TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI TP.HCM
VIỆN CÔNG NGHỆ THÔNG TIN VÀ ĐIỆN, ĐIỆN TỬ

PHƯƠNG PHÁP TOÁN CHO MÁY HỌC

CHƯƠNG VI: CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

TS. Trần Thế Vinh

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Đặc trưng mẫu (Features):
Là các thuộc tính hoặc biến số được sử dụng để mô tả dữ liệu đầu vào. Đặc trưng là yếu tố cốt lõi giúp mô hình máy học hiểu và đưa ra dự đoán.

Ví dụ: Trong bài toán dự đoán giá nhà, đặc trưng có thể là diện tích, số phòng ngủ, vị trí, tuổi nhà.
Đặc trưng có thể là số, phân loại, văn bản, hình ảnh, tùy thuộc vào bài toán.

Phương pháp ước lượng (Estimation Methods):
Là các kỹ thuật toán học để xác định tham số của mô hình hoặc dự đoán giá trị dựa trên dữ liệu.

Ví dụ: Ước lượng hợp lý cực đại (MLE) tìm tham số tối ưu cho phân phối, hoặc Gradient Descent tối ưu hóa hàm mất mát.

Tầm quan trọng

Đặc trưng mẫu:
Quyết định chất lượng đầu vào, ảnh hưởng trực tiếp đến khả năng học của mô hình. Đặc trưng tốt giúp mô hình học nhanh và chính xác hơn.

Phương pháp ước lượng:

Đảm bảo mô hình tìm ra tham số tối ưu, cân bằng giữa độ chính xác và tính tổng quát hóa, tránh overfitting.

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Ứng dụng trong máy học

1. Đặc trưng mẫu:

- **Xây dựng mô hình**: Đặc trưng là đầu vào trực tiếp cho các thuật toán máy học như hồi quy tuyến tính, SVM, mạng nơ-ron.
- **Feature Engineering**: Tạo đặc trưng mới (ví dụ, tỷ lệ diện tích/số phòng) hoặc xử lý đặc trưng (chuẩn hóa, mã hóa) để cải thiện hiệu suất mô hình .
- **Giảm chiều dữ liệu**: Các kỹ thuật như PCA sử dụng đặc trưng để giảm số lượng biến, giữ lại thông tin quan trọng.

- **Ứng dụng cụ thể**:
    - Phân loại: Nhận diện email spam dựa trên đặc trưng như tần suất từ, độ dài email.
    - Dự đoán: Dự báo thời tiết dựa trên đặc trưng như nhiệt độ, độ ẩm, áp suất.
    - Thị giác máy tính: Trích xuất đặc trưng từ ảnh (cạnh, màu sắc) để nhận diện đối tượng.

2. Phương pháp ước lượng:

✓ Học có giám sát:

Hồi quy: Ước lượng tham số để dự đoán giá trị liên tục (ví dụ, giá nhà).

Phân loại: Ước lượng ranh giới quyết định (decision boundary) để phân loại (ví dụ, bệnh nhân có bệnh hay không).

✓ Học không giám sát:

Phân cụm: Ước lượng tâm cụm trong k-Means để nhóm dữ liệu.

Giảm chiều: Ước lượng không gian đặc trưng chính trong PCA.

✓ Học tăng cường:

Ước lượng giá trị hành động (Q-value) để tối ưu hóa chính sách trong môi trường động.

✓ Ứng dụng cụ thể:

Gradient Descent: Tối ưu hóa trọng số trong mạng nơ-ron để nhận diện hình ảnh.

Bayesian Estimation: Dự đoán xác suất trong hệ thống gợi ý.

Ensemble Methods (Random Forest, XGBoost): Kết hợp nhiều ước lượng để tăng độ chính xác trong dự đoán tài chính, y học.

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Mẫu và các đặc trưng trên mẫu:
Các khái niệm về mẫu

Tổng thể điều tra:
Khi nghiên cứu một hiện tượng hay giải quyết một công việc, ta cần có những thông tin về hiện tượng hay công việc cần giải quyết. Tập hợp tất cả các đối tượng mà ta quan tâm điều tra, xem xét để có được thông tin nói trên được gọi là tổng thể điều tra, (gọi tắt là tổng thể),ký hiệu là: Ω

Tiêu chuẩn điều tra:
Những thông tin về phần tử của tổng thể phục vụ cho mục đích điều tra, nghiên cứu gọi là tiêu chuẩn điều tra, hay gọi tắt là tiêu chuẩn.

Biến quan sát:
Tập hợp tất cả các giá trị của một tiêu chuẩn điều tra gọi là một biến quan sát, ký hiệu là: X, Y, Z,…
Nếu tập giá trị của biến quan sát X là một tập hợp các số, tức là trên mỗi phần tử điều tra thông tin được cho dưới dạng con số thì tiêu chuẩn điều tra gọi là tiêu chuẩn số lượng. Khi đó gọi X là biến định lượng.
Nếu trên mỗi phần tử được điều tra, thông tin được thể hiện không phải là số (chẳng hạn: xanh, đỏ,...hay: trung bình, khá, giỏi,…) thì tiêu chuẩn điều tra được gọi là tiêu chuẩn chất lượng. Khi đó gọi biến quan sát X tương ứng là biến định tính.

