Bài giảng 4: Bài toán Phân Loại

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

1. Bài toán phân loại
2. Naive Bayes
3. Discriminant Analysis
4. Hồi quy logistic
5. Đánh giá mô hình phân loại
6. Dữ liệu không cân bằng

1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

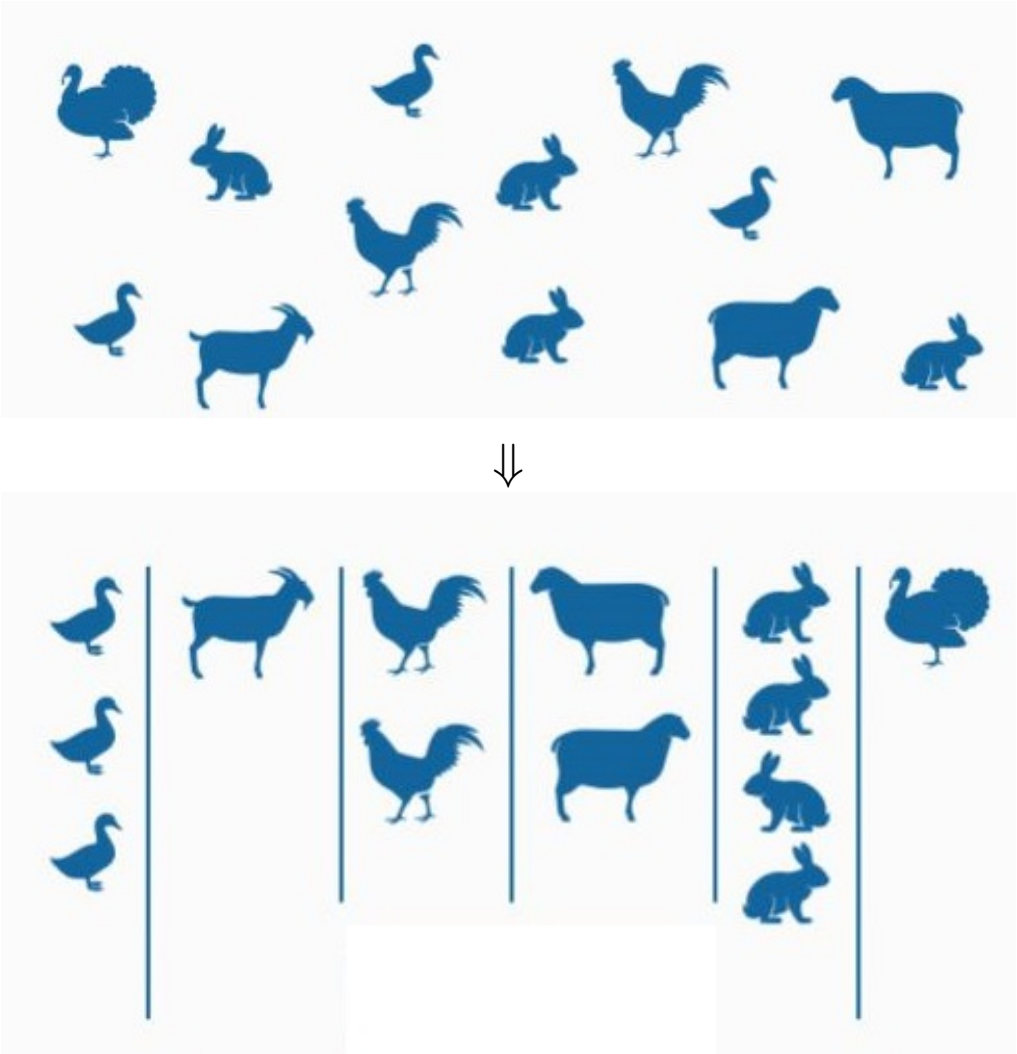
Giới thiệu

Phân loại (**Classification**) là một kỹ thuật được dùng để xử lý dữ liệu thành một số nhóm (lớp) nhất định.

Mục tiêu chính: xác định danh mục hoặc nhóm mà dữ liệu mới sẽ thuộc vào, dựa trên thông tin của một số biến giải thích độc lập:

- Một người đến phòng cấp cứu với một loạt các triệu chứng có thể do một trong ba tình trạng bệnh lý. Bác sĩ cần xác định/dự đoán bệnh trạng nào của bệnh nhân dựa trên thông tin của người đó?
- Dịch vụ ngân hàng trực tuyến phải có khả năng xác định liệu một giao dịch đang được thực hiện trên trang web là gian lận, dựa trên địa chỉ IP của người dùng, lịch sử giao dịch trong quá khứ, v.v.
- Trên cơ sở dữ liệu trình tự DNA của một số bệnh nhân mắc và không mắc một căn bệnh nhất định, một nhà sinh vật học muốn tìm ra đột biến DNA nào là có hại (gây bệnh) và đột biến nào thì không.
- Gmail lọc hộp thư đến của bạn thành "chính", "xã hội", "quảng cáo" hoặc "diễn đàn".

Giới thiệu



Giới thiệu

Có nhiều phương pháp/kỹ thuật phân loại khác nhau đã được phát triển:

- **Naive Bayes**
- **phân tích phân biệt - discriminant analysis**
- **hồi quy logistic - logistic regression**
- **hồi quy multinomial logistic - multinomial logistic regression**
- **phương pháp K-lân cận - K-nearest neighbors**
- **Cây quyết định - Decision Tree**
- **Random Forest**
- **Support Vector Machine**

Phân loại nhị phân

Trong bộ dữ liệu, ta có:

- biến phản hồi $Y = 0$ hoặc 1 (đại diện cho hai nhóm cần phân loại);
- $X_1, X_2, \ldots, X_p$ là p biến giải thích, độc lập nhau, cùng có mối liên hệ với biến phản hồi Y.

Câu hỏi

Dựa vào các giá trị của các biến giải thích $X_1, X_2, \ldots, X_p$ của một đối tượng, một cách nào để phân đối tượng đó vào nhóm $Y = 1$ hoặc nhóm $Y = 0$?

Câu trả lời

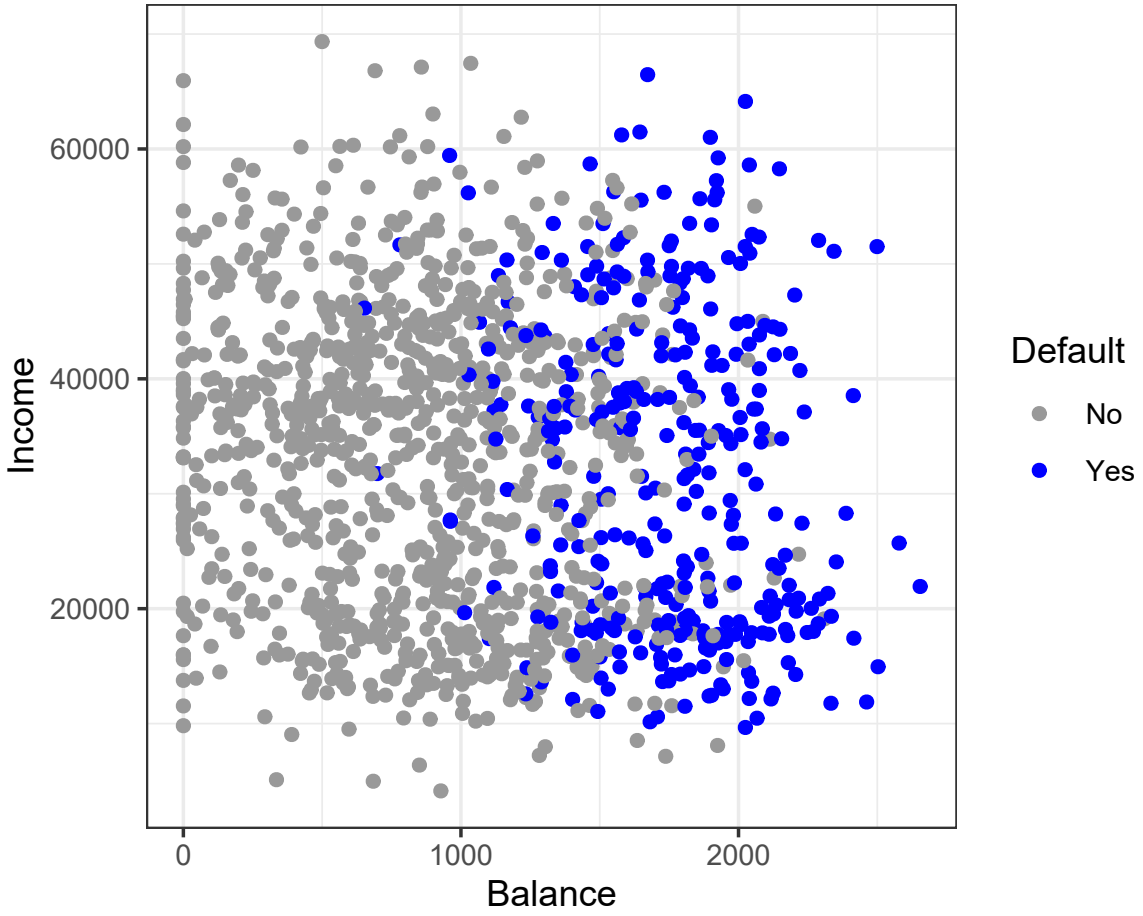
Để phân một đối tượng vào nhóm $Y = 1$, ta có thể tiếp cận theo hướng xác suất, tức là:

$$p(\boldsymbol{x}) = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

với $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$.

- nếu $p(\boldsymbol{x})$ là lớn (tối đa 1), ta có thể chỉ định Y là 1;
- ngược lại, nếu $p(\boldsymbol{x})$ là nhỏ (tối thiểu 0), ta có thể chỉ định Y là 0.

Phân loại nhị phân



Phân loại 1 người có thể bị phá sản tín dụng (`Default`) dựa trên thông tin

- dư nợ tín dụng (`Balance`);
- thu nhập hàng tháng (`Income`).

Phân loại đa nhóm

Trong bộ dữ liệu, ta có:

- biến phản hồi $Y = j$ với $j = 1, \ldots, K$ (đại diện cho K nhóm cần phân loại);
- $X_1, X_2, \ldots, X_p$ là p biến giải thích, độc lập nhau, cùng có mối liên hệ với biến phản hồi Y.

Câu hỏi

Dựa vào các giá trị của các biến giải thích $X_1, X_2, \ldots, X_p$ của một đối tượng, một cách nào để phân đối tượng đó vào nhóm $Y = 1$ hoặc nhóm $Y = j$?

Câu trả lời

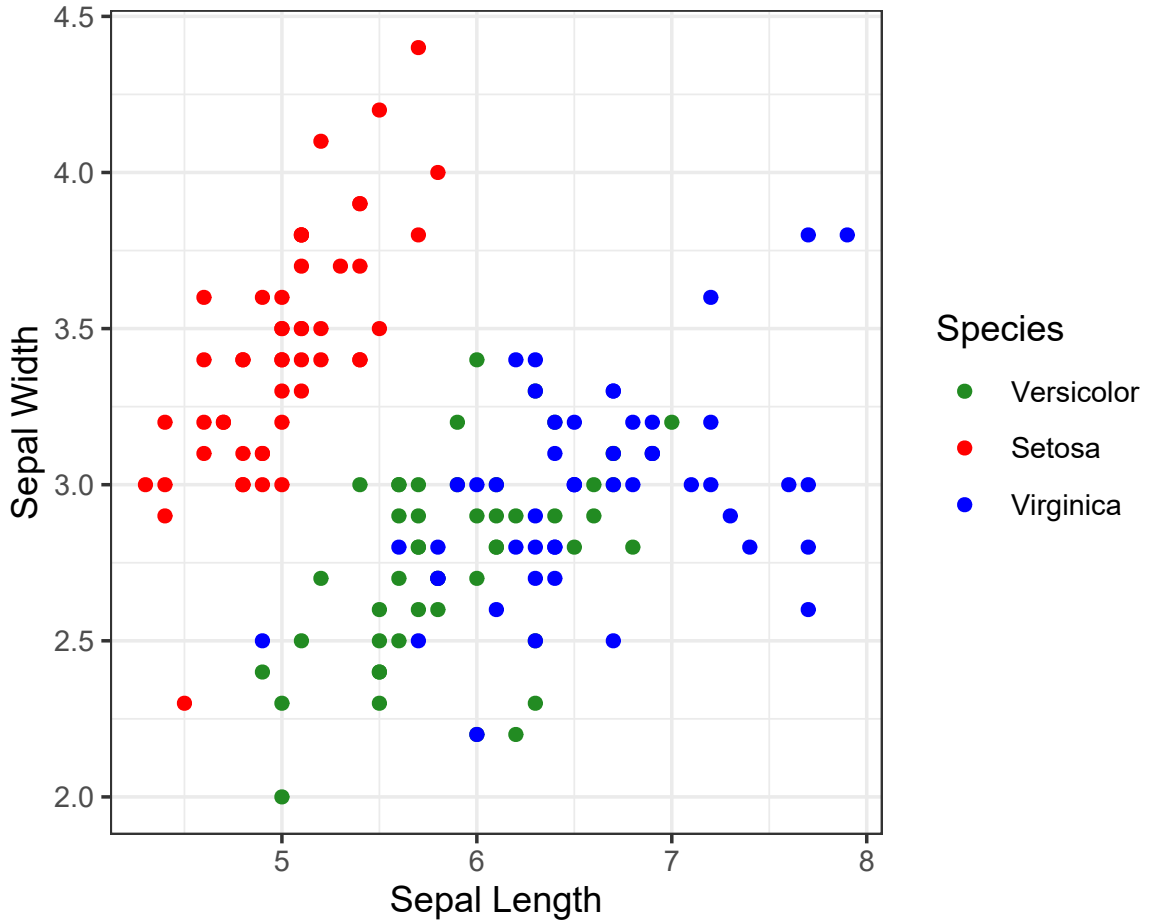
Để phân một đối tượng vào nhóm $Y = j$, ta có thể tiếp cận theo hướng xác suất, tức là:

$$p_j(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

với $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$.

- nếu $p_j(\boldsymbol{x})$ là lớn (tối đa 1), ta có thể chỉ định Y là j;
- ngược lại, nếu $p_j(\boldsymbol{x})$ là nhỏ (tối thiểu 0), ta có thể chỉ định Y là 0.

Phân loại đa nhóm



Phân loại 1 bông hoa vào 1 trong 3 loài hoa Iris dựa trên thông tin

- chiều dài lá đài (`Sepal Length`);
- chiều rộng lá đài (`Sepal Width`).

1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

Công thức Bayes

Công thức Bayes cổ điển

Công thức xác suất Bayes - được đề xuất bởi nhà thống kê học người Anh, Thomas Bayes (1702 - 1761).

