CHƯƠNG 1:
GIỚI THIỆU VỀ HỌC MÁY THỐNG KÊ

Khoa Khoa học và Kỹ thuật thông tin
Bộ môn Khoa học dữ liệu

NỘI DUNG

1. Học máy là gì ?
2. Kiểm thử & thẩm định.
3. Phân loại học máy.
4. Thách thức.

ĐỊNH NGHĨA HỌC MÁY

CÁC ĐỊNH NGHĨA

- Định nghĩa của Arthur Samuel, 1959:
  "Machine Learning is the field of study that gives computers the ***ability to learn without being explicitly programmed****."*
- Định nghĩa của Tom Mitchell, 1997:

  "A computer program is said to learn from experience E, with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Định nghĩa của Tom Mitchell là định nghĩa chuẩn nhất về Học máy, và được sử dụng rộng rãi.

Các thành phần chính trong định nghĩa MH của Tom Mitchell

- **Task T**: tác vụ – hay bài toán mà máy tính cần giải quyết.
  + Các tác vụ có thể là: nhận diện, phân loại, phát hiện, theo dõi, tạo sinh, ...
- **Kinh nghiệm E**: là đối tượng máy dựa vào để có thể ”học”. E ở đây có thể là dữ liệu, hoặc là tri thức.
  + Dữ liệu: dữ liệu văn bản, dữ liệu ảnh, ….
  + Tri thức.
- **Độ đo P**: đo lường tính hiệu quả của việc “học” của máy.
  + Các độ đo: Accuracy, Precision, Recall, F1-score, Bleu, ...

Ví dụ

Nhu cầu dự đoán nên đặt hay rút cổ phiếu (khi chơi chứng khoán).

- Tác vụ T: dự đoán (predict) giá cổ phiếu.
- Kinh nghiệm E: giá của các cổ phiếu trước đó trong một khoảng thời gian nhất định.
- Độ đo P: đo khả năng sai lệch giữa giá trị dự đoán với giá trị thực tế - dùng hiệu giữa 2 giá trị.

Bài tập

- Hãy cho một ví dụ khác về học máy mà bạn biết?
  - Task T?
  - Kinh nghiệm E?
  - Độ đo P?

Các dạng bài toán – tác vụ phổ biến

1. Phân loại (Classification)

   Dự đoán có phải thư rác hay không.

2. Hồi quy (Regression)

   Dự đoán giá nhà.

3. Xếp hạng (Ranking)

   Xếp hạng các link kết quả tìm kiếm Google search.

4. Phát hiện bất thường (Anomaly/Fraud Detection)

   Tình hình tiêu thụ điện có bất thường gì?

5. Tìm kiểu mẫu (Finding Patterns)

   Hầu như 80% khách hàng mua "khẩu trang y tế" và "nước rửa tay sát khuẩn" chung một đơn hàng trong mùa dịch COVID-19.

Tại sao cần học máy?

Làm thế nào để phân loại thư rác.

- Input: Email.
- Output: Spam/Not spam.

Giải quyết bài toán này theo lập trình bình thường làm như thế nào?

Chương trình lọc thư rác truyền thống

Hãy xem chương trình lọc **thư rác** sử dụng kỹ thuật lập trình truyền thống sau:

- **Bước 1**: Chúng ta tìm những điểm nổi bật để nhận diện thư rác. Chúng ta quan sát được một số từ, cụm từ, câu văn phổ biến trong thư rác như "*free*", "*4U*", "*amazing*", "*credit card*", v.v....
- **Bước 2**: Chúng ta viết một thuật toán dựa trên các kiểu mẫu "pattern" ở Bước 1: NẾU chứa 4U hoặc free THÌ gán là thư rác.
- **Bước 3**: Kiểm thử chương trình và lặp lại **Bước 1** và **Bước 2** đến khi chương trình đủ tốt để ứng dụng được.

Chuyện gì sẽ xảy ra nếu như người spam thay đổi cách viết để tránh phát hiện? ("for you" thay vì "4U"?)