Nhận xét: Giả sử X là biến q.sát trên tổng thể Ω. Việc chọn một phần tử từ tổng thể để quan sát, điều tra thực chất là việc thực hiện một phép thử, mỗi phần tử của tổng thể là một kết cục và giá trị của X phụ thuộc vào phần tử được chọn, tức là phụ thuộc vào kết cục của phép thử. Vậy mỗi biến quan sát X là một biến ngẫu nhiên.

Mẫu: Mỗi tập hợp con của tổng thể được lấy ra để quan sát gọi là mẫu điều tra hay gọi tắt là mẫu.

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Yêu cầu đặt ra đối với mẫu là mẫu phải có tính chất đại diện cho tổng thể. Muốn vậy khi chọn mẫu, phải đảm bảo tính khách quan, không cố ý, không thiên vị,…, những yêu cầu này gọi một cách đơn giản là chọn ngẫu nhiên. Mẫu thu được theo cách đó gọi là mẫu ngẫu nhiên. Số lượng phần tử được chọn vào mẫu gọi là kích thước mẫu hay cỡ mẫu, ký hiệu: n, m,…

Ở đây ta chỉ xét mẫu n.nhiên và khi nói tới mẫu mà không giải thích gì thêm thì hiểu rằng đó là mẫu ngẫu nhiên.

Mẫu ngẫu nhiên về biến quan sát: Xét mẫu ngẫu nhiên kích thước n để quan sát X.
Ký hiệu X_j là g.trị của X ở phần tử thứ j của mẫu, (j = 1, 2,…, n) thì X_j là biến ngẫu nhiên cùng phân phối xác suất với X. Hơn nữa do các phần tử được chọn vào mẫu một cách độc lập nhau nên là các biến ngẫu nhiên độc lập nhau.
Gọi $(X_1, X_2, \dots, X_n)$ là *mẫu ngẫu nhiên kích thước n về biến quan sát X*: đó là một véc tơ ngẫu nhiên n chiều với các thành phần độc lập nhau và có cùng phân phối xác suất với biến quan sát X.

Một số đặc trưng quan trọng trên mẫu
Giả sử $(X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên kích thước n về biến quan sát X trên tổng thể Ω. Để phân biệt với các đặc trưng mẫu dưới đây, ta gọi:
$\mu = EX$ là trung bình tổng thể (về biến quan sát X); $\sigma^2 = DX$ là phương sai tổng thể;
$p = P(A)$ là tỷ lệ tính chất A trong tổng thể hay tần suất tổng quát về A

Trung bình mẫu

Định nghĩa: Trung bình mẫu của biến quan sát X là đại lượng: $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

Các tính chất của phép lấy trung bình mẫu:

1. Nếu $X = c = const$ thì: $\bar{X} = c$;
2. Nếu $X \geq 0$ thì: $\bar{X} \geq 0$;
3. Với mọi hằng số k thì: $\overline{k.X} = k.\bar{X}$;
4. Với hai biến q.sát X, Y ta có:
   $\overline{X + Y} = \bar{X} + \bar{Y}$;
   Nếu X và Y độc lập nhau thì: $\overline{X.Y} = \bar{X}.\bar{Y}$

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

5. Với g(x) là một hàm số thì: $\overline{g(X)} = \frac{1}{n}\sum_{j=1}^{n} g(X_j)$. Đặc biệt: $\overline{X^2} = \frac{1}{n}\sum_j X_j^2$
6. $\bar{X}$ là biến ngẫu nhiên có các tính chất:

$$E\bar{X} = \mu; D\bar{X} = \frac{\sigma^2}{n};$$

Khi n đủ lớn ta có:

$\bar{X} \approx \mu$, và $\bar{X}$ có phân phối xấp xỉ chuẩn $N(\mu, \frac{\sigma^2}{n})$ hay: $\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}$ có phân phối xấp xỉ chuẩn $N(0,1)$

Phương sai mẫu:

Phương sai mẫu của biến quan sát X là: $\boldsymbol{S^2(X) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2}$

Các tính chất của phép lấy phương sai mẫu:

1. $S^2(X) \geq 0$, gọi: $S(X) = \sqrt{S^2(X)}$ là độ lệch mẫu của X
2. $\boldsymbol{S^2(X) = \overline{X^2} - (\bar{X})^2}$
3. Với mọi hằng số a, ta có: $S^2(a.X) = a^2.S^2(X)$
4. Nếu là các biến q.sát X, Y độc lập nhau thì:

$$\boldsymbol{S^2(X \pm Y) = S^2(X) + S^2(Y)}$$

5. $S^2(X)$ là một biến ngẫu nhiên mà: $\boldsymbol{ES^2(X) = \frac{n-1}{n}\sigma^2}$

Khi kích thước mẫu n đủ lớn ta có: $\boldsymbol{S^2(X) \approx \sigma^2}$

Tần suất mẫu:

Xét mẫu ngẫu nhiên kích thước n để quan sát tính chất A nào đó của các phần tử.

Định nghĩa: Ký hiệu m(A) là số phần tử có tính chất A trên mẫu. Khi đó tỷ lệ tính chất A trên mẫu là $\boldsymbol{f(A) = \frac{m(A)}{n}}$, còn gọi là tần suất mẫu của A.

