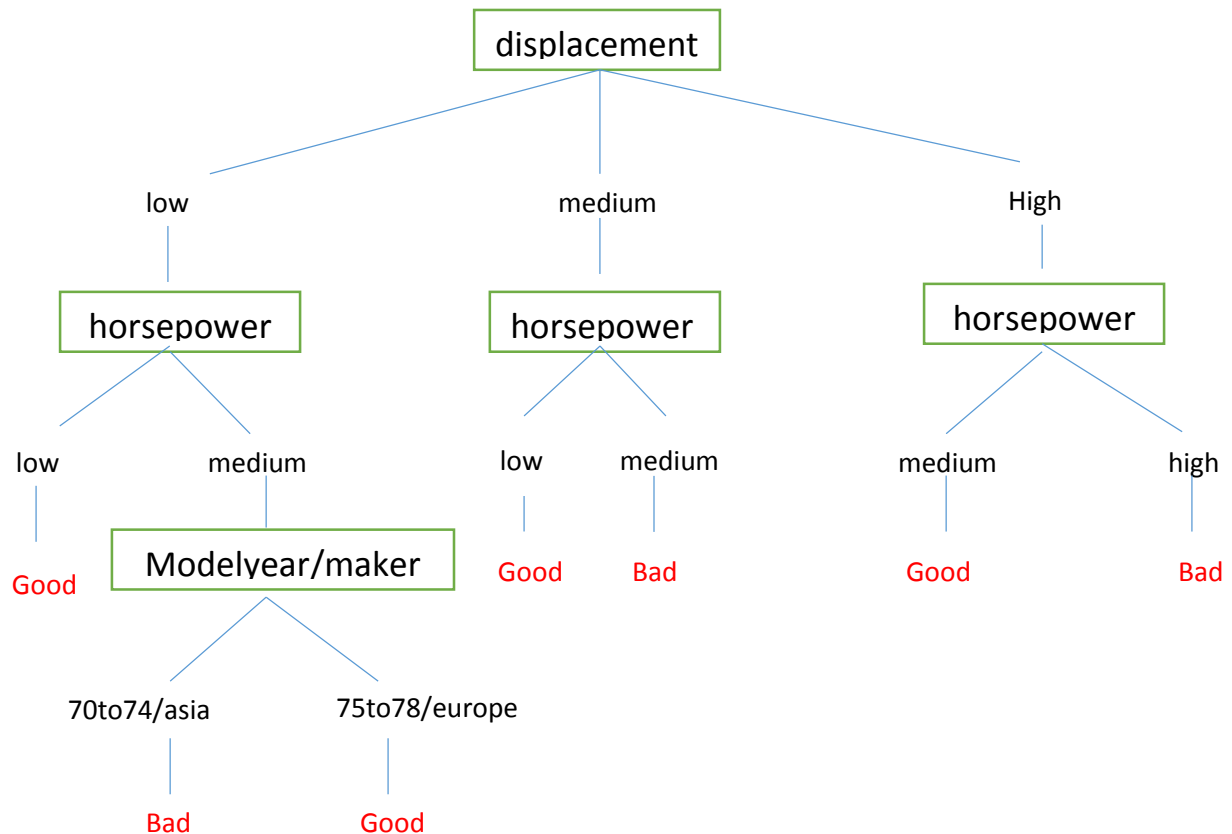


## Decision Trees – Vũ Xuân Luyện

- **Case 1:** Nếu chọn bắt đầu ‘cylinders’ thì với giá trị cylinders = 4: tiếp tục đến ‘horsepower’ với medium -> ‘acceleration’ với medium -> ‘modelyear’ mới đến được ‘pmg’. Như vậy cần 4 cấp feature.
- **Case 2:** Bắt đầu với ‘displacement’.



Tính mức độ hỗn loạn các phân tử trong đối tượng: Với good = 6, bad = 12.

$$\text{Entropy}(S) = -p+\log_2 p+ -p-\log_2 p- = -(6/18)\log_2(6/18) - (12/18)\log_2(12/18) = 0.5261 + 0.3896 = 0.9157$$

Để xác định feature nào là tốt nhất tính:

$$\text{Gain}(S, \text{displacement})$$

$$= \text{Entropy}(S) - \sum_{v \in \{ \text{displacement-low}, \text{displacement-medium}, \text{displacement-high} \}} (|S_v|/|S|) \text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - [(6/18) \cdot \text{Entropy}(S \text{ displacement-low}) + (6/18) \cdot \text{Entropy}(S \text{ displacement-medium}) + (6/18) \cdot \text{Entropy}(S \text{ displacement-high})]$$

$$= \text{Entropy}(S) - [(6/18) \cdot \text{Entropy}(S \text{ displacement-low}) + 0 + 0]$$