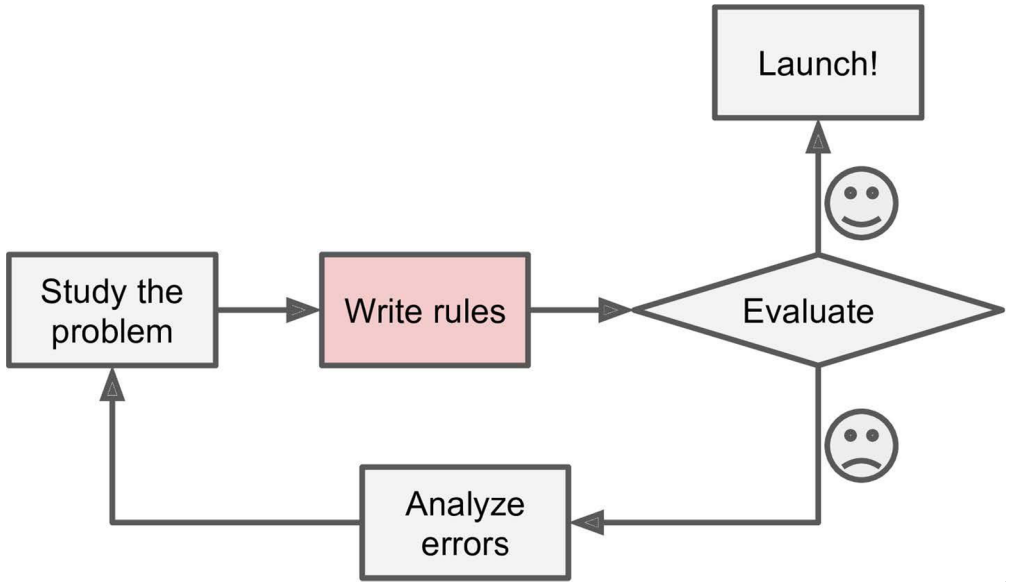
Lập trình truyền thống

Dựa trên Rule-based

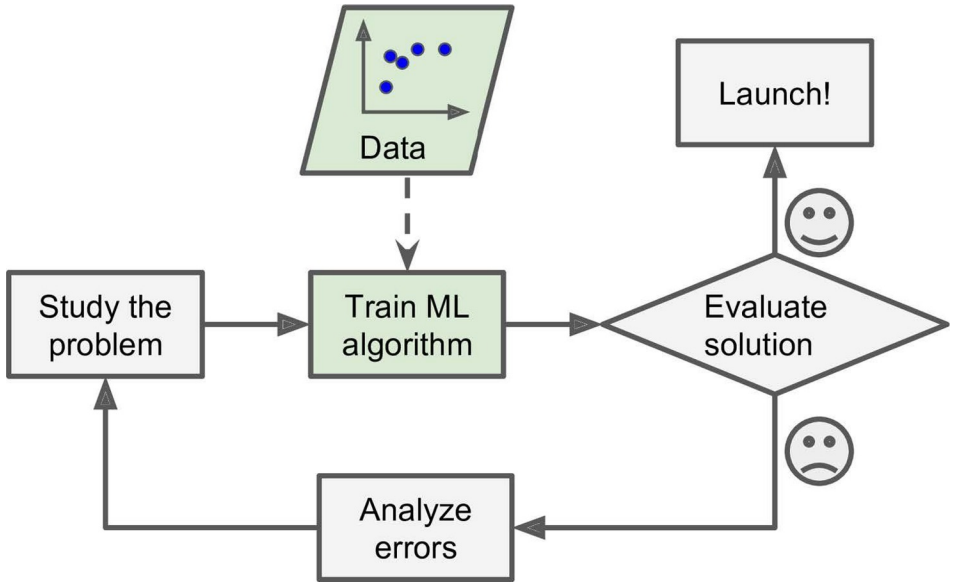
- **Tốn chi phí và thời gian** để thiết kế, thực thi, vận hành, bảo trì
- Cần có **chuyên gia (expert)** can thiệp
- Phù hợp với những **task đơn giản**
- …

Điểm lợi của học máy

Máy học giúp cho máy tính có thể tự động rút ra các “pattern/rule”.



Rule-based



Machine learning-based

Dùng học máy khi nào

- Bài toán phải giải bằng một danh sách các luật: máy học thường đơn giản hóa mã nguồn và cho kết quả tốt hơn.
- Bài toán phức tạp không có cách giải tốt bằng những cách truyền thống (rule-based): máy học có thể giúp tìm lời giải gần đúng.

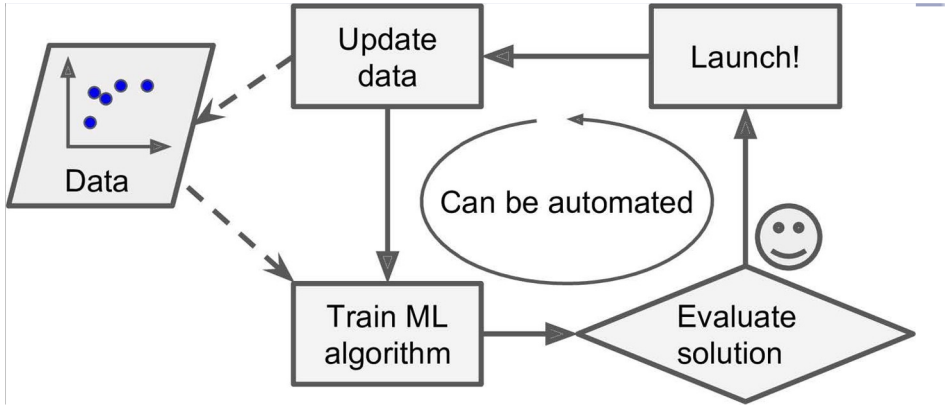
Ví dụ: nhận dạng tiếng nói (nhiều giọng nói khác nhau, môi trường âm thanh nhiễu, nhiều thứ tiếng khác nhau).

- Ngữ cảnh bài toán biến động: một hệ thống máy học có thể thích ứng với dữ liệu mới.
- Giúp con người hiểu về dữ liệu lớn

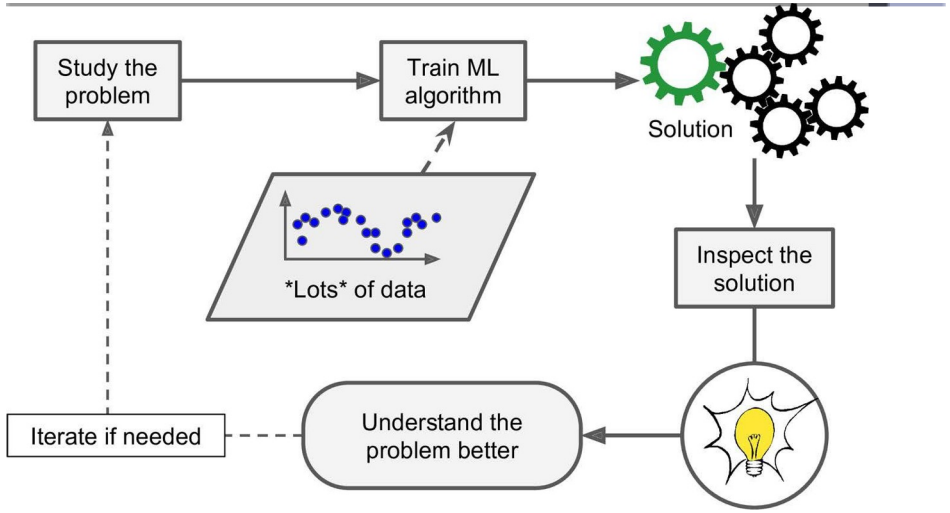
Ví dụ: Từ dữ liệu kinh doanh bán lẻ có thể rút ra là người mua hàng mua mặt hàng X hay mua mặt hàng Y.

Đặc điểm của học máy

Thích ứng với dữ liệu mới



Rút ra quy luật tự động từ dữ liệu



MỘT SỐ VÍ DỤ

– Các bài toán sau đây cần dùng học máy hay không ? Vì sao?

1. Rút trích thông tin số điện thoại di động của Việt Nam từ văn bản?
2. Rút trích ra các thực thể tên trong văn bản?

MỘT SỐ VÍ DỤ

– Các bài toán sau đây cần dùng học máy hay không ? Vì sao?

1. Rút trích thông tin số điện thoại di động của Việt Nam từ văn bản?

Dùng luật!!!

Quy luật: Số điện thoại VN có 10 chữ số, bắt đầu bằng 0, hoặc +84

9 chữ số còn lại từ 0 đến 9.

Đầu số các mạng di động: 090, 091, 092, 038, ...

.....

→ Dùng luật để trích xuất được. Công cụ sử dụng: biểu thức chính quy (Regular expression – RegExp).

