|\mathcal{D})$. Mỗi hình cho thấy ranh giới quyết định $\sigma(\mathbf{w}^T\mathbf{x}) = \mathbf{0.5}$ cho một vector tham số là một đường cứng, và hai đường coutour được cho bởi $\mathbf{w}^T\mathbf{x} = \mathbf{1}$ và $\mathbf{w}^T\mathbf{x} = \mathbf{0}$.



Hình 7.2. Mẫu thu được với phương pháp Langevin Monte Carlo.

Các trục trong hình trên đại diện cho hai đặt tính đầu vào x_1 và x_2. Mô hình bao gồm một tham số chệch, và các tham số trong mô hình được lấy mẫu từ phân phối hậu nghiệm được cho bởi dữ liệu từ hai lớp như hình minh họa ở trên. Đường mũi tên, vuông góc với ranh giới quyết định, nó minh họa hướng và độ lớn của vector trọng số.

Giả sử rằng dữ liệu của chúng ta được mô hình hóa tốt bằng hồi quy logistic, chính xác hơn là chúng ta không biết chính xác các tham số ở đây là gì. Có nghĩa là chúng ta sẽ không biết các tham số nào mà chúng ta sẽ phù hợp khi chúng ta làm việc với nhiều dữ liệu hơn. Các dự đoán đưa ra các vector trọng số hợp lí khác nhau đáng kể.

Dự đoán Bayesian cho một nhãn y là $p(y|\mathbf{x}, \mathcal{D})$, chính là xác suất nhận được từ bộ dữ liệu có sẵn $\mathcal{D}$ và vị trí thử nghiệm $\mathbf{x}$. Các trọng số $\mathbf{w}$ không xuất hiện trong biểu thức này vì chúng chưa được biết đến. Chúng ta có thể đưa các trọng số của chúng ta vào công thức trên bằng cách biến chúng thành các "biến số giả" và sau đó, sử dụng các lý thuyết xác suất để đưa ra công thức dự đoán:

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y, \mathbf{w}|\mathbf{x}, \mathcal{D}d\mathbf{w}), \qquad \text{luật tổng}$$

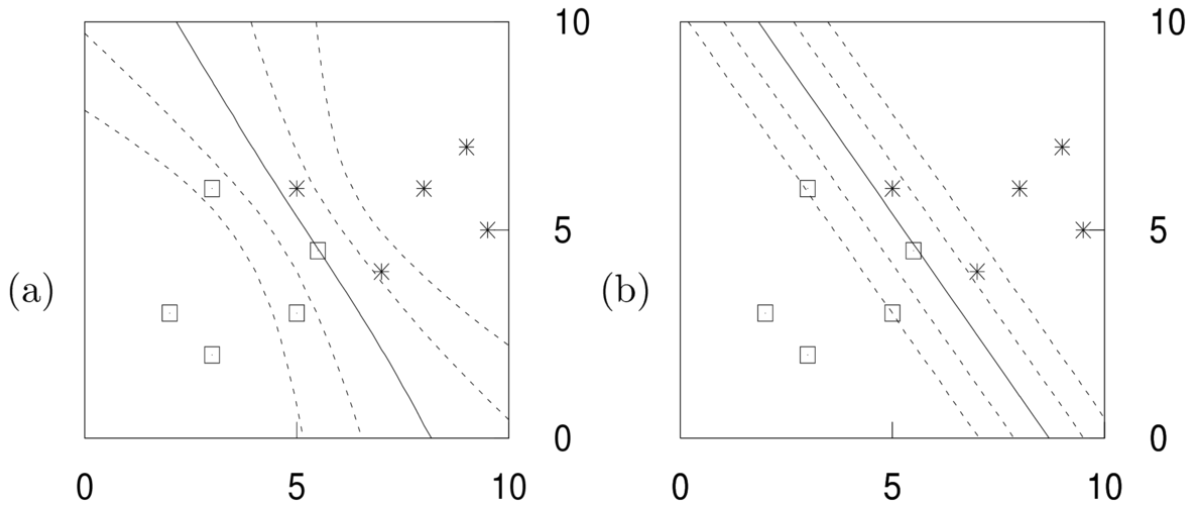
Chúng ta có thể chia biểu thức này thành các thuật ngữ mà chúng ta đã biết bằng cách sử dụng luật xác suất chính khác:

$$\begin{aligned} p(y|\mathbf{x}, \mathcal{D}) &= \int p(y|\mathbf{w}, \mathbf{x}, \mathcal{D}) p(\mathbf{w}|\mathbf{x}, \mathcal{D}) d\mathbf{w}, \qquad \text{luật tích} \\ &= \int p(y|\mathbf{w}, \mathbf{x}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w} \end{aligned}$$

Trong trường hợp này, dữ liệu $\mathcal{D}$ có thể được bỏ qua trong thuật ngữ đầu tiên, bởi vì khi chúng ta lấy điều kiện trên trọng số $\mathbf{w}$, chúng ta biết mọi thứ mà dữ liệu có thể đã cho chúng ta biết về mô hình. Kiểm tra vị trí đầu vào $\mathbf{x}$ có thể được bỏ qua trong thuật ngữ thứ hai, vì mô hình hồi quy logistic chỉ mô hình hóa nhãn, và việc biết nơi chúng ta kiểm tra không làm thay đổi niềm tin của chúng ta về các trọng số.

