CSE: Faculty of Computer Science and Engineering
Thuyloi University

Cây phân loại và hồi quy

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

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, Reduction in Variance**

Gini Index

- Gini phát biểu rằng, nếu lấy hai quan sát từ 1 tập dữ liệu đồng nhất thì xác suất hai quan sát đó cùng lớp là bằng 1.
  - Nó áp dụng cho biến mục tiêu là biến phân loại có dạng “Success” và “Failure”
  - Nó chỉ được dùng cho phân nhánh nhị phân.
  - Giá trị của Gini càng cao thì tính đồng nhất càng cao
  - CART(Classification and Regression Tree) sử dụng Gini cho phân lớp nhị phân

Gini Index

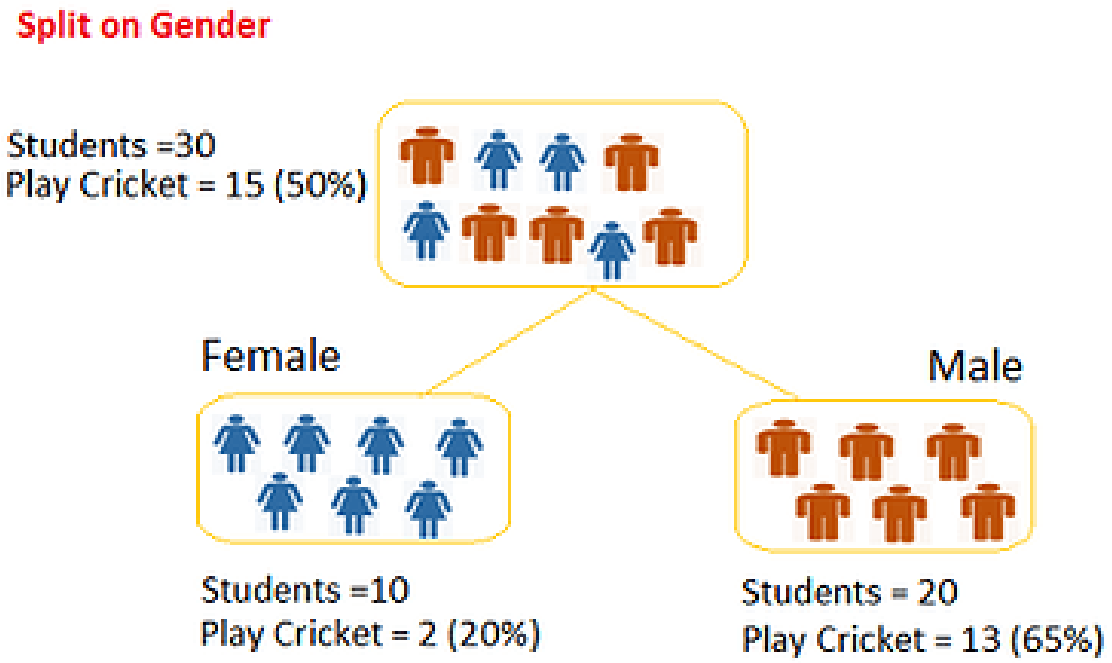
Các bước để tính chỉ số Gini

- Tính Gini cho sub-node bằng cách tính tổng bình phương xác suất của "succsess" và "failure" $(p^2 + q^2)$
- Tính trọng số Gini cho việc phân nhánh

Gini Index

Ví dụ:

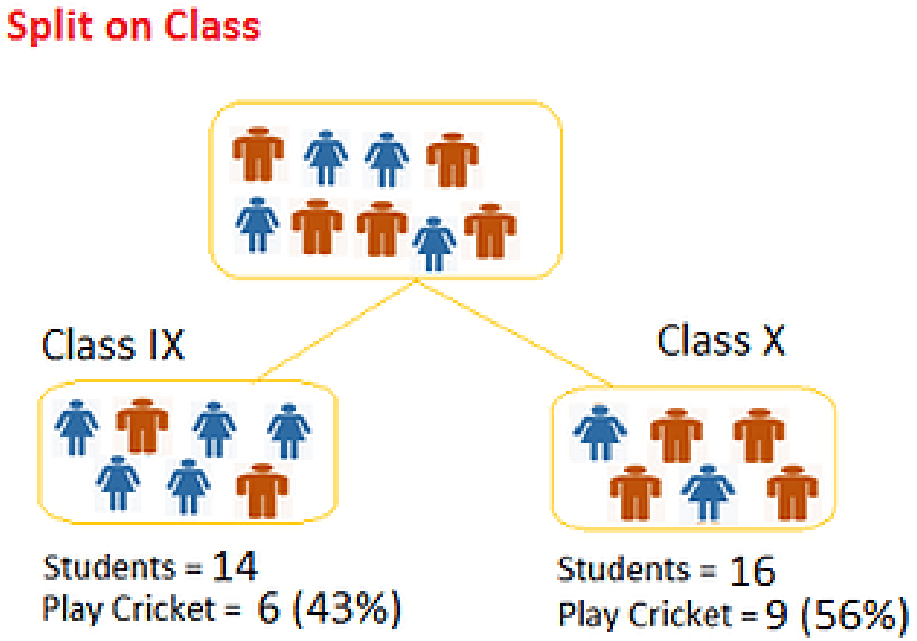
- Phân nhánh theo thuộc tính Gender
  - Tính Gini cho node Female : $0.2^2 + 0.8^2 = 0.68$
  - Gini cho node Male : $0.65^2 + 0.35^2 = 0.55$
  - Tính trọng số Gini cho việc phân nhánh theo Gender : $0.68 * \frac{10}{30} + 0.55 * \frac{20}{30} = 0.59$