Cho không gian mẫu Ω, xét dãy các biến cố $B_1, B_2, \ldots, B_k$ xung khắc nhau từng đôi một, và $B_1 \cup B_2 \cup \ldots \cup B_k = \Omega$ (tức là hệ đầy đủ). Khi đó, xác suất điều kiện của biến cố B_j khi biết một biến cố A bất kỳ, được xác định bởi:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

với j là một số bất kỳ trong $1, 2, \ldots, k$.

Công thức Bayes

Xét trường hợp dữ liệu có hai biến ngẫu nhiên X và Y. Đặt:

- sự kiện A là $X = x$;
- sự kiện B là $Y = j$ (với $j = 1, \ldots, k$).

Khi đó, công thức Bayes được viết lại

$$P(Y = j|X = x) = \frac{P(X = x|Y = j)P(Y = j)}{\sum_{i=1}^{k} P(X = x|Y = i)P(Y = i)}.$$

Một cách tổng quát,

$$P(Y = j|X = x) = \frac{f_j(x)\pi_j}{\sum_{i=1}^{k} f_i(x)\pi_i},$$

với

- $f_j(x) = P(X = x|Y = j)$ là hàm mật độ xác suất của X trên nhóm j;
- $\pi_j = P(Y = j)$ là xác suất (tỷ lệ) thuộc nhóm j.

Phân loại Bayes - Bayes classifier

Trong bộ dữ liệu, ta có:

- biến phản hồi $Y = j$ với $j = 1, \ldots, K$ (đại diện cho K nhóm cần phân loại);
- $X_1, X_2, \ldots, X_p$ là p biến giải thích, độc lập nhau, cùng có mối liên hệ với biến phản hồi Y.

Xác suất phân một đối tượng vào nhóm $Y = j$:

$$p_j(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, \ldots, X_p = x_p) = \frac{f_j(\boldsymbol{x})\pi_j}{\sum_{i=1}^{k} f_i(\boldsymbol{x})\pi_i}$$

với $\boldsymbol{x} = (x_1, \ldots, x_p)$. Ta gọi

- $f_j(\boldsymbol{x})$ là mật độ xác suất đồng thời của $X_1, \ldots, X_p$ trên nhóm j.
- $\pi_j = \Pr(Y = j)$ là xác suất tiên nghiệm (prior probability);
- $p(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, \ldots, X_p = x_p)$ là xác suất hậu nghiệm (posterior probability).

Phân loại Bayes - Bayes classifier

Dựa vào dữ liệu, ta định nghĩa ước lượng $\widehat{p}(\boldsymbol{x})$

$$\widehat{p}_j(\boldsymbol{x}) = \frac{\widehat{f}_j(\boldsymbol{x})\widehat{\pi}_j}{\sum_{i=1}^{k} \widehat{f}_i(\boldsymbol{x})\widehat{\pi}_i},$$

với

- $\widehat{f}_j(\boldsymbol{x})$ là ước lượng mật độ xác suất đồng thời của $X_1, \ldots, X_p$ trên nhóm j.
- $\widehat{\pi}_j$ là ước lượng xác suất tiên nghiệm

$$\widehat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(Y_i = j).$$

Vấn đề: ước lượng $\widehat{f}_j(\boldsymbol{x}) = \widehat{f}_j(x_1, x_2, \ldots, x_p)$?

Hàm mật độ p-chiều (của p biến $X_1, \ldots, X_p$), tính toán phức tạp.

Phương pháp Naive Bayes

Giả sử $X_1, \ldots, X_p$ là độc lập nhau, trong nhóm j, khi đó hàm mật độ p-chiều

$$f_j(\boldsymbol{x}) = f_{j1}(x_1) \times f_{j2}(x_2) \times \ldots \times f_{jp}(x_p)$$

và do đó

$$\widehat{f_j}(\boldsymbol{x}) = \widehat{f_{j1}}(x_1) \times \widehat{f_{j2}}(x_2) \times \ldots \times \widehat{f_{jp}}(x_p).$$

$\hookrightarrow$ ta cần ước lượng từng hàm mật độ thành phần $f_{jl}(x_l)$ trên nhóm j.

Xác suất hậu nghiệm

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f_{j1}}(x_1) \times \widehat{f_{j2}}(x_2) \times \ldots \times \widehat{f_{jp}}(x_p) \times \widehat{\pi}_j}{\sum_{i=1}^{k} \widehat{f_{i1}}(x_1) \times \widehat{f_{i2}}(x_2) \times \ldots \times \widehat{f_{ip}}(x_p) \times \widehat{\pi}_i}$$

Các phương pháp ước lượng

- tham số (parametric estimation): $X_l | Y = j \sim \mathcal{N}(\mu_{jl}, \sigma_{jl}^2)$ - phân phối chuẩn;
- ước lượng histogram;
- ước lượng kernel;
- ước lượng tỷ lệ (khi X_l là biến định tính).

Phương pháp Naive Bayes

Dựa trên ước lượng xác suất hậu nghiệm $\widehat{p}_{j,NB}(\boldsymbol{x})$, ta có ước lượng nhóm như sau:

- $K = 2$

$$\widehat{Y}(\boldsymbol{x}) = \begin{cases} 1 & \text{nếu } \widehat{p}_{1,NB}(\boldsymbol{x}) > p_0, \\ 0 & \text{nếu } \widehat{p}_{1,NB}(\boldsymbol{x}) \leq p_0, \end{cases}$$

  chú ý, $\widehat{p}_{0,NB}(\boldsymbol{x}) = 1 - \widehat{p}_{1,NB}(\boldsymbol{x})$

- $K \geq 3$

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ \widehat{p}_{j,NB}(\boldsymbol{x}),$$

  tức là nhóm j có xác suất hậu nghiệm lớn nhất. Cách làm này được gọi tên là phương pháp MAP (maximum a posterior).

Ví dụ phân loại tín dụng

Phân loại 1 người có thể bị phá sản tín dụng (`Default`) dựa trên thông tin

- dư nợ tín dụng (`Balance`) - X_1;
- thu nhập hàng tháng (`Income`) - X_2;

biến đáp ứng $Y = 0$ (không phá sản), $Y = 1$ (phá sản).

Áp dụng phân loại Bayes

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \widehat{\pi}_j}{\sum_{i=1}^{2} \widehat{f}_{i1}(x_1) \times \widehat{f}_{i2}(x_2) \times \widehat{\pi}_i},$$

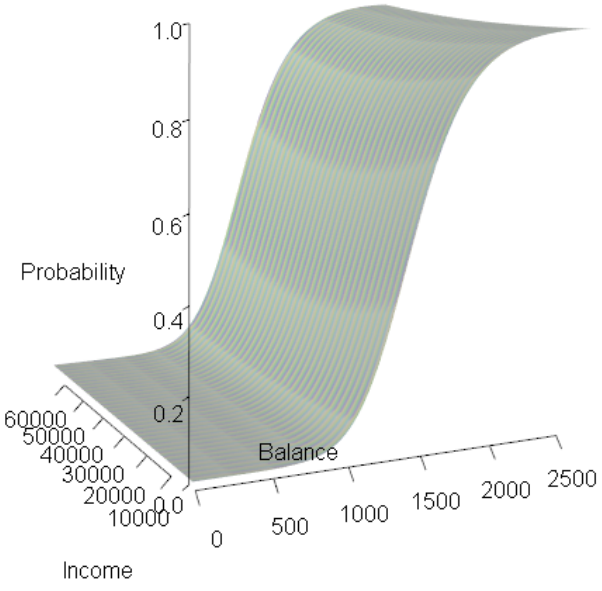
trong đó, $\widehat{f}_{jl}(x_l)$ là ước lượng hàm mật độ của phân phối chuẩn $\mathcal{N}(\widehat{\mu}_{jl}, \widehat{\sigma}^2_{jl})$

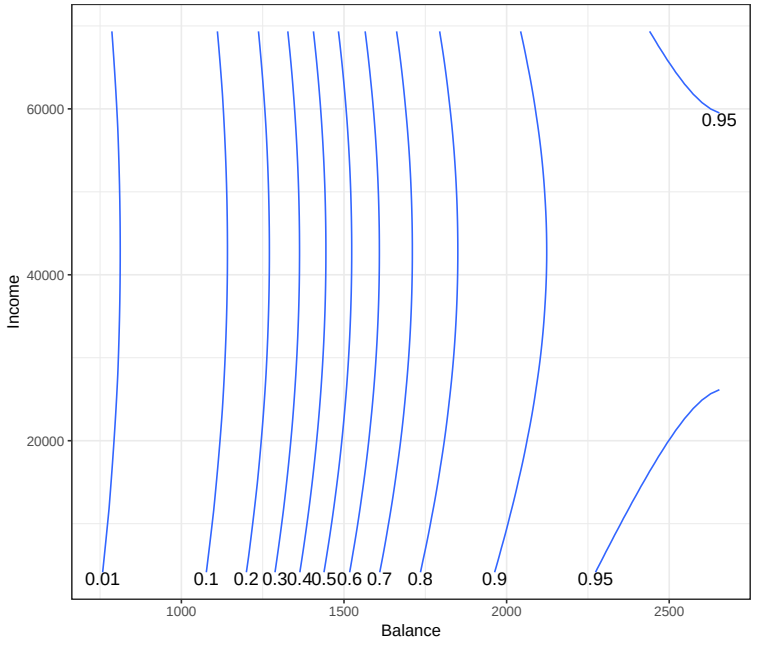
$$\begin{aligned} \widehat{\mu}_{jl} &= \frac{1}{n_j} \sum_{k=1}^{n_j} X_{jlk} \\ \widehat{\sigma}^2_{jl} &= \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (X_{jlk} - \widehat{\mu}_{jl})^2 \end{aligned}$$

Ví dụ phân loại tín dụng

Kết quả của quá trình ước lượng cho ta:

- mặt ước lượng $\widehat{p}_{1,NB}(\boldsymbol{x})$ (hình bên trái);
- đường đồng mức của $\widehat{p}_{1,NB}(\boldsymbol{x})$ (hình bên phải).

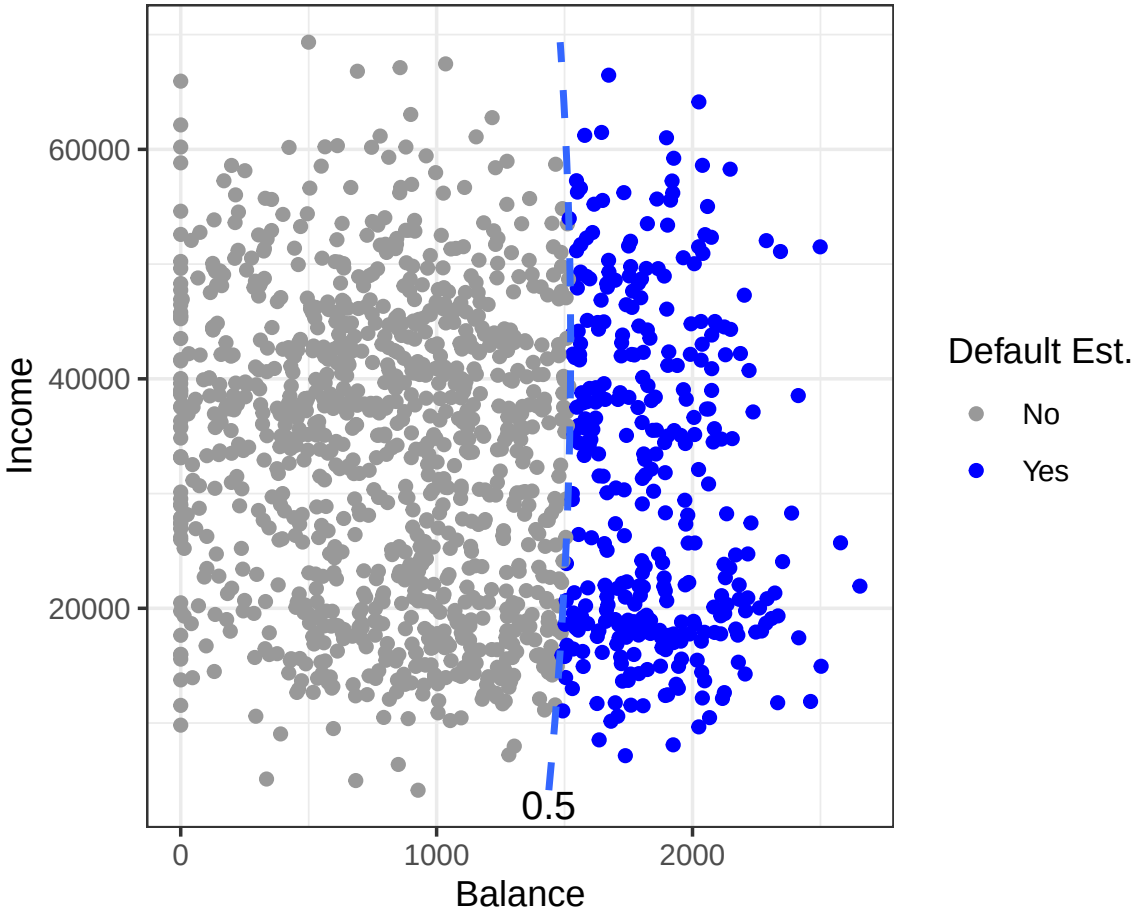




Các đường đồng mức này, gợi ý tới ngưỡng đưa ra quyết định.

