

Học Máy

(Machine Learning)

Thân Quang Khoát

khoattq@soict.hust.edu.vn

Viện Công nghệ thông tin và Truyền thông
Trường Đại Học Bách Khoa Hà Nội
Năm 2017

Cấu trúc môn học

- Số tuần: 15
 - Lý thuyết: 10-12 tuần
 - Sinh viên trình bày đồ án môn học: 03-04 tuần
- Thời gian và địa điểm
 - **Lớp CNTT&TT K58:** Thứ 6 hàng tuần; 9:20—11:50; Phòng TC-208
- Thời gian gặp sinh viên
 - **Hẹn trước qua e-mail**
 - DSLab, Viện CNTT&TT (Nhà B1, Phòng 1002)
- Thư mục bài giảng:
<http://is.hust.edu.vn/~khoattq/lectures/ML-8-2017/>

Nội dung môn học

- Lecture 1: Giới thiệu về Học máy
- Lecture 2: Phân cụm
(Clustering with K-means, Online K-means)
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4: Phân lớp với Naïve Bayes
- Lecture 5: Láng giềng gần nhất (KNN)
- Lecture 6: Rừng ngẫu nhiên (Random forests)
- Lecture 7: Máy vector hỗ trợ (SVM)
- Lecture 8: Mạng nơon (Neural networks)
- Lecture 9: Đánh giá hiệu năng và lựa chọn mô hình
(Model assessment and selection)
- Lecture 10: thảo luận bài toán thực tiễn

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
- Làm quen và sử dụng được công cụ Weka hoặc thư viện Scikit-learn
- Có kinh nghiệm về thiết kế, cài đặt, và đánh giá hiệu năng của hệ thống học máy
 - Thông qua đồ án môn học

Đánh giá

- Đề án môn học (**P**): Tối đa 10 điểm
 - Mỗi đề án được thực hiện bởi một nhóm gồm 3-5 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** = $0,3 \times \mathbf{P} + 0,7 \times \mathbf{E}$

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.
- Đăng ký **trước 22/10/2017**
 - Qua Google form (sẽ thông báo sau trên FB group)
 - Đề xuất đề tài của nhóm (tên đề tài, và mô tả)
 - Thông tin các thành viên của nhóm: Tên, Mã số sinh viên, Email, lớp

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 môn 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*
 - 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.
 - E. Alpaydin. *Introduction to Machine Learning*. MIT press, 2010.
 - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.
- 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/>
 - WEKA rep.: www.cs.waikato.ac.nz/ml/weka/index_datasets.html

Gợi ý



pandas

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

