



ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

CƠ SỞ TRÍ TUỆ NHÂN TẠO
LAB02: Decision Tree with scikit-learn

Sinh viên thực hiện : 21127083_Hoàng Đức Kiên

Lớp : 21CLC03

Giáo viên hướng dẫn : Nguyễn Ngọc Thảo

Lê Ngọc Thành

Nguyễn Trần Duy Minh

Nguyễn Hải Đăng

Phần 3 : Evaluating the decision tree classifiers

Câu hỏi : How do you interpret the classification report and the confusion matrix? From that, make your own comments on the performances of those decision tree classifiers.

_Độ chính xác (Precision): Độ chính xác là tỷ lệ giữa số lần dự đoán đúng (mẫu được dự đoán đúng) và tổng số lần dự đoán thuộc lớp dương. Nó đo lường khả năng của mô hình dự đoán đúng các mẫu thuộc lớp dương. Độ chính xác cao cho thấy tỷ lệ dương giả thấp.

_Độ phủ (Recall): Độ phủ là tỷ lệ giữa số lần dự đoán đúng và tổng số lần mẫu thực tế thuộc lớp dương. Nó đo lường khả năng của mô hình phát hiện các mẫu thuộc lớp dương. Độ phủ cao cho thấy tỷ lệ âm giả thấp.

_F1-score: F1-score là trung bình điều hòa của độ chính xác và độ phủ. Nó cung cấp một đánh giá cân bằng về hiệu suất của bộ phân loại bằng cách xem xét cả độ chính xác và độ phủ. Nó hữu ích khi các lớp không cân bằng.

_Support: Support biểu thị số lượng mẫu trong mỗi lớp.

_Hiệu suất của các bộ phân loại cây quyết định dựa trên các ma trận nhầm lẫn và báo cáo phân loại được cung cấp:

- Tỷ lệ huấn luyện/kiểm tra: 40/60 :

Bộ phân loại này đạt được độ chính xác và độ phủ cao cho lớp 0 (Không có) và lớp 1 (Một cặp). Nó gặp khó khăn trong việc phân loại các lớp khác một cách chính xác, đặc biệt là lớp 4 (Thùng) và lớp 9 (Thùng phá sảnh). Độ chính xác tổng thể của bộ phân loại này tương đối thấp, cho thấy nó có thể không tổng quát tốt với dữ liệu chưa được nhìn thấy.

- Tỷ lệ huấn luyện/kiểm tra: 60/40 :

Bộ phân loại này cho thấy hiệu suất cải thiện so với bộ phân loại trước đó. Nó đạt được độ chính xác và độ phủ cao cho lớp 0 (Không có), lớp 1 (Một cặp) và lớp 6 (Cặp ba). Tuy nhiên, nó vẫn gặp khó khăn trong việc phân loại lớp 4 (Thùng) và lớp 9 (Thùng phá sảnh). Độ chính xác tổng thể cải thiện nhẹ so với bộ phân loại trước đó.

- Tỷ lệ huấn luyện/kiểm tra: 80/20 :

Bộ phân loại này cho thấy sự cải thiện tiếp theo về hiệu suất. Nó đạt được độ chính xác và độ phủ cao cho lớp 0 (Không có), lớp 1 (Một cặp), lớp 4 (Thùng) và lớp 9 (Thùng phá sảnh). Độ chính xác tổng thể tăng đáng kể so với các bộ phân loại trước đó. Bộ phân loại này cho thấy hiệu suất triển vọng trong việc phân loại các bộ bài poker khác nhau.

- Tỷ lệ huấn luyện/kiểm tra: 90/10 :

Bộ phân loại này đạt được độ chính xác và độ phủ cao cho lớp 0 (Không có), lớp 1 (Một cặp), lớp 4 (Thùng) và lớp 9 (Thùng phá sảnh). Nó thể hiện độ chính xác tổng thể cao nhất so với tất cả các bộ phân loại khác. Bộ phân loại này cho thấy hiệu suất tốt nhất trong việc phân loại chính xác các bộ bài poker khác nhau.

Nhìn chung, khi kích thước tập huấn luyện tăng lên, các bộ phân loại cây quyết định đã cải thiện hiệu suất. Các bộ phân loại với tập huấn luyện lớn hơn đã đạt được độ chính xác, precision và recall cao hơn đối với hầu hết các lớp bài poker. Tuy nhiên, chúng vẫn gặp khó khăn trong việc phân loại chính xác một số lớp như lớp 4 (Straight) và lớp 9 (Royal flush), có thể khó phân biệt dựa trên các đặc trưng có sẵn. Để nâng cao hiệu suất của các bộ phân loại, việc khám phá các thuật toán khác hoặc kỹ thuật kỹ thuật trích xuất đặc trưng có thể mang lại lợi ích.