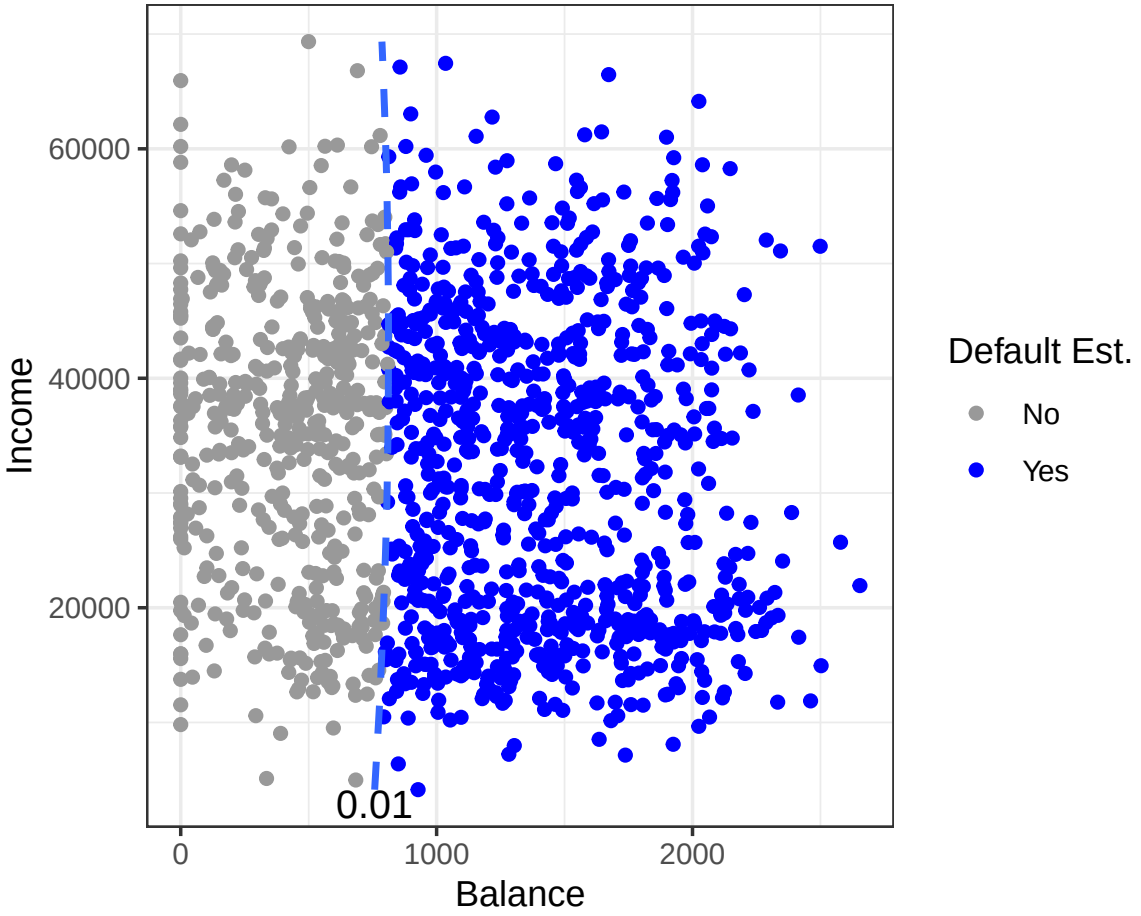
Ví dụ phân loại tín dụng

Ví dụ, chọn $p_0 = 0.5$



Ví dụ phân loại tín dụng

Thay đổi ngưỡng $p_0 \in (0, 1) \Rightarrow \widehat{Y}$ sẽ thay đổi



Ví dụ phân loại loài hoa iris

Phân loại 1 bông hoa iris vào 1 trong 3 loài:

- Versicolor ($j = 1$);
- Setosa ($j = 2$);
- Virginica ($j = 3$).

dựa trên thông tin

- chiều dài lá đài (`Sepal Length`) - X_1;
- chiều rộng lá đài (`Sepal Width`) - X_2.

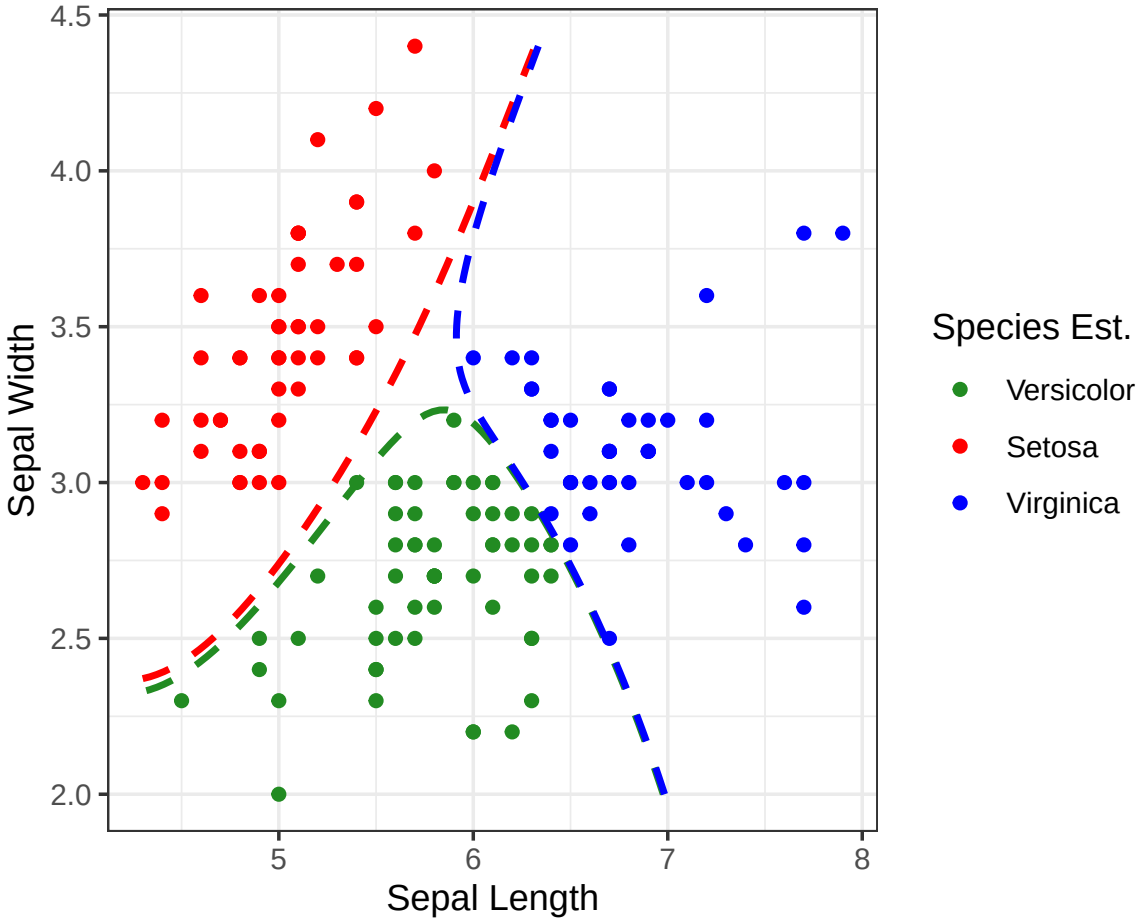
Áp dụng phân loại Bayes

$$\widehat{p}_{j,NB}(\boldsymbol{x}) = \frac{\widehat{f}_{j1}(x_1) \times \widehat{f}_{j2}(x_2) \times \widehat{\pi}_j}{\sum_{i=1}^{3} \widehat{f}_{i1}(x_1) \times \widehat{f}_{i2}(x_2) \times \widehat{\pi}_i},$$

trong đó, $\widehat{f}_{jl}(x_l)$ là ước lượng hàm mật độ của phân phối chuẩn $\mathcal{N}(\widehat{\mu}_{jl}, \widehat{\sigma}^2_{jl})$.

Ví dụ phân loại loài hoa iris

Áp dụng quy tắc MAP, ta có kết quả ước lượng nhóm được minh họa bởi hình



Vấn đề của phương pháp Naive Bayes

Như đã trình bày trong lý thuyết, phương pháp Naive Bayes dựa trên giả định

$$f_j(\boldsymbol{x}) = f_{j1}(x_1) \times f_{j2}(x_2) \times \ldots \times f_{jp}(x_p),$$

tức các $X_1, \ldots, X_p$ là độc lập, trong nhóm j.

$\hookrightarrow$ nếu dữ liệu không ủng hộ cho giả thuyết này $\Rightarrow$ phương pháp Naive Bayes hoạt động tệ.

Ví dụ: xét bài toán phân loại loài hoa iris bằng hai đặc tính:

- chiều dài lá đài (`Sepal Length`) - X_1;
- chiều rộng lá đài (`Sepal Width`) - X_2;

hệ số tương quan của hai đặc tính này: -0.1176.

phân loại loài hoa iris bằng hai đặc tính:

- chiều dài cánh hoa (`Petal Length`) - X_1;
- chiều rộng cánh hoa (`Petal Width`) - X_2;

hệ số tương quan của hai đặc tính này: 0.8718.

1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

Ý tưởng

Nhắc lại phân loại Bayes

$$p_j(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, \ldots, X_p = x_p) = \frac{f_j(\boldsymbol{x})\pi_j}{\sum_{i=1}^{k} f_i(\boldsymbol{x})\pi_i}.$$

Đối với phương pháp Naive Bayes, ta giả định rằng $X_1, \ldots, X_p$ là độc lập trong nhóm j, khi đó

$$f_j(\boldsymbol{x}) = f_{j1}(x_1) \times f_{j2}(x_2) \times \ldots \times f_{jp}(x_p),$$

$\hookrightarrow$ ước lượng $f_j(\boldsymbol{x})$ là tích của các mật độ thành phần $f_{jp}(x_p)$.

Tuy nhiên, giả định này thường bị vi phạm với dữ liệu thực tế.

$\hookrightarrow$ do đó, cần phải ước lượng hàm mật độ xác suất đồng thời $f_j(\boldsymbol{x})$ (biến nhiều chiều).

Ý tưởng

Phương pháp cho bài toán ước lượng hàm mật độ xác suất nhiều chiều:

- ước lượng tham số - phân phối chuẩn nhiều chiều (multivariate normal distribution)

$$f_j(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)^\top \Sigma_j^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)\right),$$

  với Σ_j là ma trận hiệp phương và $\boldsymbol{\mu}_j$ là vector trung bình của $X_1, \ldots, X_p$ trong nhóm j;

$$\Sigma_j = \begin{pmatrix} \sigma_{j1}^2 & \sigma_{j12} & \cdots & \sigma_{j1p} \\ \sigma_{j12} & \sigma_{j2}^2 & \cdots & \sigma_{j2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{j1p} & \sigma_{j2p} & \cdots & \sigma_{jp}^2 \end{pmatrix}$$

  với $\sigma_{j12} = \mathrm{Cov}(X_1, X_2)$ trong nhóm j.

- ước lượng phi tham số - kernel nhiều chiều (multivariate kernel density estimation)

$$\widehat{f}_{j,\mathbf{H}}(\boldsymbol{x}) = \frac{1}{n_j}\sum_{i=1}^{n_j} |\mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{x}_{ji})\right),$$

  với $K(\cdot)$ là hàm Kernel nhiều chiều, ví dụ, multivariate Gaussian kernel.

Ý tưởng

Ước lượng tham số - phân phối chuẩn nhiều chiều

Ta cần ước lượng

- trung bình $\boldsymbol{\mu}_j$;
- ma trận hiệp phương sai Σ_j.

Ước lượng phi tham số - kernel nhiều chiều

Ta cần xác định

- hàm kernel nhiều chiều;
- ma trận bandwidth $\mathbf{H}$.

Khó khăn

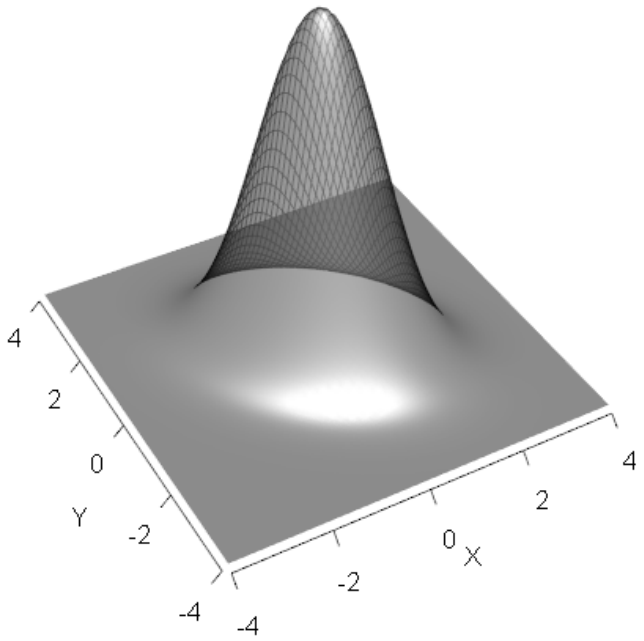
- việc xác định ma trận bandwidth nhiều chiều không đơn giản;
- hiệu quả và độ chính xác của ước lượng bị giảm xuống khi p tăng.

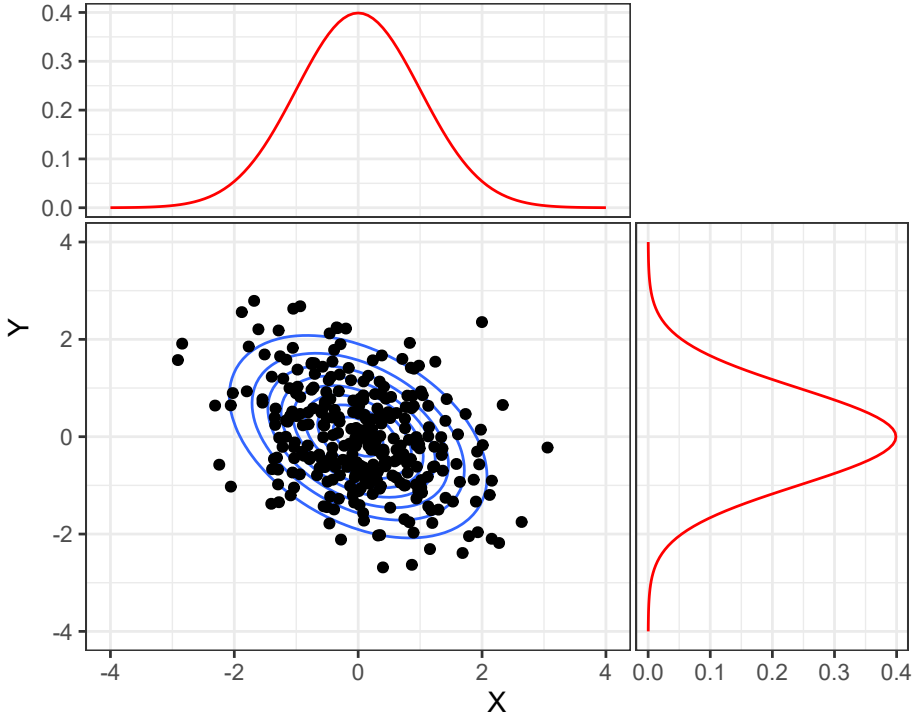
$\hookrightarrow$ sử dụng cách tiếp cận dựa trên phân phối chuẩn nhiều chiều.