Gini Index

Ví dụ:

- **Phân nhánh theo thuộc tính Class**
  - Tính Gini cho node Class IX : $0.43^2 + 0.57^2 = 0.51$
  - Tính Gini cho node Class X : $0.56^2 + 0.44^2 = 0.51$
  - Tính trọng số Gini cho việc phân nhánh theo Class : $0.51 * \frac{14}{30} + 0.51 * \frac{16}{30} = 0.51$



- Ta nhận thấy Gini Score cho **Gender** cao hơn **Class** , do đó việc phân nhánh sẽ dựa trên thuộc tính **Gender**

Reduction in Variance

- Thuật toán Gini áp dụng cho Categorical Decision Tree.
- Reduction in Variance là thuật toán sử dụng cho Regression Decision Tree:
  - Thuật toán sử dụng phương sai để chọn việc phân nhánh
  - Phân nhánh nào có phương sai nhỏ hơn thì sẽ được chọn.
- Công thức tính như sau

$$Variance = \frac{\Sigma(X-\bar{X})^2}{n}$$

- Trong đó $\bar{X}$ là giá trị trung bình, X là giá trị thực tế và n là số phần tử

Reduction in Variance

Các bước tính Variance :

- Tính variance trên mỗi node
- Variance cho mỗi phân nhánh bằng trung bình variance của các node con

Reduction in Variance

Ví dụ : Để đơn giản cho bài toán Regression , chúng ta tiến hành chuẩn hóa biến mục tiêu như sau : giá trị 1 đại diện cho Play và giá trị 0 đại diện cho Not play

- **Phân nhánh theo thuộc tính Gender**

1. Tính Variance của node cha: $\overline{X} = \frac{15*1+15*0}{30} = 0.5$
   $Variance = \frac{(1-0.5)^2+...+(1-0.5)^2+(0-0.5)^2+....+(0-0.5)^2}{30} = \mathbf{0.25}$
2. Tính Variance cho node Female: $\overline{X} = \frac{2*1+8*0}{10} = 0.2$
   $Variance = \frac{(1-0.2)^2+(1-0.2)^2+(0-0.2)^2+.....+(0-0.2)^2}{10} = 0.16$
3. Tính Variance cho node Male: $\overline{X} = \frac{13*1+7*0}{20} = 0.65$
   $Variance = \frac{(1-0.65)^2+....+(1-0.65)^2+(0-0.65)^2+.....+(0-065)^2}{20} = 0.23$
4. Tính Variance cho việc phân nhánh theo Gender $\frac{10}{30} * 0.16 + \frac{20}{30} * 0.23 = \mathbf{0.21}$

Reduction in Variance

- **Phân nhánh theo thuộc tính Class**

1. Tính Variance cho node IX: $\bar{X} = \frac{6*1+8*0}{14} = 0.43$

$$Variance = \frac{(1-0.43)^2+...+(1-0.43)^2+...+(0-0.43)^2+....+(0-043)^2}{14} = 0.24$$

2. Tính Variance cho node X: $\bar{X} = \frac{9*1+7*0}{16} = 0.56$

$$Variance = \frac{(1-0.56)^2+...+(1-0.56)^2+(0-0.56)^2+...+(0-0.56)^2}{16} = 0.25$$

3. Tính Variance cho việc phân nhánh theo Class: $\frac{14}{30} * 0.24 + \frac{16}{30} * 0.25 = \mathbf{0.25}$

- Ta nhận thấy Variance của **Gender** thấp hơn so với **Class** , do đó việc phân nhánh sẽ dựa trên thuộc tính **Gender**