CSE: **Faculty of Computer Science and Engineering**
Thuyloi University

Học kết hợp

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Giới thiệu

- Tư tưởng của học nhóm là khai thác sức mạnh của tập thể:
- Nếu không dùng học kết hợp:
  - Ta có một model nhưng đầu ra của model đó không tốt nên ta phải thử các model khác.
  - Sau khi tìm được model ưng ý, ta lại phải chỉnh chỉnh sửa sửa từ thuật toán đến hyperparameter để mô hình đạt độ chính xác cao nhất.
  - Hai việc này sẽ ngốn một đống thời gian bởi ta phải chạy từng model một,
- Để nhanh hơn, ta kết hợp những model "học yếu" này lại để tạo ra một model "học mạnh" hơn, không những thế kết quả thu được cũng tốt hơn so với từng model một.

Giới thiệu

Mô hình "yếu" và "mạnh“

- Khi làm các bài toán về phân loại (classification) hay hồi quy (regression), phần quan trọng nhất là lựa chọn model:
  - Việc chọn này phụ thuộc nhiều yếu tố: số lượng data, đặc điểm data (số chiều, phân phối), v.v...
  - Từ đó ta sẽ có tương quan giữa data và model (bias-variance tradeoff)
- Nói chung, không có một model nào hoàn hảo khi đi riêng lẻ, các model này có điểm yếu rõ rệt như:
  - có model bị high bias (model dự đoán sai so với giá trị thực tế rất nhiều)
  - có model bị high variance (đoán đúng trên bộ dữ liệu train nhưng kém với bộ dữ liệu chưa gặp bao giờ)

  nên chúng đều bị gọi là "yếu".
- Vậy tại sao ta không kết hợp các model "yếu" để tạo ra một model "mạnh"

Bootstrapping

- Bootstrap method là phương pháp lấy mẫu có hoàn lại (sampling with replacement).
- Phương pháp lấy mẫu có hoàn lại có nghĩa là một cá thể có thể xuất hiện nhiều lần trong một lần lấy mẫu.
- Giả sử ta có 5 quan sát (observation) được đánh nhãn A,B,C,D và E trên 5 quả bóng và bỏ tất cả chúng vào trong 1 cái giỏ.

Bootstrapping

- Từ 5 quan sát này:
  - ta lấy ra 1 quả bóng từ giỏ một cách ngẫu nhiên và ghi lại nhãn của chúng, sau đó bỏ lại quả bóng vừa bốc được vào giỏ và
  - tiếp tục lấy ra một quả bóng một cách ngẫu nhiên, ghi lại nhãn của bóng và bỏ lại quả bóng vào trong giỏ
  - và tiếp tục thực hiện việc lấy mẫu như vậy cho đến khi kết thúc.
  - Việc lấy mẫu này gọi là lấy mẫu có hoàn lại.
- Kết quả của việc lấy mẫu như trên có thể như sau (giả sử kích thước mẫu là 12): C, D, E, E, A, B, C, B, A, E, A, D

Bootstrapping

Các bước chính của Bootstrap method:

- Sinh ra các mẫu (Bootstrap sampling) ngẫu nhiên có hoàn lại kích thước n từ tổng thể (từ mẫu ban đầu).
- Tính các thông số thống kê đặc trưng cho của mẫu được sinh ra (mean, Confident interval, Standard Deviation, Inter Quartile,…)
- Lặp lại bước 1 và bước 2 với số lần lớn (thường trên 1000)
- Sử dụng các ước lượng thống kê của Bootstrap sampling đã tính ở bước 2 để đánh giá độ chính xác các ước lượng thống kê của mẫu ban đầu (Original sample, Training Data).