Phân phối chuẩn nhiều chiều

Xét

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix} \right)$$





Linear discriminant analysis - LDA

Nếu ta giả sử rằng $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_p$

$\hookrightarrow$ chỉ cần ước lượng một Σ chung cho tất cả p nhóm. Khi này,

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

và phân phối hậu nghiệm

$$p_j(\boldsymbol{x}) = \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)^\top \Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)\right)\pi_j}{\sum_{i=1}^{k} f_i(\boldsymbol{x})\pi_i}.$$

Áp dụng MAP và lấy log() hai vế, ta xác định được quy tắc phân loại

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ \delta_j(\boldsymbol{x}),$$

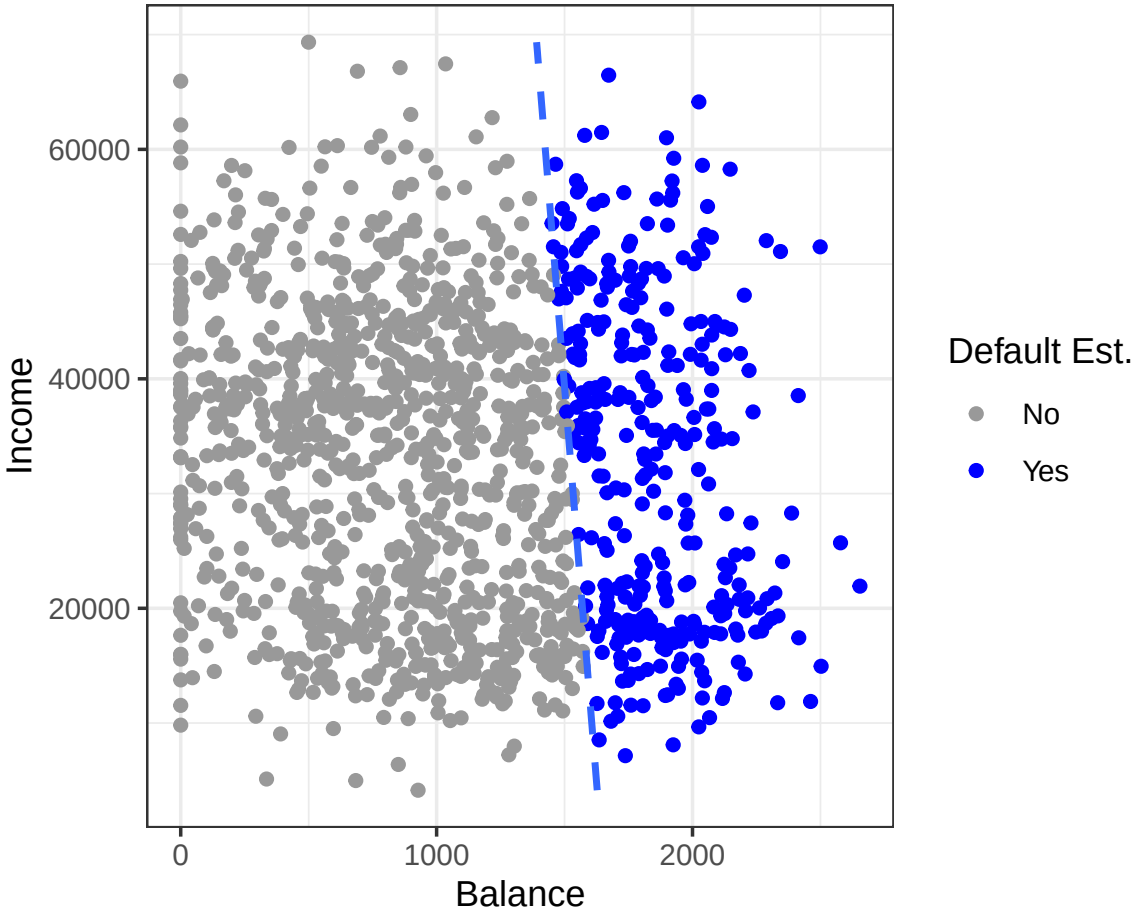
với

$$\delta_j(\boldsymbol{x}) = \boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^\top \Sigma^{-1}\boldsymbol{\mu}_j + \log(\pi_j)$$

tuyến tính theo $\boldsymbol{x}$ $\Rightarrow$ Linear discriminant analysis (LDA).

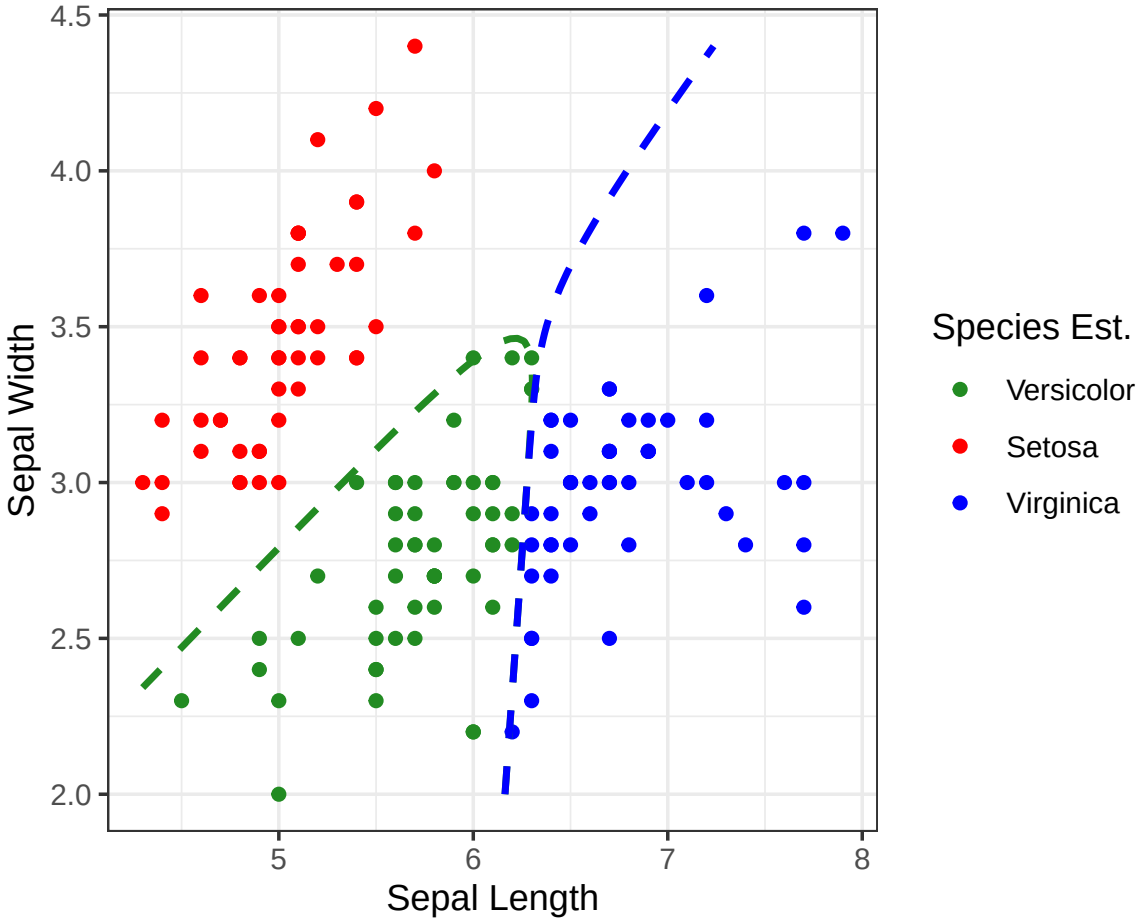
Linear discriminant analysis - LDA

Ví dụ: phân loại 1 người có thể bị phá sản tín dụng



Linear discriminant analysis - LDA

Ví dụ: phân loại 1 bông hoa iris vào 1 trong 3 loài



Quadratic discriminant analysis - QDA

Xét tổng quát, $\Sigma_i \neq \Sigma_k$, khi này, phân phối hậu nghiệm

$$p_j(\boldsymbol{x}) = \frac{\frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)^\top \Sigma_j^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_j\right)\right)\pi_j}{\sum_{i=1}^{k} f_i(\boldsymbol{x})\pi_i}.$$

Áp dụng MAP và lấy log() hai vế, ta xác định được quy tắc phân loại

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ \delta_j(\boldsymbol{x}),$$

với

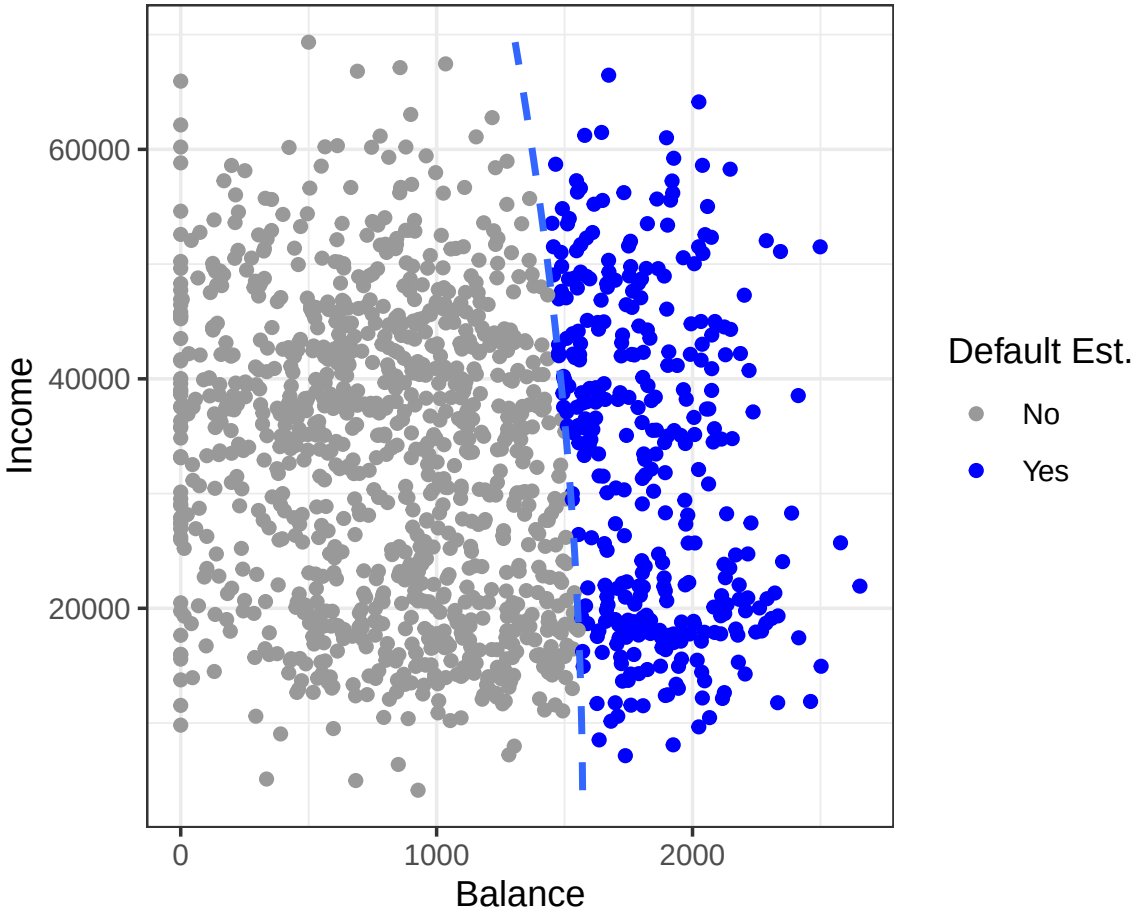
$$\delta_j(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^\top\Sigma_j^{-1}\boldsymbol{x} + \boldsymbol{x}^\top\Sigma_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^\top\Sigma_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\log(|\Sigma_j|) + \log(\pi_j)$$

đa thức bậc hai theo $\boldsymbol{x} \Rightarrow$ Quadratic discriminant analysis (QDA).

LDA là một trường hợp đặc biệt của QDA.

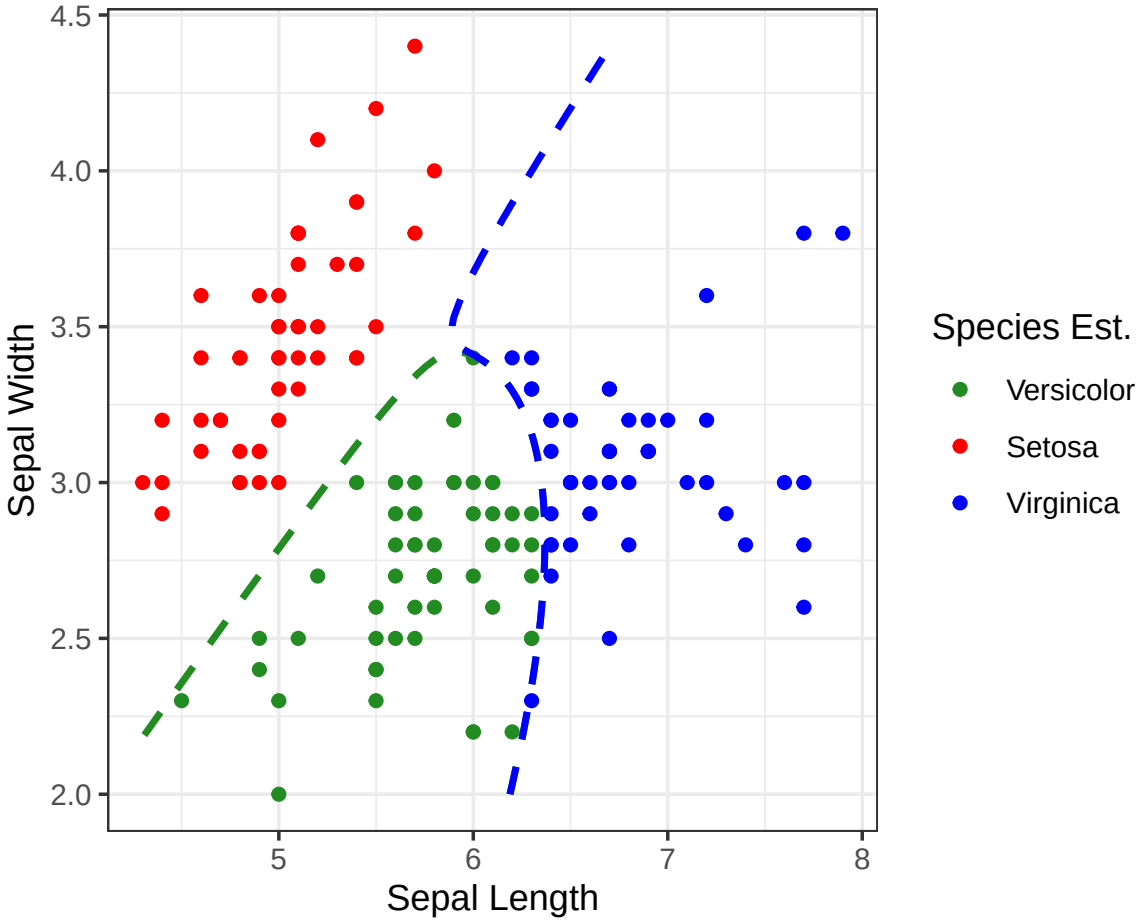
Quadratic discriminant analysis - QDA

Ví dụ: phân loại 1 người có thể bị phá sản tín dụng



Quadratic discriminant analysis - QDA

Ví dụ: phân loại 1 bông hoa iris vào 1 trong 3 loài



1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

Mô hình hồi quy logistic

Nhắc lại rằng, mục tiêu chính trong phân loại là xác định

$$p_j(\boldsymbol{x}) = \Pr(Y = j | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

Trong các phần trước, ta đã xét các phương pháp ước lượng dựa trên

- ý tưởng của công thức Bayes,
- và ước lượng hàm mật độ xác suất.

