Classification

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

May 22, 2023

phân loại

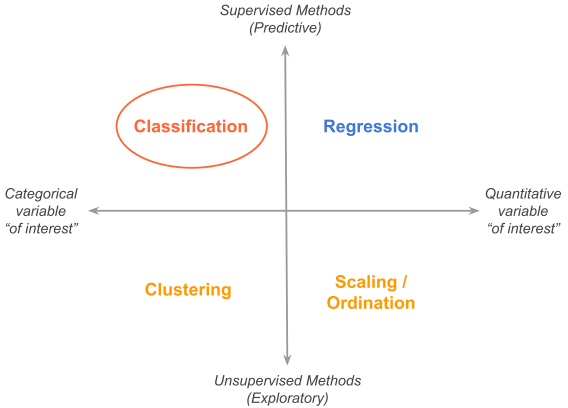
Lương Ngọc Hoàng

Trường Đại học Công nghệ Thông tin (UIT), ĐHQG-HCM

22 Tháng Năm, 2023

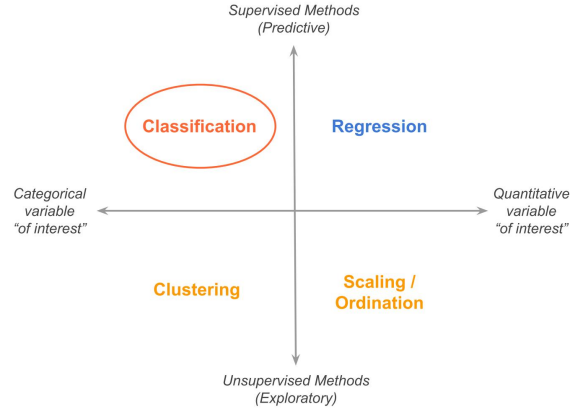
VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Introduction



- The goal in classification is to take an input vector $\mathbf{x}$ and to assign it into one of K discrete classes or groups C_k where $k = 1, 2, \ldots, K$.
- The classes are assumed to be disjoint, i.e., each input is assigned to one and only one class.

Giới thiệu



- Mục tiêu của phân loại là lấy một vectơ đầu vào x và gán nó vào một trong K lớp hoặc nhóm rời rạc Ck trong đó k = 1, 2, . . . , K. • Các lớp được coi là rời rạc, nghĩa là mỗi đầu vào được gán vào một và chỉ một lớp.

Classification - Example

- Consider a credit application with p predictors $X = [X_1, \ldots, X_p]$: Age, Salary, Residential Status, Marital Status, Debt, etc.
- A **credit score** is computed for each application to relate how like each applicant can pay the debt.
- Customers are divided into two classes: *good* and *bad*:
  - Good customers are those that payed their loan back.
  - Bad customers are those that defaulted on their loan

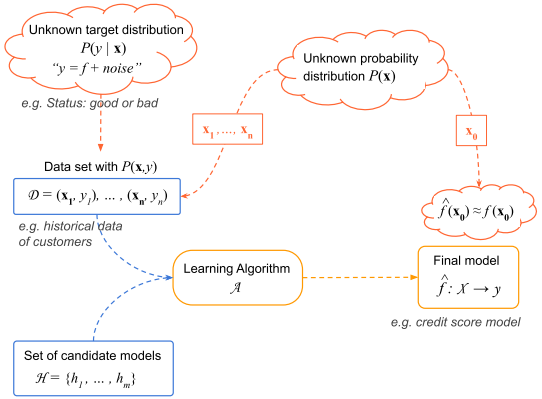
Phân loại - Ví dụ

- Xem xét một ứng dụng tín dụng với p dự đoán X = [X1, . . . , Xp]: Tuổi, Mức lương, Tình trạng cư trú, Tình trạng hôn nhân, Khoản nợ,

v.v. • Điểm tín dụng được tính cho mỗi đơn đăng ký để liên quan đến việc mỗi người đăng ký có thể trả nợ như thế nào.