Mô hình học kết hợp

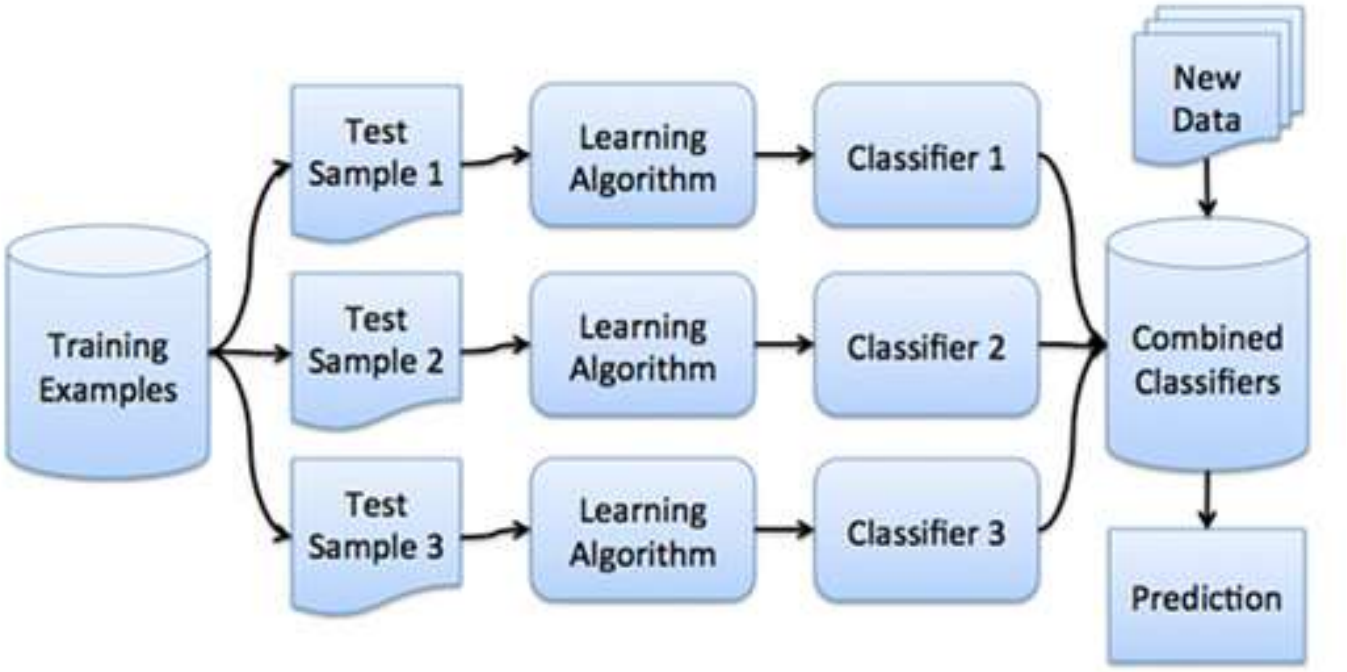
Bagging

- Xây dựng một lượng lớn các model (thường là cùng loại) trên những subsamples khác nhau từ tập training dataset (random sample trong 1 dataset để tạo 1 dataset mới).

- Những model này sẽ được train độc lập và song song với nhau nhưng đầu ra của chúng sẽ được tính là trung bình cộng (hoặc bỏ phiếu) để cho ra kết quả cuối cùng.

Mô hình học kết hợp

- Minh họa Bagging

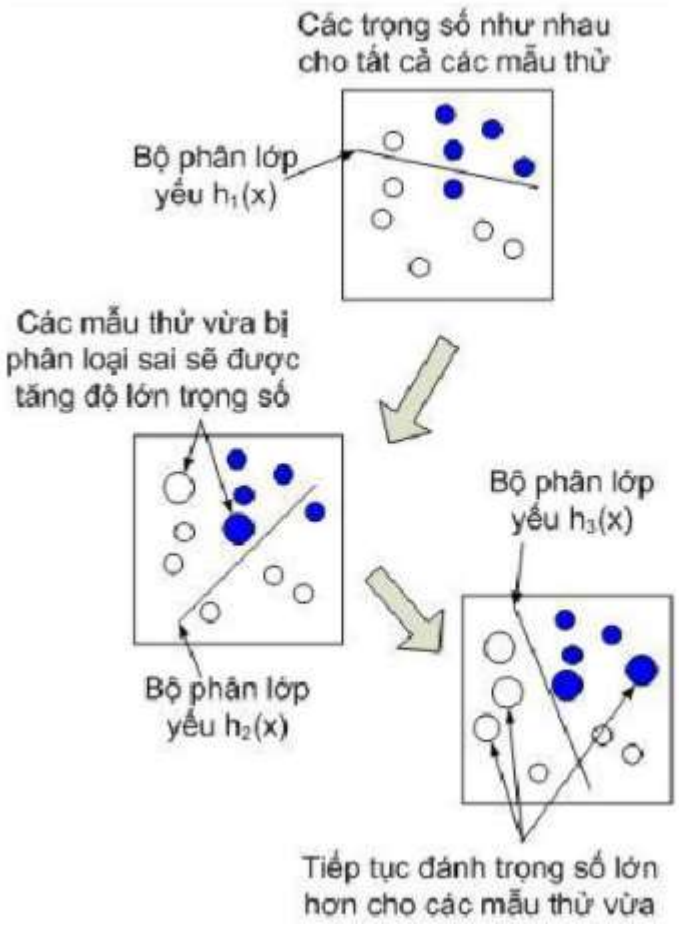


Mô hình học kết hợp

Boosting:

- Xây dựng một lượng lớn các model (thường là cùng loại).
- Mỗi model sau sẽ học cách sửa những error của model trước (dữ liệu mà model trước dự đoán sai)
- Tạo thành một chuỗi các model mà model sau sẽ tốt hơn model trước bởi trọng số được update qua mỗi model:
  - trọng số của những dữ liệu dự đoán đúng sẽ không đổi,
  - còn trọng số của những dữ liệu dự đoán sai sẽ được tăng thêm
- Chúng ta sẽ lấy kết quả của model cuối cùng trong chuỗi model này làm kết quả trả về (vì model sau sẽ tốt hơn model trước nên tương tự kết quả sau cũng sẽ tốt hơn kết quả trước).

Mô hình học kết hợp



Mô hình học kết hợp

Stacking:

- Xây dựng một số model (thường là khác loại) và một meta model (supervisor model),
- Train những model này độc lập,
- Sau đó meta model sẽ học cách kết hợp kết quả dự báo của một số mô hình một cách tốt nhất.

Mô hình học kết hợp