Các tính chất:

1. $\boldsymbol{n.f(A) \sim B(n,p)}$ và vì thế: $\boldsymbol{Ef(A) = p; Df(A) = \frac{p(1-p)}{n}}$
2. Khi n đủ lớn thì: a/ $\boldsymbol{f(A) \approx p}$; b/ $f(A)$ có p.phối x.xỉ chuẩn $\boldsymbol{N(p; \frac{p(1-p)}{n})}$

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Mẫu cụ thể - Trình bày số liệu điều tra

Cho X là biến quan sát trên tổng thể Ω và $(X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên kích thước n về X. Khi ta chọn n phần tử cụ thể vào mẫu thì n phần tử đó gọi là một mẫu cụ thể hay một giá trị xác định của mẫu. Với một mẫu cụ thể thì $(X_1, X_2, \dots, X_n)$ có một giá trị xác định là n giá trị của X quan sát được, ta gọi đó là tập số liệu hay kết quả điều tra.

Trình bày bảng số liệu điều tra

Trường hợp 1: Khi số liệu điều tra là n số liệu xác định gồm k giá trị khác nhau: $x_1, x_2, \dots, x_k$. Ký hiệu: n_i là số số liệu có giá trị là x_i. Ta gọi n_i là tần số của x_i. Khi đó số liệu được trình bày bằng bảng sau:

X	x_1	x_2	⋯	x_k	Σ
Tần số n_i	n_1	n_2	⋯	n_k	n

(1)

Trường hợp 2: Khi n số liệu không được cho cụ thể mà được chỉ ra trong k khoảng rời nhau: $I_1, I_2, \dots, I_k$, trong đó khoảng I_i chứa n_i số liệu.Ta thay tất cả các số liệu có trong khoảng I_i bởi trị trung tâm x_i (hay trị đại diện) là điểm giữa của khoảng .

Khi đó số liệu điều tra được trình bày bởi bảng sau:

Các khoảng X	x_1	x_2	⋯	x_k	Σ
Trị đại diện	x_1	x_2	⋯	x_k	Σ
Tần số n_i	n_1	n_2	⋯	n_k	n

(2)

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Tính giá trị các đặc trưng mẫu từ tập số liệu

Giả sử biến quan sát X có kết quả điều tra được cho bởi bảng (1) hoặc (2). Khi đó từ công thức của các đặc trưng mẫu ta nhận được:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i;\ \overline{X^2} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2; S^2(X) = \overline{X^2} - (\bar{X})^2$$

Ví dụ: Điều tra về mức thu nhập X(triệu đ/ tháng) của các hộ gia đình ở đường phố B, có các khoảng thu nhập: [4,8 – 5), [5 – 5,2), [5,2 – 5,4), [5,4 – 5,6), [5,6 – 5,8), [5,8 – 6), [6 – 6,2), [6,2 – 6,4], với số hộ tương ứng: 5, 20, 65, 105, 100, 50, 15, 5. Hãy tính mức thu nhập bình quân và độ lệch mẫu về mức thu nhập của các hộ này.

X	[4,8 - 5)	[5 – 5,2)	[5,2-5,4)	[5,4-5,6)	[5,6-5,8)	[5,8-6)	[6-6,2)	[6,2-6,4}	Σ
Tần số	5	20	65	105	100	50	15	5	365
x_i	4,9	5,1	5,3	5,5	5,7	5,9	6,1	6,3	

Chú ý: Nếu biến quan sát X là vector k chiều thì dữ liệu thu được ở mẫu kích thước n là n vector k chiều: $X_1, X_2, \dots, X_n$. Khi đó vector trung bình mẫu $\bar{X}$ và ma trận hiệp phương sai mẫu $S^2(X)$ có công thức tính tương tự như một chiều:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i\ ;\ S^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}).(X_i - \bar{X})^T = \frac{1}{n}\hat{X}.(\hat{X})^T$$

Trong đó các vector được viết theo ma trận cột, các phép toán cộng, nhân là cộng, nhân các ma trận, $(\hat{X})^T = (\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n)$, với $\hat{X}_i = X_i - \bar{X}$

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Ước lượng khoảng tin cậy (KTC) cho các tham số:

Các khái niệm

✓ **Đặt vấn đề:** Giả sử θ là một giá trị chân thực chưa biết, nhưng cần biết mà lại không thể biết chính xác. Khi đó phải tìm một đại lượng $\hat{\theta}$ để xấp xỉ cho θ, ta nói $\hat{\theta}$ là một ước lượng cho θ. Vậy tìm $\hat{\theta}$ như thế nào? Yêu cầu: $\hat{\theta}$ phải phù hợp với vai trò và ý nghĩa cũng như những thông tin có được về θ .

✓ **Phương pháp chung giải quyết vấn đề:** Trước hết cần phải có thông tin về θ, nên ta cần chỉ ra biến quan sát, ta ký hiệu là X, có liên quan đến θ mà θ đóng vai trò là tham số của phân phối xác suất. Việc lấy thông tin ở đây có nghĩa là lập mẫu điều tra về biến quan sát X, những thông tin này cũng là những thông tin về θ. Trên mẫu ta có đại lượng $\hat{\theta}$ tương ứng với θ(Vai trò của $\hat{\theta}$ trên mẫu tương tự như vai trò của θ đối với biến q.sát X). Ta dùng đại lượng $\hat{\theta}$ để ước lượng cho θ.
Chú ý: $\hat{\theta}$ phụ thuộc vào mẫu: $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ nên là biến ngẫu nhiên phụ thuộc cỡ mẫu n.

