Decision Tree Learning

Impurity

- **Entropy** đo mức độ **không đồng nhất** (impurity) của một tập dữ liệu.
- **Công thức entropy:**

$$H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

Information gain

- Tại mỗi **nút (node)**, ta chọn đặc trưng (feature) nào **giảm entropy nhiều nhất** – tức là **tăng độ tinh khiết (purity)**.
- Trong học máy, giảm entropy được gọi là **Information Gain** (IG).
- **Information Gain** được tính theo công thức:

$$\text{IG} = H(p_{\text{root}}) - \left(w^{\text{left}} H(p_1^{\text{left}}) + w^{\text{right}} H(p_1^{\text{right}})\right)$$

- **Nhánh trái (left)**: xác suất dương p_1^{left}, tỷ lệ dữ liệu w^{left}
- **Nhánh phải (right)**: xác suất dươngp_1^{right}, tỷ lệ dữ liệu w^{right}
- **IG càng lớn** thì đặc trưng càng tốt để phân nhánh.

Step

Ý tưởng chính

- Dùng **information gain** để chọn **feature tốt nhất** để tách tại mỗi **nút** của cây.
- Quá trình này **lặp lại đệ quy** trên từng nhánh con đến khi **thỏa mãn tiêu chí dừng (stopping criteria)**.

Quy trình tổng quát

1. **Bắt đầu từ root node** với toàn bộ tập huấn luyện.
2. Với mỗi đặc trưng:

- Tính **information gain**.

3. **Chọn feature có IG lớn nhất** để phân tách dữ liệu → tạo nhánh trái và phải.
4. **Chuyển mỗi tập con** sang cây con tương ứng (trái/phải).
5. **Lặp lại quá trình phân tách** tại mỗi node con cho đến khi đạt tiêu chí dừng.

Tiêu chí dừng có thể gồm:

- Node chứa **toàn bộ ví dụ cùng lớp** (entropy = 0).
- **Đạt độ sâu tối đa** cho phép (max depth).
- **Information gain < ngưỡng** định sẵn.
- **Số lượng ví dụ quá ít** tại node.

Continuous value

- Với **continuous feature** (như weight, age,...), cây quyết định:
  - **Tạo nhiều ngưỡng chia tiềm năng**
  - **Tính IG** cho mỗi ngưỡng
  - **Chọn ngưỡng cho IG cao nhất**
- Nếu IG từ đặc trưng liên tục này **cao hơn tất cả đặc trưng khác**, thì dùng nó để tách dữ liệu tại node.
- **Quy trình đệ quy lặp lại như thường** cho các nhánh con sau khi chia.

Regression Tree

Quá trình huấn luyện Regression Tree

Tại mỗi node:

1. **Xét các đặc trưng** có thể chia tách
2. Với mỗi đặc trưng:
   - Tách tập dữ liệu thành 2 nhánh.
   - **Tính phương sai** của nhãn mục tiêu (Y) ở mỗi nhánh.
   - Tính **trung bình trọng số của phương sai** sau khi chia

$$\text{Weighted Variance} = w^{\text{left}} \cdot \text{Var}_{\text{left}} + w^{\text{right}} \cdot \text{Var}_{\text{right}}$$

3. Tính **Giảm phương sai**

$$\text{Var}_{\text{root}} - \text{Weighted Variance}$$

4. **Chọn đặc trưng** chia tách có **giảm phương sai lớn nhất**.