SVM là một trong những phương pháp phân loại được dùng phổ biến nhất hiện nay do có độ chính xác cao, nhiều phần mềm và thư viện có hỗ trợ. Nhìn chung, khi có bài toán phân loại, nếu không có thêm thông tin gì đặc biệt thì thuật toán phân loại đầu tiên nên thử là SVM.

Boosting

Khác với những phương pháp trình bày ở trên, boosting là một dạng meta learning, tức là làm việc dựa trên những phương pháp học khác. Để giải quyết bài toán phân loại, boosting kết hợp nhiều bộ phân loại đơn giản với nhau để tạo ra một bộ phân loại có độ chính xác cao hơn.

Ví dụ, ta có thể xây dựng bộ phân loại đơn giản bằng cách sử dụng cây quyết định chỉ có một nút – nút gốc, còn được gọi là gốc cây quyết định. Cây quyết định như vậy sẽ có độ chính xác không cao. Thuật toán boosting khi đó kết hợp các gốc cây như sau:

- Mỗi ví dụ huấn luyện được gán một trọng số, đầu tiên trọng số bằng nhau.
- Thuật toán lặp lại nhiều vòng
  - Tại mỗi vòng, lựa chọn một gốc cây có độ chính xác tốt nhất. Gốc cây chính xác nhất là gốc cây có lỗi nhỏ nhất với lỗi tính bằng tổng trọng số những ví dụ bị phân loại sai.
  - Những ví dụ bị phân loại sai được tăng trọng số trong khi những ví dụ đúng bị giảm trọng số. Nhờ việc thay đổi trọng số như vậy, thuật toán sẽ chú ý nhiều hơn tới ví dụ bị phân loại sai trong những vòng sau.
- Bộ phân loại cuối được tạo ra bằng tổng các cây quyết định xây dựng tại mỗi vòng lặp.

Thuật toán boosting có một số ưu điểm như:

- Độ chính xác cao.
- Ít bị ảnh hưởng bởi hiện tượng quá vừa dữ liệu.
- Có thể sử dụng với nhiều phương pháp phân loại đơn giản khác nhau.

5.8. CÂU HỎI VÀ BÀI TẬP CHƯƠNG

1. Cho dữ liệu huấn luyện như trong bảng (f là nhãn phân loại).

X	Y	Z	f
1	0	1	1
1	1	0	0
0	0	0	0
0	1	1	1
1	0	1	1
0	0	1	0
0	1	1	1
1	1	1	0

a) Hãy xây dựng cây quyết định sử dụng thuật toán ID3. Trong trường hợp có hai thuộc tính tốt tương đương thì chọn theo thứ tự bảng chữ cái.

b) Giả sử không biết nhãn phân loại của ví dụ cuối cùng, hãy xác định nhãn cho ví dụ đó bằng phương pháp Bayes đơn giản (chỉ rõ các xác suất điều kiện thành phần) và k láng giềng gần nhất với k = 5.

2. Cho dữ liệu huấn luyện dưới đây với 16 ví dụ.

A	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
B	0	0	0	1	1	1	1	1	0	0	0	0	0	1	1	1
C	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	1

Sử dụng phân loại Bayes để tính giá trị của C nếu biết A = 0, B = 1. Yêu cầu viết chi tiết các bước.

3. Cho dữ liệu huấn luyện như trong bảng (f là nhãn phân loại).
   a) Hãy xác định nhãn cho ví dụ (Màu: Trắng, Hình dạng: Tròn, KL: Nặng) bằng phương pháp Bayes đơn giản (chỉ rõ các xác suất điều kiện thành phần)
   b) Hãy xác định nút gốc cho cây quyết định sử dụng thuật toán ID3

Màu	**Hình dạng**	**KL**	**f**
Xanh	Tròn	Nặng	+
Đỏ	Tròn	Nhẹ	-
Xanh	Méo	Nhẹ	+
Trắng	Méo	Nặng	+
Đỏ	Méo	Nặng	-
Trắng	Tròn	Nhẹ	-
Trắng	Méo	Nhẹ	+

4. Cho dữ liệu huấn luyện như trong bảng sau, trong đó mỗi cột (trừ cột ngoài cùng bên trái) ứng với một mẫu, dòng dưới cùng (T) chứa giá trị đích:

1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
2	0	0	0	1	1	0	0	1	1	0	1	0	1	1
3	1	1	1	0	1	0	0	1	1	0	0	0	1	1
4	0	1	0	0	1	0	0	1	0	1	1	1	0	1
5	0	0	1	1	0	1	1	0	1	1	0	0	1	0
6	0	0	0	1	0	1	0	1	1	0	1	1	1	0
T	1	1	1	1	1	1	0	1	0	0	0	0	0	0

a) Hãy thực hiện thuật toán giảm gradient cho hồi quy tuyến tính với dữ liệu trên, ghi lại giá trị trọng số sau mỗi bước.

b) Xây dựng cây quyết định cho dữ liệu trong bảng.

5. Hãy vẽ cây quyết định để biểu diễn các biểu thức logic sau: $\wedge$, $\vee$, $\neg$, $\Rightarrow$, $\Leftrightarrow$

a) $A \wedge \neg B$

b) $A \vee (B \wedge C)$

c) $A \Leftrightarrow B$

d) $(A \wedge B) \wedge (C \wedge D)$