✓ **Ước lượng không chệch và ước lượng vững**
Tính không chệch: Ước lượng $\hat{\theta}$ được gọi là không chệch cho θ nếu: $E\hat{\theta} = \theta$
Tính vững: $\hat{\theta}$ được gọi là vững cho θ nếu: $\forall \varepsilon > 0, \lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$
Chú ý: $S^2(X)$ là ước lượng chệch cho σ^2, vì: $ES^2(X) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. Tuy nhiên, nếu xét đại lượng: $S'^2(X) = \frac{n}{n-1}S^2(X)$, thì $S'^2(X)$ là ước lượng vững, không chệch cho phương sai tổng thể σ^2. Ta gọi $S'^2(X)$ là *phương sai mẫu điều chỉnh* của biến quan sát X.
* Lưu ý, trong các phần mềm và một số tài liệu lại định nghĩa phương sai mẫu của biến quan sát X là $S^2(X) = \frac{n}{n-1}\{\overline{X^2} - (\bar{X})^2\}$, tuy nhiên đại lượng này lại không có tính chất tương tự như phương sai tổng thể.

✓ **Khoảng tin cậy cho tham ẩn**
Nếu ta chỉ ra một khoảng ngẫu nhiên: $(\hat{\theta}_1, \hat{\theta}_2)$ mà tham ẩn θ có thể rơi vào với xác suất $\gamma = P\{\theta \in (\hat{\theta}_1, \hat{\theta}_2)\}$ đủ lớn, thì ta gọi $(\hat{\theta}_1, \hat{\theta}_2)$ là KTC cho θ với độ tin cậy γ và bán kính của KTC được gọi là độ chính xác của ước lượng

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Ước lượng KTC cho các tham số của tổng thể

Giả sử $(X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên kích thước n về biến quan sát X trên tổng thể Ω. Trong đó trung bình tổng thể $\mu = EX$, phương sai tổng thể $\sigma^2 = \mathrm{Var}X$ và tỷ lệ tính chất A trong tổng thể là $p = P(A)$ chưa biết, cần tìm KTC với độ tin cậy $\gamma = 1 - \alpha$.

KTC cho trung bình tổng thể $\mu = EX$

KTC một phía: $(-\infty;\ \overline{\boldsymbol{X}} + \boldsymbol{\varepsilon})$; (KTC bên trái, $\overline{\boldsymbol{X}} + \boldsymbol{\varepsilon}$ gọi là ước lượng tối đa cho μ)

$(\overline{\boldsymbol{X}} - \boldsymbol{\varepsilon};\ +\infty)$ (KTC bên trái, $\overline{\boldsymbol{X}} - \boldsymbol{\varepsilon}$ gọi là ước lượng tối thiểu cho μ)

Trong đó: $\boldsymbol{\varepsilon} = \begin{cases} \boldsymbol{u(\alpha)}\frac{\sigma}{\sqrt{n}}, \text{nếu biết phương sai tổng thể } \sigma^2 \\ \boldsymbol{t_{n-1}(\alpha)}\frac{S(X)}{\sqrt{n-1}}, \boldsymbol{n}\text{ếu chưa biết } \sigma^2 \text{ và cỡ mẫu n} < 30 \\ \boldsymbol{u(\alpha)}\frac{S(X)}{\sqrt{n-1}}, \text{nếu chưa biết } \sigma^2 \text{ và cỡ mẫu n} \geq 30 \end{cases}$

với: $\boldsymbol{u(\lambda)}$ là phân vị mức $1 - \boldsymbol{\lambda}$ của phân phối chuẩn; $\boldsymbol{t}_{\mathrm{k}}(\boldsymbol{\lambda})$: phân vị mức $1 - \boldsymbol{\lambda}$ của phân phối Student với k bậc tự do. $S(X) = \sqrt{S^2(X)}$ ($S^2(X)$ *là phương sai mẫu*)

KTC đối xứng:

$$(\overline{\boldsymbol{X}} - \boldsymbol{\varepsilon}, \overline{\boldsymbol{X}} + \boldsymbol{\varepsilon})\ (\boldsymbol{\varepsilon} = \begin{cases} u\left(\frac{\alpha}{2}\right).\frac{\sigma}{\sqrt{n}}, \textit{nếu biết phương sai tổng thể } \sigma^2 \\ t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S(X)}{\sqrt{n-1}}, \textit{nếu chưa biết } \sigma^2, \textit{với cỡ mẫu } n < 30 \\ u\left(\frac{\alpha}{2}\right).\frac{S(X)}{\sqrt{n-1}}, \textit{nếu chưa biết } \sigma^2, \textit{với cỡ mẫu } n \geq 30 \end{cases})$$

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Chú ý: Nếu dùng phương sai mẫu điều chỉnh: $S'^2(X) = \frac{n}{n-1}.S^2(X)$ (hay có: $S^2(X) = \frac{n-1.}{n}S'^2(X)$ thì trong công thức của ε, biểu thức $\frac{S(X)}{\sqrt{n-1}} = \frac{S'(X)}{\sqrt{n}}$

KTC cho tỷ lệ:

Lập mẫu ngẫu nhiên kích thước n để quan sát tính chất A, trên đó ta có: f = $f(A)$ là tỷ lệ tính chất A trên mẫu với độ tin cậy $\gamma = 1 - \alpha$, bài toán tìm KTC cho p = P(A) là tỷ lệ t/c A trong tổng thể Ω chỉ xét dựa trên mẫu có kích thước n khá lớn và được giải quyết trong các trường hợp như sau:

1. Khi n lớn và f không quá gần 0 hoặc 1: KTC cho p là:

$$(\boldsymbol{f - \varepsilon, f + \varepsilon}), (\text{với } \varepsilon = u\left(\frac{\alpha}{2}\right).\sqrt{\frac{f(1-f)}{n}})$$

2. *Khi n lớn và f gần 0 hoặc 1,với độ tin cậy* 95%*:*

KTC cho $\lambda = np$ là: $(\boldsymbol{np_1; np_2})$: Với m từ mẫu, tra bảng p.lục, tìm được: $np_1; np_2$. Suy ra KTC cho p là: $(\boldsymbol{p_1; p_2})$

KTC cho phương sai tổng thể σ^2: Với độ tin cậy $\gamma = 1 - \alpha$, KTC cho σ^2 là:

KTC đối xứng cho σ^2 là**:**

$$\left(\frac{\boldsymbol{nS^2(X)}}{\chi^2_{n-1}\left(\frac{\alpha}{2}\right)}, \frac{\boldsymbol{nS^2(X)}}{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)}\right)$$

Các KTC một phía cho phương sai là:

$$\left(\boldsymbol{0}, \frac{\boldsymbol{nS^2(X)}}{\chi^2_{n-1}(1-\alpha)}\right)(\textit{bên trái}); \left(\frac{\boldsymbol{nS^2(X)}}{\chi^2_{n-1}(\alpha)}; +\infty\right)(\text{bên phải})$$

trong đó $\chi^2_k(\lambda)$ là giá trị tới hạn mức λ của phân phối khi – bình phương với k bậc tự do, tra từ bảng phụ lục của phân phối khi – bình phương

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Một số phương pháp ước lượng tham số của mô hình

Nhiều mô hình Machin Learning được xây dựng dựa trên các mô hình thống kê. Các mô hình thống kê dựa vào các mô hình phân phốixác suất được đề cập trong chương 4. Gọi θ tập các tham số của một mô hình thống kê. Chẳng hạn với phân phối Poisson, tham $\theta = \lambda$ là trung bình tổng thể, với phân phối chuẩn n chiều, tham θ gồm vector kì vọng μ và ma trận hiệp phương sai. Learning chính là quá trình ước lượng các tham số θ sao cho mô hình ước lượng có được phù hợp nhất với phân phối của dữ liệu.

Phương pháp bình phương tối thiểu (OLS: *Ordinary Least Squares*)

Được biết biến Y có quan hệ hàm số với X (một biến số hoặc một biến vector): Y = Y(X) (Y(X) chưa biết) và qua quan sát có dữ liệu: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Ta muốn xấp xỉ Y(X) với một dạng hàm $f(x, a_1, a_2, \dots, a_m)$ đã biết, trong đó $a_1, a_2, \dots, a_m$ là các tham số chưa biết (ví như dạng đa thức bậc 3: $f(x) = ax^3 + bx^2 + cx + d$ có 4 tham số a, b, c, d chưa biết). Pháp pháp OLS dựa vào dữ liệu quan sát để tìm ước lượng cho các tham số $a_1, a_2, \dots, a_m$ sao cho tổng bình phương các sai số giữa các giá trị ước lượng và các giá trị quan sát được của hàm Y(x) là bé nhất, tức là:

$$F(a_1, a_2, \dots, a_m) = \sum_{j=1}^{n} \{f(x_j, a_1, a_2, \dots, a_m) - y_j\}^2 \longrightarrow \min$$

Vậy các ước lượng cho các tham số $a_1, a_2, \dots, a_m$ là nghiệm của hệ phương trình:

$$\frac{\partial F}{\partial a_k} = 0, \forall k = 1,2, \dots, m \qquad (1)$$

VD1. Giả sử từ bảng n dữ liệu quan sát về mối quan hệ hàm của biến Y và biến số X được cho ở trên, ta tìm hàm số f(x) = a + bx x ấp xỉ cho hàm Y(x).

Giải. Có: $F(a, b) = \sum_{j=1}^{n} \{f(x_j, a, b) - y_j\}^2 = \sum_{j=1}^{n} \{a + bx_j - y_j\}^2$. Theo OLS, ta tìm a, b từ hệ phương trình:

$$\begin{cases} \frac{\partial F}{\partial a} = 0 \\ \frac{\partial F}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{j=1}^{n} \{a + bx_j - y_j\} = 0 \\ \sum_{j=1}^{n} \{a + bx_j - y_j\} x_j = 0 \end{cases} \Leftrightarrow \begin{cases} a = \hat{a} = \overline{Y} - \hat{b}.\overline{X} \\ b = \hat{b} = \frac{\overline{X.Y} - \overline{X}.\overline{Y}}{\overline{X^2} - (\overline{X})^2} \end{cases} \quad (*)$$

Thay a, b từ (*) vào, ta có: $A = \frac{\partial^2 F}{\partial a^2} > 0$; $\Delta = \frac{\partial^2 F}{\partial a^2}.\frac{\partial^2 F}{\partial b^2} - \frac{\partial^2 F}{\partial a \partial b} > 0$, nên (a,b) là điểm cực tiểu duy nhất của F, tức là F đạt trị nhỏ nhất tại (a,b) xác định bởi (*). Do đó hàm $f(x) = a + bx$, với (a,b) xác định từ (*) là hàm ước lượng cần tìm.