Phần 4 : The depth and accuracy of a decision tree

Max depth	None	2	3	4	5	6	7
Accuracy	0.647	0.506	0.508	0.527	0.558	0.558	0.559

Dựa trên các số liệu về độ chính xác (accuracy) với các giá trị maximum depth (độ sâu tối đa) khác nhau của bộ phân loại cây quyết định trên tỷ lệ train/test 80/20, chúng ta có thể nhận thấy những xu hướng sau:

- None (không giới hạn độ sâu): Độ chính xác đạt được khi không giới hạn độ sâu (unlimited depth) là 0.6472. Điều này cho thấy mô hình cây quyết định đang overfit dữ liệu huấn luyện và không tổng quát hóa tốt cho dữ liệu kiểm tra chưa được nhìn thấy trước.
- Độ sâu 2: Độ chính xác giảm xuống còn 0.5058 khi giá trị maximum depth được đặt là 2. Điều này cho thấy cây quyết định không đủ sức mạnh để mô phỏng độ phức tạp của vấn đề với một cây quyết định sâu như vậy.
- Độ sâu 3: Độ chính xác tăng nhẹ lên 0.5079 khi giá trị maximum depth là 3. Tuy nhiên, sự cải thiện này không đáng kể, cho thấy mô hình vẫn không đủ mạnh để mô phỏng dữ liệu một cách tốt.
- Độ sâu 4: Độ chính xác tiếp tục cải thiện lên 0.5271 khi giá trị maximum depth được tăng lên 4. Điều này cho thấy việc cho phép cây quyết định phát triển sâu hơn giúp mô hình nắm bắt được nhiều mẫu và cải thiện khả năng dự đoán.
- Độ sâu 5, 6 và 7: Độ chính xác tiếp tục tăng lên, đạt giá trị 0.5585 cho tất cả ba giá trị maximum depth. Điều này cho thấy cây quyết định dần dần học được các mẫu phức tạp hơn trong dữ liệu và cải thiện khả năng dự đoán của mô hình.

_Nhìn chung, chúng ta có thể nhìn thấy một sự cải thiện ban đầu về độ chính xác khi maximum depth được tăng từ 2 lên 4. Tuy nhiên, sau khi đạt độ sâu 4, sự cải thiện về độ chính xác giảm dần khi tăng độ sâu.

_Tuy nhiên, chỉ sử dụng độ chính xác để đánh giá mô hình không đủ để có cái nhìn toàn diện về hiệu suất của cây quyết định. Để đánh giá mô hình một cách chi tiết hơn, chúng ta cần xem xét các yếu tố khác như precision, recall và F1-score. Độ chính xác (accuracy) chỉ cho chúng ta biết tỷ lệ dự đoán chính xác trên toàn bộ tập dữ liệu, nhưng nó không cho biết mô hình có thể phân loại đúng các lớp cụ thể hay không.

Để có cái nhìn chi tiết hơn, chúng ta cần xem xét các yếu tố khác như precision, recall và F1-score cho từng lớp. Precision cho biết tỷ lệ dự đoán chính xác trong số các dự đoán thuộc một lớp cụ thể. Recall cho biết tỷ lệ các mẫu thuộc một lớp được phân loại đúng bởi mô hình. F1-score là trung bình điều hòa của precision và recall, cung cấp một phép đo cân bằng giữa hai yếu tố này. Để có cái nhìn tổng quan về hiệu suất của cây quyết định với độ sâu khác nhau, chúng ta cần xem xét không chỉ độ chính xác mà còn các yếu tố khác như precision, recall và F1-score.

Nguồn tham khảo :

- https://github.com/wwyc/decision_tree_visualization/blob/master/decistion_tree_hcart_example.ipynb?fbclid=IwAR1kPIBgrJONbe5wKn6McAYEbymlYEQnU4cOq9LHzHR0EjG6bp6-irRV_pc
- https://www.datacamp.com/tutorial/decision-tree-classification-python?fbclid=IwAR2R6hHk_AY38ERHnVwXD9JmaOC-TvjfM0RYxwktpdT6trZxIoFz0bOxp3k
- <https://www.stackvidhya.com/plot-confusion-matrix-in-python-and-why/>