Chúng ta nên lấy trung bình các phân phối dự đoán $p(y|\mathbf{x}, \mathbf{w})$ cho các tham số khác nhau, được cân nhắc bởi mức độ hợp lí của các tham số đó $p(\mathbf{w}|\mathcal{D})$. Contour của phân phối dự đoán này là $p(y = 1|\mathbf{x}, \mathcal{D}) \in \{0.5, 0.27, 0.73, 0.12, 0.88\}$, được minh họa bằng hình bên dưới. Predictions at some constant distance away from the decision boundary are less certain when further away from the training inputs. That's because the different predictors above disagreed in regions far from the data.

Một lần nữa, các trục biểu thị đặc trưng đầu vào x_1 và x_2. Hình bên phải biểu thị $p(y = 1|\mathbf{x}, \mathbf{w}^*)$ đối với một vài trọng số $\mathbf{w}^*$ đã được khớp. Cho dù các trọng số được khớp được chọn như thế nào, thì các đường contour phải là tuyến tính. Các đường contour song song nhau có nghĩa là không chắc chắn của các dự đoán giảm cùng



Hình 7.3. Bayesian predictions found by the Langevin Monte Carlo method compared with the predictions using the optimized parameters

một tốc độ khi di chuyển ra khỏi ranh giới quyết định, no matter how far we are from the training inputs.

Khi nghiên cứu các phương pháp Monte Carlo được minh họa ở (7.2), chúng ta thấy rằng chúng ta có thể tính gần đúng mức trung bình này với mức trung bình theo kinh nghiệm cho các ví dụ về tham số hợp lý, chúng ta cũng có thể xấp xỉ $p(y|\mathbf{x}, \mathcal{D})$ bằng cách sử dụng xấp xỉ phân phối hậu nghiệm trên các trọng số $p(\mathbf{w}|\mathcal{D})$. Thông thường chúng ta có thể ước lượng các trọng số $\mathbf{w}$ bằng MLE hay MAP. Tuy nhiên, nếu chúng ta dự đoán các trọng số với các ước lượng này, thì giống như là chúng ta đã biết các trọng số trong khi chúng ta hoàn toàn không biết. Phân phối dự đoán của chúng ta sẽ có dạng là các đường contour song song thẳng (hình bên phải). Do đó, MAP không phải là phương pháp Bayes, mặc dù nó sử dụng phân phối tiên nghiệm nhưng các quy tắc của lý thuyết xác suất không nói cho chúng ta về fix một vector tham số thành một ước lượng. Chúng ta có thể xem MAP là một xấp xỉ với phương pháp Bayes, nhưng hình trên cho thấy rằng đó là một điều crude bởi vì các dự đoán Bayes (hình bên trái) khác biết về chất lượng đối với MAP.

Thật không may, chúng ta không thể đánh giá tích phân cho các dự đoán $p(y|\mathbf{x}, \mathcal{D})$ ở dạng đóng. Việc đưa ra các lựa chọn cho hồi quy Logistic Bayes cũng là một thử thách tính toán. Xác suất cận biên của dữ liệu $p(\mathcal{D})$ là Likelihood cận biên của mô hình, mà chúng ta có thể viết là $p(\mathcal{D}|\mathcal{M})$ khi chúng ta đang đánh giá một vài

mô hình chọn $\mathcal{M}$(chẳng hạn như các hàm cơ sở và siêu tham số). Chúng ta cũng không thể đánh giá tích phân cho $p(\mathcal{D})$ ở dạng đóng.
Đối với các dạng, chúng ta sẽ xét một Gaussian tiên nghiệm đa chiều, giống như trong trường hợp hồi quy tuyến tính:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mu, \mathbf{\Sigma}).$$

Bây giờ chúng ta áp dụng định lý Bayes để tìm phân phối hậu nghiệm:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}.$$

Thật không may, tích của phân phối tiên nghiệm Gaussian và likelihood (**??**) (cho một trong hai lựa chọn về hàm sigmoid) không dẫn đến kết quả cho phân phối hậu nghiệm trong một họ các tham số đủ tốt mà chúng ta biết. Một cách tương tự, tích phân trong hằng số chuẩn hóa (the evidence) $p(\mathbf{y}|\mathbf{X})$ is intractable as well.
Làm thế nào chúng ta có thể tiến hành quá trình này? Có hai cách tiếp cận chính để tiếp tục suy luận Bayessian trong cùng một tình huống. Đầu tiên, sử dụng phương pháp deterministic để tìm xấp xỉ cho phân phối hậu nghiệm (thường sẽ lấy các tham số trong họ các tham số đã chọn). Thứ hai, là ta bỏ qua biểu thức dạng đóng cho phân phối hậu nghiệm và thay vào đó, chọn ra một thuật toán để lấy mẫu từ phân phối hậu nghiệm, mà chúng ta có thể dùng, chẳng hạn, ước lượng Monte Carlo theo kỳ vọng. Bây giờ, chúng ta sẽ xét xấp xỉ Laplace, đầy là một ví dụ cho dạng tiếp cận đầu tiên.