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Phương pháp Hợp lý cực đai ML *(MLE: Maximum Likelihood Estimate)*

Giả sử biến X có phân phối xs phụ thuộc vào tham vector $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ chưa biết, cần tìm ước lượng cho θ. Với $W_n = (X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên của biến quan sát X. Ký hiệu $q(x, \theta)$ là xác suất để $W_n = (X_1, X_2, \dots, X_n)$ nhận giá trị $x = (x_1, \dots, x_n)$: $q(x, \theta) = P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1). \dots P(X_n = x_n)$ nếu X là biến ngẫu nhiên rời rạc, và $q(x, \theta)$, là giá trị hàm mật độ xác suất của vector $W_n = (X_1, \dots, X_n)$ tại x nếu X là biến liên tục.
Ta gọi hàm $L(\theta) = q(X_1, X_2, \dots, X_n, \theta)$ là hàm hợp lý và ước lượng $\hat{\theta}$ mà tại đó hàm hợp lý $L(\theta) = q(X_1, X_2, \dots, X_n, \theta)$ đạt trị lớn nhất, là ước lượng hợp lý cực đại hay ước lượng hợp lý nhất, và nó là nghiệm hệ phương trinh sau mà ta gọi là hệ p.trình hợp lý:

$$\frac{\partial l(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_j} = 0, \forall j = 1,2, \dots, k \qquad (2a)$$

Phương pháp hợp lý cực đại, ký hiệu là p.pháp ML (Maximum Liklihood) là phương pháp tìm ước lượng $\hat{\theta}$ cho tham ẩn θ là tìm nghiệm $\hat{\theta}$ của hệ phương trình hợp lý sao cho tại đó hàm hợp lý đạt trị lớn nhất.

Vì $L(\theta)$ cùng tính đơn điệu với hàm Log $L(\theta)$, nên hệ p.trình hợp lý được thay bởi:

$$\frac{\partial \text{Log } l(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_j} = 0, \forall j = 1,2, \dots, k \qquad (2b)$$

VD2. Với $W_n = (X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên của biến quan sát $X \sim N(\mu, \sigma^2)$, trong đó μ và σ^2 là các tham số chưa biết. Cần tìm ước lượng ML cho μ và σ^2
Giải. Đặt $t = \sigma^2$, cần tìm ước lượng hợp lý cực đại cho tham vector: $\theta = (\mu, t)$.

X có hàm mật độ xác suất là : $f(x) = \frac{1}{\sqrt{2\pi t}} . e^{-\frac{1}{2}\left(\frac{x-\mu}{t}\right)^2}$, do đó hàm hợp lý là:

$$L(\theta) = L(\mu, t) = \prod_{j=1}^{n} f(x_j) = (2\pi)^{-\frac{n}{2}} t^{-\frac{n}{2}} . \exp\left\{-\frac{1}{2t}\sum_{j=1}^{n}(x_j - \mu)^2\right\}$$

$$\text{Log}(L(\theta)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log t - \frac{1}{2t}\sum_{j=1}^{n}(x_j - \mu)^2$$

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Hệ phương trình hợp lý:

$$\begin{cases} \frac{\partial \mathrm{Log}\, l(\theta)}{\partial \mu} = 0 \\ \frac{\partial \mathrm{Log}\, l(\theta)}{\partial t} = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{1}{t}\sum_{j=1}^{n}(x_j - \mu) = 0 \\ -\frac{n}{2t} + \frac{1}{2t^2}\sum_{j=1}^{n}(x_j - \mu)^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \mu = \hat{\mu} = \overline{X} = \frac{1}{n}\sum_{j=1}^{n} x_j \\ t = \hat{t} = S^2(X) = \overline{X^2} - (\overline{X})^2 \end{cases}$$

Như vậy đối với biến ngẫu nhiên có phân phối chuẩn thì trung bình mẫu $\overline{X}$ là ước lượng hợp lý nhất cho giá trị trung bình tổng thể $\mu = EX$ và phương sai mẫu $S^2(X)$ là ước lượng hợp lý nhất cho phương sai tổng thể $\sigma^2 = \mathrm{Var}(X)$.

VD3. Tìm ước lượng ML cho véc tơ kỳ vọng μ và ma trận hiệp phương sai Λ của véc tơ ngẫu nhiên X có phân phối chuẩn k chiều.

Giải. X có mật độ chuẩn k chiều:

$$f(x) = f(x_1, x_2, \dots, x_k) = \frac{1}{(2\pi)^{k/2}.(\det\Lambda)^{1/2}} exp\left\{\frac{1}{2}.(x-\mu)^T\Lambda^{-1}.(x-\mu)\right\}$$

Khi đó từ mẫu ngẫu nhiên kích thước n $W = (X_1, X_2, \dots, X_n)$, ta có hàm hợp lý:

$$p(X_1, X_2, \dots, X_n) = \prod_{j=1}^{n} f(X_j) = \frac{1}{(2\pi)^{kn/2}.(\det\Lambda)^{n/2}} \prod_{j=1}^{n} exp\left\{\frac{1}{2}.(X_j-\mu)^T\Lambda^{-1}.(X_j-\mu)\right\}$$

$$\ln p(X_1, X_2, \dots, X_n) = \ln\frac{1}{(2\pi)^{kn/2}.(\det\Lambda)^{n/2}} + \sum_j \left\{\frac{1}{2}.(X_j-\mu)^T\Lambda^{-1}.(X_j-\mu)\right\}$$