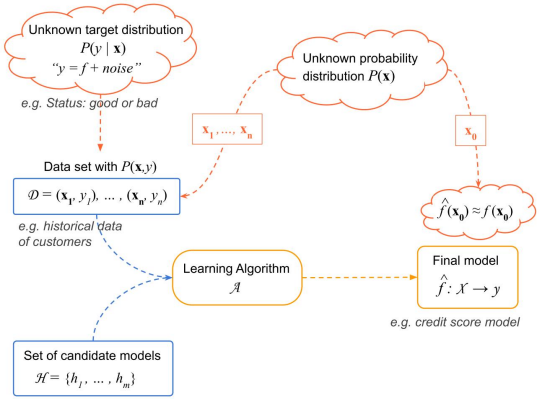
- Khách hàng được chia thành hai loại: tốt và xấu: •
  Khách hàng tốt là những khách hàng đã trả nợ. •
  Khách hàng xấu là những khách hàng không trả được nợ

Classification - Supervised Learning Diagram



- Joint distribution of data: $P(\mathbf{x}, y)$
- Conditional distribution of target, given inputs $P(y \mid \mathbf{x})$
- Marginal distribution of inputs $P(\mathbf{x})$

Phân loại - Sơ đồ học tập có giám sát

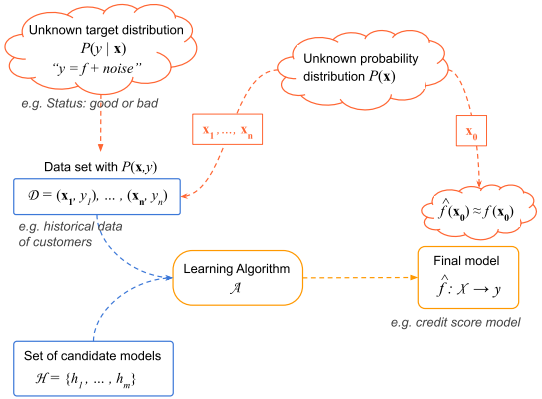


• Phân phối dữ liệu chung: P(x, y) •

Phân phối có điều kiện của mục tiêu, cho trước đầu vào P(y

x) • Phân phối cận biên của đầu vào P(x)

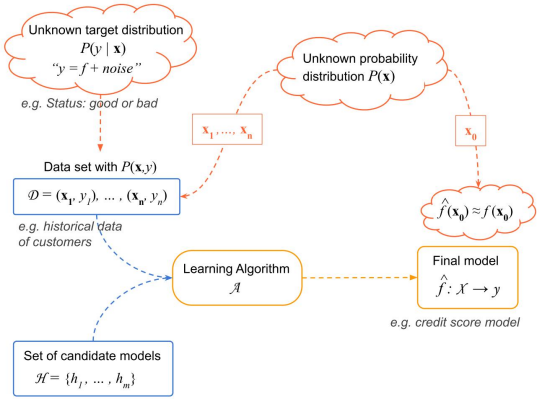
Classification - Supervised Learning Diagram



- The idea in classification problems is: Given a customer's attributes $X = \mathbf{x}$, to which class y we should assign this customer?
- We would like to know what is **the conditional probability**:

$$P(y \mid X = \mathbf{x})$$

Phân loại - Sơ đồ học tập có giám sát



- Ý tưởng trong các bài toán phân loại là: Cho trước một khách hàng thuộc tính X = x, chúng ta nên gán khách hàng này cho lớp y nào?
- Chúng tôi muốn biết xác suất có điều kiện là gì:

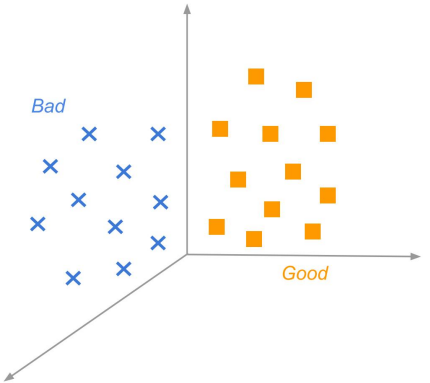
P(y X = x)

Classification - Example

- Suppose we have n individuals in a p-dimensional space.
- Suppose each class of customers forms its own cloud: the good customers, the bad customers.

