Hoạt Động của XGBoost

MỤC TIÊU BÀI HỌC

Sau bài học này, học viên sẽ:

- Hiểu rõ cơ chế hoạt động của XGBoost.
- Biết cách XGBoost sử dụng trọng số để cải thiện dự đoán.
- Hiểu khái niệm **ensemble learning** và cách XGBoost xử lý **đa cộng tuyến (multicollinearity)**.
- Biết cách áp dụng XGBoost qua ví dụ minh họa.

I. CÁCH XGBOOST HOẠT ĐỘNG

1. Ví dụ Minh Họa Cơ Bản

Giả sử bạn có bảng dữ liệu nhỏ:

Quan sát	Đặc trưng (X)	Kết quả thật (y)
1	0.2	1
2	0.8	0
3	0.5	1
4	0.4	0

2. Vòng Lặp 1 (Cây đầu tiên)

- Mô hình khởi đầu: Mỗi quan sát có **trọng số bằng nhau**.
- Mô hình dự đoán: đúng quan sát 1 và 2, sai quan sát 3 và 4.
- **Kết quả**:
  - Quan sát đúng → giảm trọng số.
  - Quan sát sai → tăng trọng số.

XGBoost học từ sai lầm bằng cách ưu tiên học tốt hơn ở các điểm đã sai.

3. Vòng Lặp 2 (Cây thứ hai)

- Dựa trên trọng số mới.

- Mô hình học tập trung vào các điểm sai (ví dụ: quan sát 3 và 4).
- Lặp lại quá trình: cập nhật mô hình → đánh giá → điều chỉnh trọng số.

Mỗi cây mới **sửa lỗi của cây trước**.

II. ENSEMBLE & SUBSAMPLING

1. Không dùng toàn bộ dữ liệu

- XGBoost **không sử dụng toàn bộ quan sát** trong mỗi vòng lặp → gọi là **subsampling**.
- Ví dụ: Cây 1 bỏ qua quan sát số 3, cây 2 bỏ qua số 4…

2. Không dùng toàn bộ đặc trưng

- Mỗi cây học với một tập con của các đặc trưng.
- Gọi là **column subsampling**.

Nhờ đó, XGBoost tạo ra **nhiều mô hình nhỏ**, mỗi mô hình học trên dữ liệu khác nhau → tổng hợp lại tạo mô hình mạnh mẽ.

Đây chính là **Ensemble Learning**.

III. ƯU ĐIỂM CỦA XGBOOST

Tự động sửa lỗi qua từng vòng
Tránh quá khớp (overfitting) nhờ dùng một phần dữ liệu
Xử lý tốt đa cộng tuyến (Multicollinearity)
Hiệu quả cao cả về tốc độ và độ chính xác

IV. VÍ DỤ THỰC TẾ ĐƠN GIẢN

Bài toán: Dự đoán khách hàng có mua hàng không (1 = mua, 0 = không)

Tuổi	Số lần truy cập	Kết quả
25	3	1
40	5	0
30	2	1
45	6	0

1. **Cây đầu tiên**:
   - Dự đoán đúng 2, sai 2.

- Cập nhật trọng số.

2. **Cây thứ hai**:
   - Tập trung học các điểm sai.
   - Giảm sai sót tổng thể.
3. **Cây thứ ba trở đi**:
   - Lặp lại quá trình.

Kết quả cuối cùng: Dự đoán chính xác cao hơn nhờ tổ hợp các cây nhỏ.

VI. KẾT LUẬN

- XGBoost học qua từng vòng → cải thiện kết quả dần dần.
- Sử dụng kỹ thuật **tổ hợp** và **lấy mẫu ngẫu nhiên** để tăng hiệu suất.
- Là một trong những thuật toán mạnh mẽ nhất hiện nay cho bài toán phân loại và hồi quy.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 15 tháng 5 2025 03:09:12 bởi Đỗ Ngọc Tú
Được cập nhật 15 tháng 5 2025 03:27:15 bởi Đỗ Ngọc Tú