MỘT SỐ VÍ DỤ

– Các bài toán sau đây cần dùng học máy hay không ? Vì sao?

2. Rút trích ra các thực thể tên trong văn bản?

Không thể dùng luật do:

– Tên thực thể rất đa dạng: tên người, tên tổ chức, tên địa danh, ...

– Tên thực thể phụ thuộc vào ngữ cảnh.

VD:

Long An – có thể là tên tỉnh, hoặc tên người !!!

Đồng Nai– có thể là tên tỉnh, hoặc là tên đường !! (Đường Đồng Nai, Cư xá Bắc Hải, Q10).

KIỂM THỬ VÀ THẨM ĐỊNH

Kiểm thử là gì

- Thẩm định độ tin cậy (hay độ chính xác) của mô hình khi dự đoán một dữ liệu ngoài thực tế.
- Phương pháp đánh giá: đánh giá dựa trên các tập dữ liệu đã được gán nhãn sẵn.
- Các tập dữ liệu cần để xây dựng hệ thống máy học:
  + Huấn luyện (training dataset).
  + Thẩm định (validation/development dataset).
  + Kiểm thử (test dataset).

Lưu ý

- Không bao giờ đánh giá một hế thống máy học trện tập dữ liệu dành để phát triển hệ thống (dữ liệu dùng cho huấn luyện, tinh chỉnh tham số)!
- Đánh giá chính thức hệ thống máy học một lần duy nhất trên tập dữ liệu kiểm thử (tập test)!
- Nhãn của các dữ liệu test thường sẽ được che giấu (đối với các cuộc thi về học máy).

Ví dụ thực tế

- Mô hình học máy: học sinh.
- Tác vụ: Thi đại học
- Đánh giá: điểm (0-10).
- Kinh nghiệm (dữ liệu huấn luyện):
  + Tập huấn luyện: Tập các đề ôn thi. Học sinh giải các đề ôn thi để lấy kiến thức → quá trình học.
  + Tập thẩm định: Tập các đề thi thử để lấy điểm. Học sinh dựa vào điểm thi thử để biết khả năng hiện tại mà điều chỉnh → tinh chỉnh mô hình.
  + Tập kiểm tra: Kỳ thi đại học thực sự. Điểm số quyết định cuối cùng là **đậu hay tạch** → kết quả mô hình.

PHÂN LOẠI HỌC MÁY

Các dạng học máy

Dựa trên có sự giám sát của con người hay không

- Học có giám sát (supervised learning)
- Học không giám sát (unsupervised learning)
- Học bán giám sát (semi-supervised learning)
- Học tăng cường (**Reinforcement Learning**)

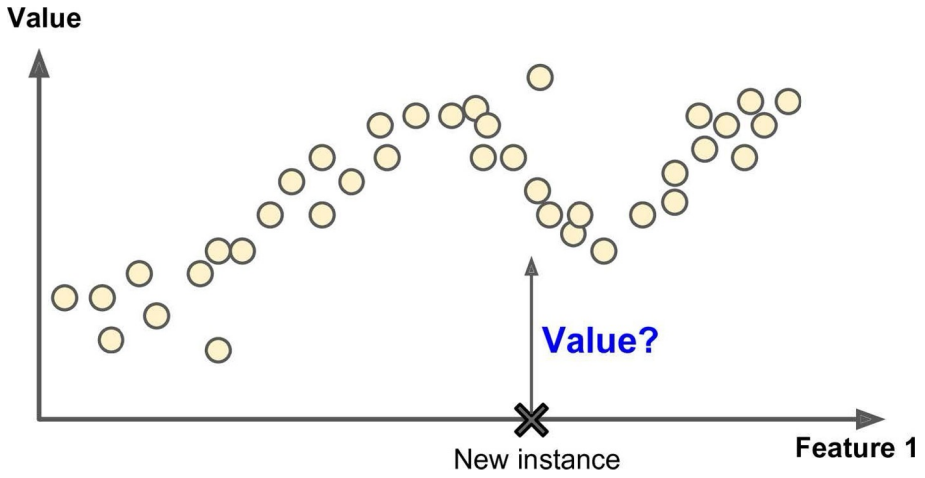
1. Học máy có giám sát

- Dữ liệu huấn luyện bao gồm nhãn (label).
- Có sự giám sát của con người thông qua việc con người gán nhãn cho dữ liệu, và học máy dựa trên dữ liệu.
- Các loại bài toán phổ biến trên học có giám sát:
  + Phân loại (classification): dữ liệu được xác định thuộc vào lớp (hay nhãn) cụ thể trong tập các nhãn của bộ dữ liệu.
  + Hồi quy (regression): dữ liệu được dự đoán ra một giá trị cụ thể.
- Phân biệt hồi quy và phân lớp → dựa vào nhãn của dữ liệu.

1. Học máy có giám sát

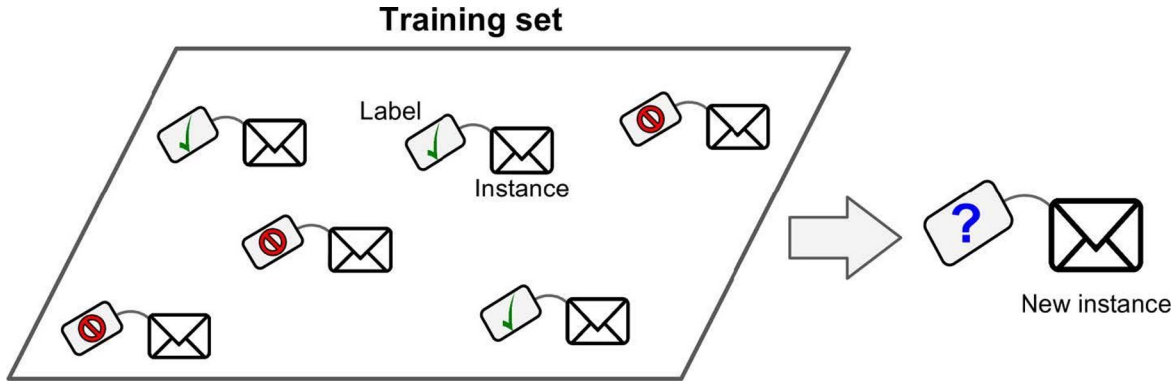
Bài toán hồi quy

- Dự đoán giá nhà đất.



Bài toán phân lớp

- Phân loại email rác. Mỗi email được gán một trong 2 nhãn: thư rác và không phải thư rác.



