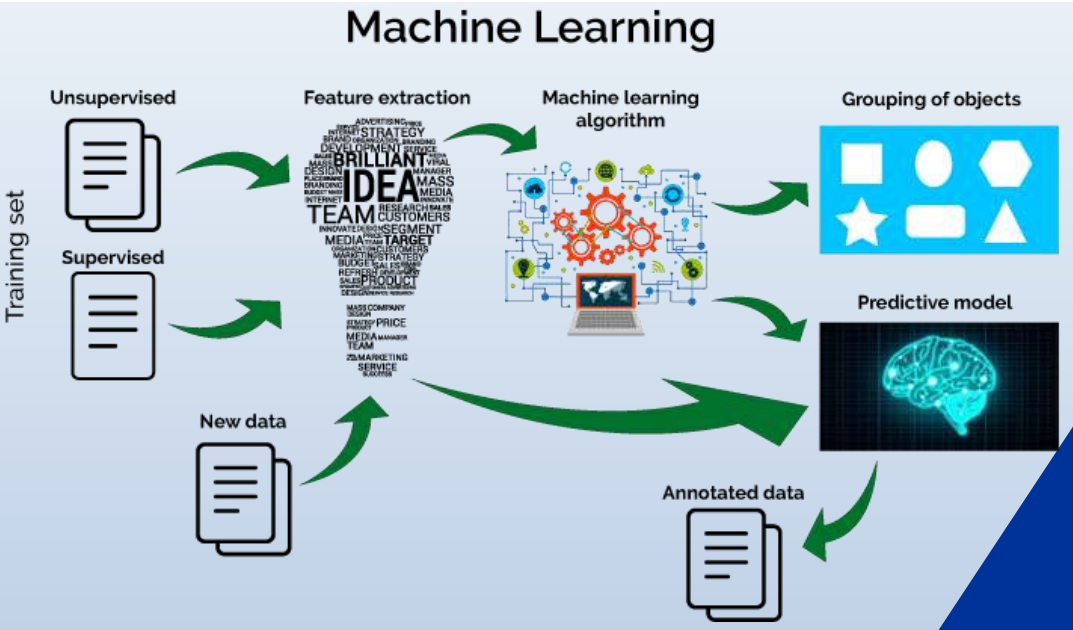
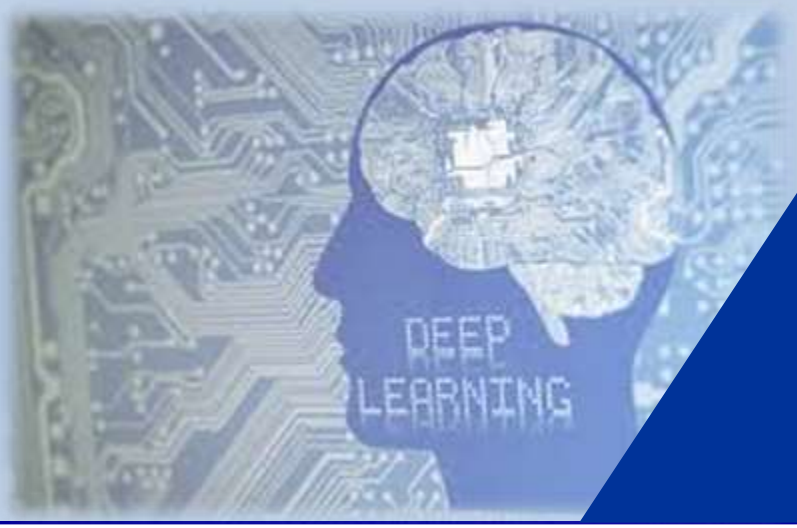
TRƯỜNG ĐẠI HỌC THUỶ LỢI
THUYLOI UNIVERSITY





KHOA CÔNG NGHỆ THÔNG TIN

HỌC MÁY NÂNG CAO

TS. Tạ Quang Chiểu
Email:
quangchieu.ta@tlu.edu.vn

Hà Nội ★ 01.2023

Lecturer information

Dr. TẠ QUANG CHIỂU

Position & Education

- Lecturer of Compter Scicence Engineering faculty (CSE) , TLU
- Ph.D in Computer Science, Polytech Tours, France

Field of research

- Big data, Data science, machine learing, AI
- Programing language C/C++/Python/...
- Scheduling and planning probleme, matheuristic and metaheristic method

Contact:

