Bài tập về nhà

Buổi số: 05

Bài 1: Khi sử dụng thuật toán để xây dựng cây quyết định, giả sử em tiến hành phân chia các nút cho tới khi có một cây quyết định rất phức tạp với nhiều nút, nhiều nhánh và nhiều nút lá với chỉ rất ít các điểm dữ liệu (vấn đề quá khớp xảy ra). Dựa trên những kiến thức đã học trên lớp và đọc thêm tài liệu ở nhà, em hãy trình bày và giải thích những cách có thể giải quyết được vấn đề quá khớp (overfitting) gặp phải khi xây dựng các cây quyết định.

Bài 2[1]: (*Giải bài toán bằng bút và máy tính cầm tay*)

Xét lại bộ dữ liệu huấn luyện như trình bày trong Bảng 1. Dựa vào ví dụ đã trình bày trên lớp, hãy thực hiện đầy đủ các bước giúp lựa chọn được đặc trưng để xây dựng nút gốc trong cây quyết định sử dụng thuật toán ID3.

Bảng 1*: Bộ dữ liệu huấn luyện gồm 14 mẫu; mỗi mẫu có 04 đặc trưng (Outlook, Temperature, Humidity, Wind) và nhãn (PlayTennis).*

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

[1] Dựa trên ví dụ tại cuốn sách "Machine Learning" của tác giả Tom M. Mitchell

Bài 3: (*Thực hành với Python*)

Trong ví dụ về phân lớp các loài hoa diên vĩ của mình, tác giả Andreas Mueller đã xây dựng một mô hình ML sử dụng *k*-Nearest Neighbors với $k = 1$ (KNeighborsClassifier(n_neighbors=1)). Ví dụ tham khảo có thể xem và download tại:

https://github.com/amueller/introduction_to_ml_with_python/blob/main/01-introduction.ipynb

Dựa trên ví dụ của tác giả, em hãy sử dụng một cách tiếp cận khác (ví dụ như Support Vector Machine, Decision Trees, Random Forests, …) để phân lớp các loài hoa diên vĩ. So sánh kết quả em đạt được với kết quả của tác giả.