Các mô hình học máy có giám sát phổ biến

Phân lớp

1. k-Nearest Neighbors.
2. Logistic Regression.
3. Support Vector Machines (SVMs).
4. Decision Trees & Random Forests.
5. Neural networks.

Hồi quy

1. Linear Regression.
2. Support Vector Machines (SVMs).
3. Multivariate Regression.
4. Lasso Regression.

2. Học máy không giám sát

- Dữ liệu huấn luyện không được gán nhãn bởi con người.
- Mô hình tìm kiếm các cấu trúc ẩn/thú vị trong dữ liệu.
- Các dạng bài toán học máy không giám sát:
  + Gom cụm (clustering): Gom dữ liệu có sự giống nhau về một số khía cạnh thành các cụm mong muốn.
  + Trực quan hóa và Giảm số chiều của dữ liệu (Visualization & Dimensionality Reduction).
  + Học luật kết hợp (Association rule learning) → dùng nhiều trong khai phá dữ liệu (Data mining).

Các thuật toán học không giám sát

Gom cụm

- k-Means.
- Hierarchical Cluster Analysis (HCA).
- Expectation Maximization.

Giảm chiều dữ liệu

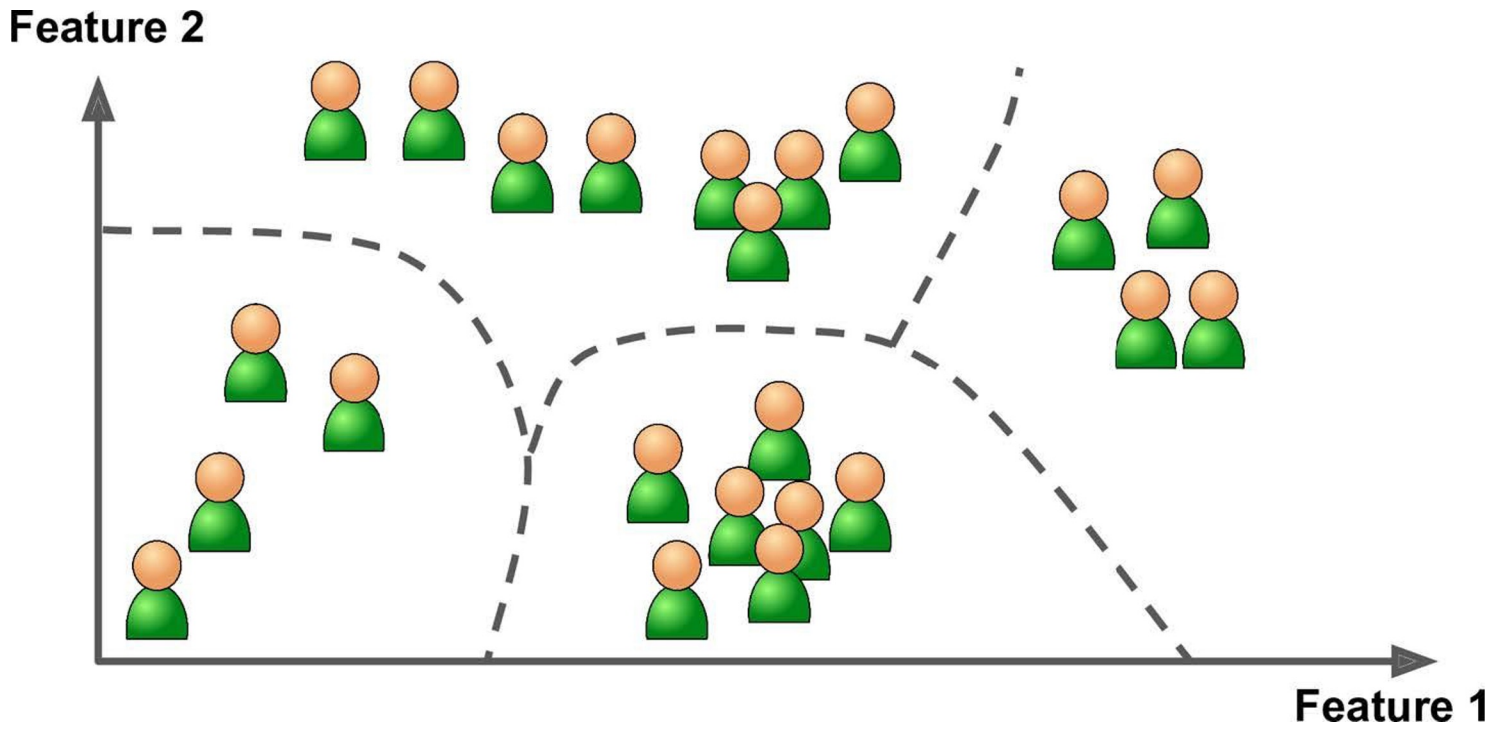
- Principal Component Analysis (PCA).
- Kernel PCA.
- Locally-Linear Embedding (LLE).
- t-distributed Stochastic Neighbor Embedding (t-SNE).

Các thuật toán học máy không giám sát

- Các giải thuật khai phá luật kết hợp sử dụng trong lĩnh vực data mining
  + Apriori.
  + Eclat.