[E]: quangchieu.ta@gmail.com/quangchieu.ta@tlu.edu.vn

[M]: 0913 522 275

Tài liệu tham khảo

- Vũ Hữu Tiệp, *Machine Learning cơ bản*, *2018*. Link download https://github.com/tiepvupsu/ebookMLCB
- Blog: https:// machinelearningcoban.com
- Facebook Page: https:// www.facebook.com/machinelearningbasicvn/
- Facebook Group: https:// www.facebook.com/ groups/machinelearningcoban/
- Interactive Learning: https:fundaml.com
- Bài giảng Học máy của PGS.TS Nguyễn Hữu Quỳnh
- Bài giảng Học máy của PGS.TS Nguyễn Thanh Tùng
- Bài giảng Học máy của TS Nguyễn Thị Kim Ngân

CSE: **Faculty of Computer Science and Engineering**
Thuyloi University

Cây phân loại và hồi quy
(Classification and regression tree, CART)

Giới thiệu

- Dựa vào đặc điểm của biến mục tiêu, có thể chia Decision Tree thành hai dạng:
  - Classification Tree: nếu biến mục tiêu thuộc dạng categorical variable
  - Regression Tree: nếu biến mục tiêu thuộc dạng continous variable
- Sự khác nhau giữa **Classification Tree** và **Regression Tree**
  - Regression Tree có biến mục tiêu là biến liên tục, trong khi Classification Tree có biến mục tiêu la biến phân loại.
  - Trong Regression Tree, khi huấn luyện, giá trị tại nút lá bằng trung bình các giá trị biến mục tiêu của các điểm dữ liệu có trong nút đó. Nên khi đưa tập test vào, nếu các điểm dữ liệu rơi vào nút lá nào, kết quả trả ra sẽ là giá trị trung bình.
  - Với Classification Tree, khi huấn luyện, giá trị tại nút lá(phân lớp) bằng giá trị có tần suất cao nhất(Mode) của các dữ liệu trong nút đó. Nên khi đưa tập test vào, nếu các điểm dữ liệu rơi vào nút lá nào, kết quả trả ra sẽ là Mode.

Giới thiệu

- **Làm sao Decision Tree quyết định khi nào sẽ phân nhánh**
  - Các quyết định phân nhánh sẽ ảnh hưởng đến độ chính xác của Cây.
  - Cây hồi quy và cây phân lớp có các thuật toán phân nhánh khác nhau.
  - Có nhiều thuật toán phân nhánh, tùy vào kiểu của biến mục tiêu mà sử dụng thuật toán như thế nào.
  - Có thuật toán chính : **Gini Index (CART), Reduction in Variance**

Gini Index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Trong đó,

- C: số lớp cần phân loại
- $p_i = n_i / N$,
- n_i là số lượng phần tử ở lớp thứ i
- N là tổng số lượng phần tử ở node đó

Gini Index

$$gini\_index = gini(p) - \sum_{i=1}^{K} \frac{m_k}{M} gini(c_k)$$

Trong đó,

- gini(p): chỉ số gini ở node cha
- K: số node con được tách ra
- gini(c_k): chỉ số gini ở node con thứ k
- M: số phần tử ở node p
- m_i: là số phần tử ở node con thứ i

$$\sum_{i=1}^{K} m_i = M$$

Gini split

Chọn thuộc tính có hệ số $\text{Gini}_{\text{split}}$ nhỏ

- $\text{Gini}_{\text{split}} = \sum_{i=1}^{K} \frac{m_k}{M} gini(c_k)$

Example

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example

$$G(sunny) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$

$$G(overcast) = 1 - (\frac{4}{4})^2 = 0$$

$$G(rainy) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$

Từ đó có được $Gini$ của thuộc tính *outlook* sẽ bằng:

$$G_{split}(outlook) = \frac{5}{14}G(sunny) + \frac{4}{14}G(overcast) + \frac{5}{14}G(rainy) = \frac{5}{14}0.48 + \frac{4}{14}0 + \frac{5}{14}0.48 \approx 0.34$$

Lần lượt, sẽ được giá trị GiniGini của các thuộc tính còn lại:

$$G_{split}(temperature) \approx 0.43$$

$$G_{split}(humidity) \approx 0.365$$

$$G_{split}(wind) \approx 0.43$$

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

Example

id	outlook	temperature	humidity	wind	play
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainy	mild	high	weak	yes
5	rainy	cool	normal	weak	yes
6	rainy	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainy	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainy	mild	high	strong	no

$G(hot)=1-(2/4)^2-(2/4)^2$
$G(mild)=1-(4/6)^2-(2/6)^2$
$G(cool)=1-(3/4)^2-(1/4)^2$

$G_{split}(temperature)=(4/14)G(hot)+(6/14)G(mild) + (4/14)G(cool)$