Hệ phương trình hợp lí: $\begin{cases} \nabla_\mu \ln p(X_1, X_2, \dots, X_n) = 0 \\ \nabla_{\Lambda^{-1}} \ln p(X_1, X_2, \dots, X_n) = 0 \end{cases}$ (*)

Tính: $\nabla_\mu \left\{(X_j-\mu)^T\Lambda^{-1}.(X_j-\mu)\right\} = -(\Lambda^{-1} + (\Lambda^{-1})^T).(X_j-\mu)$

$$\nabla_{\Lambda^{-1}} \left\{(X_j-\mu)^T\Lambda^{-1}.(X_j-\mu)\right\} = (X_j-\mu)(X_j-\mu)^T \qquad \nabla_{\Lambda^{-1}} \left\{\ln\frac{1}{(2\pi)^{kn/2}.(\det\Lambda)^{n/2}}\right\} = \nabla_{\Lambda^{-1}}\left(\frac{n}{2}.\ln(\det\Lambda^{-1})\right) = \frac{n}{2}.\Lambda$$

$$(*) \Leftrightarrow \begin{cases} (\Lambda^{-1} + (\Lambda^{-1})^T)\{\sum_j X_j - n\mu\} = 0 \\ \frac{n}{2}.\Lambda + \frac{1}{2}\sum_j (X_j-\mu)(X_j-\mu)^T = 0 \end{cases} \Leftrightarrow \begin{cases} \mu = \hat{\mu} = \frac{1}{n}\sum_j X_j = \overline{X} \\ \Lambda = \hat{\Lambda} = \frac{1}{n}\sum_j (X_j-\mu)(X_j-\mu)^T \end{cases} (**)$$

Vậy ước lượng ML cho véc tơ kì vọng μ và ma trận hiệp phương sai Λ của phân phối chuẩn nhiều chiều là $\hat{\mu}$ và $\hat{\Lambda}$ được cho bởi (**)

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Ví dụ 4. Tìm ước lượng ML cho các tham số $\lambda_1, \lambda_2, \dots, \lambda_k$ của phân phối phân loại (categorical distribution)

Giải. Xét X là biến ngẫu nhiên có phân phối phân loại với các tham số $\lambda_1, \lambda_2, \dots,$

λ_k: $\boldsymbol{P(X = i) = \lambda_i, i = 1,2, \dots, k.}$

Việc quan sát biến ngẫu nhiên X có thể được mô hình hóa bởi việc tung một khối đa diện lồi có k mặt được đánh số từ 1 đến k với xác suất để mặt i (mặt có chữ số i) tiếp xúc với mặt phẳng nền khi rơi xuống là λ_i, $i = 1,2, \dots, k$. Tiến hành n lần tung khối đa diện này. kí hiệu X_i là số lần mặt i tiếp xúc với mặt phẳng sàn, thì $X_i \sim B(n, \boldsymbol{\lambda_i})$ và vector $(X_1, X_2, \dots, X_k)$ có phân phối đa thức với các tham số n, $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$, tức là ta có hàm hợp lí:

$$L(\lambda_1, \lambda_2, \dots, \lambda_{k-1}) = \frac{n!}{X_1! X_2! \dots, X_k!} . \lambda_1^{X_1} \lambda_2^{X_2} \dots \lambda_k^{X_k} \text{ (Lưu ý: } \sum_{j=1}^{k} \lambda_j = 1, \sum_{j=1}^{k} X_j = n)$$

$$\log L(\lambda_1, \lambda_2, \dots, \lambda_{k-1}) = \log \frac{n!}{X_1! X_2! \dots, X_k!} + \sum_{j=1}^{k-1} X_j . \log \lambda_j + X_k . \log \left(1 - \sum_{j=1}^{k-1} \lambda_j\right)$$

$\frac{\partial \log L(\lambda_1, \lambda_2, \dots, \lambda_{k-1})}{\partial \lambda_i} = \frac{X_i}{\lambda_i} - X_k . \frac{1}{\left(1 - \sum_{j=1}^{k-1} \lambda_j\right)}, i = 1,2, \dots, k-1.$ Có hệ phương trình hợp lí:

$\frac{\partial \log L(\lambda_1, \lambda_2, \dots, \lambda_{k-1})}{\partial \lambda_i} = 0, \left(\forall i = \overline{1, k-1}\right) \Leftrightarrow X_k . \lambda_i = X_i . \lambda_k, \forall i = \overline{1, k-1}$ **(*)**

Từ (*) lấy tổng 2 vế theo i, nhận đươc: $X_k . (1 - \lambda_k) = (n - X_k) . \lambda_k$, nhận được: $\lambda_k = \frac{X_k}{n}$ **(**).** Thay (**) vào (*), nhận được:

$$\lambda_i = \frac{X_i}{n}, \forall i = \overline{1, k}$$

Bài tập. Tìm ước lượng hợp lý cực đại (ML) cho tham số p trong phân phối nhị thức B(n, p), tham số λ của phân phối P_λ, tham số λ của phân phối mũ E_λ

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Phương pháp ước lượng hậu nghiệm cực đại