7.4.1 Xấp xỉ Laplace cho phân phối Hậu nghiệm

Xấp xỉ Laplace sẽ tìm một xấp xỉ Gaussian cho phân phối có điều kiện của một tập hợp các biến liên lục. Giả sử chúng ta có một tham số tiên nghiệm tùy ý $p(\boldsymbol{\theta})$ và một likelihood tùy ý $p(\mathcal{D}|\boldsymbol{\theta})$, và chúng ta mong muốn xấp xỉ hậu nghiệm. Giả sử $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \frac{1}{Z}e^{-E(\boldsymbol{\theta})},$$

với hằng số chuẩn hóa Z là bằng chứng chưa biết và $Z = p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ và $E(\boldsymbol{\theta})$ được gọi là hàm *energy* và bằng với logarit âm của logarit hậu nghiệm chưa

chuẩn hóa, $E(\boldsymbol{\theta}) = -\ln p(\boldsymbol{\theta}, \mathcal{D}) = -\ln p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

Bây giờ, chúng ta định nghĩa hàm (có thể được gọi là Energy) như sau:

$$\Psi(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}).$$

Khi đó, Ψ là logarit của phân phối hậu nghiệm chưa chuẩn hóa. Xấp xỉ Laplace dựa trên khai triển Taylor mở rộng của Ψ quanh giá trị tối đa của nó. Đầu tiên, ta tìm giá trị cực đại của Ψ:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta})$$

Lưu ý rằng, điểm $\hat{\boldsymbol{\theta}}$ là xấp xỉ MAP với các tham số. Vấn đề tìm kiếm $\hat{\boldsymbol{\theta}}$ có thể được thực hiện bằng nhiều cách, nhưng trong thực tế $\hat{\boldsymbol{\theta}}$ thường được tìm bằng cách lấy Gradient và ma trận Hessian của Ψ theo $\boldsymbol{\theta}$ và dùng những phương pháp tối ưu thông thường.
Như đã đề cập ở trên, khai triển Taylor bậc 2 tốt cho các vấn đề tìm cực đại địa phương. Do đó, một khi chúng ta sẽ tìm thấy $\hat{\boldsymbol{\theta}}$, chúng ta sẽ lấy mở rộng Taylor bậc hai cho Ψ xung quanh điểm $\hat{\boldsymbol{\theta}}$ (i.e, trạng thái energy là thấp nhất), ta nhận được:

$$\Psi(\boldsymbol{\theta}) \approx \Psi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{g} - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

với $\mathbf{g}$ là gradient và $\boldsymbol{H}$ là ma trận Hessian của hàm energy tại điểm mốt $\hat{\boldsymbol{\theta}}$:

$$\mathbf{g} = \nabla\Psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \qquad \boldsymbol{H} = -\nabla\nabla\Psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Khi $\hat{\boldsymbol{\theta}}$ là điểm mốt thì graident bằng 0. Do đó,

$$\Psi(\boldsymbol{\theta}) \approx \Psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Theo định nghĩa hàm Ψ như trên. Bây giờ, ta thực hiện biến đổi sau để xấp xỉ

phân phối hậu nghiệm. Khi đó, ta có:

$$\Psi(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$
$$e^{\Psi(\boldsymbol{\theta})} = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$
$$e^{\Psi(\boldsymbol{\theta})} = Zp(\boldsymbol{\theta}|\mathcal{D}).$$

Từ đó, suy ra:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{e^{\Psi}}{Z},$$

Với

$$\Psi(\boldsymbol{\theta}) \approx \Psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Nên phân phối hậu nghiệm lúc này có dạng sau:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \frac{1}{Z} \exp\big(\Psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\big), \tag{7.6}$$
$$\propto \frac{1}{Z} \exp\big(\Psi(\hat{\boldsymbol{\theta}})\big) \exp\big(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\big), \tag{7.7}$$
$$\propto \frac{1}{Z} exp(-E(\hat{\boldsymbol{\theta}})) \exp\big(-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\big). \tag{7.8}$$

Ở đây, ta nhận thấy rằng $p(\boldsymbol{\theta}|\mathcal{D})$ dường như tỷ lệ thuận với phân phối Gaussian. Do đó, xấp xỉ Laplace dẫn đến một xấp xỉ chuẩn cho phân phối hậu nghiệm:

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}, \mathbf{H}^{-1}).$$

Ví dụ:

Chúng ta có thể tìm một tối ưu và độ cong cho Gaussian trong trường hợp một chiều $\mathcal{N}(\mu, \sigma^2)$ có dạng:

$$\Psi_{\mathcal{N}}(w) = \Psi_{\mathcal{N}}(\hat{w}) - \frac{(w-\mu)^2}{2\sigma^2},$$

Hàm trên đạt giá trị cực tiểu tại $\hat{w} = \mu$ và đạo hàm riêng cấp hai là $\mathbf{H} = \frac{1}{\sigma^2}$ nên phương sai $\sigma^2 = \frac{1}{H}$.

Tương tự, ta có thể tổng quát hóa cho trường hợp Gaussian với số chiều cao hơn $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$:

$$\Psi_{\mathcal{N}}(\mathbf{w}) = \Psi_{\mathcal{N}}(\hat{\mathbf{w}}) - \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{w} - \boldsymbol{\mu})$$

với $\hat{\mathbf{w}} = \boldsymbol{\mu}$ và $\mathbf{H} = \Sigma^{-1}$ thì ma trận hiệp phương sai là $\Sigma = \mathbf{H^{-1}}$

Phép tính xấp xỉ là một Gaussian tập trung vào hình thức của hậu thế, θ, với hiệp phương sai bắt buộc logarit của phép xấp xỉ đúng với độ cong của logarit hậu nghiệm đúng tại điểm đó.