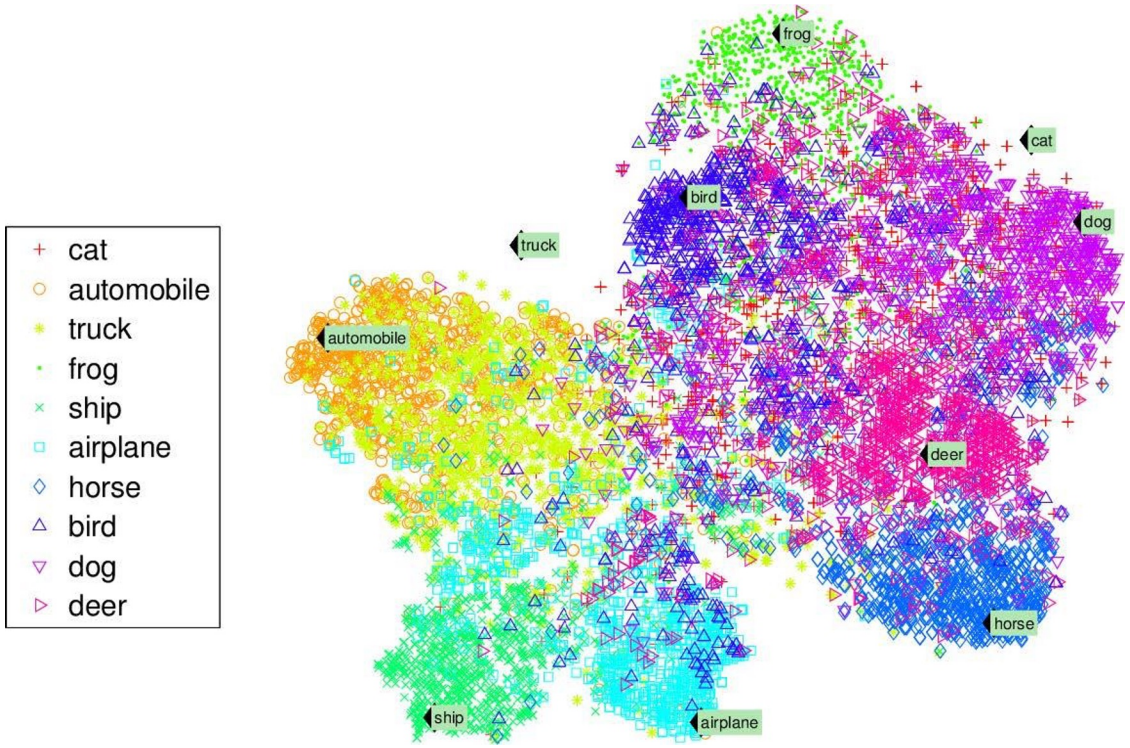
Ví dụ

– Bài toán gom nhóm khách hàng:



Ví dụ

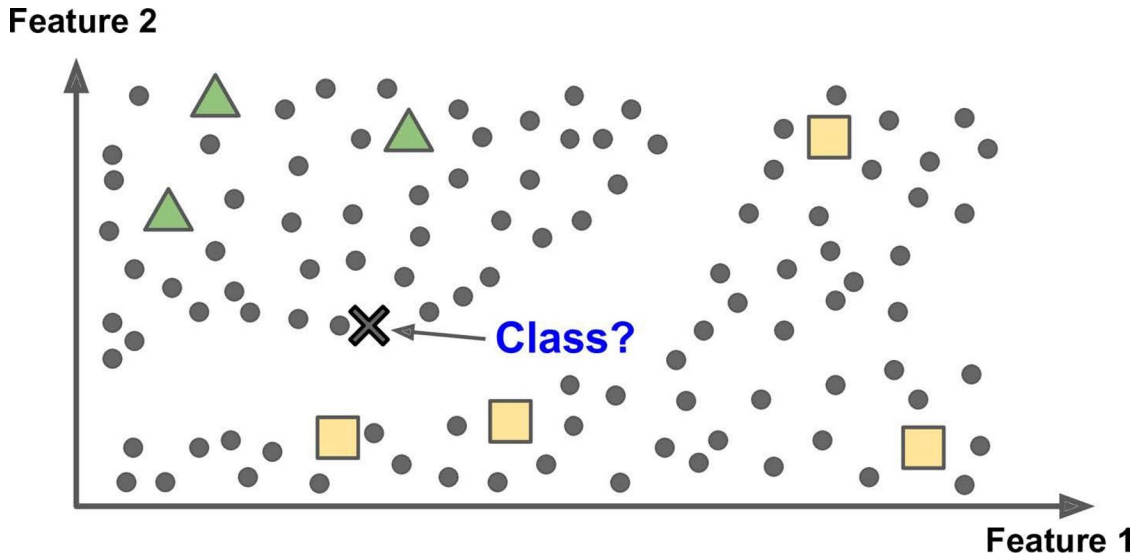
– Bài toán trực quan hoá dữ liệu: Chuyển dữ liệu có cấu trúc phức tạp thành dạng biểu diễn 2D hoặc 3D để có thể vẽ biểu đồ để quan sát trực quan.



3. Học bán giám sát

– Dữ liệu huấn luyện được gán nhãn một phần (thường là nhiều dữ liệu không nhãn và một ít dữ liệu có nhãn).

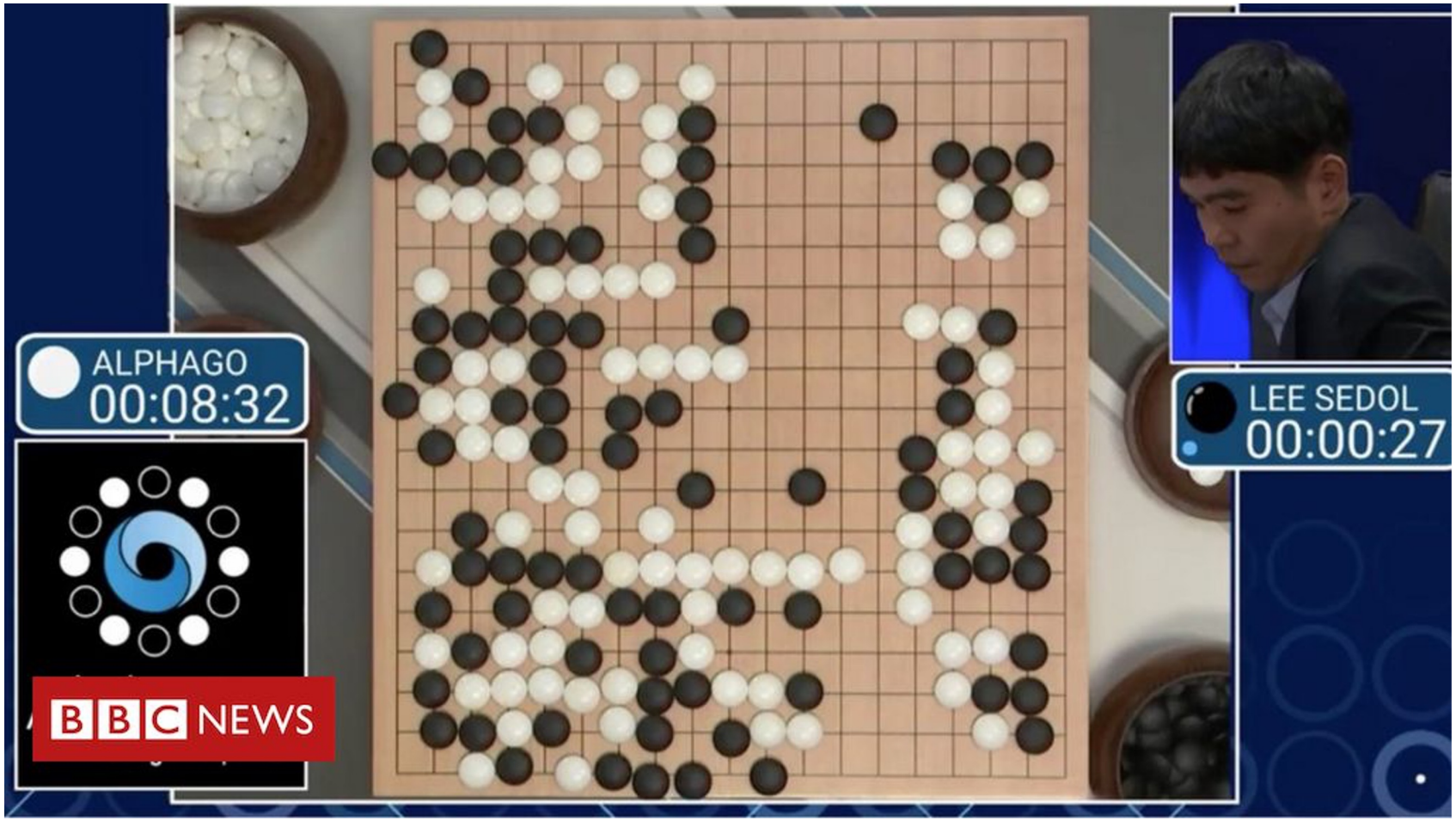
VD: Bài toán phân 2 lớp: hình vuông và hình tam giác. Những điểm hình tròn là dữ liệu không gán nhãn.



