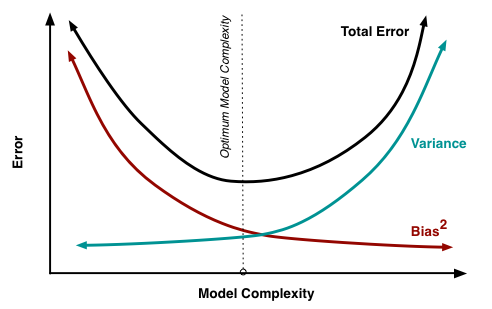
**Bias, Variance và Trade-off**

****

*Hình 1: biểu diễn giữa Bias, Variance và Trade-off*

Mục đích của bất kỳ thuật toán học máy có giám sát là cố gắng ước tính tốt nhất hàm ánh xạ ***(mapping function)* (f)** cho biến đầu ra **(Y)** cho dữ liệu đầu vào **(X)**. Hàm ánh xạ thường được gọi là ***target function***. Việc dự đoán lỗi dự đoán cho mô hình máy học được chia làm ba loại:

* Bias Error (thiên vị)
* Variance Error (phương sai)
* Irreducible Error

**Bias Error:** nếu Bias lớn thì model sẽ bị underfitting, do model chọn quá đơn giản dẫn đến model chưa có hết thông tin data.

Một số thuật toán máy học có chỉ số ***bias*** thấp như: Decision Trees, *k-*Nearest Neighbors và Support Vector Machine.

Các thuật toán máy học có chỉ số ***bias*** cao như: Linear Regression, Linear Discriminant, Analysis và Logistic Regression.

**Variance Error:** nếu Variance lớn thì model sẽ bị overfitting, do model chọn quá phức tạp, ghi nhớ quá nhiều thông tin trong training set.

Một số thuật toán máy học có chỉ số ***variance*** thấp như: Linear Discriminant, Analysis và Logistic Regression.

Một số thuật toán máy học có chỉ số ***variance*** cao như: Decision Trees, *k-*Nearest Neighbors và Support Vector Machine.

**Irreducible Error:** đây là lỗi do một số nhân tố, ví dụ như một số biến unknown mà ảnh hưởng đến đầu vào của hàm ánh xạ.

**Bias-variance trade-off**

Mục đích của các thuật toán máy học có giám sát là cố gắng đạt được chỉ số Bias thấp và Variance thấp. Một số xu hướng chung:

* Thuật toán học máy có tham số (parametric) hoặc tuyến tính (linear) thường có bias cao nhưng variance thấp.
* Thuật toán học máy không tham số (nonparametric) hoặc phi tuyến (non-linear) thường có bias thấp nhưng variance cao.

Việc tham số hóa các thuật toán học máy thường là một “trận chiến” để cân bằng sự thiên vị và phương sai. Dưới đây là hai ví dụ về việc điều chỉnh tham số bias-variance trade-off cho cụ thể thuật toán:

* Thuật toán k-láng giềng gần nhất (*k-*Nearest Neighbors) có bias thấp và variance cao, nhưng sự cân bằng có thể được thay đổi bằng cách tăng giá trị của k.
* Thuật toán support vector machine có bias thấp và variance cao, nhưng sự cân bằng có thể được thay đổi bằng cách tăng tham số C.

Tóm lại ta sẽ có mối quan hệ giữa ***bias*** và ***variance*** như sau:

* Tăng ***bias*** sẽ làm giảm ***variance***.
* Tăng ***variance*** sẽ làm giảm ***bias****.*

**Residual sum of squares (tổng số dư bình phương)**

Trong thống kê, residual sum of squares (RSS) hoặc the sum of squared residuals (SSR) hoặc the sum of squared errors of prediction (SSE), là tổng số dư bình phương (độ lệch giữa giá trị dự đoán và giá trị thực tế của dữ liệu). Đây là thước đo sự khác biệt giữa dữ liệu thực và mô hình ước tính. Nếu giá trị RSS nhỏ thì có nghĩa là độ **fit** (khít) giữa mô hình dự đoán và mô hình ước tính cao. Nó được dùng như một tiêu chí tối ưu trong việc lựa chọn tham số và lựa chọn mô hình.