Ta lưu ý rằng phép gần đúng Laplace cũng đưa ra một xấp xỉ với hằng số chuẩn hóa Z. Trong trường hợp này, nó chỉ là một câu hỏi về hằng số chuẩn hóa mà chúng ta phải sử dụng để lấy (7.6) để chuẩn hóa. Một tính toán khá đơn giản cho.

$$\begin{aligned}
Z &= \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \\
&= \int e^{\ln p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}d\boldsymbol{\theta}, \\
&= \int \exp \Psi(\boldsymbol{\theta})d\boldsymbol{\theta}, \\
&\approx \int \exp\big(\Psi(\hat{\boldsymbol{\theta}})\big) \exp\bigg(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\mathbf{H}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\bigg)\mathbf{d}\boldsymbol{\theta}, \\
&= \exp\big(\Psi(\hat{\boldsymbol{\theta}})\big) \int \exp\bigg(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\mathbf{H}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\bigg)\mathbf{d}\boldsymbol{\theta}, \\
&= \exp \Psi(\hat{\boldsymbol{\theta}})\sqrt{\frac{(2\pi)^d}{det\mathbf{H}}}, \\
&= \exp(-E(\hat{\boldsymbol{\theta}}))\sqrt{\frac{(2\pi)^d}{det\mathbf{H}}}.
\end{aligned} \tag{7.9}$$

với d là số chiều $\boldsymbol{\theta}$. Và Z là hằng số chuẩn hóa của phân phối Gaussian nhiều chiều. Phương trình (7.9) được coi là xấp xỉ Laplace của marginal likelihood. Phương trình (7.6) đôi khi được gọi là xấp xỉ Laplace của hậu nghiệm. Một xấp xỉ Gaussian thường là một xấp xỉ hợp lí, tức là khi hậu nghiệm trở nên nên giống Gaussian hơn khi kích thước mẫu tăng tương tự như định lý giới hạn trung tâm.

Khi một số người nói rằng "Xấp xỉ Laplace" , chúng có liên hệ với sự gần đúng của chuẩn hóa $p(\mathcal{D})$,chứ không phải là xấp xỉ Gaussian trung gian cho phân phối. Vậy xấp xỉ này có hợp lý không?

Nếu chúng ta nghĩ rằng *Energy* hoạt động tốt và chóp nhọn xung quang mốt của phân phối, chúng ta có thể nghĩ rằng chúng ta có thể xấp xỉ nó với chuỗi Taylor như phần trên:

$$\Psi(\boldsymbol{\theta}) \approx \Psi(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{H}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}).$$

Một *energy* bậc hai (xác suất logarit âm) của phân phối Gaussian. Phân phối gần với sự phù hợp của Gaussian khi chuỗi Taylor chính xác.
Đối với các mô hình có số lượng tham số nhận dạng cố định, hậu nghiệm trở nên cực đại trong giới hạn của bộ dữ liệu lớn. Khi đó, mở rộng Taylor của logarit hậu nghiệm không cần phải được ngoại suy xa và sẽ chính xác.
Thuật ngữ tìm kiếm để biết thêm thông tin: Định lý giới hạn trung tâm Bay Bayesian.

7.4.2 Tiêu chuẩn thông tin Bayesian (BIC)

Chúng ta có thể sử dụng xấp xỉ Gaussian để viết logarit likelihood cận biện như sau, loại bỏ các hằng số không liên quan:

$$\begin{aligned}
\ln Z &= \ln p(\mathcal{D}), \\
&= \ln \exp(-E(\hat{\boldsymbol{\theta}}))\sqrt{\frac{(2\pi)^d}{det\mathbf{H}}}, \\
&= \ln \exp(-E(\hat{\boldsymbol{\theta}})) + \ln \sqrt{\frac{(2\pi)^d}{det\mathbf{H}}}, \\
&= -E(\hat{\boldsymbol{\theta}}) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \mathbf{H}, \\
&= \ln p(\mathcal{D}|\hat{\theta}) + \ln p(\hat{\theta}) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln \mathbf{H}, \\
&= \ln p(\mathcal{D}|\hat{\theta}) + \ln p(\hat{\theta}) - \frac{1}{2}\ln \mathbf{H}.
\end{aligned}$$

Thuật ngữ *phạt* được thêm vào $\ln p(\mathcal{D}|\hat{\boldsymbol{\theta}})$ đôi khi được gọi là **Occam factor** và là thước đo độ phức tạp của mô hình. Nếu chúng ta có tiên nghiệm đều, $p(\boldsymbol{\theta}) \propto 1$, có thể bỏ qua thuật ngữ thứ hai, và thay $\hat{\boldsymbol{\theta}}$ bằng ước lượng MLE là θ^*. Bây giờ chúng ta tập trung vào xấp xỉ thuật ngữ thứ ba. Chúng ta có $\mathbf{H} = \sum_{i=1}^{N} \mathbf{H}_i$ với $\mathbf{H}_i = -\nabla\nabla \ln p(\mathcal{D}_i|\theta)$ và N là số các điểm dữ liệu. Bây giờ chúng ta xấp xỉ mỗi $\mathbf{H}_i$ bằng một ma trận cố định $\hat{\mathbf{H}}$. Khi đó, chúng ta có:

$$\ln det(\mathbf{H}) = \ln(det(N\hat{\mathbf{H}})) = \ln(N^d et(\hat{\mathbf{H}}) = d\ln N + \ln \hat{\mathbf{H}}.$$

với $d = dim(\theta)$ và chúng ta giả định rằng $\mathbf{H}$ là ma trận hạng đầy đủ. Chúng ta có thể bỏ qua sự giảm xuống của $\ln \hat{\mathbf{H}}$ vì nó độc lập với N, và do đó sẽ bị áp đảo bởi

likelihood.

