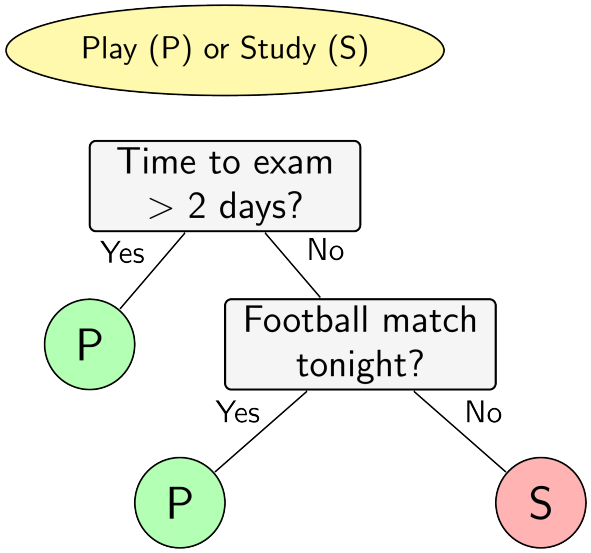
Decision Tree Quizz

Hoàng-Nguyên Vũ

1. **Mô tả:**

   - **Decision Tree** là một trong những thuật toán supervised-learning đơn giản nhất trong Machine Learning. Thuật toán này dựa trên các node được xây dựng từ trước và rẽ nhánh phù hợp để nhằm đưa ra kết quả cho bài toán.



Hình 1: Ví dụ về Decision Tree

2. **Bài tập: Lưu ý:** Một số câu có trên 2 đáp án

Câu 1. Hãy nêu ra sự khác biệt chính của việc áp dụng Decision Tree vào bài toán Classification và Regression ?

A) Thuật toán Decision Tree sử dụng ý tưởng Entropy và GINI cho bài toán Classification và ý tưởng Mean Square Error cho bài toán Regression

B) Thuật toán Decision Tree sử dụng ý tưởng Mean Square Error cho bài toán Classification và ý tưởng Entropy cho bài toán Regression

C) Thuật toán Decision Tree sử dụng ý tưởng tính khoảng cách Euclidean cho bài toán Classification và ý tưởng tính khoảng cách Mahattan cho bài toán Regression

D) Thuật toán Decision Tree sử dụng ý tưởng tính khoảng cách Mahattan cho bài toán Classification và ý tưởng tính khoảng cách Euclidean cho bài toán Regression

- **Đáp Án:** A - Vì ý tưởng chính của giải thuật Decision Tree: Entropy và Gini cho bài toán Classification và ý tưởng Mean Square Error cho bài toán Regression.

Câu 2. Quan sát đoạn code sau:

```
# Paragraph B
df = pd.read_csv('Salary_Data_simple.csv')

# Paragraph A
import numpy as np
import pandas as pd
import matplotlib . pyplot as plt
from sklearn.tree import DecisionTreeRegressor

# Paragraph D
dt_regressor = DecisionTreeRegressor(max_depth=2)
dt_regressor.fit(X, y)
y_pred_train = dt_regressor.predict(X)
y_pred = dt_regressor.predict(X_test)

# Paragraph C
X = df.iloc[:, -1]
y = df.iloc[:, -1]
X_train , X_test , y_train , y_test = train_test_split (X , y ,
    random_state = 0)

```

Thứ tự đúng của các đoạn trên là:

A) A - C - B - D

B) A - D - B - C

C) A - B - C - D

D) A - B - D - C

- **Đáp Án:** C - Vì theo thứ tự khi code: import thư viện → đọc file dataset → Chia dữ liệu train/test → dựng mô hình và kiểm tra trên tập test.

Câu 3. Để giảm tỉ lệ overfitting cho Decision Tree, chúng ta sử dụng kĩ thuật gì ?

A) Rebuilding Trees

B) Prunning

C) Boosting

D) None of the above

- **Đáp Án:** B - để giảm thiểu overfitting trong Decision Tree, kỹ thuật Prunning sẽ giúp chúng ta chặt bớt nhánh của cây, giúp giải thuật không bị overfit.

Câu 4. Entropy trong Machine Learning là gì ?

A) Là thuật ngữ đánh giả sự hỗn loạn của các phần tử trong vũ trụ.

B) Không đáp án chính xác.

C) Là một thuật ngữ được xài trong Decision Tree bởi cái tên bí ẩn.

D) Là thuật ngữ đo lường về thông tin đánh giá mức độ chắc chắn của một dataset.

- **Đáp Án:** D.

Câu 5. Đâu là lý do khiến chúng ta sử dụng hàm logarithm trong khi tính toán Entropy?