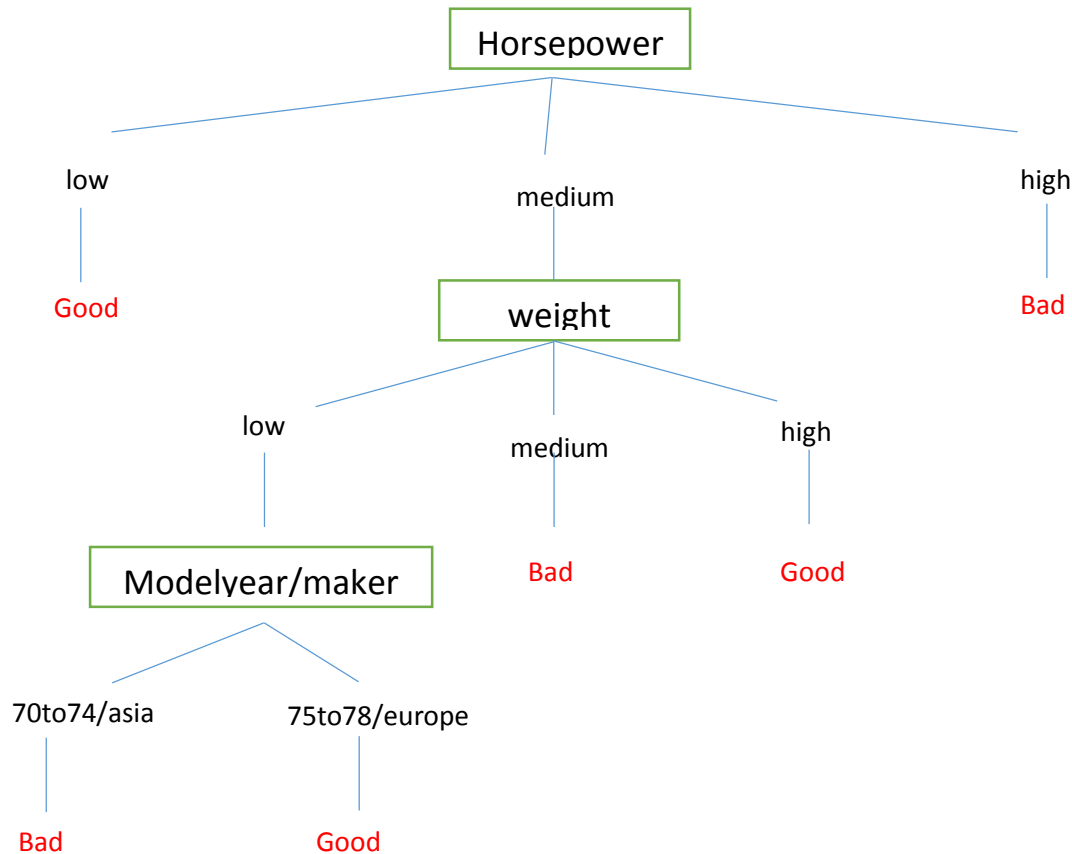
$$= \text{Entropy}(S) - [(6/18) \cdot [(3/6) \text{Entropy}(S \text{ horsepower-low}) + (3/6) \text{Entropy}(S \text{ horsepower-medium})]]$$

$$= \text{Entropy}(S) - [(6/18) \cdot [0 + (3/6) \text{Entropy}(S_{\text{horsepower-medium}})]]$$

$$= \text{Entropy}(S) - [(6/18) \cdot (3/6) [ \text{Entropy}(S_{\text{Modelyear -70to74}}) + \text{Entropy}(S_{\text{Modelyear -75to78}}) ] ]$$

$$= 0.9157 - (6/18)(3/6)((1/2)\log_2(1/2) - (1/2)\log_2(1/2)) = 0.9157 - 0.1667 = 0.794$$

### **Case 3:** Bắt đầu với ‘Horsepower’



$$\text{Gain}(S, \text{horsepower}) = \text{Entropy}(S) - \sum_{v \in \{ \text{horsepower-low}, \text{horsepower-medium}, \text{horsepower-high} \}} (|S_v|/|S|) \text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - [(4/18) \cdot \text{Entropy}(S_{\text{horsepower-low}}) + (9/18) \cdot \text{Entropy}(S_{\text{horsepower-medium}}) + (5/18) \cdot \text{Entropy}(S_{\text{horsepower-high}})]$$

$$= \text{Entropy}(S) - [0 + (5/18) \cdot \text{Entropy}(S_{\text{horsepower-medium}}) + 0]$$

$$= \text{Entropy}(S) - [(5/18) \cdot [ (3/9) \text{Entropy}(S_{\text{weight-low}}) + (5/9) \text{Entropy}(S_{\text{weight-medium}}) + (1/9) \text{Entropy}(S_{\text{weight-high}}) ] ]$$

$$= \text{Entropy}(S) - [(5/18) \cdot [ (3/9) \text{Entropy}(S_{\text{weight-low}}) + 0 + 0 ] ]$$