4. Học củng cố

- Hệ thống máy học gọi là “tác tử” (agent) học bằng cách quan sát môi trường, chọn và thực hiện “các hành động” và nhận được “phần thưởng” hoặc “hình phạt”.
- Hệ thống này phải học **chiến lược** tốt nhất để nhận được nhiều phần thưởng nhất sau một thời gian.
- Chiến lược xác định hành động gì sẽ được chọn trong một hoàn cảnh nhất định.

Ví dụ



THÁCH THỨC CỦA HỌC MÁY

THÁCH THỨC LỚN CỦA HỌC MÁY

Dữ liệu không tốt

- Thiếu dữ liệu huấn luyện
- Dữ liệu huấn luyện thiếu tính đại diện.
- Dữ liệu có chất lượng kém.
- Đặc trưng không phù hợp.

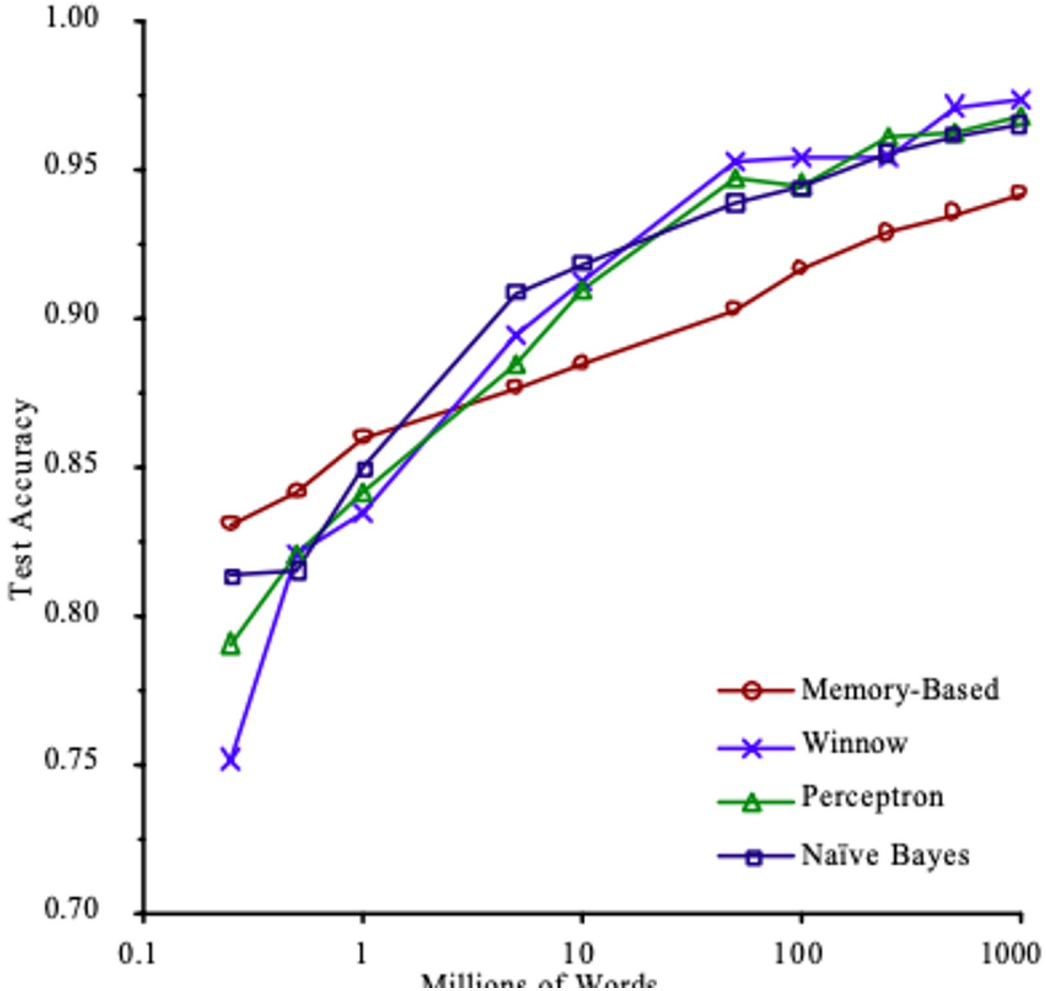
Thuật toán không tốt

- Quá khớp với dữ liệu huấn luyện (over-fitting)
- Chưa khớp với dữ liệu huấn luyện (under-fitting).