A) Để mô hình phân biệt với cách tính Gini
B) Bởi vì hàm logarithm được lập trình trong máy tính dễ dàng.
C) Để quy chuẩn thông tin về mặt độ lớn về cùng một tham chiếu.
D) Bởi vì nếu không sử dụng thì các con số được xử lý sẽ rất lớn.

- **Đáp Án:** C.

Câu 6. Cho biết Big-O Notation, ký hiệu là O() là công cụ đánh giá thời gian chạy của một thuật toán. Ví dụ: Thuật toán cộng các giá trị từ 1 tới n vào một biến sẽ có Big-O Notation là O(N).
Cho biết N là số lượng mẫu cho thuật toán, k là số lượng features, d là độ sâu của cây, hãy tính toán Big-O Notation thuật toán Decision Tree được xây dựng ?

A) O(N**2kd)
B) O(Nkd)
C) O(N)
D) O(Nkd**2)

- **Đáp Án:** B - Độ phức tạp thời gian và bộ nhớ của thuật toán Decision Tree phụ thuộc vào:
  + N: Số lượng mẫu trong tập dữ liệu
  + k: Số lượng features (thuộc tính)
  + d: Độ sâu của cây

Giai đoạn dự đoán: Duyệt cây từ gốc đến node lá: O(d) →Tính toán kết quả cho mỗi mẫu: O(k) → Tổng thời gian dự đoán trên toàn tập N mẫu data: O(n* d * k)

Câu 7. Lý do **chính** khi tính GINI tổng, chúng ta cần nhân thêm hệ số cho mỗi nhánh của node chính ?

A) Bởi để node chính trong trường hợp này không bị thua thiệt khi so sánh với các trường hợp khác.
B) Bởi nếu không thì GINI tổng của chúng ta sẽ vượt quá giá trị tối đa có thể.
C) Bởi để phân biệt sự khác nhau giữa mỗi nhánh
D) Bởi để đảm bảo sự đóng góp cho mỗi nhánh của node chính.

- **Đáp Án:** D.

Câu 8. Tiếp tục quan sát đoạn code dưới đây:

```
# Paragraph A
def gini_split_a(attribute_name):
    attribute_values = df1[attribute_name].value_counts()
    gini_A = 0
    for key in attribute_values.keys():
        df_k = df1[class_name][df1[attribute_name] == key].
   value_counts()
        n_k = attribute_values[key]
```

```
        n = df1.shape[0]
        gini_A = gini_A + ((n_k / n) * gini_impurity(df_k))
    return gini_A

gini_attribute = {}

# Paragraph B
def gini_impurity(value_counts):
    n = value_counts.sum()
    p_sum = 0
    for key in value_counts.keys():
        p_sum = p_sum + (value_counts[key] / n ) * (value_counts[key] / n)
    gini = 1 - p_sum
    return gini

class_value_counts = df1[class_name].value_counts()
gini_class = gini_impurity(class_value_counts)

# Paragraph C
min_value = min(gini_attribute.values())
selected_attribute = min(gini_attribute.keys())

```

Hãy đặt tên tương ứng cho nhiệm vụ ở mỗi đoạn:

A.
Paragraph A: Calculate Gini
Paragraph B:: Calculate Gini Impurity for the attributes
Paragraph C: Compute Gini gain values to find the best split, an attribute has maximum Gini gain is selected for splitting.

B.
Paragraph A:: Calculate Gini Impurity for the attributes.
Paragraph B: Calculate Gini.
Paragraph C: Compute Gini gain values to find the best split, an attribute has maximum Gini gain is selected for splitting.

C.
Paragraph A: Calculate Gini Impurity for the attributes, an attribute has maximum Gini gain is selected for splitting.
Paragraph B: Calculate Gini.
Paragraph C: Compute Gini gain values to find the best split.

D.
Paragraph A: Calculate Gini.
Paragraph B: Calculate Gini Impurity for the attributes, an attribute has maximum Gini gain is selected for splitting.
Paragraph C: Compute Gini gain values to find the best split.

- **Đáp Án:** B.

Câu 9. Các loại Decision Tree phổ biến là gì ?

A. SVM, KNN, Naive Bayes.

B. Linear Regression, Logistic Regression, Decision Tree.

C. Bởi vì thuật toán Decision Tree được xây dựng giống với các ra quyết định của con người hơn.

D. ID3, C4.5, CART.

- **Đáp Án:** D. ID3, C4.5 và CART đều là các thuật toán cây quyết định, là một loại mô hình học máy sử dụng cấu trúc dạng cây để phân loại hoặc dự đoán điểm dữ liệu. Chúng hoạt động bằng cách chia dữ liệu thành các tập hợp con ngày càng nhỏ hơn dựa trên các tính năng (thuộc tính) nhất định của dữ liệu, cuối cùng đi đến nút lá đại diện cho phân loại hoặc dự đoán.