Theo quan điểm thống kê, có một cách khác để ước lượng (tiên đoán) đó là dự vào mô hình hồi quy.

Xét trường hợp $K = 2$, khi đó, xác suất phân loại mục tiêu là

$$p(\boldsymbol{x}) = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

Một cách tự nhiên (và đơn giản), ta thiết lập ngay mô hình hồi quy tuyến tính

$$p(\boldsymbol{x}) = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k$$

Mô hình hồi quy logistic

Tuy nhiên, ước lượng này sẽ vượt ra khỏi khoảng (0, 1) $\Rightarrow$ không thỏa mãn tính chất xác suất.

Ta cần một hàm biến đổi $\beta_0 + \sum_{k=1}^{p} \beta_k x_k$ từ $\mathbb{R}$ vào trong $(0, 1)$.

- hàm chỉ

$$\mathbb{I}(x) = \begin{cases} 0, & \text{nếu } x < 0 \\ x, & \text{nếu } 0 \leq x \leq 1 \\ 1, & \text{nếu } x > 1 \end{cases}$$

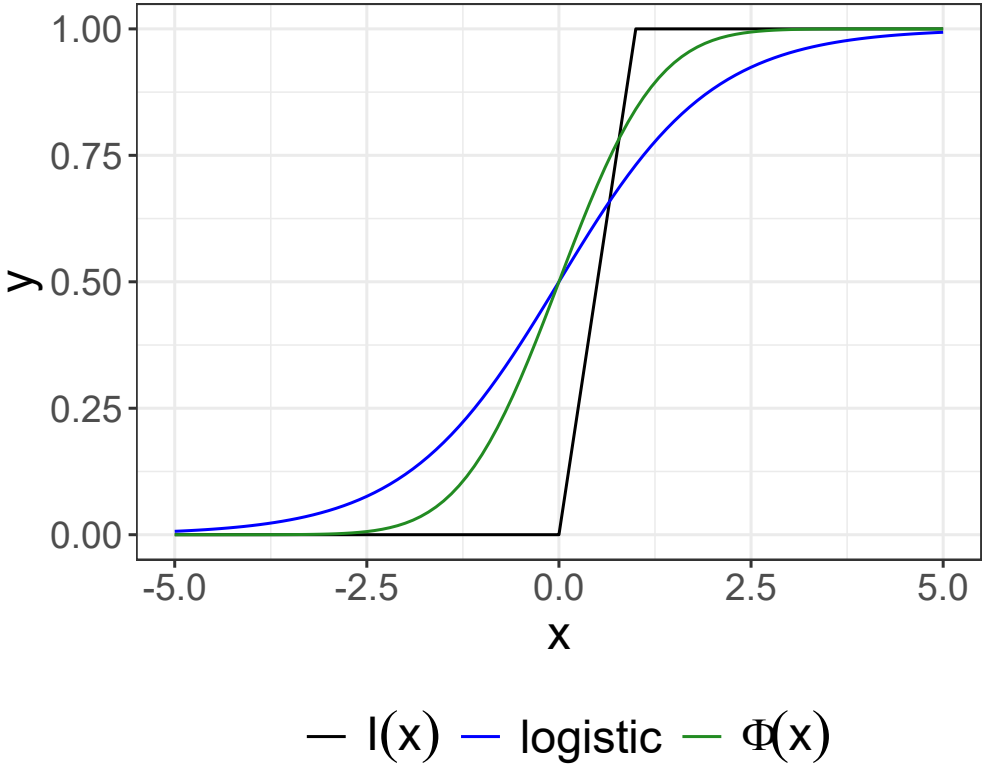
- hàm logistic

$$g(x) = \frac{\exp(x)}{1 + \exp(x)}$$

- hàm probit

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds$$

Mô hình hồi quy logistic



- Hàm chỉ $\mathbb{I}(x)$ là hàm không trơn, do đó, không tồn tại đạo hàm.
- Hàm logistic và probit là các hàm trơn, liên tục tại mọi x.

Mô hình hồi quy logistic

Nếu ta lựa chọn

- hàm logistic, thì

$$p(\boldsymbol{x}) = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \frac{\exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_k)}$$

khi đó, ta có mô hình **hồi quy logistic**. Hơn nữa, sử dụng biến đổi ta được

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k;$$

- hàm probit, thì

$$p(\boldsymbol{x}) = \Pr(Y = 1 | X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \Phi\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)$$

khi đó, ta có mô hình **hồi quy probit**. Hơn nữa, sử dụng biến đổi ta được

$$\Phi^{-1}(p(\boldsymbol{x})) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k.$$

Mô hình hồi quy logistic

Như vậy, xác suất $p(\boldsymbol{x})$ có thể được ước lượng thông qua mô hình hồi quy logistic:

$$p(\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}.$$

Sau khi xác định được các ước lượng $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ bằng phương pháp hợp lý cực đại, ta tính được

$$\widehat{p}(\boldsymbol{x}) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_p x_p)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_p x_p)}$$

Quy tắc phân loại nhị phân ($K = 2$):

$$\widehat{Y}(\boldsymbol{x}) = \begin{cases} 1 & \text{nếu } \widehat{p}(\boldsymbol{x}) > p_0, \\ 0 & \text{nếu } \widehat{p}(\boldsymbol{x}) \leq p_0, \end{cases}$$

Ví dụ

Xét ví dụ phân loại một người có khả năng bị phá sản tín dụng (`Default`) dựa trên mô hình

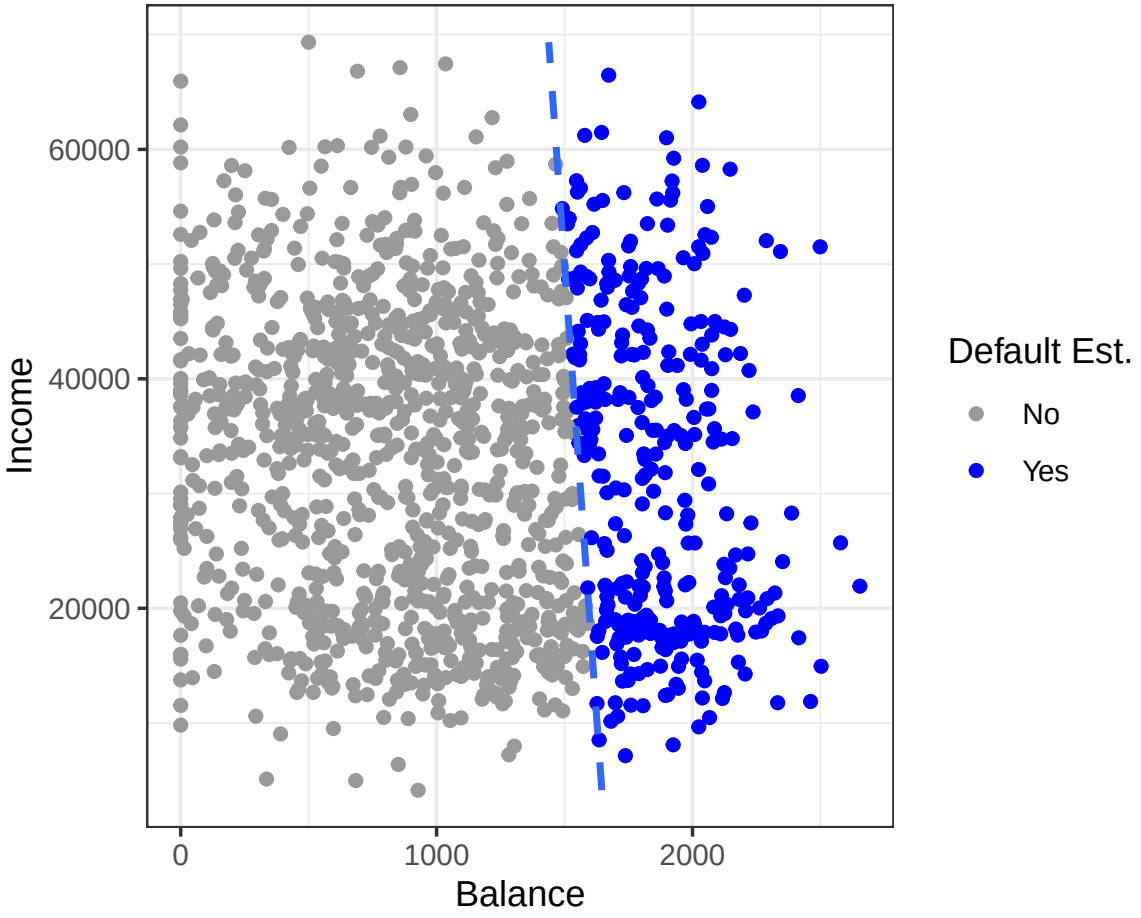
$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 \texttt{Balance} + \beta_2 \texttt{Income}$$

Ước lượng mô hình cho ta kết quả

	Estimate	Std. Error	Pr(> |z|)
(Intercept)	-8.414	0.5325	< 0.0001
balance	5.072×10^{-3}	2.990×10^{-4}	< 0.0001
income	1.614×10^{-5}	7.105×10^{-6}	0.0232

Ví dụ

Chọn $p_0 = 0.5$



Hồi quy multinominal logistic

Xét lại mô hình hồi quy logistic

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k;$$

Nhận xét rằng $1 - p(\boldsymbol{x}) = \Pr(Y = 0|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$. Từ đây, mô hình được biểu diễn theo

$$\log\left(\frac{\Pr(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}{\Pr(Y = 0|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k;$$

Như vậy, hồi quy logistic là một mô hình miêu tả log tỷ lệ xác suất của hai nhóm.

Điều này giúp ta mở rộng được mô hình lên trường hợp $K \geq 3$.

Hồi quy multinominal logistic

Với $K \geq 3$, mở rộng ý tưởng của mô hình hồi quy logistic, ta sẽ mô hình log tỷ lệ của hai nhóm bất kỳ

$$\log\left(\frac{\Pr(Y = j|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}{\Pr(Y = l|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}\right)$$

$\Rightarrow$ cần C_K^2 mô hình!

Thay vì so sánh tất cả cặp nhóm, ta sẽ so sánh một nhóm bất kỳ với một nhóm tham chiếu (reference group)

$$\log\left(\frac{\Pr(Y = j|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}{\Pr(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)}\right),$$

với $j = 2, \ldots, K$.

$\Rightarrow$ chỉ cần $K - 1$ mô hình.

Hồi quy multinominal logistic

Với cách làm này, ta có mô hình

$$\log\left(\frac{p_j(\boldsymbol{x})}{p_1(\boldsymbol{x})}\right) = \beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{jk},$$

với $j = 2, \ldots, K$. Từ đây ta thu được

$$p_j(\boldsymbol{x}) = \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{jk}\right)}{1 + \sum_{l=2}^{K} \exp\left(\beta_{l0} + \sum_{k=1}^{p} \beta_{lk} x_{lk}\right)}$$

và

$$p_1(\boldsymbol{x}) = \frac{1}{1 + \sum_{l=2}^{K} \exp\left(\beta_{l0} + \sum_{k=1}^{p} \beta_{lk} x_{lk}\right)}$$

Quy tắc phân loại

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ p_j(\boldsymbol{x}),$$

Ví dụ

Xét bài toán phân loại loài hoa iris vào 1 trong 3 loài hoa

- `versicolor` (nhóm tham chiếu) - 1;
- `setosa` - 2;
- `virginica` - 3.

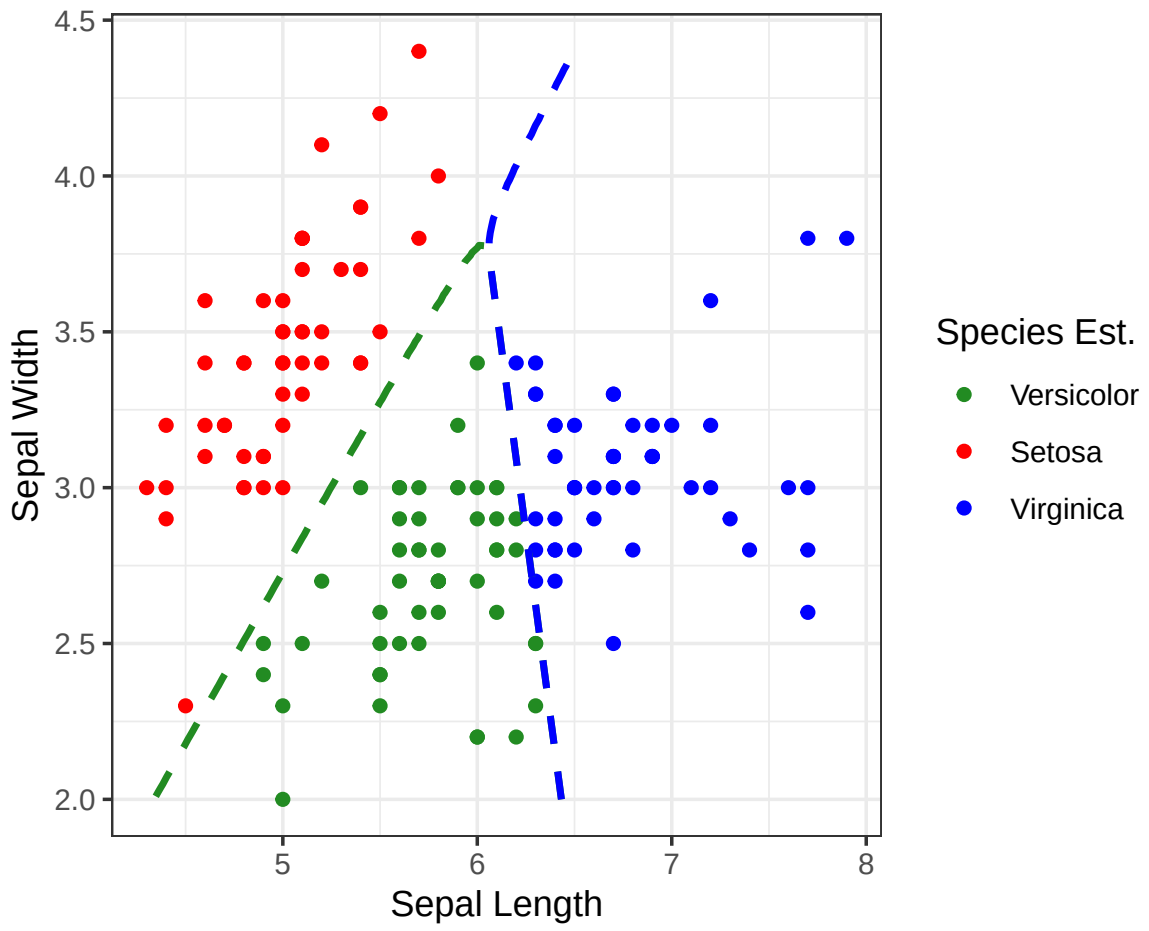
Mô hình

$$\log\left(\frac{p_j(x)}{p_1(x)}\right) = \beta_{j0} + \beta_{j1}\texttt{Sepal.L}_j + \beta_{j1}\texttt{Sepal.W}_j,$$

Ước lượng mô hình cho ta kết quả

	2 vs. 1		3 vs. 1	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	104.1164	141.0653	−13.0477	3.0977
Sepal.L	−41.2011	55.0576	1.9025	0.5170
Sepal.W	37.2945	59.7691	0.4048	0.8628

Ví dụ


4.5
4.0
3.5
3.0
2.5
2.0
Sepal Width
5
6
7
8
Sepal Length
Species Est.
Versicolor
Setosa
Virginica

Các vấn đề đối với mô hình

Ưu điểm

- Không phụ thuộc vào giả định phân phối của các $X_1, \ldots, X_p$ trong nhóm j.
- Giải thích được sự biến động của Y thông qua mô hình.
- Tìm được biến giải thích phù hợp (có ý nghĩa) thông qua mô hình.

