**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**- - - 🙞 🕮** **🙜 - - -**

Icon

Description automatically generated

**BÁO CÁO**

MÔN: **NHẬP MÔN KHOA HỌC DỮ LIỆU**

CÂU 3: **Trình bày và xử lý bài toán phân lớp cây quyết định dựa vào học máy bằng ngôn ngữ Pyhon**

**Giảng viên hướng dẫn: TS. Tân Hạnh**

**Sinh viên thực hiện:**

Nguyễn Trần Đức Thuận – N19DCCN203

**TP. HCM, 20/06/2023**

# **Giới thiệu**

Mô hình cây quyết định là một mô hình được sử dụng khá phổ biến và hiệu quả trong cả hai lớp bài toán phân loại và dự báo của học có giám sát. Khác với những thuật toán khác trong học có giám sát, mô hình cây quyết định không tồn tại phương trình dự báo. Mọi việc chúng ta cần thực hiện đó là tìm ra một cây quyết định dự báo tốt trên tập huấn luyện và sử dụng cây quyết định này dự báo trên tập kiểm tra.

Hệ số entropy sẽ được tính bằng công thức:



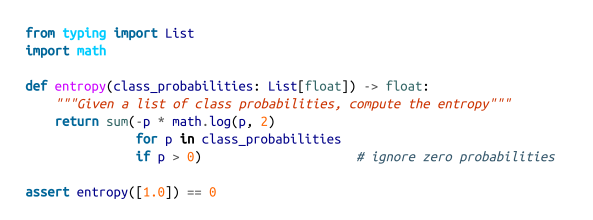
Hệ số entropy của từng thuộc tính sẽ tính bằng công thức:



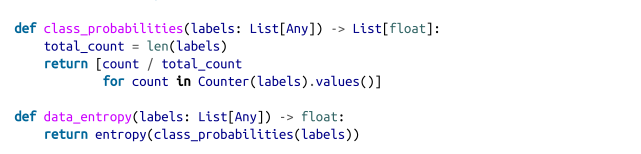
Với q là tỷ lệ số lượng của giá trị trên tổng số lượng giá trị của thuộc tính

1. **Hàm tính Entropy:**

Hàm **entropy** tính dựa trên các sác xuất của các nhãn

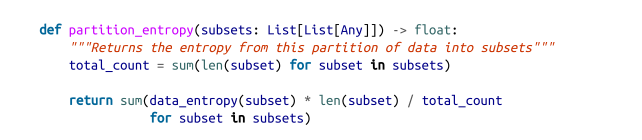


Hàm **class\_probabilities** tính xác suất của các nhãn xuất hiện trong cột nhãn



1. **Hàm tính entropy của một thuộc tính:**

Một thuộc tính sẽ có nhiều giá trị ta sẽ chia các giá trị thành từng subset và dung hàm **partition\_entropy** để tính entropy của thuộc tính đó

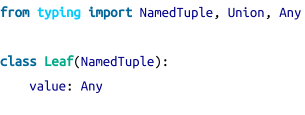


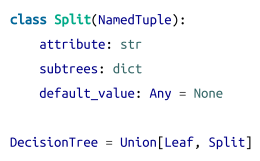
1. **Tạo cấu trúc dữ liệu cho cây:**

Sử dụng cấu trúc và dữ liệu sau đây:



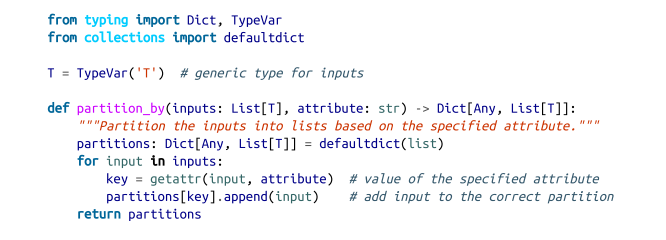
Node Leaf và Node Split



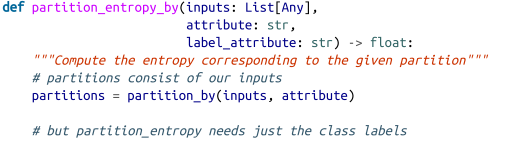


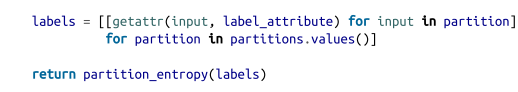
1. **Xư lý dữ liệu cho việc tạo thành cây**

Hàm **partition\_by** trả về dict có key là giá trị của một thuộc tính và value là một list chứa những dòng trong bộ dư liệu có chứa giá trị của thuộc tính đó.



Hàm **partition\_entropy\_by** tính entropy của một thuộc tính

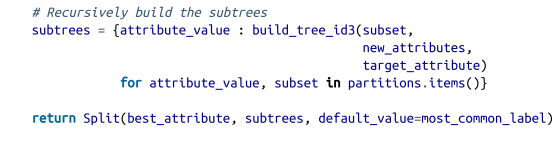
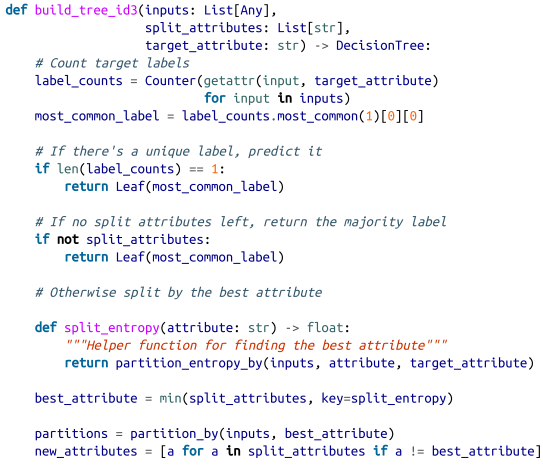




1. **Tạo cây**

Tạo cây sử dụng phép đệ quy:

Với các giá trị không có trong tập dữ liệu ta trả về nhãn có số lượng lớn nhất trong tập dữ liệu.



1. **Phân loại**

Dùng đệ quy đưa vào một dữ liệu để phân loại dựa theo cây có sẵn.