Câu 10. Cho một tập Dataset như hình dưới đây:
Bạn hãy xây dựng cây quyết định và lần lượt chọn các cột theo thứ tự **Love Art, Love Nature, Love Math, Love Physics** làm node gốc để quyết định tỉ lệ **Love AI**. Sau đó, hãy tính tổng **GINI Impurity** cho từng lựa chọn và xem xét nên lựa chọn thông số nào làm node gốc

A) Love Art: 0.417, Love Nature: 0.476, Love Math: 0.5, Love Physics: 0.5, Root: Love Art

B) Love Art: 0.5, Love Nature: 0.5, Love Math: 0.417, Love Physics: 0.476, Root: Love Math

C) Love Art: 0.5, Love Nature: 0.5, Love Math: 0.476, Love Physics: 0.417, Root: Love Physics

D) Love Art: 0.476, Love Nature: 0.417, Love Math: 0.5, Love Physics: 0.5, Root: Love Nature

E) Love Art: 0.416, Love Nature: 0.476, Love Math: 0.5, Love Physics: 0.5, Root: Love Art

F) Love Art: 0.367, Love Nature: 0.492, Love Math: 0.394, Love Physics: 0.412, Root: Love Art

NO	LOVE ART	LOVE NATURE	LOVE MATH	LOVE PHYSICS	LOVE AI
1	TRUE	FALSE	TRUE	FALSE	FALSE
2	FALSE	FALSE	TRUE	FALSE	TRUE
3	FALSE	FALSE	TRUE	TRUE	TRUE
4	TRUE	TRUE	FALSE	TRUE	TRUE
5	TRUE	FALSE	TRUE	FALSE	FALSE
6	FALSE	TRUE	FALSE	TRUE	FALSE
7	FALSE	FALSE	TRUE	FALSE	TRUE
8	FALSE	TRUE	FALSE	TRUE	FALSE
9	TRUE	FALSE	TRUE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	TRUE

Hình 2: Dataset cho sẵn

- **Đáp Án:** E. Các bạn có thể xem lại công thức tính Gini: G = 1 - $\sum_{i=1}^{c}(p_i)^2$.

Câu 11. Cross-validation là gì ?

A) Kỹ thuật đánh giá hiệu suất của mô hình trên nhiều tập dữ liệu khác nhau.

B) Kỹ thuật huấn luyện mô hình trên nhiều tập dữ liệu khác nhau.

C) Kỹ thuật chọn lựa các thuộc tính tốt nhất để xây dựng cây quyết định.

D) Tất cả đáp án trên.

- **Đáp Án:** A - Cross-validation là một kỹ thuật được sử dụng để đánh giá hiệu suất của mô hình học máy trên nhiều tập dữ liệu khác nhau. Kỹ thuật này giúp giảm thiểu sai số và tăng độ tin cậy của kết quả đánh giá.

Câu 12. Chúng ta đã biết Gini và Entropy và hay cách để xây dựng Decision Tree cho bài toán Classification. Vậy đâu là lý do lý giải cho việc Gini được sử dụng thường xuyên hơn trong các bài toán thực tế?

A) Do Entropy là một khái niệm phức tạp và khó hiểu hơn.

B) Do thuật toán sử dụng Entropy có thời gian tính toán chậm hơn (bởi việc sử dụng hàm logarithm).

C) Do trong bài báo tác giả đã thử nghiệm trên rất nhiều trường hợp, và thực tế cho thấy rằng việc sử dụng Entropy lại bất ngờ cho kết quả thấp hơn.

D) Do Gini được khám phá gần đây hơn. Cùng với sự bùng nổ của Machine Learning gần đây thì Gini cũng được ưu chuộng hơn. (1912 so với 1850)

- **Đáp Án:** B.

Câu 13. Đâu là lý do **chính** mà chúng ta không chia Decision Tree tới Gini bằng 0 ?

A) Thuật toán sẽ chạy rất lâu

B) Điều này sẽ khiến thuật toán được sinh ra có tỉ lệ overfitting rất cao.

C) Tốn nhiều thời gian chia cây ở các tập dữ liệu lớn.

D) Chúng ta không thể đạt được trường hợp có GINI bằng 0.

- **Đáp Án:** B.

Câu 14. Pruning là gì?

A. Kỹ thuật tăng kích thước của cây quyết định để cải thiện độ chính xác.

B. Kỹ thuật chọn lựa các thuộc tính tốt nhất để xây dựng cây quyết định.

C. Kỹ thuật cắt tỉa các nhánh của cây quyết định để giảm thiểu overfitting.

D. Tất cả đáp án trên.

- **Đáp Án:** C.