DỮ LIỆU KHÔNG TỐT

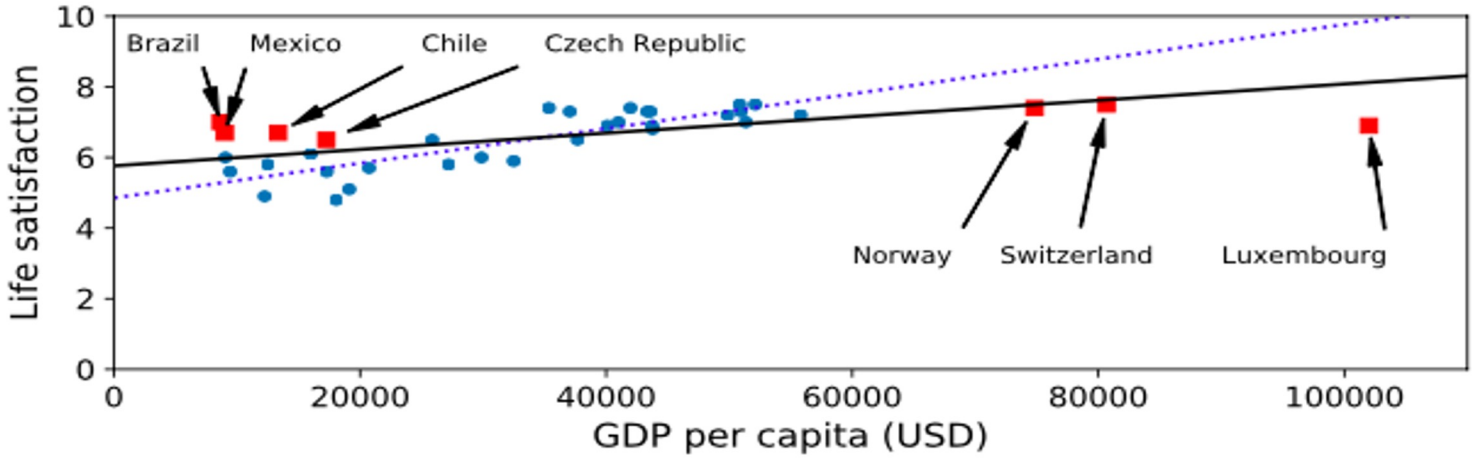
Thiếu dữ liệu huấn luyện

- Với một đứa trẻ chập chững học nhận biết quả táo là gì, chúng ta chỉ vào một quả táo và nói rằng "táo" vài lần là trẻ có thể nhận biết quả táo với nhiều màu sắc và hình dạng khác nhau.
- Máy học đến thời điểm hiện tại cần nhiều dữ liệu huấn luyện hơn. Với một bài toán rất đơn giản, thông thường bạn cần **hàng nghìn mẫu** dữ liệu, và đối với các bài toán phức tạp như nhận diện hình ảnh hoặc giọng nói bạn có thể cần **hàng triệu mẫu** dữ liệu.



Dữ liệu huấn luyện thiếu tính đại diện

- Để hệ thống máy học có thể tổng quát hóa tốt, tập dữ liệu huấn luyện phải có **tính đại diện** cho tất cả dữ liệu muốn dự đoán.
- Nếu kích thước tập dữ liệu quá nhỏ, dữ liệu sẽ có những mẫu không đại diện cho phần nhiều dữ liệu, gọi là **nhiễu do lấy mẫu** “sampling noise”.
- Thậm chí một tập dữ liệu kích thước rất lớn cũng có thể có tính đại diện không tốt nếu phương pháp lấy mẫu có sai sót, gọi là **lệch do lấy mẫu** “sampling bias”.



Dữ liệu có chất lượng kém

- Dữ liệu huấn luyện có chất lượng kém là dữ liệu có nhiều lỗi, nhiều giá trị ngoại biên (outlier), nhiễu (noise).
- Cần **làm sạch dữ liệu huấn luyện**. Sự thật là, hầu hết các nhà khoa học dữ liệu dành phần lớn thời gian chỉ để làm điều này.
  + Quy trình làm sạch dữ liệu chiếm 80% trong một dự án học máy.
- Một số cách làm sạch dữ liệu:
  + Bỏ dữ liệu mang giá trị ngoại biên hoặc sửa lại bằng tay.
  + Nếu một số mẫu thiếu một vài đặc trưng (VD: thiếu thông tin độ tuổi) thì có 3 cách: bỏ đặc trưng này, bỏ những mẫu này hoặc điền giá trị cho đặc trưng này.

Đặc trưng không phù hợp

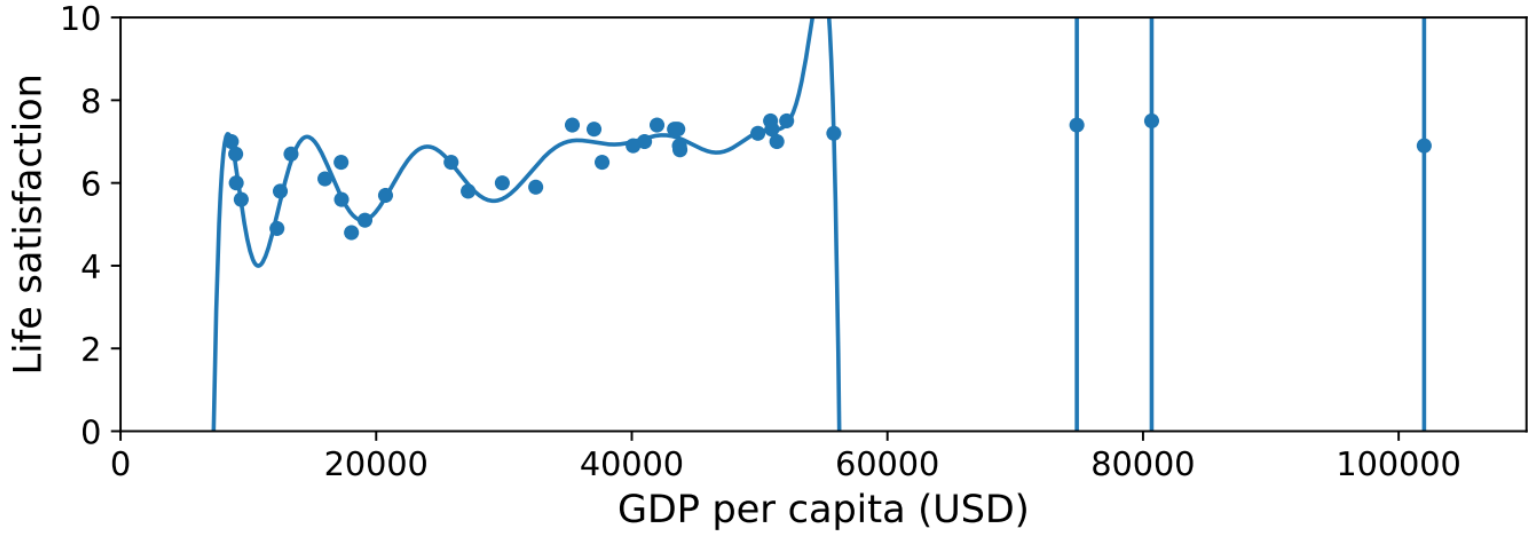
- Hệ thống của bạn sẽ tệ nếu dữ liệu huấn luyện chứa quá nhiều đặc trưng không hữu ích.
- Nhiều phương pháp đòi hỏi phải xác định được tập đặc trưng hữu ích. Công đoạn này gọi là **chế tác đặc trưng** "feature engineering", gồm các bước sau:
  + Lựa chọn đặc trưng (Feature selection): lựa chọn các đặc trưng hữu ích nhất từ một tập hợp các đặc trưng.
  + Rút trích đặc trưng (Feature extraction): kết hợp các đặc trưng đã có để tạo thêm các đặc trưng mới hữu ích hơn.
  + Tạo đặc trưng mới bằng cách thu thập thêm dữ liệu.

