

Cấu trúc môn học

- Số tiết: 45 tiết
- Thời gian và địa điểm

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 BTL môn học

Đánh giá

- BTL Lý thuyết + Bài tập (**P**): Tối đa 10 điểm
 - Mỗi BTL 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 báo cáo (**E**)
- Điểm học phần (**G**)
 - **G** = $10\% + 10\% + 20\% + 60\%$

BTL 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 hoặc lựa chọn trong các bài cô gợi ý
- 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.

BTL môn học: các yêu cầu

- Kết quả của BTL phải được trình bày ở cuối môn học
 - Tất cả các thành viên phải tham gia vào việc thực hiện và trình bày BTL
- Báo cáo kết quả của BTL 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 BTL môn học (lưu trong file .doc):
 - 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 BTL, và cách thức được dùng để giải quyết (vượt qua)
 - Các khám phá mới hoặc kết luận

BTL môn học: đánh giá

- Công việc BTL 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 BTL
 - Chất lượng của tài liệu báo cáo kết quả BTL
 - 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