Overfit và Underfit

Khi xây dựng và huấn luyện một mô hình machine learning, ta luôn cần ít nhất 2 bộ dữ liệu, bộ dữ liệu huấn luyện (training data) và bộ dữ liệu kiểm thử (test data). Trong đó, bộ dữ liệu train đại diện cho những dữ liệu mà mô hình được phép thấy và học, bộ dữ liệu test đại diện cho những dữ liệu dùng để đánh giá cuối cùng về mô hình, thậm chí, ta có thể coi bộ dữ liệu test là bộ dữ liệu thực tế trong tương lai nếu đưa mô hình machine learning này vào hoạt động thực tế.

Việc lựa chọn mô hình machine learning phù hợp để học bộ dữ liệu train và cho kết quả dự đoán tốt trên bộ dữ liệu test là điều quan trọng nhất nhưng ko dễ. Với những bộ dữ liệu train có phân bố đơn giản, ta có thể sử dụng một mô hình machine learning đơn giản để xử lý. Với những bộ dữ liệu có mức độ phức tạp tăng dần, ta cần những mô hình machine learning phức tạp hơn, tuy nhiên, câu hỏi đặt ra là như thế nào là một mô hình machine learning đủ phức tạp để xử lý bộ dữ liệu? Nếu một mô hình không đủ phức tạp so với bộ dữ liệu hoặc quá phức tạp so với bộ dữ liệu thì sao?

1. Underfit

1.1. Khi nào xảy ra hiện tượng underfit?

Ví dụ với bộ dữ liệu trong bài toán định giá nhà đơn giản, ta chỉ cần một mô hình machine learning đơn giản như linear regression để xử lý.

Tuy nhiên, với bộ dữ liệu có độ phức tạp cao hơn, linear regression không đủ sức để mô tả được, điều này dẫn đến độ chính xác của mô hình trên bộ test thấp.

Khi đó, ta gọi mô hình đang bị underfit.

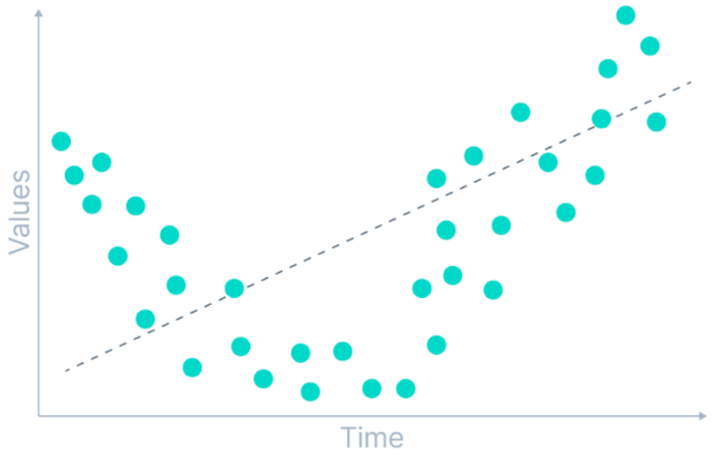
Cụ thể, khi mô hình không đủ sức để học ra được những quy luật, xu hướng trong bộ train và dẫn đến chất lượng của các dự đoán trên bộ test thấp, nói cách khác, giá trị loss của mô hình trên cả bộ train và bộ test đều thấp, ta gọi đây là hiện tượng underfit.

1.2. Giải pháp xử lý hiện tượng underfit

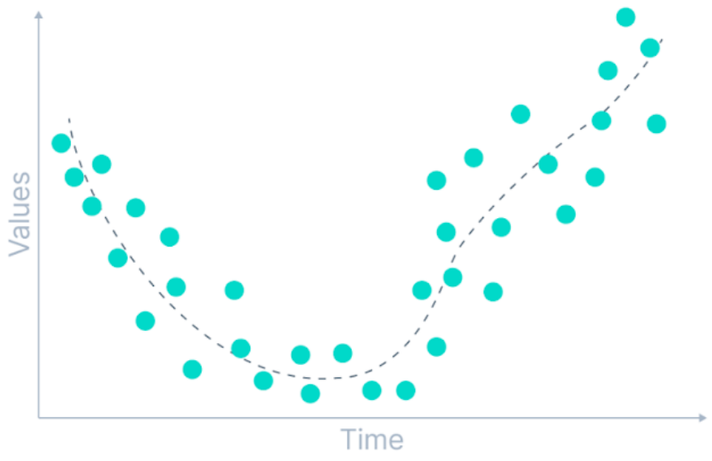
Hiện tượng underfit có thể xảy ra khi bộ dữ liệu train chứa quá nhiều nhiễu (noise) và các điểm dữ liệu ngoại lai (outlier), điều này khiến cho mô hình gặp khó khăn để khái quát hoá được bộ dữ liệu. Một lý do khác là mô hình có độ phức tạp quá thấp (mô hình quá đơn giản) so với bộ dữ liệu.

Từ đó, để giải quyết hiện tượng underfit, cách đơn giản nhất là tăng độ phức tạp của mô hình hay nói cách khác là tăng kích thước của mô hình. Thông thường, ta sẽ thiết kế mô hình machine sao cho nó có thể dễ dàng vượt qua hiện tượng underfit và dễ dàng overfit với bộ dữ liệu, sau đó ta sẽ cố gắng sử dụng các kỹ thuật để giảm hiện tượng overfit.