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{d}{2}\ln N$$

Khi biểu diễn mô hình chọn, một đại diện thay thế của mô hình hậu nghiệm (nó có thể khó tính toán khi phải đối nặt với vấn đề nội suy) được gọi là *Tiêu chuẩn thông tin Bayesian (BIC)*. Cho tập hợp các mô hình $\{\mathcal{M}_i\}$ và dữ liệu được quan sát $\mathcal{D}$, chúng ta có thể tính thống kê bên dưới cho mỗi BIC_i, theo trên:

$$BIC_i = \ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}_i) - \frac{d}{2}\ln N,$$

với d là số chiều của $\boldsymbol{\theta}_i$ và $\hat{\boldsymbol{\theta}}_i$ là các tham số MAP cho $\mathcal{M}_i$. Mô hình với BIC cao nhất được cho là hoàn hảo.
Chúng ta có thể rút ra điều này thông qua phép tính xấp xỉ Laplace cho phân phối hậu nghiệm. Lấy logarit cho ước lượng Z ở trên, ta có:

$$\ln p(\mathcal{D}|\mathcal{M}_i) = \ln Z \approx \ln p(\mathcal{D}|\hat{\boldsymbol{\theta}}_i) + \ln p(\hat{\boldsymbol{\theta}}_i) + \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln det\mathbf{H}.$$

Nếu chúng ta giả sử tiên nghiệm rất rộng, chúng ta có thể bỏ qua thuật ngữ $\ln p(\hat{\boldsymbol{\theta}}_i)$, nếu chúng ta giả sử các điểm dữ liệu là độc lập với các tham số (như trong hồi quy tuyến tính và hồi quy logistic) và $\mathbf{H}$ có hạng đầy đủ, sau đó chúng ta có thể xấp xỉ tiệm cận này với điểm BIC, khi loại bỏ hằng số.

Bản chất của điểm BIC, là nó cho phép cho các mô hình giải thích dữ liệu tốt nhưng lại phạt chúng vì quá phức tạp, chính xác là sự đánh đổi được xem xét trong lựa chọn mô hình đầy đủ của Bayesian.

7.4.3 Xấp xỉ Gaussian cho hồi quy logistic

Bây giờ, chúng ta áp dụng xấp xỉ Gaussian cho hồi quy logistic. Chúng ta sẽ sử dụng tiên nghiệm Gaussian dưới dạng $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; 0, \mathbf{V}_0)$, giống như chúng ta đã làm trong ước lượng MAP. Xấp xỉ hậu nghiệm được cho bởi

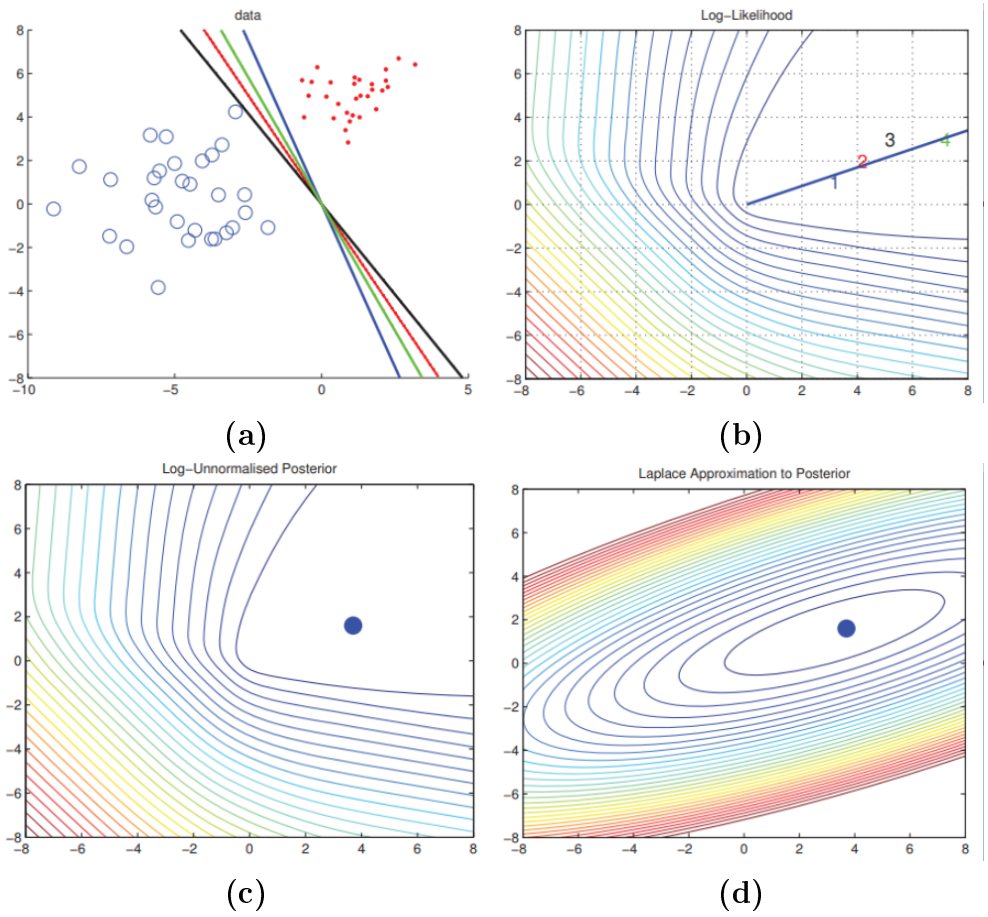
$$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \mathbf{H}^{-1}),$$