Phương pháp MLE ước lượng tập tham số θ trong phân bố của biến quan sát X là cực đại hóa hàm hợp lí dựa trên mẫu điều tra về X. Vì chỉ phụ thuộc mẫu về X nên khi mẫu không đủ tính đại diện (chẳng hạn cỡ mẫu bé) thì ước lượng MLE kém hiệu quả. Trong khi phương pháp ước lượng hậu nghiệm (MAP: *Maximum A Posteriori Esitimation*) coi θ là một biến ngẫu nhiên,không những dựa vào mẫu về X mà còn dựa vào thông tin đã biết về phân bố của θ là $p(\theta)$. Khi đã có mẫu $(X_1, X_2, \dots, X_n)$ thì thông tin về θ là $p(\theta|(X_1, X_2, \dots, X_n))$, biểu thức này gọi là xác suất hậu nghiệm. Phương pháp tìm ước lượng cho θ là cực đại xác suất hậu nghiệm gọi là phương pháp ước lượng hậu nghiệm cực đại:

$$\theta = \arg\max p(\theta|(X_1, X_2, \dots, X_n))$$

Ta có:
$$p(\theta|(X_1, X_2, \dots, X_n)) \to \max \Leftrightarrow \frac{p((X_1, X_2, \dots, X_n)|\theta)p(\theta)}{(X_1, X_2, \dots, X_n)} \to \max$$
$$\Leftrightarrow p((X_1, X_2, \dots, X_n)|\theta)p(\theta) \to \max \Leftrightarrow \prod_{i=1}^{n} p(X_i|\theta) \,.\, p(\theta) \to \max$$

Chú ý:
$p((X_1, X_2, \dots, X_n)|\theta) = L(\theta)$ là hàm hợp lí, còn $p(\theta)$ được gọi là xác suất tiên nghiệm.
Khác biệt giữa MAP và MLE là hàm mục tiêu của MAP có thêm phân phối $p(\theta)$ của θ, phân phối này chính là những thông tin biết trước về θ, nên gọi là tiên nghiệm.

Chọn tiên nghiệm: Để chọn tiên nghiệm, cần lưu ý các k/niệm sau:

1. Tiên nghiệm liên hợp: Nếu phân phối hậu nghiệm $p(\theta|(X_1, X_2, \dots, X_n))$ có cùng dạng với phân phối tiên nghiệm $p(\theta)$ thì hai phân phối này được gọi là cặp phân phối liên hợp và $p(\theta)$ được gọi là tiên nghiệm liên hợp của hàm hợp lí $p((X_1, X_2, \dots, X_n)|\theta)$

- Nếu hàm hợp lí và tiên nghiệm cho vector kỳ vọng là các phân phối chuẩn thì phân phối hậu nghiệm cũng là phân phối chuẩn. Ta nói phân phối chuẩn liên hợp với chính nó (tự liên hợp).
- Nếu hàm hợp lí là phân phối chuẩn và tiên nghiệm cho phương sai là phân phối Gamma, phân phối hậu nghiệm cũng là phân phối chuẩn thì ta nói p.phối Gamma là tiên nghiệm liên hợp cho phương sai của phân phối chuẩn.
- Phân phối Beta là liên hiệp của phân phối Bernoulli.
- Phân phối Dirichlet là liên hợp của phân phối Catergorical.

CÁC ĐẶC TRƯNG MẪU VÀ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG

Phương pháp ước lượng hậu nghiệm cực đại

Siêu tham số: Khi phân phối tiên nghiệm p(θ) lại phụ thuộc vào các tham số $\alpha, \beta, \ldots$ khác thì các tham số mới này được gọi là các siêu tham số

Ví dụ 5. Cần ước lượng cho tham số θ là xác suất thành công trong của biến X có p.phối Bernoulli: $\theta = P(X = 1)$. Với mẫu cỡ n về biến quan sát X, gọi m là số thành công (số lần xuất hiện (X = 1)). MLE cho ta ước lượng của θ là $\hat{\theta} = \frac{m}{n}$. Ta cần tìm ước lượng MAP cho θ.
Phân phối Bernoulli có mật độ xác suất: $p(x|\theta) = \theta^x(1-\theta)^{1-x}, x \in \{0,1\}$. Liên hợp của nó là phân phối Beta có hàm mật độ xác suất $p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)}.\theta^{\alpha-1}.(1-\theta)^{\beta-1}$.

Như vậy thành phần chứa θ của p(θ) cùng dạng với phân phối Bernoulli và nếu dung phân phối Beta làm tiên nghiệm cho tham θ, không kể hằng số $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)}$ thì $p(\theta|x)$ tỉ lệ với $p(x|\theta).p(\theta)$ hay tỉ lệ với $\theta^{x+\alpha-1}.(1-\theta)^{(1+\beta-x)-1}$ có dạng một phân phối Bernoulli. Vậy phân phối Beta là một tiên nghiệm liên hợp của phân phối Bernoulli và α, β là các siêu tham số.

Để ước lượng cho θ trong phân phối Bernoulli của biên X theo MAP, với tiên nghiệm p(θ) là hàm mật độ phân phối Beta (với 2 siêu tham số α, β), ta có bài toán tối ưu:

$$p(x_1, x_2, \ldots, x_n|\theta).p(\theta) \to \max \Leftrightarrow \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.\theta^{\alpha-1}.(1-\theta)^{\beta-1} \to \max$$

$$\Leftrightarrow \theta^{m+\alpha-1}.(1-\theta)^{n-m+\beta-1} \to \max \Leftrightarrow \theta = \hat{\theta} = \frac{m+\alpha-1}{n+\alpha+\beta-2}$$

Nhận xét: Khi $\alpha = \beta = 1$, nhận được $\hat{\theta} = \frac{m}{n}$ là ước lượng MLE của θ.