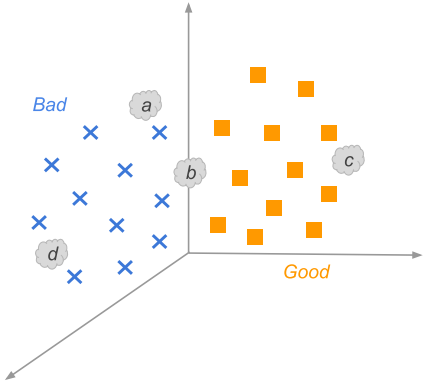
Phân loại - Ví dụ

- Giả sử chúng ta có n cá nhân trong không gian p chiều.
- Giả sử mỗi lớp khách hàng hình thành đám mây của riêng mình: khách hàng tốt, khách hàng xấu.



Classification - Example

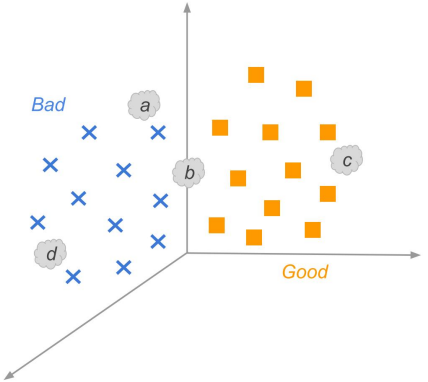
- Now, assume there are four individuals a, b, c, d that we want to predict their classes.
- We want to have a mechanism or **rule** to classify observations.



Phân loại - Ví dụ

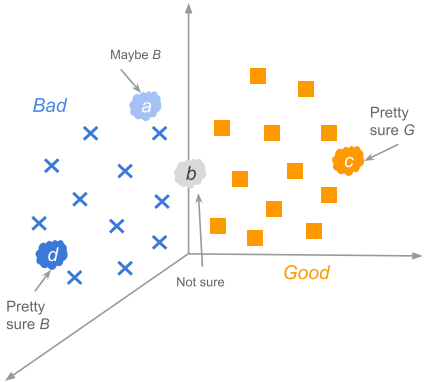
- Bây giờ, giả sử có bốn cá nhân a, b, c, d mà chúng ta muốn dự đoán lớp của họ.

Chúng tôi muốn có một cơ chế hoặc quy tắc để phân loại các quan sát.



Classification - Example

- Customer a could be assigned to class bad.
- Customer d could also be assigned to class bad with high confidence.
- Customer c could be assigned with high confidence to class good.
- We could be uncertain to which class customer b belongs.



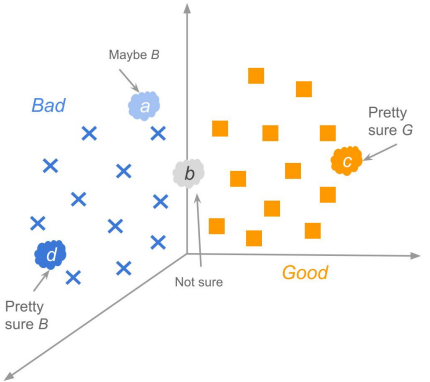
Phân loại - Ví dụ • Khách hàng

a có thể được xếp vào loại xấu. • Khách hàng

d cũng có thể được chỉ định vào lớp xấu với mức cao sự tự tin.

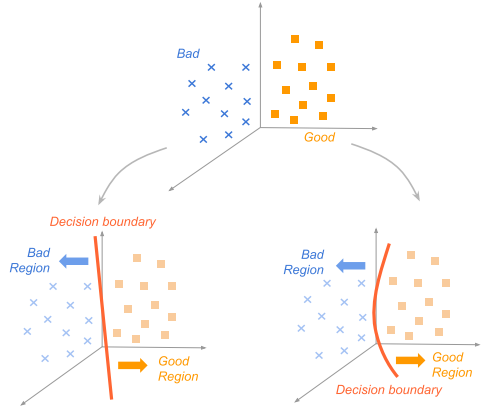
• Khách hàng c có thể được xếp vào loại tốt với độ tin cậy cao. •

Chúng ta có thể không chắc khách hàng b thuộc loại nào.