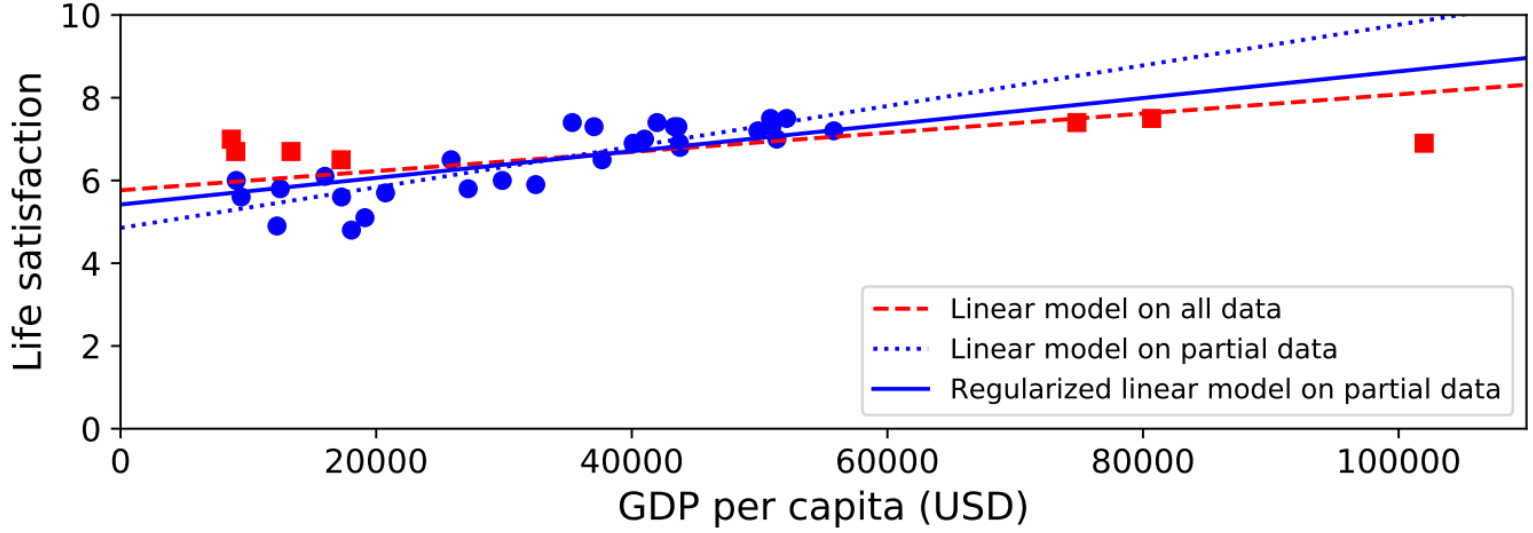
THUẬT TOÁN KHÔNG TỐT

Quá khớp dữ liệu huấn luyện

- Hiện tượng quá khớp dữ liệu huấn luyện **over-fitting** xảy ra khi mô hình máy học quá chi tiết đến nỗi nó học ra những kiểu mẫu từ mẫu nhiễu.
- Một số giải pháp:
  + Đơn giản hóa mô hình bằng cách chọn một mô hình có **ít tham số** hơn (ví dụ: một mô hình tuyến tính hơn là một mô hình đa thức bậc cao)
  + Đơn giản hóa mô hình bằng cách giảm số lượng đặc trưng của dữ liệu huấn luyện, hoặc bằng cách thêm ràng buộc vào mô hình.
  + Thu thập thêm nhiều dữ liệu huấn luyện.
  + Giảm nhiễu trong dữ liệu huấn luyện (ví dụ: sửa lỗi dữ liệu bằng tay hoặc xóa giá trị ngoại biên).

Quá khớp dữ liệu huấn luyện (tt)





Chưa khớp dữ liệu huấn luyện

- Hiện tượng chưa khớp **under-fitting** có tính chất ngược lại với quá khớp **over-fitting,** Xảy ra khi mô hình quá đơn giản so với cấu trúc cơ bản của dữ liệu.
- Một số giải pháp:
  + Chọn mô hình máy học mạnh mẽ hơn với nhiều tham số hơn.
  + Thêm đặc trưng tốt hơn (chế tác đặc trưng).
  + Giảm thiểu ràng buộc cho mô hình (giảm tham số của phương pháp chính quy hóa).

Tổng kết

- Học máy giúp máy tính giải quyết tốt hơn một số bài toán bằng cách học từ dữ liệu mà không phải viết luật.
- Có nhiều loại hệ thống máy học: giám sát, không giám sát, bán giám sát, học tăng cường.
- Hệ thống máy học sẽ không hoạt động tốt nếu dữ liệu huấn luyện quá nhỏ, không có tính đại diện, chứa nhiều nhiễu, hoặc nhiều đặc trưng không liên quan.
- Mô hình học không được quá đơn giản mà cũng không được quá phức tạp.
- Phải **đánh giá** hệ thống máy học để biết chất lượng của nó.

Tài liệu tham khảo

- Chương 1 của sách: *Hands-on Machine Learning with ScikitLearn, Keras & TensorFlow.*