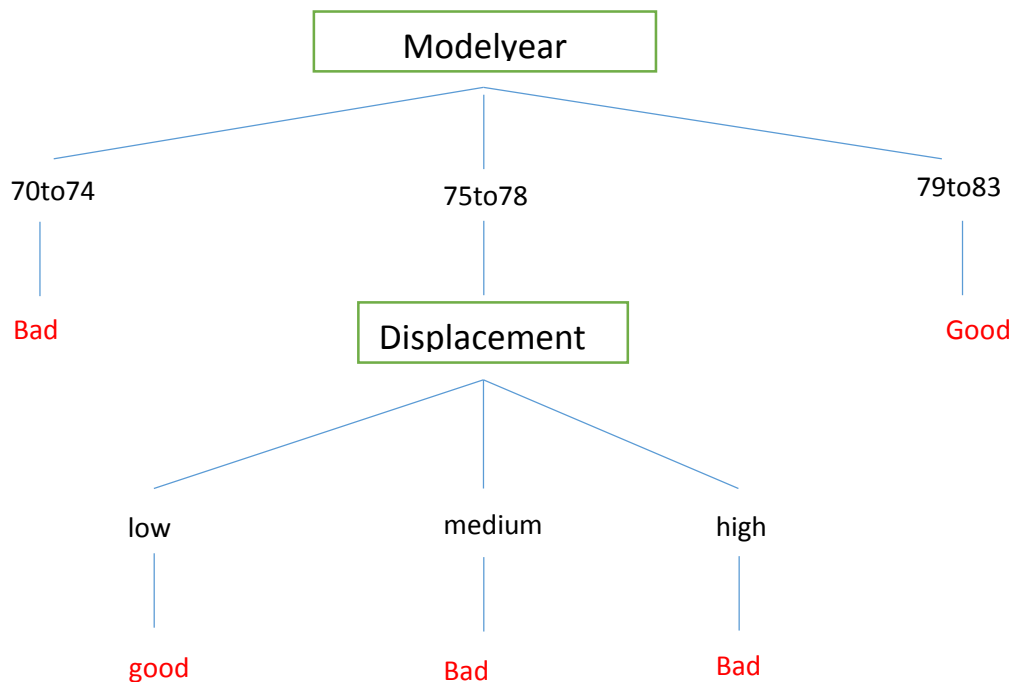
$$= 0.9157 - (5/18)(3/9)((1/2)\log_2(1/2) - (1/2)\log_2(1/2))$$

$$= 0.9157 - 0.0922 = 0.8235$$

**Case 4:** Nếu chọn bắt đầu từ ‘weight’ -> ‘horsepower’ thì sẽ lặp lại giống như case 2.

**Case 5:** Nếu chọn bắt đầu từ ‘acceleration’ với low -> ‘displacement’ với low -> ‘horsepower’/’modelyear’ -> ‘pmg’ cần 3 cấp feature giống với case 3.

**Case 6:** modelyear -> ‘weight’ hoặc ‘placement’



Gain(S, Modelyear)

$$= \text{Entropy}(S) - \sum_{v \in \{ \text{Modelyear -70to74}, \text{Modelyear -75to78}, \text{Modelyear -79to83} \}} (|S_v|/|S|) \text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - [(7/18) \cdot \text{Entropy}(S \text{ Modelyear -70to74}) + (7/18) \cdot \text{Entropy}(S \text{ Modelyear -75to78}) + (4/18) \cdot \text{Entropy}(S \text{ Modelyear -79to78})]$$

$$= \text{Entropy}(S) - [0 + (7/18) \cdot \text{Entropy}(S \text{ Modelyear -75to78}) + 0]$$

$$= \text{Entropy}(S) - [(7/18) \cdot \text{Entropy}(S \text{ displacement})]$$

$$= \text{Entropy}(S) - [(7/18) \cdot [(3/7) \text{Entropy}(S \text{ displacement -low}) + (2/7) \cdot \text{Entropy}(S \text{ displacement -medium}) + (2/7) \cdot \text{Entropy}(S \text{ displacement -high})]]$$

$$= 0.9157 - 0.3889 = 0.5268$$

**Case 7:** ‘maker’ với america -> ‘weight’ với high -> ‘horsepower’ -> ‘pmg’. Cần 3 cấp feature để đến ‘pmg’.

**Nhận xét:** qua giá trị Gain(S,feature) ta thấy case 6 với modelyear là Gain(S, Modelyear) nhỏ nhất nên tốt nhất.