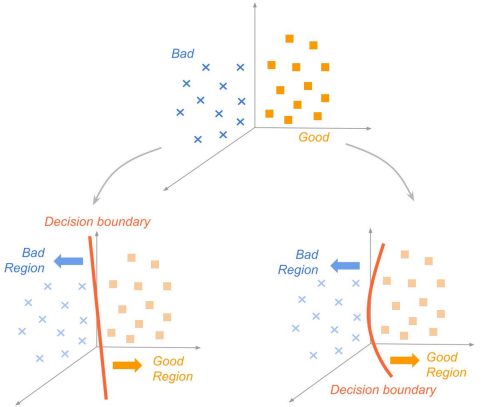
Classification - Example

- Classification rules allow us to divide the input space into regions $\mathcal{R}_k$ called **decision regions** (one for each class).
- The boundaries between decision regions establish the decision boundaries or decision surfaces.

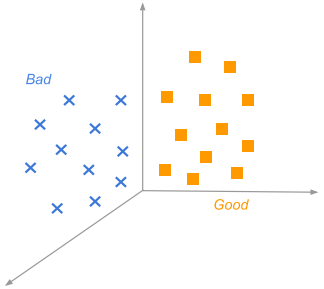


Phân loại - Ví dụ

- Các luật phân loại cho phép chúng ta chia không gian đầu vào thành các vùng Rk được gọi là các vùng quyết định (một vùng cho mỗi lớp).
- Ranh giới giữa các vùng quyết định thiết lập quyết định ranh giới hoặc bề mặt quyết định.



Classification - Two-class Example



- We have customers belonging to one of two classes C_1 = good and C_2 = bad.
- We can first investigate how X values vary according to a given class C_k - the class-conditional distribution:

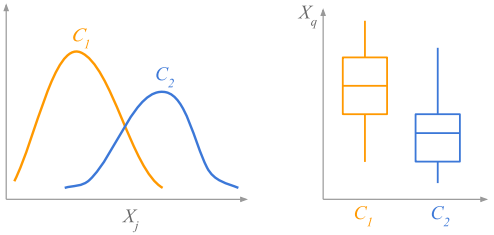
$$P(X = \mathbf{x} \mid y = k)$$

Phân loại - Ví dụ hai lớp

- Chúng tôi có khách hàng thuộc một trong hai loại C1 = tốt và C2 = xấu.
- Trước tiên, chúng ta có thể điều tra xem các giá trị X thay đổi như thế nào theo một lớp Ck đã cho - phân phối có điều kiện của lớp:

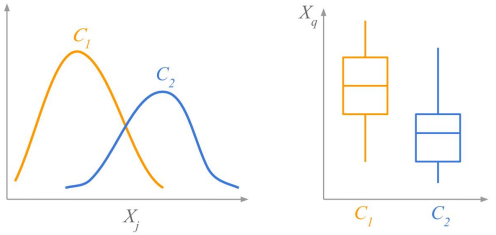
P(X = x y = k)

Classification - Exploring Conditional Distributions



- How does $X_j \mid y = 1$ compare with $X_j \mid y = 2$?
- How does $X_q \mid y = 1$ compare with $X_q \mid y = 2$?
- From data, we can have descriptive information about $X \mid y = k$. We calculate summary statistics, compare visual displays of these distributions.

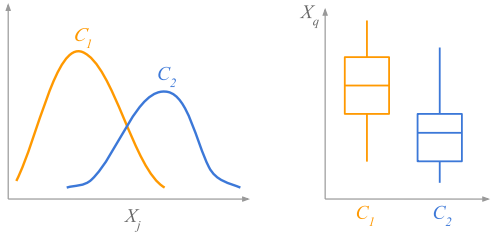
Phân loại - Khám phá phân phối có điều kiện



• Xj y = 1 so với Xj y = 2 như thế nào ? • Xq y = 1 so với Xq y = 2 như thế nào ? • Từ dữ liệu, chúng ta có thể có thông tin mô tả về X y = k.

Chúng tôi tính toán số liệu thống kê tóm tắt, so sánh hiển thị trực quan của các bản phân phối này.

