Giới thiệu về thuật toán Decision Tree

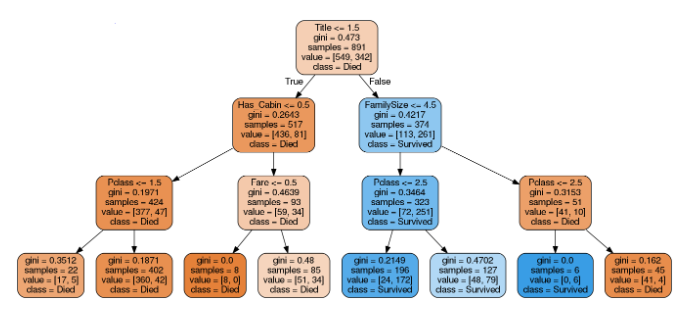
Một thuật toán Machine Learning thường có 2 bước:

1.Huấn luyện: Từ dữ liệu thuật toán sẽ học ra model

2.Dự đoán: Dùng model học được từ bước trên dự đoán ra giá trị mới

Bước huấn luyện ở thuật toán Decision Tree sẽ xây dựng ra một cây quyết định

Ví DỤ:



Thông tin Title được lấy ra từ trường Name. Sau đó trường Title, Sex được chuyển về dạng số:

Title\_mapping = {“Mr”: 1, “Miss”: 2, “Mrs”: 3, “Master”: 4, “Rare”: 5},

Sex\_mapping = {‘female’: 0, ‘male’: 1}

Tiếp theo, ở bước dự đoán, thuật toán dựa vào thông tin của khách hang và đi theo các điều kiện của cây từ trên xuống để dự đoán người đó đã sống hay chết.

Ví dụ với thông tin khách hang như sau:

| **PassengerId** | **Survived** | **Pclass** | **Name** | **Sex** | **Age** | **SibSp** | **Parch** | **Ticket** | **Fare** | **Cabin** | **Embarked** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.25 | NaN | S |

Từ trường Name, mình sẽ lấy được Title là “Mr” rồi chuyển về dạng số 1.

1.Title = 1 <= 1.5, điều kiện đúng nên ta tiếp tục kiểm tra điều kiện ở node con phía dưới bên trái.

2.Has\_Cabin = 0 <= 0.5, điều kiện đúng nên kiểm tra tiếp điều kiện ở node con ở dưới bên trái vì trường Cabin là NaN nên Has\_Cabin = 0.

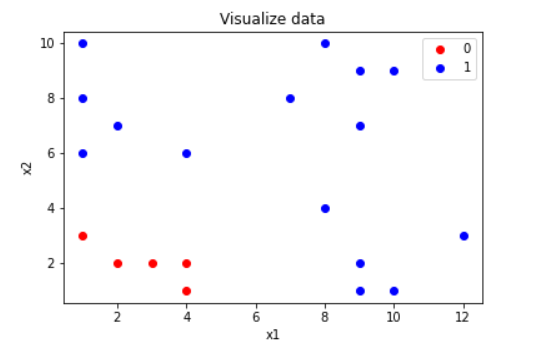
3.Pclass = 3 >= 1.5, điều kiện sai nên đi xuống node con bên phải, node này là node lá nên không cần kiểm tra nữa.

4.Lấy thông tin từ node lá, dự đoán class = Died.

**Xây dựng cây quyết định**

Ví dụ:

Giả sử mình có bài toán phân loại 2 lớp và mỗi dữ liệu có 2 thuộc tính là x1 và x2.



Xét điều kiện x1 > 5, chia dữ liệu làm 2 phần, 1 phần thỏa mãn điều kiện và 1 phần không thỏa mãn điều kiện.

Chart, scatter chart

Description automatically generated

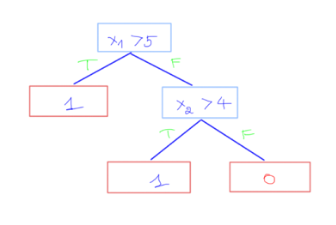
Ta thấy, nếu x1 > 5 đúng thì tất cả các dữ liệu thuộc lớp 1, thế nên ta sẽ dung lớp là để dự đoán đây là lớp 1. Ngược lại, ta thấy dữ liệu có cả lớp 1 và lớp 0, nên ta sẽ tiếp tục thêm điều kiện x2 > 4.

Chart, scatter chart

Description automatically generated

Nếu điều kiện x2 > 4 đúng, ta thấy các dữ liệu thuộc lớp 1, ngược lại các dữ liệu thuộc lớp 0. Do đó 2 node con của node điều kiện trên đều là node lá để cho ra kết quả dự đoán.

