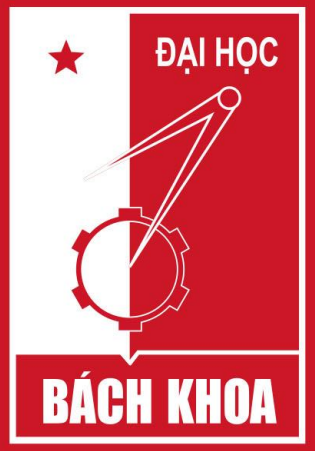


ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nhập môn
Học máy và Khai phá dữ liệu
(IT3190)

Cấu trúc môn học

- Số tuần: 15
  - Lý thuyết: 11-13 tuần
  - Sinh viên trình bày đồ án môn học: 02-03 tuần
- Thời gian và địa điểm

- Thời gian gặp sinh viên
  - **Hẹn trước qua e-mail**
  - Viện CNTT&TT, Nhà B1

Nội dung môn học

- Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- Lecture 9: Học dựa trên xác suất
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp
- Lecture 13: Thảo luận ứng dụng học máy và khai phá dữ liệu trong thực tế

Mục tiêu của môn học

- Có kiến thức cơ bản về học máy
- Có hiểu biết về các phương pháp học máy, các điểm mạnh (ưu điểm) và các điểm yếu (nhược điểm) của các giải thuật học máy và khai phá dữ liệu
- Làm quen và sử dụng được thư viện Scikit-learn
- Có kinh nghiệm về thiết kế, cài đặt, và đánh giá hiệu năng của một phương pháp học máy hoặc khai phá dữ liệu
  - Thông qua đồ án môn học

Đánh giá

- Đồ án môn học (**P**): Tối đa 10 điểm
  - Mỗi đồ án được thực hiện bởi một nhóm sinh viên
  - Chọn một phương pháp học máy được giới thiệu trong môn học để giải quyết một bài toán thực tế
  - Cài đặt và đánh giá hiệu năng của phương pháp đó dựa trên dữ liệu thực tế
- Thi viết (**E**): Tối đa 10 điểm
- Điểm học phần (**G**)
  - **G** = 0,4 x **P** + 0,6 x **E**

Đồ án môn học: đề tài

- Tự do đề xuất bài toán thực tế, (các) giải thuật học máy để giải quyết bài toán, và (các) tập dữ liệu được sử dụng
- Đề xuất đề tài phải được **diễn giải cụ thể**
  - **Mô tả bài toán thực tế** sẽ được giải quyết (mục đích, yêu cầu, kịch bản ứng dụng, …)
  - Xác định rõ **giải thuật học máy** dùng để giải quyết bài toán.
  - Trình bày các thông tin về **đầu vào (input)** và **đầu ra (output)** của hệ thống học máy sẽ được cài đặt, và **cách thức biểu diễn dữ liệu.**
  - Xác định rõ **(các) tập dữ liệu (datasets)** sẽ được sử dụng.

Đồ án môn học: các yêu cầu

- Kết quả của đồ án phải được trình bày ở cuối môn học

  Tất cả các thành viên phải tham gia vào việc thực hiện và trình bày đồ án

- Báo cáo kết quả của đồ án bao gồm:
  - **Mã nguồn** (source codes): lưu trong một file nén
  - **File hướng dẫn** (readme.txt) mô tả chi tiết cách thức cài đặt/biên dịch/chạy chương trình (và các gói phần mềm được sử dụng kèm theo)
  - **Tài liệu báo cáo** kết quả đồ án mô học (lưu trong file .pdf):
    - Giới thiệu và mô tả về bài toán thực tế được giải quyết
    - Các chi tiết của (các) phương pháp học máy và (các) tập dữ liệu được sử dụng
    - Các kết quả thí nghiệm đánh giá hiệu năng của hệ thống học máy đối với (các) tập dữ liệu được sử dụng
    - Các chức năng chính của hệ thống (và cách sử dụng)
    - Cấu trúc của mã nguồn chương trình, vai trò của các lớp (classes) và các phương thức (methods) chính/quan trọng
    - Các vấn đề/khó khăn gặp phải trong quá trình thực hiện công việc của đồ án, và cách thức được dùng để giải quyết (vượt qua)
    - Các khám phá mới hoặc kết luận

Đồ án môn học: đánh giá

- Công việc đồ án được đánh giá theo các tiêu chí sau:
  - *Mức độ phức tạp / khó khăn của bài toán thực tế được giải quyết*
  - *Chất lượng (sự đúng đắn và phù hợp) của phương pháp được dùng để giải quyết bài toán*
  - *Đánh giá và lựa chọn kỹ lưỡng mô hình*
  - Chất lượng của bài trình bày (presentation) kết quả đồ án
  - Chất lượng của tài liệu báo cáo kết quả đồ án
  - Cài đặt hệ thống thử nghiệm (các chức năng, dễ sử dụng, …)
- Bài trình bày trong khoảng 15 phút, và phù hợp với những gì được nêu trong tài liệu báo cáo
- **Nếu sử dụng lại / kế thừa / khai thác các mã nguồn / các gói phần mềm / các công cụ sẵn có, thì phải nêu rõ ràng và chính xác trong tài liệu báo cáo (và đề cập trong bài trình bày)**

Tài liệu học tập

- Các bài giảng trên lớp (Lecture slides)
- Sách tham khảo:
  - T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.
  - Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.
  - E. Alpaydin. *Introduction to Machine Learning*. MIT press, 2020.
  - Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques (3rd Edition). Morgan Kaufmann, 2011.
- Công cụ phần mềm:
  - Scikit-learn (http://scikit-learn.org/)
  - WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
- Các tập dữ liệu (datasets):
  - UCI repository: *http://archive.ics.uci.edu/ml/*

Thư viện hoặc ngôn ngữ

python™

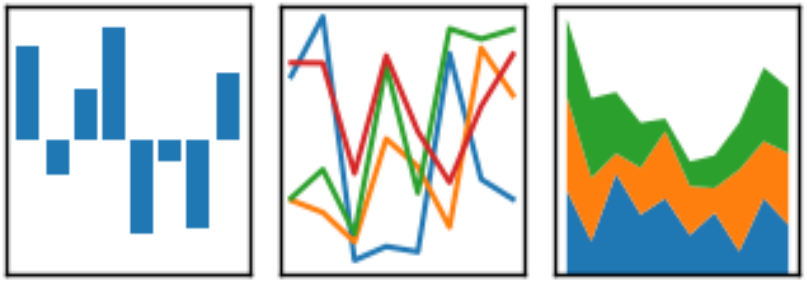
ANACONDA®
Powered by Continuum Analytics

TensorFlow

PyTorch

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict