

Giới thiệu về môn Nhập môn Khai phá dữ liệu

Trịnh Anh Phúc¹

¹Bộ môn Khoa Học Máy Tính, Viện CNTT & TT, Trường Đại Học Bách Khoa Hà Nội

Ngày 10 tháng 4 năm 2014

Giới thiệu

1 Thông tin về học phần

- Khối lượng giờ học
- Nhiệm vụ của sinh viên
- Các tài liệu tham khảo chính
- Các kiến thức cơ bản cần có của học phần

Khối lượng giờ học

- Lý thuyết: 45 giờ
- Bài tập (Bài tập lớn): 15 giờ
- Tự học: 60 giờ

Nhiệm vụ của sinh viên

- Dự lớp: đầy đủ theo quy chế
- Bài tập: hoàn thành các bài tập của học phần

Nội dung tóm tắt học phần

Khai phá dữ liệu (data mining)

Khai phá dữ liệu đề cập đến các kỹ thuật nhằm khám phá các mối quan hệ và các tri thức có ích tiềm năng trong những nguồn dữ liệu lớn. Học phần này cung cấp các kiến thức tổng quan về các nguyên tắc, kỹ thuật và ứng dụng của khai phá dữ liệu, đồng thời cho phép sinh viên thực hành những bài tập thực tế.

Nội dung tóm tắt học phần

Nội dung các chương của học phần

- 1 Giới thiệu chung về khai phá dữ liệu
- 2 Chuẩn bị dữ liệu và tiền xử lý dữ liệu
- 3 Luật kết hợp
- 4 Phân loại và dự báo
- 5 Phân cụm
- 6 Hồi quy
- 7 Khai phá dữ liệu với các kiểu dữ liệu phức tạp

Các tài liệu tham khảo chính

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2/e, 2006, Morgan Kaufmann Publishers, ISBN 1-55860-901-6.
- Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2/e, 2005, Morgan Kaufmann Publishers, ISBN 0-12-088407-0.
- David Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining, 2001, MIT Press, ISBN 0-262-08290-X.
- Daniel T. Larose, Discovering Knowledge in Data: An introduction to Data Mining, 2005, John Wiley & Sons, Inc., ISBN 0-471-66657-2.
- Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth (1996). From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining 1996: 1-34

Các thư viện, code trong lĩnh vực khai phá dữ liệu

- Weka (The University of Waikato, New Zealand)
<http://www.cs.waikato.ac.nz/ml/weka/>
- Matlab : Neural Network Toolbox (Bản quyền, trả tiền)
<http://www.mathworks.com/products/neural-network/>

Các thư viện, code trong lĩnh vực khai phá dữ liệu

- UCI Machine Learning Repository (University of California, Irvine)
<http://archive.ics.uci.edu/ml/>
- <http://www.sdss.org>
- <http://www.kdnuggets.com/>

Các kiến thức cơ bản cần có của học phần

- Dữ liệu số, độ lớn và kích thước dữ liệu
- Hệ quản trị cơ sở dữ liệu và truy vấn
- Giải thuật và cấu trúc dữ liệu
- Không gian đại số tuyến tính
- Thống kê và xác suất

Dành cho trả lời câu hỏi

