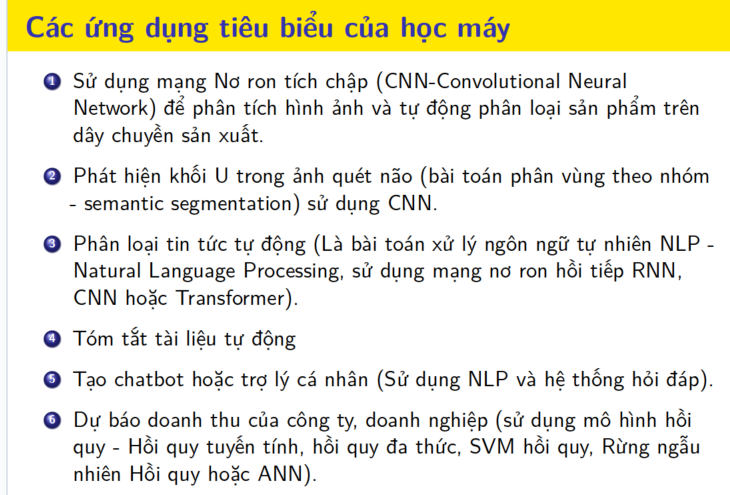
# Chapter1

1. Nêu khái niệm học máy

* Một chương trình máy tính được gọi là học tập từ kinh nghiệm E để hoàn thành nhiệm vụ T với hiệu quả được đo bằng phép đánh giá P, nếu hiệu quả của nó khi thực hiện nhiệm vụ T, khi được đánh giá bởi P có cải thiện theo kinh nghiệm E.

1. **Hãy nêu một số ví dụ ứng dụng tiêu biểu trong thực tế sử dụng học máy?**



A white text with black text

Description automatically generated

1. **Theo em, phương pháp tiếp cận theo lập trình truyền thống và tiếp cận theo phương pháp học máy có gì khác nhau, hãy trình bày chi tiết?**

* Tiếp cận theo truyền thống cần hiểu quy tắc và logic nghiệp vụ để viết code thực hiện.
* Tiếp cận theo pp học máy: máy tính học các quy tắc thực hiện từ dữ liệu, đưa ra dự đoán hoặc quyết định từ các tập dữ liệu thay vì phải lập trình rõ ràng

1. **Dựa theo phương thức học, học máy chia làm mấy dạng, hãy kể tên và nêu chi tiết các bài toán trong các dạng đó?**

* Học có giám sát: phân loại thư rác, nhận dạng văn bản, hình ảnh
* Học không giám sát: phân tích giỏ hàng kinh doanh, phân cụm dữ liệu
* Học bán giám sát: dịch vụ lưu trữ Google Photos
* Học máy tăng cường: robot tự hành, huấn luyện máy tính chơi game

1. **Học tăng cường là gì? Hãy trình bày về học tăng cường?**

* Học tăng cường (RL) là phương pháp có cấu trúc rất khác. Hệ thống học trong RL được gọi là tác nhân (agent). Nó có thể quan sát môi trường xung quanh và chọn việc hành động, sau đó nhận về điểm thưởng (reward) hoặc lượng phạt (penalty) dưới dạng điểm thưởng âm, hệ thống sau đó cần tự học chiến lược tốt nhất để nhận về nhiều điểm thưởng nhất qua thời gian, và chiến lược này được gọi là chính sách (policy).

1. **Các bài toán cơ bản trong học máy là gì? Liệt kê chi tiết?**

* Bài toán phân lớp
* Bài toán hồi quy
* Bài toán máy dịch
* Bài toán phân cụm
* Bài toán hoàn thiện dữ liệu

1. **Tại sao cần sử dụng hàm mất mát?**

* Giúp đánh giá mức độ chính xác của mô hình dự đoán. Nó đo lường sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế mà chúng ta muốn đạt được. Mục tiêu của quá trình huấn luyện mô hình là tối ưu hóa hàm mất mát này, tức là làm giảm thiểu giá trị của nó, để mô hình dự đoán càng chính xác càng tốt.

1. **Overfitting và underfitting là gì?**

* Overfitting: hiện tượng mô hình tìm được quá khớp dữ liệu huấn luyện, tuy nhiên với dữ liệu kiểm tra lại dự đoán nhiều sai sót. Thường xảy ra khi dữ liệu huấn luyện quá nhỏ trong khi mô hình lại quá phức tạp. Khắc phục với phương pháp K-cross validation.
* Underfitting: hiện tượng khi mô hình xây dựng chưa có độ chính xác cao trong tập dữ liệu huấn luyện cũng như tổng quát hóa với tổng thể dữ liệu, do đó mô hình đó sẽ không tốt với bất kì bộ dữ liệu nào. Nó ít xảy ra, tuy nhiên khi xảy ra có thể khắc phục bằng cách thay đổi thuật toán hoặc là bổ sung thêm dữ liệu đầu vào.

