Tập huấn luyện càng lớn(\*) thì hiệu năng của hệ thống học càng tốt.  
Tập kiểm thử càng lớn thì việc đánh giá càng chính xác.  
Rất khó (ít khi) có thể có được các tập dữ liệu (rất) lớn.

**Cross-validation (kiểm tra chéo) là gì?**

Cross-validation là một phương pháp kiểm tra độ chính xác của 1 máy học dựa trên một tập dư liệu học cho trước. Thay vì chỉ dùng một phần dữ liệu làm tập dữ liệu học thì cross-validation dùng toàn bộ dữ liệu để dạy cho máy. Nó thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán. Kỹ thuật này dễ hiểu, dễ thực hiện và cho ra các ước lượng tin cậy hơn so với các phương pháp khác.

3 phương pháp cross-validation phổ biến:

1. Hold-out: phương pháp đơn giản nhất. Dữ liệu được chia một cách ngẫu nhiên thành một tập dữ liệu học và một tập dữ liệu kiểm tra. Dùng tập đầu tiên để dạy máy rồi dùng ngay tập còn lại để kiểm tra.  kỹ thuật này nói chung không hoạt động tốt trong trường hợp chúng ta không có một bộ dữ liệu lớn. Khi dữ liệu hạn chế, việc chia tập dữ liệu thành các tập Train và Test có thể khiến một số điểm dữ liệu có thông tin hữu ích bị loại trừ khỏi quy trình đào tạo và mô hình không thể tìm hiểu phân bổ dữ liệu đúng cách.
2. K-fold: đây là phương pháp nâng cấp của hold-out. Toàn bộ dữ liệu được chia thành K tập con. Quá trình học của máy có K lần. Trong mỗi lần, một tập con được dùng để kiểm tra và K-1 tập còn lại dùng để dạy. Ví dụ với k=10, phương pháp sẽ mang tên **10-fold**[**cross-validation**](https://trituenhantao.io/tu-dien-thuat-ngu/cross-validation/)**.** Các lựa chọn thông thường của k: 10, hoặc 5. Phù hợp khi ta có một tập dữ liệu nhỏ
3. Leave-one-out: Tương tự như K-Fold nhưng tối đa hóa số tập con (K = số dữ liệu). Phù hợp khi ta có một tập dữ liệu (rất) nhỏ

Stratified k-fold