Nhược điểm

- Yêu cầu $X_1, \ldots, X_p$ là độc lập, hoặc có sự tương quan thấp.
- Cần một quy trình xây dựng mô hình (tương tự mô hình hồi quy tuyến tính).
- Hoạt động kém hiệu quả nếu mô hình thực sự không có dạng tuyến tính.

1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

Các chỉ số đánh giá tổng quát

Độ chính xác của một mô hình phân loại nhị phân có thể được đánh giá thông qua:

- confusion matrix: một bảng 2×2 thể hiện số lượng phân loại đúng và sai của hai nhóm;
- accuracy: phần trăm phân loại chính xác của cả hai nhóm;
- sensitivity: phần trăm phân loại chính xác nhóm "1";
- specificity: phân trăm phân loại chính xác nhóm "0";
- precision: phần trăm dự đoán nhóm "1" mà thực chất là nhóm "1";
- ROC curve: đường cong thể hiện sensitivity và specificity khi thay đổi ngưỡng xác suất;
- AUC: diện tích dưới đường cong ROC.

Các chỉ số đánh giá tổng quát

Để đánh giá một mô hình phân loại, ta chia dữ liệu thành hai phần

- training set: dữ liệu dùng để xây dựng mô hình;
- test set: dữ liệu dùng để đánh giá mô hình.

Nhắc lại rằng, đối với một mô hình phân loại nhị phân:

- xác suất dự đoán $\widehat{p}_1(\boldsymbol{x})$ được xác định dựa trên mô hình
- tiên đoán $\widehat{Y}(\boldsymbol{x})$ được xác định bởi

$$\widehat{Y}(\boldsymbol{x}) = \begin{cases} 0 & \text{nếu } \widehat{p}_1(\boldsymbol{x}) \leq p_0 \\ 1 & \text{nếu } \widehat{p}_1(\boldsymbol{x}) > p_0 \end{cases}$$

  với p_0 là một ngưỡng xác suất.

Việc đánh giá độ chính xác mô hình phân loại, dựa trên ý tưởng về việc so sánh

- số lượng đối tượng thuộc nhóm "1" được phân loại vào nhóm "1" / "0";
- số lượng đối tượng thuộc nhóm "0" được phân loại vào nhóm "0" / "1";

tại một ngưỡng xác suất p_0.

Ma trận phân loại - Confusion matrix

Khi đó, với p_0 cho trước, mỗi giá trị quan sát Y_i và giá trị ước lượng $\widehat{Y}_i$, ta quan sát:

$$Z_i = \begin{cases} 0, & \text{nếu } (Y_i = 0, \widehat{Y}_i = 1) \text{ hoặc } (Y_i = 1, \widehat{Y}_i = 0) \\ 1, & \text{nếu } (Y_i = 0, \widehat{Y}_i = 0) \text{ hoặc } (Y_i = 1, \widehat{Y}_i = 1) \end{cases}$$

Tổng hợp tất cả các giá trị của Z_i , ta có bảng sau:

		Ước lượng	
		$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$
Quan sát	$Y_i = 0$	n_{00}	n_{01}
	$Y_i = 1$	n_{10}	n_{11}

với $n_{00} + n_{10} = n_0$ (số lượng quan sát $Y = 0$), $n_{01} + n_{11} = n_1$ (số lượng quan sát $Y = 1$). Trong đó,

- n_{00} và n_{11} là hai giá trị cho thấy số lượng ước lượng khớp của mô hình;
- n_{01} và n_{10} là hai giá trị cho thấy số lượng ước lượng sai của mô hình.

$\hookrightarrow$ n_{00} càng gần n_0 và n_{11} càng gần n_1, thì mô hình là dự đoán tốt.

Ma trận phân loại - Confusion matrix

- n_{00} là true negative (TN)
- n_{11} là true positive (TP)
- n_{01} là false positive (FP)
- n_{10} là false negative (FN)

Ví dụ: ma trận phân loại của mô hình Naive Bayes phân loại 1 người bị phá sản tín dụng,

- tại ngưỡng $p_0 = 0.5$ là

		Ước lượng	
		$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$
Quan sát	$Y_i = 0$	918	82
	$Y_i = 1$	80	253

- tại ngưỡng $p_0 = 0.3$ là

		Ước lượng	
		$\widehat{Y}_i = 0$	$\widehat{Y}_i = 1$
Quan sát	$Y_i = 0$	853	147
	$Y_i = 1$	43	290

Accuracy

Accuracy là tổng số lượng phân loại chính xác nhóm "1" và "0" của một mô hình phân loại tại một ngưỡng xác suất p_0:

$$\text{Accuracy} = \frac{n_{00} + n_{11}}{n}.$$

Giá trị của Accuracy thay đổi từ 0 tới 1, trong đó,

- 0: mô hình phân loại hoàn toàn sai tại ngưỡng p_0;
- 1: mô hình phân loại hoàn toàn chính xác tại ngưỡng p_0;
- càng gần 1 thì có nghĩa mô hình càng phân loại đúng hai nhóm.

Ví dụ: Accuracy của mô hình Naive Bayes tại

- $p_0 = 0.5$:
$$\text{Accuracy}_{0.5} = \frac{918 + 253}{1333} = 0.8785;$$
- $p_0 = 0.3$:
$$\text{Accuracy}_{0.3} = \frac{853 + 290}{1333} = 0.8575.$$

Accuracy không cho biết góc nhìn về tỷ lệ phân loại sai.

Precision, Sensitivity và Specificity

Từ bảng kết quả so sánh nhị phân với một ngưỡng p_0, ta định nghĩa được

- tỷ lệ dự đoán đúng "0", được gọi là **Specificity** - Sp:

$$\mathrm{Sp}(p_0) = \frac{n_{00}}{n_{00} + n_{01}} = \frac{1}{n_0} \sum_{i=1}^{n} \mathrm{I}(Y_i = 0, \widehat{p}_i(\boldsymbol{x}) \leq p_0)$$

- tỷ lệ dự đoán đúng "1", được gọi là **Sensitivity** - Se:

$$\mathrm{Se}(p_0) = \frac{n_{11}}{n_1} = \frac{1}{n_1} \sum_{i=1}^{n} \mathrm{I}(Y_i = 1, \widehat{p}_i(\boldsymbol{x}) > p_0)$$

  Sensitivity còn được gọi là **Recall** - thường được dùng trong machine learning;

- tỷ lệ dự đoán "1" của mô hình, được gọi là **Precision**:

$$\mathrm{Precision}(p_0) = \frac{n_{11}}{n_{01} + n_{11}}$$

  Precision cho biết tỷ lệ dương tính thật (True positive) so với dương tính giả (False positive) của mô hình.

  Precision càng cao $\Rightarrow$ tỷ lệ dương tính thật càng cao $\Rightarrow$ mô hình phân loại nhóm 1 tốt.

Precision, Sensitivity và Specificity

Với $p_0 = 0.5$, ta tính được

- tỷ lệ dự đoán đúng "0"

$$\mathrm{Sp}(0.5) = \frac{918}{918 + 82} \approx 0.918$$

- tỷ lệ dự đoán đúng "1"

$$\mathrm{Se}(0.5) = \frac{253}{253 + 80} \approx 0.760$$

- tỷ lệ dương tính thật trong nhóm được dự đoán "1":

$$\mathrm{Precision}(0.5) = \frac{253}{253 + 82} \approx 0.755$$

Precision, Sensitivity và Specificity

Với $p_0 = 0.3$, ta tính được

- tỷ lệ dự đoán đúng "0"

$$\mathrm{Sp}(0.3) = \frac{853}{853 + 147} \approx 0.853$$

- tỷ lệ dự đoán đúng "1"

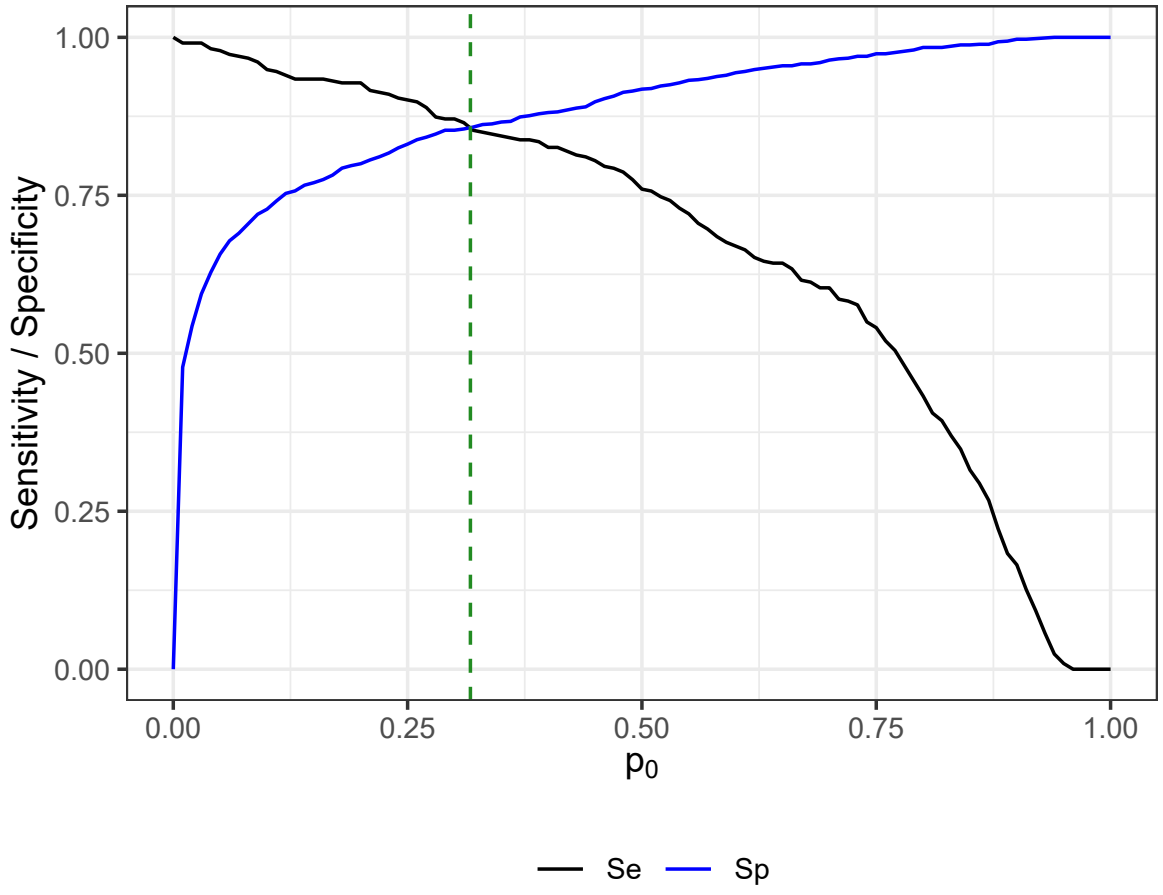
$$\mathrm{Se}(0.3) = \frac{290}{290 + 43} \approx 0.871$$

- tỷ lệ dương tính thật trong nhóm được dự đoán "1":

$$\mathrm{Precision}(0.3) = \frac{290}{290 + 147} \approx 0.664$$

Precision, Sensitivity và Specificity

Thay đổi ngưỡng xác suất p_0 từ 0 tới 1, ta thu được biểu đồ sau:



Đường nét đứt màu xanh tại $p_0 = 0.317$ và $\mathrm{Sp}(0.317) = \mathrm{Sp}(0.317) = 0.858$

Phân tích đường cong ROC

- Qua hai ví dụ trên, ta thấy rằng specificity và sensitivity chỉ đo lường hiệu suất của sự dự đoán liên quan đến một ngưỡng cụ thể p_0.
- Bằng cách thay đổi p_0, ta sẽ có các giá trị khác nhau của specificity và sensitivity, và do đó, hiệu suất cũng khác nhau!

$\Rightarrow$ Chúng ta cần đánh giá specificity và sensitivity ở tất cả các ngưỡng p_0 có thể có để tổng hợp độ chính xác của mô hình.

Receiver Operating Characteristic (ROC) curve

Một đường cong ROC (receiver operating characteristic) là đường cong mô tả tất cả các cặp điểm $(1 - \mathrm{Sp}(p_0), \mathrm{Se}(p_0))$ với tất cả ngưỡng $p_0 \in (0, 1)$ có thể:

$$\mathrm{ROC} = \{(1 - \mathrm{Sp}(p_0), \mathrm{Se}(p_0)), p_0 \in (0, 1)\}$$

Đường cong ROC được biểu diễn trên đồ thị giới hạn bởi hình vuông đơn vị

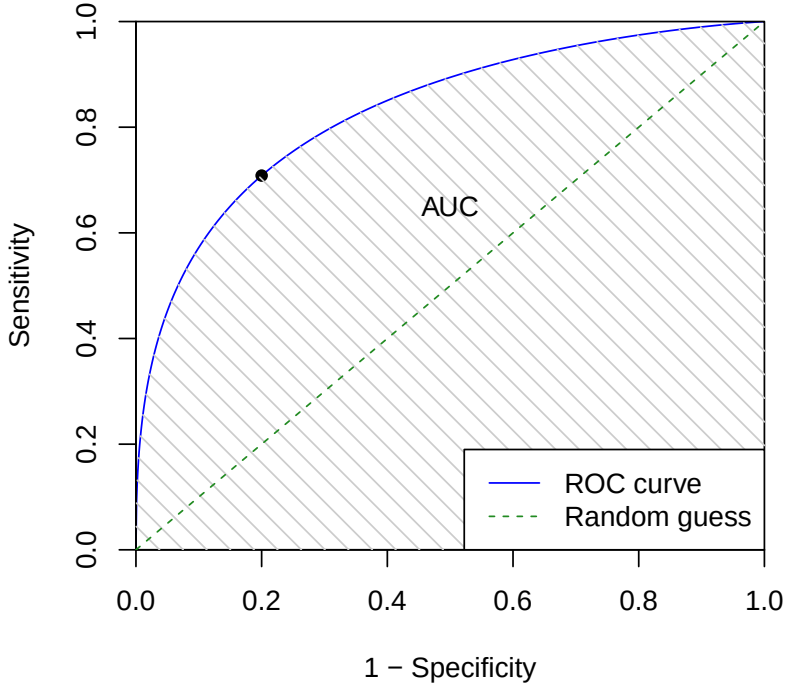
Diện tích dưới đường cong (AUC)

Diện tích dưới đường cong ROC (AUC) là một độ đo tổng hợp cho độ chính xác của phép phân loại.