Classification - Exploring Conditional Distributions



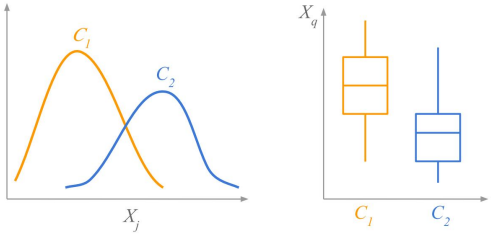
- If we have the class-conditional distribution $P(X \mid y = k)$, we can compute:

$$P(X = \mathbf{x} \mid \text{Good}) = \frac{\text{applicant is Good and has attributes } \mathbf{x}}{\text{applicant is Good}}$$

or

$$P(X = \mathbf{x} \mid \text{Bad}) = \frac{\text{applicant is Bad and has attributes } \mathbf{x}}{\text{applicant is Bad}}$$

Phân loại - Khám phá phân phối có điều kiện



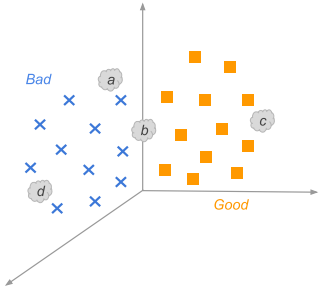
- Nếu chúng ta có phân phối loại có điều kiện P(X y = k), chúng ta có thể tính toán:

$$P(X = x \quad \text{Tốt}) = \text{ứng} \frac{\text{ứng viên là Tốt và có thuộc tính x}}{\text{viên Tốt}}$$

hoặc

$$P(X = x \quad \text{Xấu}) = \frac{\text{ứng viên là Xấu và có thuộc tính x}}{\text{ứng viên Xấu}}$$

Classification - Conditional Probability



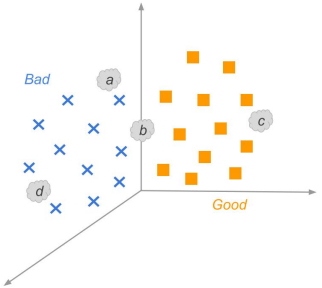
- However, we are actually interested in the conditional probability $P(y = k \mid X = \mathbf{x})$, we can compute:

$$P(\text{Good} \mid X = \mathbf{x}) = \frac{\text{applicant is Good and has attributes } \mathbf{x}}{\text{applicant has attributes } \mathbf{x}}$$

or

$$P(\text{Bad} \mid X = \mathbf{x}) = \frac{\text{applicant is Bad and has attributes } \mathbf{x}}{\text{applicant has attributes } \mathbf{x}}$$

Phân loại - Xác suất có điều kiện



- Tuy nhiên, chúng tôi thực sự quan tâm đến xác suất có điều kiện P(y = k X = x), chúng ta tính được:

$$P(\text{Tốt} \quad X = x) = \frac{\text{ứng viên là Tốt và có thuộc tính } x}{\text{ứng viên có thuộc tính } x}$$

hoặc

$$P(\text{Xấu} \quad X = x) = \frac{\text{ứng viên là Xấu và có thuộc tính } x}{\text{ứng viên có thuộc tính } x}$$

Bayes' Rule Reminder

- We have the conditional probabilities:

$$P(X = \mathbf{x} \mid y = k) = \frac{P(y = k, X = \mathbf{x})}{P(y = k)}$$

and

$$P(y = k \mid X = \mathbf{x}) = \frac{P(y = k, X = \mathbf{x})}{P(X = \mathbf{x})}$$

- We have the joint probability:

$$\begin{aligned} P(X = \mathbf{x}, y = k) &= P(y = k \mid X = \mathbf{x})P(X = \mathbf{x}) \\ &= P(X = \mathbf{x} \mid y = k)P(y = k) \end{aligned}$$

- Thus, we have:

$$P(y = k \mid X = \mathbf{x}) = \frac{P(X = \mathbf{x} \mid y = k)P(y = k)}{P(X = \mathbf{x})}$$

Nhắc nhở Quy tắc Bayes •

Ta có các xác suất có điều kiện:

$$P(X = x \quad y = k) = \frac{P(y = k, X = x)}{p(y = k)}$$

Và

$$P(y = k \quad X = x) = \frac{P(y = k, X = x)}{P(X = x)}$$

- Ta có xác suất chung là:

$$\begin{aligned} P(X = x, y = k) &= P(y = k \quad X = x)P(X = x) \\ &= P(X = x \quad y = k)P(y = k) \end{aligned}$$

