"Đoàn kết là sức mạnh". Tôi không biết câu này xuất hiện ở đâu khi nào nhưng nó lại khái quát ý tưởng của phương thức Ensemble trong Machine Learning.

Lấy ví dụ thế này, bạn có 1 model nhưng đầu ra của model đó không tốt nên bạn thử các model khác. Sau khi tìm được model ưng ý và "có lẽ chính xác", bạn lại phải chỉnh chỉnh sửa sửa từ thuật toán đến hyperparameter để mô hình đạt độ chính xác cao nhất. Tất cả những việc kể trên sẽ ngốn của bạn một đống thời gian bởi bạn phải chạy từng model một, thế nên để nhanh hơn bạn 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.

Để hiểu sâu thêm, chúng ta sẽ làm rõ khái niệm mô hình "yếu" và "mạnh".

Khi làm các bài toán về phân loại (classification) hay hồi quy (regression), chắc ai cũng biết 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) aka (mối quan hệ đánh đổi giữa bias và variance)

First, chọn model sẽ làm base model cho cả thuật toán. Thường thì một model sẽ được chọn ra (ví dụ là Decision Tree), ta lại cần nhiều mô hình "yếu" nên ta phải tăng số lượng model cần dùng lên => ta có n model Decision Tree

Một điểm quan trọng trong khi chọn model ở đây là chúng ta nên chọn model phù hợp với chiến thuật mà chúng ta lựa chọn ban đầu. Như các model có bias thấp nhưng variance cao vậy thì khi ta kết hợp chúng lại để tạo ra một model mạnh hơn, model tổ hợp này phải có xu hướng giảm variance. Ngược lại các model có bias cao nhưng variance thấp, model tổ hợp phải có xu hướng giảm bias

Second, đến phần quan trọng rồi :3. Xin giới thiệu với các bạn 3 biến thể của phương thức ensemble learning được dùng khá nhiều hiện nay:

* **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 trung bình cộng để cho ra kết quả cuối cùng.
* **Boosting**: Xây dựng một lượng lớn các model (thường là cùng loại). Mỗi model sau sẽ học cách sửa những errors 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 (cụ thể ở đây là 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) <Well, có thể hơi khó hiểu tí nhưng tôi sẽ giải thích rõ hơn cho các bạn sau>. 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).
* **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.