với $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} E(\mathbf{w}), E(\mathbf{w}) = -(\ln p(\mathcal{D}|\mathbf{w}) + \ln p(\mathbf{w})$, và $\mathbf{H} = \nabla\nabla E(\mathbf{w})|_{\hat{\mathbf{w}}}$. Ví dụ, xem xét dữ liệu 2D có thể phân tách tuyến tính trong Hình (7.4) (a). Có nhiều các thiết lập tham số tương ứng với các dòng tách biệt hoàn toàn dữ liệu huấn luyện, được hiển thị trong 4 ví dụ. Bề mặt của likelihood được thể hiện trong hình (7.4)(b), trong đó chúng ta thấy rằng likelihood không bị ràng buộc khi chúng ta di chuyển lên và sang phải trong không gian tham số, dọc theo ridge $\frac{w_2}{w_1} = 2.35$ (điều này được biểu thị bằng đường chéo).Lý do cho điều này là chúng ta có thể tối đa hóa likelihood bằng cách điều khiển $||\mathbf{w}||$ đến vô cùng, vì trọng số hồi quy lớn làm cho hàm sigmoid rất dốc, biến nó thành hàm bước. Do đó, MLE không được định nghĩa tốt khi dữ liệu được phân tách tuyến tính.
Để chính quy hóa vấn đề, chúng ta hãy dùng một hình cầu tiên nghiệm mơ hồ tập trung tại điểm gốc $\mathcal{N}(\mathbf{w}; 0, 100\mathbf{I})$. Nhân của tiên nghiệm hình cầu này với bề mặt likelihood dẫn đến hậu quả bị lệch rất cao, được thể hiện trong Hình (7.4)(c). Hậu nghiệm bị lệch bởi vì hàm likelihood rời khỏi không gian tham số, không hoạt động theo dữ liệu. Ước tính MAP được hiển thị bằng dấu chấm màu xanh. Không giống như MLE, đây không phải là vô cùng

Giá trị gần đúng của Gaussian cho hậu nghiệm này được hiển thị trong Hình (7.4)(d). Ta thấy rằng đây là một phân phối đối xứng, và do đó, nó không phải là một xấp xỉ lớn. Tất nhiên, nó có mốt chính xác (bằng cách xây dựng), và ít nhất nó đại diện cho thực tế là có nhiều sự không chắc chắn dọc theo hướng tây nam-đông bắc (tương ứng với sự không chắc chắn về hướng của các đường phân cách) hơn là trực giao với điều này. Mặc dù là một xấp xỉ thô, nhưng điều này chắc chắn tốt hơn so với xấp xỉ hậu nghiệm bởi một hàm delta, đó là những gì ước tính MAP thực hiện.

7.5 Đưa ra dự đoán

Cho một hậu nghiệm, chúng ta có thể tính các khoảng tin cậy, thực hiện các kiểm định giả thuyết,.... Nhưng trong máy học, sự quan tâm thường tập trung vào vấn



Hình 7.4. (a) Dữ liệu hai lớp trong 2d. (b) Log- Likelihood cho mô hình hồi quy logistic. Dường này được vẽ từ điểm gốc theo hướng MLE (nằm ở vô cực). Các số tương ứng với 4 điểm trong không gian tham số, tương ứng với các dòng trong (a). (c) Log-likelihood chuẩn hóa (giả sử hình cầu tiên nghiệm mờ). (d) Laplace gần đúng với hậu nghiệm.

đề dự đoán. Khi đó, phân phối dự đoán hậu nghiệm có dạng

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w}.$$

Thật không may, tích phân này khó có thể đánh giá.

Phép xấp xỉ đơn giản nhất là xấp xỉ *plug-in* trong trường hợp nhị phân có dạng như sau:

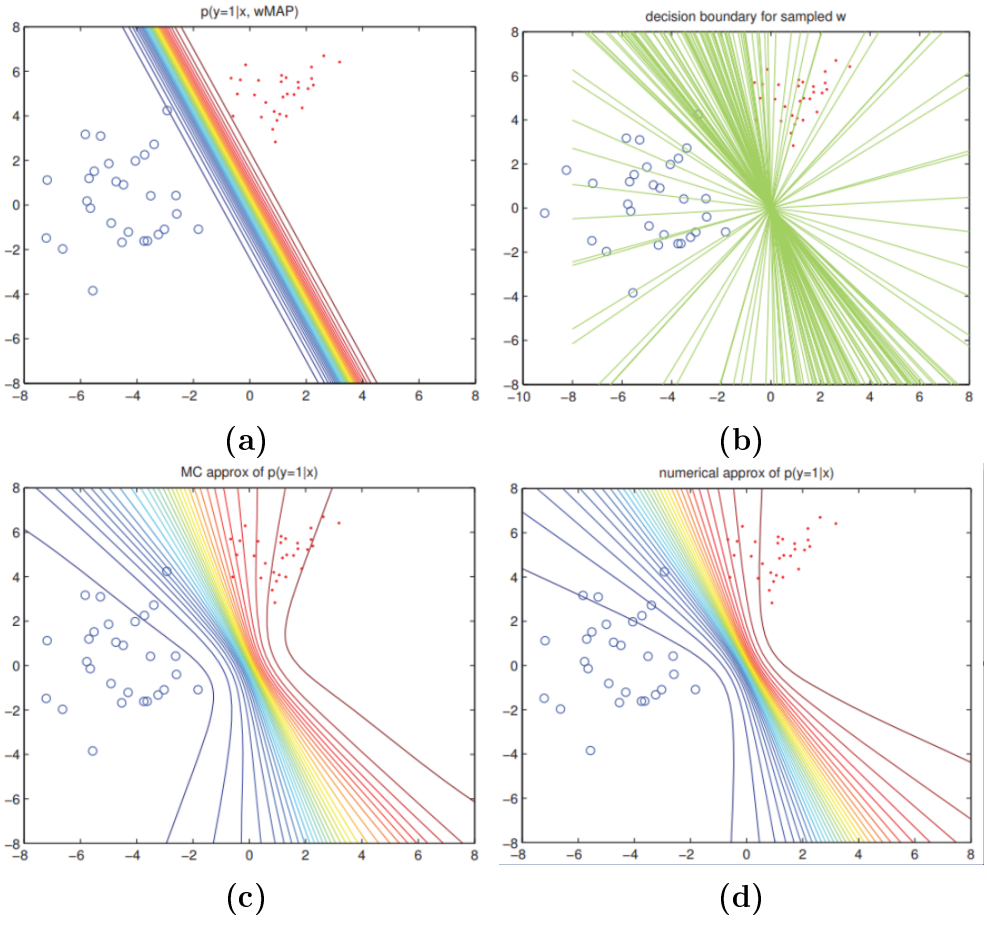
với $E(\mathbf{w})$ là trung bình hậu nghiệm. Trong ngữ cảnh này, $E(\mathbf{w})$ được gọi là **Bayes point**. Tất nhiên, ước lượng plug-in đánh giá không chắc chắn. Chúng ta có thể thảo luận một vài xấp xỉ tốt hơn bên dưới:

7.5.1 Xấp xỉ Monte Carlo

Một cách tiếp cận tốt hơn là sử dụng xấp xỉ Monte Carlo, như sau:

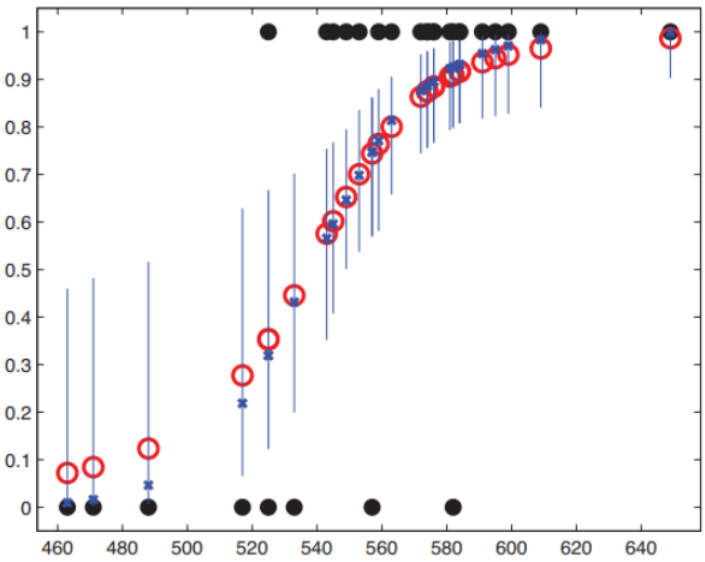
$$p(y=1|\mathbf{x},\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\sigma((\mathbf{w}^s)^\top \mathbf{x}),$$

với $\mathbf{w}^s \sim p(\mathbf{w}|\mathcal{D})$, là mẫu lấy từ phân phối hậu nghiệm.(Kĩ thuật này có thể được mở rộng trong trường hợp nhiều lớp). Nếu chúng ta đã xấp xỉ hậu nghiệm bằng Monte Carlo, chúng ta có thể sử dụng lại các mẫu cho việc dự đoán. Nếu chúng ta thực hiện xấp xỉ Gaussian cho hậu nghiệm, chúng ta có thể vẽ các mẫu độc lập từ Gaussian bằng các phương pháp tiêu chuẩn. Hình (7.5)(b) hiển thi các mẫu từ dự



Hình 7.5. Posterior predictive distribution for a logistic regression model in 2d. Top left: contours of $p(y = 1|\mathbf{x}, \hat{\mathbf{w}}_MAP)$. Top right: samples from the posterior predictive distribution. Bottom left: Averaging over these samples. Bottom right: moderated output (probit approximation). Based on a figure by Mark Girolami. Figure generated by logregLaplaceGirolamiDemo.

báo hậu nghiệm trong trường hợp 2D. Hình (7.5)(c) cho thấy mức trung bình của các mãu này. Bằng cách tính trung bình trên nhiều dự đoán, chúng ta thấy rằng sự không chắc chắn trong ranh giới quyết định "xóa tách ra khỏi (splays out)" khi chúng ta di chuyển chúng ra xa hơn dữ liệu huấn luyện. Vì vậy, mặc dù ranh giới quyết định là tuyến tính, mật độ dự báo hậu nghiệm không phải là tuyến tính. Cũng lưu ý rằng ranh giới quyết định trung bình hậu nghiệm là gần như xa cả hai lớp.



Hình 7.6. Mật độ dự báo hậu nghiệm cho dữ liệu SAT. Vòng tròn màu đỏ biểu thị giá trị trung bình sau, màu xanh chéo giữa trung vị sau và đường màu xanh biểu thị phần trăm thứ 5 và 95 của phân bố dự báo.

Hiển thị một ví dụ trong 1d. Các chấm đỏ biểu thị giá trị trung bình của tiên đoán hậu nghiệm được đánh giá tại dữ liệu huấn luyện. Các đường màu xanh thẳng đứng biểu thị khoảng tin cậy 95 % cho dự đoán hậu nghiệm; ngôi sao nhỏ màu xanh là trung vị. Chúng ta thấy rằng, với phương pháp Bayes, có thể mô hình hóa sự không chắc chắni về xác suất học sinh sẽ vượt qua bài kiểm tra dựa trên điểm SAT của mình, thay vì chỉ ước tính điểm.