- Như vậy, ta có:

$$P(y = k \quad X = x) = \frac{P(X = x \quad y = k)P(y = k)}{P(X = x)}$$

Bayes' Rule Reminder

$$P(y = k \mid X = \mathbf{x}) = \frac{P(X = \mathbf{x} \mid y = k)P(y = k)}{P(X = \mathbf{x})}$$

where the marginal probability $P(X = \mathbf{x})$ can be computed with the total probability formula:

$$P(X = \mathbf{x}) = \sum_k P(X = \mathbf{x} \mid y = k)P(y = k)$$

We can use Bayes' Theorem for classification purpose:

- $P(X = \mathbf{x} \mid y = k) = \pi_k$: the prior probability for class k.
- $P(X = \mathbf{x} \mid y = k) = f_k(\mathbf{x})$: the class-conditional density for inputs X in class k.

The **posterior probability** (the conditional probability of the response given the input) is:

$$P(y = k \mid X = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k}$$

Nhắc nhở Quy tắc Bayes

$$P(y = k \mid X = x) = \frac{P(X = x \mid y = k)P(y = k)}{P(X = x)}$$

trong đó xác suất cận biên P(X = x) có thể được tính bằng công thức xác suất tổng:

$$P(X = x) = \sum_k P(X = x \mid y = k)P(y = k)$$

Chúng ta có thể sử dụng Định lý Bayes cho mục đích phân loại:

- P(X = x | y = k) = πk: xác suất ưu tiên của lớp k.
- P(X = x | y = k) = fk(x): mật độ đầu vào có điều kiện của lớp X ở lớp k.

Xác suất sau (xác suất có điều kiện của phản hồi cho đầu vào) là:

$$P(y = k \mid X = x) = \frac{fk(x)\pi k}{\sum_{k=1}^{K} fk(x)\pi k}$$

Bayes' Rule Reminder

- The posterior probability:

$$P(y = k \mid X = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{k=1}^{K} f_k(\mathbf{x})\pi_k}$$

- By using Bayes' Theorem, we are modeling the posterior probability $P(y = k \mid X = \mathbf{x})$ in terms of likelihood densities $f_k(\mathbf{x})$ and prior probabilities π_k.

$$\texttt{posterior} = \frac{\texttt{likelihood} \times \texttt{prior}}{\texttt{evidence}}$$

Nhắc nhở Quy tắc Bayes

- Xác suất sau:

$$\texttt{P(y = k} \quad \texttt{X = x)} = \frac{\texttt{fk(x)}\pi\texttt{k}}{\sum_{\texttt{k=1}}^{\texttt{K}} \texttt{fk(x)}\pi\texttt{k}}$$

- Bằng cách sử dụng Định lý Bayes, chúng ta đang mô hình hóa hậu nghiệm xác suất P(y = k X = x) xét về mật độ khả năng fk(x) và xác suất trước πk.

$$\texttt{sau} = \frac{\texttt{khả năng} \times \texttt{trước}}{\texttt{chứng cớ}}$$

Bayes Classifiers

- In supervised learning, the goal is to find a model $\hat{f}()$ that makes good predictions.
- In a classification setting, we minimize the probability of assigning an individual $\mathbf{x}_i$ to the wrong class.
- We should classify $\mathbf{x}_i$ to the class k that makes $P(y = k \mid X = \mathbf{x})$ as large as possible, i.e., classify $\mathbf{x}_i$ to the most likely class, given its predictors.

Bộ phân loại Bayes

- Trong học có giám sát, mục tiêu là tìm một mô hình ˆf() khiến dự đoán tốt.
- Trong cài đặt phân loại, chúng tôi giảm thiểu khả năng chỉ định một cá nhân xi vào nhầm lớp.
- Chúng ta nên phân loại xi vào lớp k sao cho P(y = k X = x) càng lớn càng tốt, nghĩa là phân loại xi vào lớp có nhiều khả năng nhất, với các biến dự đoán của nó.