Phân tích đường cong ROC

Thông thường, đường cong ROC của mô hình logistic nằm phía trên đường tham chiếu, tức là đường chéo $y = x$.

- Sự cân bằng giữa sensitivity và specificity được hiển thị trên đường cong ROC.
- Nếu đường cong ROC trùng với đường tham chiếu thì AUC $= 0.5$, chúng ta có thể nói rằng việc phân loại dựa trên mô hình logistic là **sự đoán ngẫu nhiên**.
- Nếu đường cong ROC trùng góc trên bên trái thì AUC $= 1$, điều đó có nghĩa là việc phân loại dựa trên mô hình logistic có hiệu suất **hoàn hảo**.



Giá trị hợp lý của AUC thay đổi từ 0.5 đến 1. AUC càng cao thì khả năng dự đoán của mô hình càng tốt.

Phân tích đường cong ROC

Đường cong ROC có thể được biểu diễn là $\mathrm{ROC}(p)$, tức là hàm của p với $p \in (0, 1)$ - tức là $1 - \mathrm{Sp}$.

Do đó, diện tích dưới đường cong ROC (AUC) được định nghĩa bởi

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(p)\,\mathrm{d}p$$

AUC có thể được ước lượng bằng **trapezoidal rule** (quy tắc hình thang):

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{k=1}^{N-1} \frac{f(x_{k+1}) + f(x_k)}{2}(x_{k+1} - x_k).$$

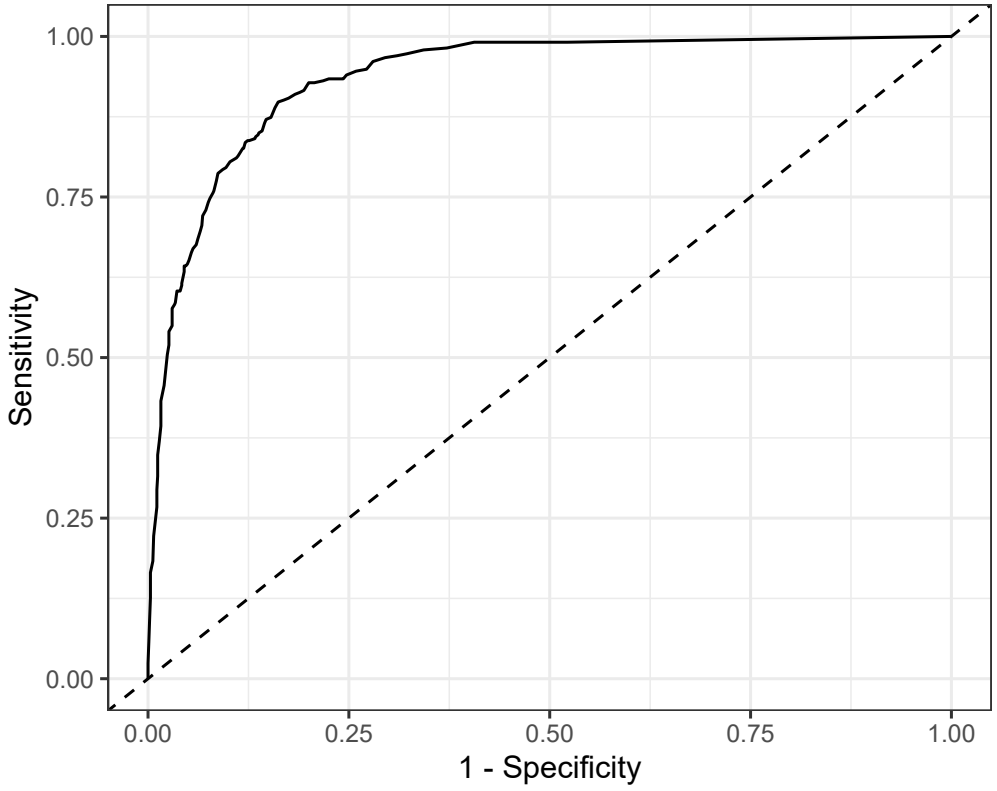
Áp dụng với đường cong ROC:

$$\widehat{\mathrm{AUC}} = \sum_{k=1}^{N-1} \frac{\mathrm{Se}(p_{0,k+1}) + \mathrm{Se}(p_{0,k})}{2} \left(\mathrm{Sp}(p_{0,k+1}) - \mathrm{Sp}(p_{0,k})\right).$$

Khoảng tin cậy cho AUC có thể được xây dựng bằng phương pháp Delong hoặc phương pháp bootstrap.

Phân tích đường cong ROC

Đối với ví dụ về dự đoán phá sản tín dụng, đường cong ROC ước lượng cho mô hình Naive Bayes được biểu diễn như sau:



Giá trị AUC ước lượng: 0.937, 95% khoảng tin cậy: (0.924, 0.95).

Phân tích đường cong ROC

Quy tắc đánh giá dự trên AUC

Chúng ta có thể sử dụng quy tắc sau để đánh giá tính chính xác của mô hình phân loại:

AUC	Tính chính xác
0.5	sự đoán ngẫu nhiên
(0.5, 0.65]	thiếu chính xác
(0.65, 0.8]	chấp nhận được
(0.8, 0.95]	tốt
(0.95, 1)	rất tốt
1	hoàn hảo

Trong ví dụ của chúng ta, AUC ước tính là 0.937, 95% khoảng tin cậy cho AUC (0.924, 0.95), do đó mô hình Naive Bayes có độ chính xác tốt để dự đoán một người có khả năng bị phá sản tín dụng.

Khi so sánh nhiều mô hình phân loại, mô hình nào có AUC lớn hơn, mô hình đó có xu hướng được lựa chọn sử dụng.

Đánh giá mô hình phân loại đa nhóm

Nhắc lại rằng, trong mô hình phân loại đa nhóm

$$\widehat{Y}(\boldsymbol{x}) = \underset{j=1,\ldots,K}{\arg\max}\ \widehat{p}_{j,*}(\boldsymbol{x}),$$

với đấu $*$ đại diện cho phương pháp phân loại được sử dụng.

Để đánh giá mô hình phân loại đa nhóm, ta so sánh $\widehat{Y}(\boldsymbol{x})$ và Y

$\Rightarrow$ kết quả là ma trận phân loại - Confusion matrix

		Ước lượng			
		$\widehat{Y}_i = 1$	$\widehat{Y}_i = 2$	$\ldots$	$\widehat{Y}_i = K$
Quan sát	$Y_i = 1$	n_{11}	n_{12}	$\ldots$	n_{1K}
	$Y_i = 2$	n_{21}	n_{22}	$\ldots$	n_{2K}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$Y_i = K$	n_{K1}	n_{K2}	$\ldots$	n_{KK}

trong đó, n_{jj} là số lượng đối tượng được phân loại đúng.

Đánh giá mô hình phân loại đa nhóm

Từ ma trận phân loại ta có thể tính các chỉ số đánh giá:

- Accuracy - tỷ lệ phân loại đúng của tất cả các nhóm

$$\text{Accuracy} = \frac{n_{11} + n_{22} + \ldots + n_{KK}}{n}$$

- Precision cho từng nhóm

$$\text{Precision}_j = \frac{n_{jj}}{n_{1j} + n_{2j} + \ldots + n_{Kj}}$$

- Recall cho từng nhóm

$$\text{Recall}_j = \frac{n_{jj}}{n_{j1} + n_{j2} + \ldots + n_{jK}}$$

Đánh giá mô hình phân loại đa nhóm

Ngoài ra, ta còn có 2 chỉ số thường được dùng trong đánh giá tổng quát hiệu quả của mô hình phân loại đa nhóm.

Marco F1-score

$$\text{Marco F1-score} = 2\frac{\overline{\text{Precision}} \times \overline{\text{Recall}}}{\frac{1}{\overline{\text{Precision}}} + \frac{1}{\overline{\text{Recall}}}},$$

trong đó, $\overline{\text{Precision}}$ và $\overline{\text{Recall}}$ là trung bình cộng của các Precision_j và của Recall_j.

Chỉ số thu được đánh giá thuật toán theo quan điểm lớp:

- giá trị Macro-F1 nằm trong $[0, 1]$;
- giá trị Macro-F1 cao chỉ ra rằng thuật toán có hiệu suất tốt trên tất cả các lớp;
- giá trị Macro-F1 thấp chỉ các lớp được dự đoán kém.

Đánh giá mô hình phân loại đa nhóm

Cohen's Kappa

$$\text{Kappa} = \frac{c \times s - \sum_{j=1}^{K} p_j \times t_j}{s^2 - \sum_{j=1}^{K} p_j \times t_j},$$

trong đó,

- $c = \sum_{j=1}^{K} n_{jj}$ (tổng số lượng phân loại chính xác);
- $s = \sum_{j=1}^{K} \sum_{i=1}^{K} n_{ji}$ (tổng tất cả phần tử);
- $p_j = \sum_{i=1}^{K} n_{ji}$ số lần lớp j được dự đoán (tổng số cột);
- $t_j = \sum_{i=1}^{K} n_{ij}$ số lần lớp j được quan sát (tổng số dòng).

Đánh giá mô hình phân loại đa nhóm

Nhận xét:

- khi Kappa $= 0 \Rightarrow$ Dự đoán của mô hình hoàn toàn độc lập với Phân loại thực tế;
- nếu Kappa $= 1 \Rightarrow$ Dự đoán của mô hình hoàn toàn phụ thuộc vào Phân loại thực tế;
- nếu Kappa $< 0 \Rightarrow$ sự tương đồng giữa thuật toán và phân phối nhãn thực tế tệ hơn sự tương đồng ngẫu nhiên

  $\Rightarrow$ không có sự phù hợp giữa Dự đoán của mô hình và Phân loại thực tế.

Đánh giá mô hình phân loại đa nhóm

Ví dụ: xét mô hỉnh multinomial logistic phân loại 3 loài hoa iris: `setosa` (1), `versicolor` (2) và `virginica` (3)

$$\log\left(\frac{p_j(x)}{p_1(x)}\right) = \beta_{j0} + \beta_{j1}\texttt{Sepal.L}_j + \beta_{j1}\texttt{Sepal.W}_j,$$

Confusion matrix của mô hình trên tập test:

		Ước lượng		
		$\widehat{Y}_i = 1$	$\widehat{Y}_i = 2$	$\widehat{Y}_i = K$
Quan sát	$Y_i = 1$	50	0	0
	$Y_i = 2$	0	38	12
	$Y_i = K$	0	13	37

Nhận xét

- mô hình phân loại hoàn toàn chính xác nhóm `setosa`;
- có sai sót trong phân loại hai nhóm `versicolor` và `virginica`.

Đánh giá mô hình phân loại đa nhóm

Dựa trên confusion matrix, ta tính được các chỉ số đánh giá

- Accuracy $= \dfrac{50 + 38 + 37}{150} = 0.8333$
- Precision và Recall cho từng nhóm

	setosa	versicolor	virginica
Precision	1	0.7451	0.7551
Recall	1	0.76	0.74

- Marco F1-score $= 0.5788$;
- Kappa $= 0.75$.

Các kết quả này cho thấy mô hình này không hiệu quả với tất cả các nhóm.

1 Bài toán phân loại

2 Naive Bayes

3 Discriminant Analysis

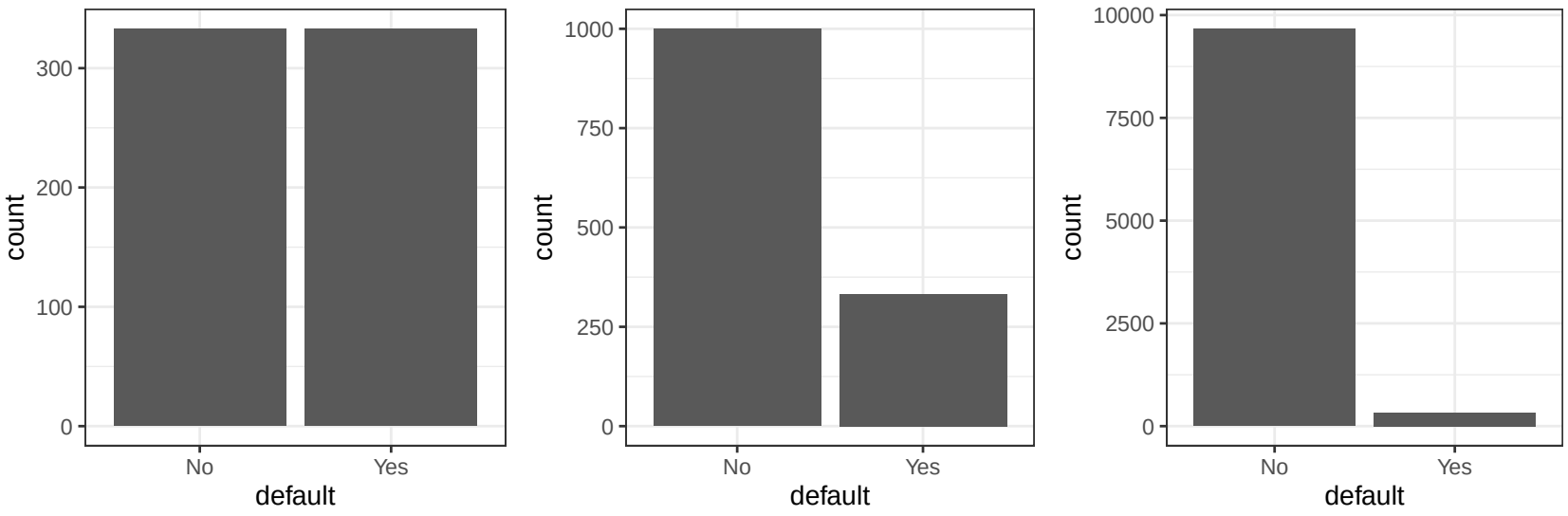
4 Hồi quy logistic

5 Đánh giá mô hình phân loại

6 Dữ liệu không cân bằng

Dữ liệu không cân bằng

Trong bài toán phân loại, dữ liệu không cân bằng được thể hiện bởi sự mất cân bằng về dữ liệu được quan sát giữa các nhóm cần phân loại.