7.5.2 Xấp xỉ Probit

Giả sử chúng ta thu được một xấp xỉ Gaussian cho phân phối hậu nghiệm $p(\mathbf{w}|\mathcal{D})$. Chẳng hạn, phần trên chúng ta đã dùng xấp xỉ Laplace và nhận được $p(\mathbf{w}|\mathcal{D}) \approx \mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \mathbf{H}^{-1})$ với $\hat{\mathbf{w}}$ là ước lượng MAP cho các tham số và $\mathbf{H}$ là ma trận Hessian

của logarit hậu nghiệm âm được đánh giá tại $\hat{\mathbf{w}}$.

Giả sử bây giờ chúng ta đã đưa ra kiểm thử đầu vào $\mathbf{x}_*$ và mong muốn dự đoán nhãn nhị phân y_*. Phép xấp xỉ Bayesian, chúng tôi đưa ra các tham số chưa biết w để tìm phân phối dự đoán (predictive distribution):

$$Pr(y_* = 1|\mathbf{x}_*, \mathcal{D}) = \int Pr(y_* = 1|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w} = \int \sigma(\mathbf{x}_*\mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w} = \mathbf{E}_{\mathbf{p}(\mathbf{w}|\mathcal{D})}[\sigma(\mathbf{x}_*\mathbf{w}].$$

Thật không may, thậm chí là với xấp xỉ Gaussian với $p(\mathbf{w}|\mathcal{D})$, tích phân này không thể đánh giá nếu chúng ta sử dụng hàm hồi quy với vai trò là hàm sigmoid σ. Tuy nhiên, chúng ta có thể tính tích phân khi ta sử dụng CDF của phân phối chuẩn cho σ:

$$Pr(y_* = 1|\mathbf{x}_*, \mathcal{D}) = \int \Psi(\mathbf{x}_*^T\mathbf{w})\mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \mathbf{H}^{-1})d\mathbf{w}.$$

Điều này trông giống như một tích phân d chiều "khó chịu". Kỳ vọng sẽ là một tích phân trên vectỏ trọng số cao $\mathbf{w}$. Tuy nhiên, hàm này chỉ quan tâm đến độ lớn của vectỏ đó theo hướng của các đặc trưng kiểm tra thông qua tích trong $a = \mathbf{w}^\top\mathbf{x}$.

$$Pr(y_* = 1|\mathbf{x}_*, \mathcal{D}) = \int \Psi(a)p(a|\mathcal{D})d\mathcal{D}.$$

Chú ý rằng, a là biến đổi tuyến tính của phân phối Gausian theo trọng số $\mathbf{w}$. Nếu chúng ta xấp xỉ hậu nghiệm với Gaussian

$$p(\mathbf{w}|\mathcal{D}) \approx \mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \mathbf{H^{-1}}).$$

khi đó $a = \mathbf{w}^T\mathbf{x}$ cũng tuân theo phân phối Gaussian, $p(a|\mathcal{D}) \approx \mathcal{N}(a; \mu_{a|\mathcal{D}}, \sigma_{a|\mathcal{D}})$ với

$$\mu_{a|\mathcal{D}} = \mathbf{x}_*^T\hat{\mathbf{w}}; \qquad \sigma^2_{a|\mathcal{D}} = \mathbf{x}_*^T\mathbf{H}^{-1}\mathbf{x}_*.$$

Ưu điểm của việc sử dụng probit là người ta có thể kết hợp nó với một phân tích Gaussian.

$$\int \Psi(\lambda a)\mathcal{N}(a; \mu, \sigma^2)da = \Psi\left(\frac{a}{(\lambda^{-2} + \sigma^2)^{\frac{1}{2}}}\right)$$

Vì vậy, chúng ta có thể tính được:

$$Pr(y_* = 1|\mathbf{x}_*, \mathcal{D}) = \int \Psi(a)\mathcal{N}(\mathbf{w}; \hat{\mathbf{w}}, \mathbf{H}^{-1})d\mathcal{D} = \Psi\left(\frac{\mu_{a|\mathcal{D}}}{\sqrt{1+\sigma^2_{a|\mathcal{D}}}}\right)$$

Chú ý, $\Psi(\mu_{a|\mathcal{D}})$ sẽ được ước lượng, chúng ta sé sử dụng MAP để tìm $\hat{\mathbf{w}}$ như một ước lượng *plug in* thông thường. Thuật ngữ $\sqrt{1+\sigma^2_{a|\mathcal{D}}}$ làm cho dự đoán của chúng ta kém chắc chắn hơn (có nghĩa là, gần đến $\frac{1}{2}$) theo sự không chắc chắn trong giá trị của $a = \mathbf{x}_*^T\mathbf{w}$. Phương pháp này đôi khi được gọi là moderation , bởi vì chúng ta buộc các dự đoán của mình phải ôn hòa hơn mức chúng ts có thể sử dụng ước tính điểm bổ trợ plug-in của $\mathbf{w}$.

Chúng ta cũng lưu ý rằng nếu chúng ta chỉ muốn đưa ra dự đoán điểm của y_* bằng cách sử dụng hàm mất mát 0-1, chúng ta chỉ cần biết lớp nào có thể xảy ra hơn (đây là kết quả chung từ thảo luận về lý thuyết quyết định Bayes). Trong trường hợp này, moderation không ảnh hưởng đến dự đoán cuối cùng và thay vào đó chúng ta chỉ cần tìm $\hat{\mathbf{w}}$. Điều này tương tự với kết quả chúng ta đã có trong hồi quy tuyến tính, nơi chúng ta có thể chỉ cần tìm ước lượng MAP cho $\mathbf{w}$ nếu cuối cùng chúng ta chỉ quan tâm đến dự đoán điểm theo bình phương mất mát.

Kết luận

Tài liệu tham khảo

[1] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, 2011.

[2] Sergios Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2015.

[3] Rencher, *Linear Mode in Statistics*, 2008.