Câu 15. Đâu là lời giải thích xác đáng cho 2 khái niệm Bias và Variance ?

A. Bias là thông số đánh giá độ lỗi trong quá trình training, Variance là thông số đánh giá độ chênh lệch giữa lỗi trong quá trình training và testing.

B. Bias là thông số đánh giá độ lỗi trong quá trình testing, Variance là thông số đánh giá độ chênh lệch giữa lỗi trong quá trình training và testing.

C. Variance là thông số đánh giá độ lỗi trong quá trình training, Bias là thông số đánh giá độ chênh lệch giữa lỗi trong quá trình training và testing.

D. Variance là thông số đánh giá độ lỗi trong quá trình testing, Bias là thông số đánh giá độ chênh lệch giữa lỗi trong quá trình training và testing.

- **Đáp Án:** A và B.

(*) **Ôn tập Toán Xác Suất cơ bản**

Câu 16. Gieo một con xúc xắc 6 mặt cân đối 2 lần. Xác suất để tổng số chấm xuất hiện trong hai lần gieo là 7 là ?

A. 1/36

B. 1/6

C. 1/12

D. 1/18

- **Đáp Án:** B.
  1. Xác định số kết quả có thể xảy ra: Khi gieo hai con xúc xắc 6 mặt cân đối, mỗi con có 6 khả năng xuất hiện (từ 1 đến 6). Do đó, có 6 * 6 = 36 kết quả có thể xảy ra.
  2. Xác định số kết quả thuận lợi: Để tổng số chấm xuất hiện trong hai lần gieo là 7, có 6 trường hợp sau: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)
  3. Xác suất để tổng số chấm xuất hiện trong hai lần gieo là 7 là: 6/36 = 1/6

Câu 17. Ba người cùng bắn vào một bia. Xác suất để người thứ nhất, thứ hai, thứ ba bắn trúng đích lần lượt là 0,8; 0,6; 0,5. Xác suất để có đúng 2 người bắn trúng đích là ?

A. 0.24

B. 0.96

C. 0.46

D. 0.92

- **Đáp Án:** C. Gọi ba người cùng bắn vào 1 bia với xác suất 0,8; 0,6; 0,5 lần lượt là A, B, C.
  + TH1: A, B bắn trúng, C không bắn trúng nên xác suất $P_1 = P_A * P_B * (1 - P_C) = 0.24$
  + TH2: A, C bắn trúng, B không bắn trúng nên xác suất $P_2 = P_A * (1 - P_B) * P_C = 0.16$
  + TH3: C, B bắn trúng, A không bắn trúng nên xác suất $P_3 = (1 - P_A) * P_B * P_C = 0.06$

Vậy xác suất cần tính là tổng xác xuất 3 TH trên: 0.46

Câu 18. Một lô hàng có 100 sản phẩm, biết rằng trong đó có 8 sản phẩm hỏng. Người kiểm định lấy ra ngẫu nhiên từ đó 5 sản phẩm. Tính xác suất của biến cố A: "Người đó lấy được đúng 2 sản phẩm hỏng" ?

A. 0.046

B. 0.084

C. 0.146

D. 0.208

- **Đáp Án:** A. Số phần tử của không gian mẫu: $\omega = C_{100}^5$. Trong 100 sản phẩm đó có 8 sản phẩm hỏng và 92 sản phẩm không hỏng nên số phần tử của biến cố A là: n(A) $C_8^2 * C_{92}^3$. Vậy xác suất như đề bài sẽ là $\frac{n(A)}{\omega} = 0.046$

Câu 19. Một hộp đựng 10 viên bi trong đó có 4 viên bi đỏ,3 viên bi xanh,2 viên bi vàng,1 viên bi trắng. Lấy ngẫu nhiên 2 bi tính xác suất biến cố : A: "2 viên bi cùng màu" ?

A. 1/9

B. 2/9

C. 1/3

D. 4/9

- **Đáp Án:** A. Số phần tử của không gian mẫu: $\omega = C_{10}^2$. Gọi các biến cố: D: "lấy được 2 viên đỏ" ; X: "lấy được 2 viên xanh" ; V: "lấy được 2 viên vàng". Ta có D, X, V là các biến cố đôi một xung khắc và C = D ∪ X ∪ V. Vậy P(C) = P(D) + P(X) + P(V) = $\frac{C_4^2}{C_{10}^2} + \frac{C_3^2}{C_{10}^2} + \frac{C_2^2}{C_{10}^2} = \frac{2}{9}$

- Hết -