Nhập môn Học máy và Khai phá dữ liệu

(IT3190)

Nguyễn Nhật Quang

quang.nguyennhat@hust.edu.vn

Trường Đại học Bách Khoa Hà Nội

Viện Công nghệ thông tin và truyền thông

Năm học 2021-2022

Cấu trúc của học phần

- Số tuần: 15
  - Lý thuyết: 11 tuần đầu
  - Sinh viên trình bày đồ án môn học: 4 tuần cuối
- Thời gian và địa điểm
  - Thứ 5 hàng tuần, 06:45-10:05, Nhà TC, Phòng 307
- Kênh trao đổi của lớp trên Microsoft Teams

  Hoc_ky_2021-1_Hoc_phan_IT3190_Ma_lop_128706

Mục tiêu của học phần

- Có kiến thức cơ bản về học máy và khai phá dữ liệu
- Có hiểu biết về các phương pháp học máy và khai phá dữ liệu, các ưu điểm và các nhược điểm của mỗi phương pháp
- Được giới thiệu về các framework và các công cụ phần mềm
- Trải nghiệm về thiết kế, cài đặt và đánh giá hiệu năng của hệ thống học máy/khai phá dữ liệu
  - Thông qua đồ án môn học (làm việc nhóm)

Nội dung của học phần

- Giới thiệu về Học máy, Khai phá dữ liệu, Các framework và công cụ phần mềm
- Tiền xử lý dữ liệu
- Đánh giá hiệu năng của hệ thống
- Hồi quy
- Phân lớp
- Phân cụm
- Phát hiện luật kết hợp

Đánh giá điểm học phần

- Đồ án môn học (**P**): Tối đa 10 điểm
  - Mỗi đồ án được thực hiện bởi **một nhóm gồm 3-4 sinh viên**
  - Tự do chọn một phương pháp học máy/khai phá dữ liệu để giải quyết một bài toán thực tế
  - Cài đặt hệ thống học máy/khai phá dữ liệu, và đánh giá hiệu năng của hệ thống sử dụng một tập dữ liệu (dataset) phù hợp
- Thi kết thúc học phần (**E**): Tối đa 10 điểm
- Điểm học phần (**G**)
  - **G** = 0,4x**P** + 0,6x**E**

Đồ án môn học: Đề xuất đề tài

- Tự do đề xuất bài toán thực tế, (các) giải thuật học máy/khai phá dữ liệu để giải quyết bài toán, và (các) tập dữ liệu được sử dụng
- Đề xuất đề tài (được lưu trong .pdf file) phải được **diễn giải cụ thể**:
  - Dài khoảng **1 hoặc 2 trang**
  - **Mô tả bài toán thực tế** được giải quyết (mục đích, yêu cầu, kịch bản ứng dụng, …)
  - Chỉ định rõ **(các) giải thuật học máy/khai phá dữ liệu** sẽ được dùng để giải quyết bài toán
  - Trình bày các thông tin về **đầu vào (input)** và **đầu ra (output)** của hệ thống học máy/khai phá dữ liệu sẽ được cài đặt, và **cách biểu diễn các ví dụ học (the representation of learning examples)**
  - Chỉ định rõ **(các) tập dữ liệu (datasets)** sẽ được sử dụng
  - **Kế hoạch thực hiện** (tên nhiệm vụ, những người tham gia, thời điểm bắt đầu, thời điểm kết thúc)
- Gửi đến địa chỉ *quang.nguyennhat@hust.edu.vn*/ quangnn@soict.hust.edu.vn **không muộn hơn 27/10/2021**
  - **Đề xuất đề tài** của nhóm
  - Thông tin các thành viên của nhóm: **Họ tên, Mã số sinh viên, Email**

Đồ án môn học: Các yêu cầu

- Kết quả của các đồ án môn học sẽ được trình bày ở 4 tuần cuối

  Tất cả các thành viên phải tham gia vào việc thực hiện công việc và trình bày!

- Báo cáo kết quả của đồ án môn học bao gồm:
  - **Mã nguồn** (source codes): Lưu trong một file nén
  - **File hướng dẫn** 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** (được lưu trong .pdf file):
    - Giới thiệu và mô tả về bài toán thực tế được giải quyết
    - Các chi tiết của (các) phương pháp học máy/khai phá dữ liệu và (các) tập dữ liệu được sử dụng
    - Các kết quả thí nghiệm đánh giá hiệu năng của hệ thống đối với (các) tập dữ liệu được sử dụng
    - Các vấn đề/khó khăn gặp phải trong quá trình thực hiện công việc của đồ án, và cách thức giải quyết (khắc phục)
    - Các tranh luận/khám phá/kết luận, và các đề cử cho việc tiếp tục phát triển và cải tiến trong tương lai

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 học máy/khai phá dữ liệu được dùng để giải quyết bài toán**
  - **Các kết quả thí nghiệm minh chứng thuyết phục, và sự thỏa đáng của các tranh luận (nhận xét, đánh giá) đối với các kết quả thí nghiệm**
  - Chất lượng của Bài trình bày (presentation) kết quả công việc
  - Chất lượng của Tài liệu báo cáo kết quả đồ án
- Nội dung của Bài trình bày (presentation) phải 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 Bài trình bày**
- **Nghiêm cấm sao chép tài liệu hoặc mã nguồn của người khác!**

Tài liệu học tập

- Các bài giảng (Lecture slides)
  - Trong thư mục Files\Class Materials trên Microsoft Teams
- Sách tham khảo:
  - Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*. Springer, 2009.
  - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. *Introduction to Data Mining (2nd Edition)*. Pearson, 2017.
  - Jiawei Han, Micheline Kamber, Jian Pei. *Data Mining: Concepts and Techniques (3rd Edition)*. Morgan Kaufmann, 2011.
- Các framework, thư viện, công cụ phần mềm dành cho Học máy và Khai phá dữ liệu
- Các tập dữ liệu (datasets):
  - Kaggle
  - UCI
  - WEKA