(a) (b) (c)

Dữ liệu không cân bằng

Vấn đề

Trong những trường hợp dữ liệu không cân bằng, trừ khi các nhóm dễ dàng được phân tách, mô hình phân loại chính xác nhất có thể là mô hình phân loại mọi thứ thành "0".

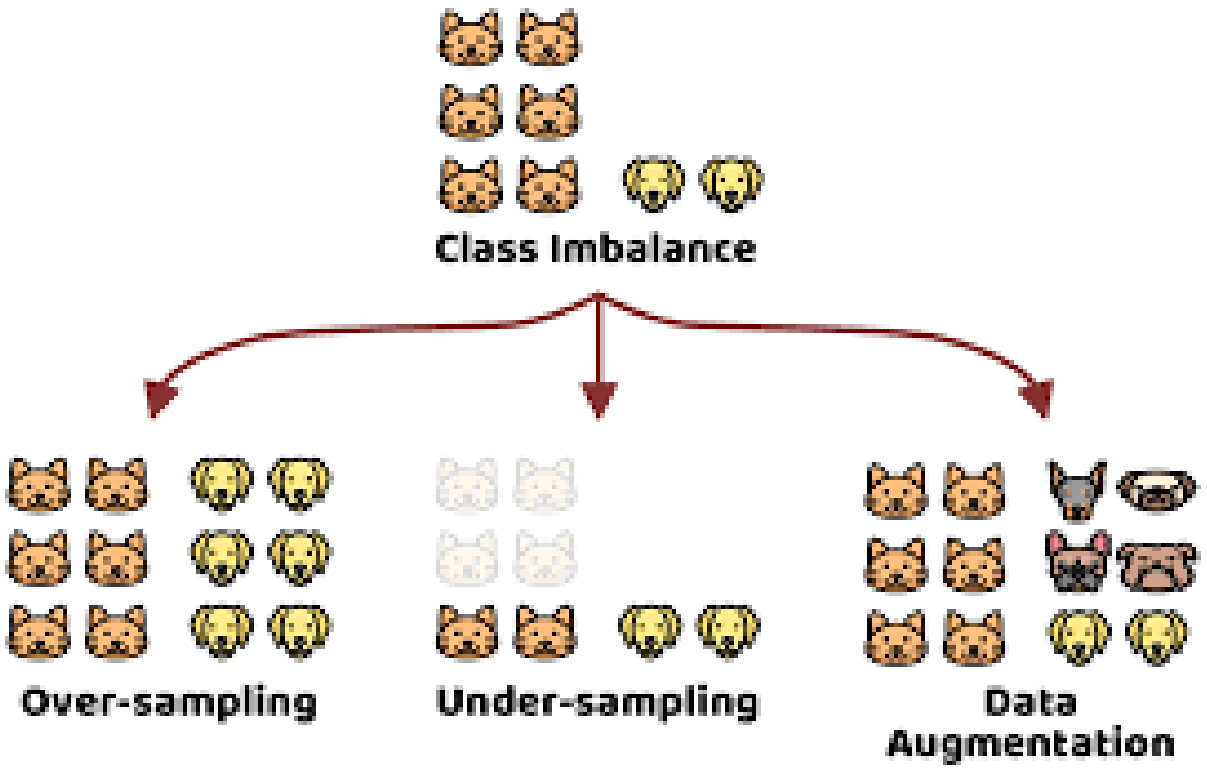
Ví dụ: nếu chỉ có 0.1% người duyệt web tại một cửa hàng trực tuyến thực sự mua hàng.

- một mô hình dự đoán rằng mỗi người duyệt web sẽ rời đi mà không mua hàng sẽ có độ chính xác 99,9%;
- tuy nhiên, mô hình này sẽ là vô dụng.

Thay vào đó, chúng ta sẽ hài lòng với một mô hình

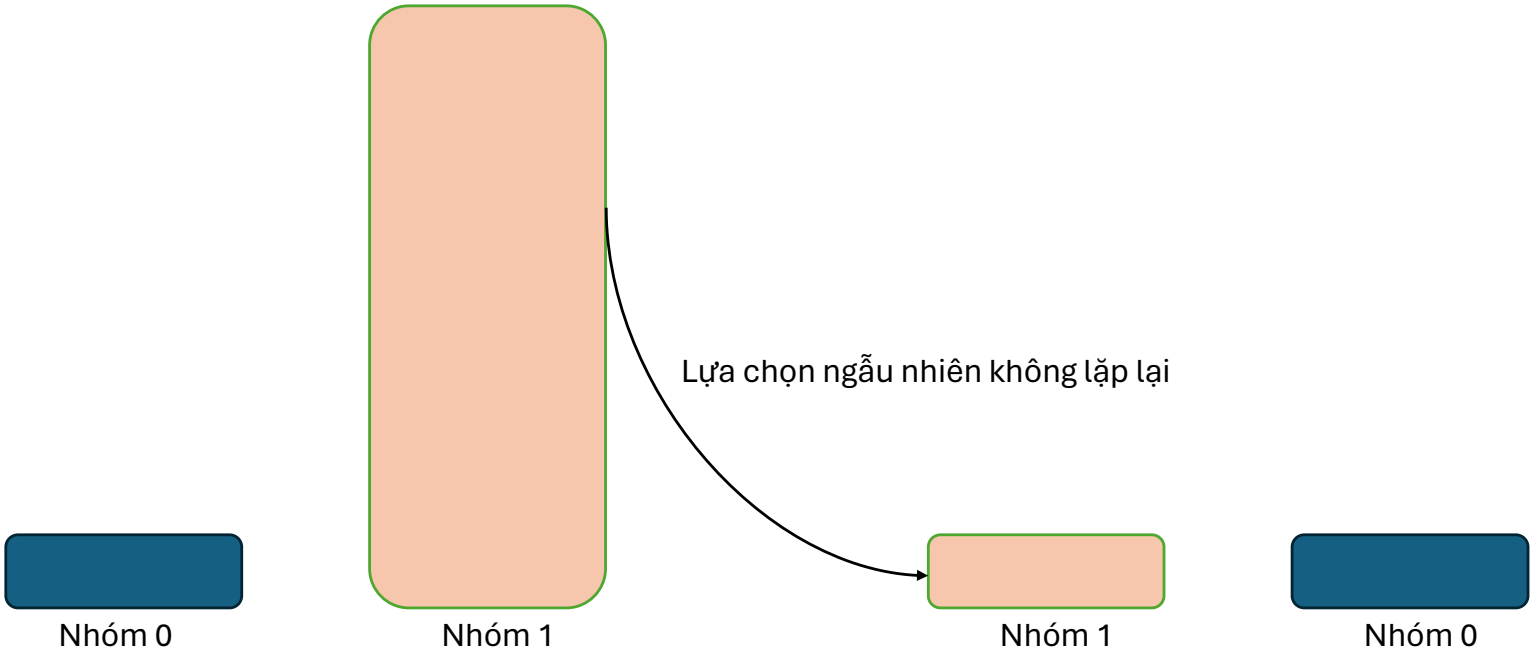
- ít chính xác hơn về tổng thể;
- nhưng tốt trong việc nhận diện những người mua hàng;
- và ta sẽ chấp nhận các trường hợp bị dương tính giả.

Một số chiến lược cơ bản



- under-sampling: cắt bỏ bớt dữ liệu trong nhóm có nhiều quan sát;
- overs-sampling: làm đầy thêm dữ liệu trong nhóm có ít quan sát, dựa trên quan sát có sẵn trong nhóm đó;
- data generation (data augmentation): làm đầy dữ liệu trong nhóm có ít quan sát bằng cách giả lập dữ liệu.

Under-sampling



Under-sampling

Yêu cầu:

- dữ liệu đủ lớn (khoảng 1000 cho lớp ít quan sát);
- các quan sát là độc lập nhau.

Ý tưởng

Ý tưởng cơ bản của giảm mẫu là dữ liệu của nhóm có nhiều quan sát (có nhiều quan sát dư thừa), thông qua một quá trình chọn mẫu ngẫu nhiên đơn giản không lặp lại.

Việc xử lý một tập dữ liệu nhỏ hơn, cân bằng hơn sẽ

- mang lại lợi ích cho hiệu suất mô hình;
- làm cho việc chuẩn bị dữ liệu dễ dàng hơn, cũng như việc khám phá và thử nghiệm các mô hình.

Lưu ý: Dữ liệu càng dễ phân biệt các trường hợp 1 với các trường hợp 0, thì càng ít quan sát cần thiết.

Over-sampling

Phương pháp under-sampling có một số nhược điểm:

- loại bỏ đi dữ liệu (có thể quan trọng);
- không tận dụng được toàn bộ thông tin;
- không thể được sử dụng khi cỡ mẫu tương đối bé.

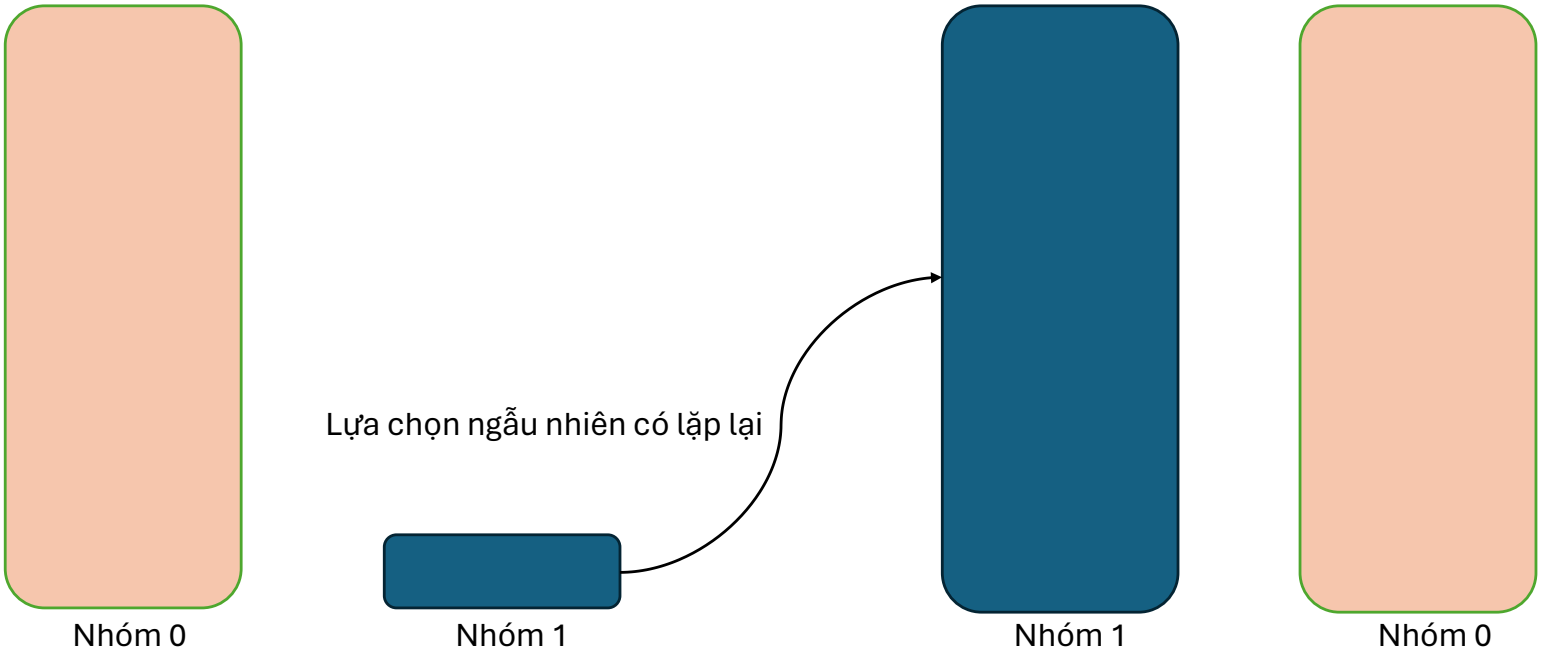
Trong trường hợp này, thay vì cắt giảm dữ liệu của nhóm có nhiều quan sát, ta sẽ làm đầy quan sát cho nhóm ít quan sát, thông qua quá trình lấy mẫu có lặp lại (bootstrap).

Một cách khác là ta có thể thêm trọng số vào trong các quan sát:

- trọng số nhỏ cho các quan sát trong nhóm nhiều;
- trọng số lớn cho các quan sát trong nhóm ít;

khi đó, hàm lượng quan sát vẫn được giữ nguyên nhưng tính quan trọng của các quan sát sẽ thay đổi.

Over-sampling



Data generation

Trong thực tế lấy mẫu:

- ta chỉ quan sát được một tập hợp hạn chế các trường hợp;

  $\hookrightarrow$ thuật toán không có một tập hợp thông tin phong phú để xây dựng các "quy tắc" phân loại.

Tuy nhiên, nếu ta có thể tạo ra các quan sát mới tương tự nhưng không giống hệt các quan sát hiện có

$\hookrightarrow$ thuật toán có cơ hội học một tập hợp các quy tắc mạnh mẽ hơn.

Data generation

Một biến thể của phương pháp tăng mẫu thông qua bootstrapping là tạo dữ liệu bằng cách làm nhiễu các quan sát hiện có để tạo ra các quan sát mới.

Data generation

SMOTE

Thuật toán SMOTE viết tắt của "Synthetic Minority Oversampling Technique".

Thuật toán SMOTE

- tìm một quan sát tương tự với quan sát đang được gia tăng mẫu;
- tạo ra một quan sát tổng hợp là trung bình có trọng số ngẫu nhiên của quan sát gốc và quan sát lân cận;
- trọng số được tạo riêng cho mỗi biến dự đoán.

Số lượng quan sát tổng hợp được tạo ra phụ thuộc vào tỷ lệ gia tăng mẫu cần thiết để đưa tập dữ liệu vào trạng thái cân bằng xấp xỉ, với các lớp kết quả.

Quy trình áp dụng thực tế

Quy trình xử lý dữ liệu cho bài toán phân loại:

1. Chuẩn bị dữ liệu:

   kiểm tra số lượng quan sát trong các nhóm;

   áp dụng under-sampling/over-sampling/SMOTE khi dữ liệu có dấu hiệu mất cân bằng lớn.
2. Xây dựng và huấn luyện mô hình: sử dụng các thuật toán phân loại.
3. Đánh giá độ chính xác của mô hình: sử dụng ROC curve, và AUC để tìm ra mô hình tốt nhất.
4. Áp dụng mô hình và giám sát.