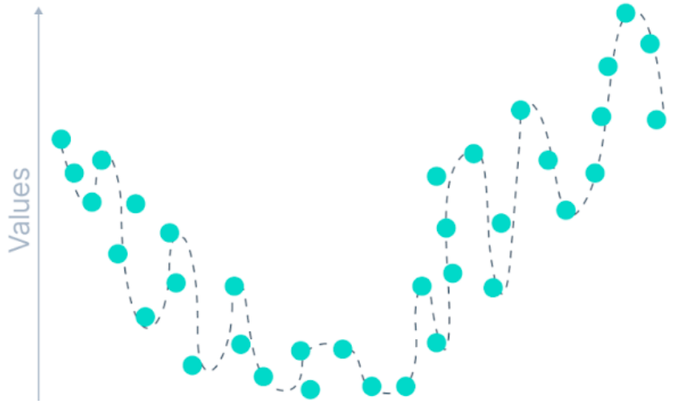
Underfitted



Good Fit/Robust



Overfitted

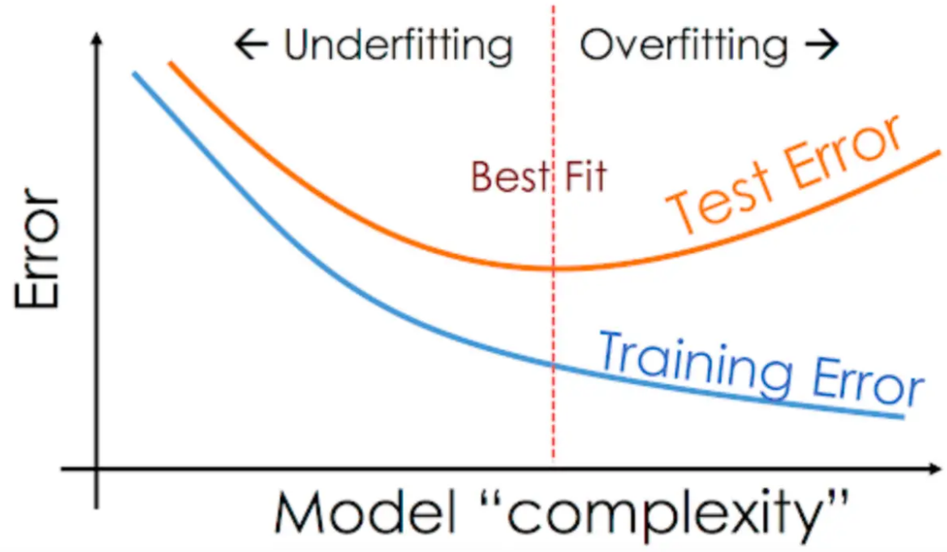


2. Overfit

2.1. Khi nào xảy ra hiện tượng overfit?

Khi ta tăng dần độ phức tạp của mô hình để giải quyết vấn đề underfit, ta sẽ gặp hiện tượng overfit.

Overfit là hiện tượng xảy ra khi ta sử dụng một mô hình machine learning quá phức tạp so với bộ dữ liệu. Lúc này, mô hình, thay vì việc khái quát hoá bộ dữ liệu train, lại học thuộc phần lớn (hoặc thậm chí toàn bộ) bộ dự liệu train. Điều này khiến cho mô hình đạt giá trị loss rất thấp trên bộ train. Nhưng đối với bộ test, độ chính xác của mô hình vẫn thấp do mô hình chỉ ghi nhớ và dự đoán tốt những điểm dữ liệu trong bộ train mà thôi.



Đến đây, câu hỏi đặt ra là làm sao để chỉ với bộ dữ liệu train, ta có thể xây dựng được một mô hình machine learning cho kết quả dự đoán tốt với bộ test?

2.2. Sự xuất hiện của bộ dữ liệu đánh giá (validation data)

Trong khi bộ test đóng vai trò là bài kiểm tra cuối cùng đối với một mô hình machine learning, ta cần một bộ dữ liệu nữa giúp đánh giá một cách khách quan và chính xác tình trạng overfit của mô hình và là căn cứ để giúp ta phần nào đó dự đoán được kết quả của mô hình trên bộ test cuối cùng.

2.3. Giải pháp xử lý hiện tượng overfit

Hiện tượng overfit xảy ra khi bộ dữ liệu train quá nhỏ và đơn giản. Điều này dẫn đến hai vấn đề:

- Bộ dữ liệu train không đủ khái quát để mô tả được bộ dữ liệu test
- Bộ dữ liệu train quá dễ để mô hình có thể học và ghi nhớ, thậm chí tới từng điểm dữ liệu

Do đó, để giải quyết vấn đề overfit, cách đơn giản nhất là tăng thêm số lượng dữ liệu cho bộ train. Tuy nhiên, trong trường hợp không thể tăng thêm được bộ dữ liệu train, ta có thể sử dụng một số kỹ thuật:

- Data augmentation: đây là cách kỹ thuật giúp tạo ra thêm dữ liệu dùng cho bộ train. Đối với từng loại dữ liệu khác nhau, ta sẽ có các kỹ thuật data augmentation khác nhau.
- Regularization: đây là kỹ thuật tác động thêm vào hàm loss của mô hình machine learning giúp mô hình giảm bớt hiện tượng overfit.

- ...