Ta có cây quyết định như sau:



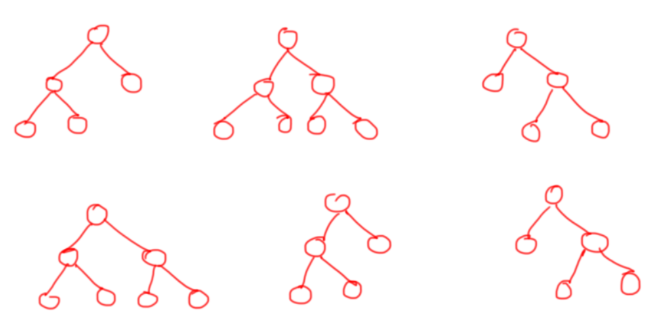
Ta đã có tiêu chí để chọn nhưng tiêu chí trên ta dựa vào cảm quan, máy tính cần số liệu để đánh giá và so sánh các điều kiện phân tách. Các chỉ số để đánh giá ra đời, bao gồm: entropy, information gain.

Với mỗi điều kiện để tách thì sẽ có chỉ số information gain tương ứng, chỉ số information gain càng cao thì việc tách càng tốt. Do đó mình sẽ duyệt qua hết các thuộc tính của dữ liệu, mỗi thuộc tính thử các giá trị để tách khác nhau, rồi chọn điều kiện có chỉ số information gain cao nhất để tách, và tiếp tục như thế cho tới node lá, chỉ gồm dữ liệu 1 lớp duy nhất.

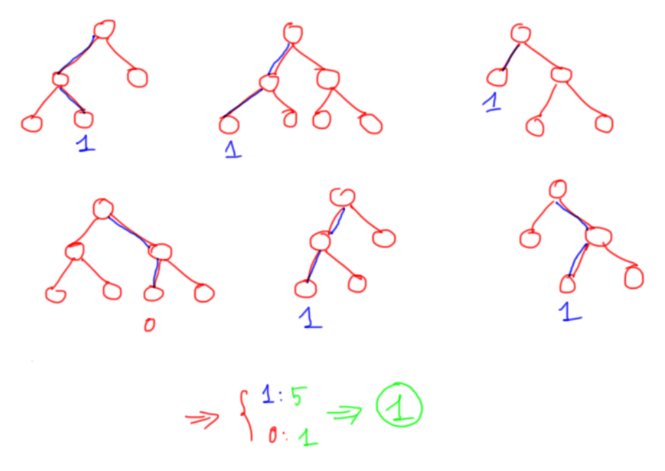
**Giới thiệu về thuật toán Random Forest**

Thuật toán Random Forest sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả được dự đoán dựa vào cây quyết định.

Ở bước huấn luyện ta sẽ xây dựng nhiều cây quyết định, các cây quyết định có thể khác nhau(phần sau mình sẽ nói mỗi cây được xây dựng như thế nào).



Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định mình sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán, sau đó kế quả cuối cùng được tổng hợp từ kết quả của các cây quyết định.



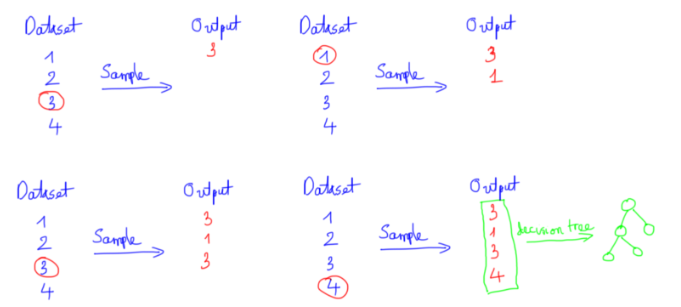
Trong ví dụ trê, thuật toán Random Forest có 6 cây quyết định, 5 cây dự đoán ra 1 và 1 cây dự đoán ra 0.

**Xây dựng thuật toán Random Forest**

Giả sử ta có bộ dữ liệu có n dữ liệu(sample) và mỗi dữ liệu có d thuộc tính(feature).

Ta sẽ xây dựng thuật toán Random Forest như sau:

1. Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kỹ thuật Bootstrapping. Tức khi mình sample được 1 dữ liệu thì mình không bỏ dữ liệu đấy ra mà vẫn giữ lại trong tập dữ liệu ban đầu, rồi tiếp tục sample cho tới khi sample đủ n dữ liệu. Khi dung kĩ thuật này thì tập n dữ liệu có thể có những dữ liệu trùng lặp nhau.



1. Sau khi sample được n dữ liệu từ bước 1, ta chọn ngẫu nhiên k thuộc tính. Ta có bộ dữ liệu mới gồm n dữ liệu và k thuộc tính.
2. Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu trên.