1. **Hãy trình bày tổng quan về phương pháp K-cross validation?**

* Là một kỹ thuật phổ biến trong học máy được sử dụng để đánh giá hiệu suất của mô hình một cách chính xác hơn so với phương pháp chia dữ liệu thành tập huấn luyện và tập kiểm tra chỉ một lần. Kỹ thuật này giúp đảm bảo rằng mô hình không quá khớp (overfitting) và có thể tổng quát hóa tốt cho dữ liệu chưa thấy.
* Cách hoạt động:
  + Chia dữ liệu thành K phần (folds): Chia dữ liệu thành K phần có kích thước xấp xỉ bằng nhau.
  + Huấn luyện và kiểm tra trên từng fold: một fold được dùng làm tập kiểm tra, và K-1 folds còn lại được dùng làm tập huấn luyện. Thực hiện K lần, sao cho mỗi lần lặp thì mỗi fold sẽ được dùng làm tập kiểm tra.
  + Tính score cho từng fold: Sau khi huấn luyện mô hình trên K-1 folds, dùng mô hình đó để dự đoán tập kiểm tra (fold còn lại) và tính điểm số (ví dụ: accuracy, precision, recall, F1-score, v.v.).
  + Tính trung bình score: Sau khi lặp K lần, tính trung bình các score để có được một ước lượng chính xác hơn về hiệu suất mô hình.

1. **Theo em mô hình như thế nào gọi là tốt (good fitting)?**

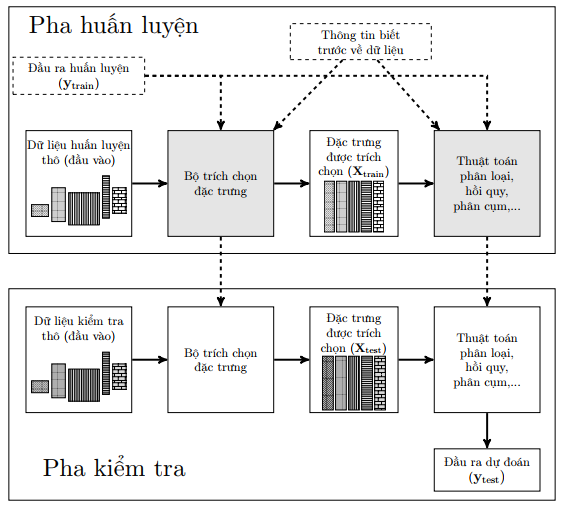
* Mô hình cho ra kết quả hợp lý với cả tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Đây là mô hình lý tưởng mang được tính tổng quát và khớp được với nhiều dữ liệu mẫu và cả các dữ liệu mới.
* Thực tế rất khó đạt được, để tìm được Good Fitting, phải theo dõi hiệu suất thuật toán học máy theo thời gian khi thuật toán thực hiện việc học trên bộ dữ liệu huấn luyện.

# Chapter2

1. Trích chọn đặc trưng là gì? Khi nào cần sử dụng trích chọn đặc trưng

* Là quá trình đưa dữ liệu thô về dữ liệu dạng chuẩn, loại bỏ đi những dữ liệu nhiễu (noise). Dữ liệu chuẩn này phải đảm bảo giữ được những thông tin đặc trưng của dữ liệu thô ban đầu. Ta còn cần phải thiết kế những phép biến đổi để có những đặc trưng phù hợp cho từng bài toán.
* Khi nào cần:
  + Đơn giản hóa các mô hình để giúp các nhà nghiên cứu/người dùng diễn dịch dễ dàng hơn.
  + Giảm thời gian huấn luyện.
  + Tránh lời nguyền chiều (curse of dimensionality).
  + Tăng cường tổng quát hóa bằng cách giảm sự quá khớp.

1. Trình bày các pha chính khi xây dựng mô hình học máy



* Pha huấn luyện: xây dựng mô hình dựa trên dữ liệu huấn luyện
* Pha kiểm tra: đánh giá hiệu quả của mô hình
  + Khác biệt: trong pha kiểm tra không có bước huấn luyện mô hình

1. Trình bày một số kỹ thuật trích chọn đặc trưng phổ biến hiện nay?

* Trực tiếp lấy dữ liệu thô
* Lựa chọn đặc trưng
* Giảm chiều dữ liệu
* Túi từ

1. Trình bày ứng dụng của túi đựng từ (BoW) trong thị giác máy tính (computer vision)

* Sử dụng vector histogram của ảnh trong computer vision
* Xét cửa sổ ảnh

1. Trình bày các phương pháp chuẩn hoá véc tơ đặc trưng

* Rescaling (Thay đổi lại kích thước)
* Standardization (Chuẩn hoá giá trị)
* Scaling to unit length (Chia tỷ lệ theo chiều dài đơn vị)