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

Các ví dụ của đồ án môn học

- Có thể chọn một trong số các ví dụ đồ án môn học, *hoặc*
- Có thể đề xuất thay đổi dựa trên một trong số các ví dụ đồ án môn học, *hoặc*
- Có thể đề xuất một đề tài hoàn toàn mới
  - *Giải quyết một bài toán ứng dụng thực tế bằng kỹ thuật học máy/khai phá dữ liệu*

Lọc thư rác

- **Mô tả bài toán**. Xác định (phân loại) những thư điện tử là thư rác (spam e-mails)
- **Đầu vào**. Biểu diễn nội dung của một e-mail (vd: một vectơ các từ khóa – có/không có trọng số)
- **Đầu ra**. Thư rác (“spam”) hoặc thư hợp lệ (“normal”)
- **Phương pháp**. Phân lớp Naïve Bayes (Naïve Bayes classification)
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn nội dung của một e-mail và nhãn lớp (“spam” hoặc “normal”)

Phân loại các trang Web

- **Mô tả bài toán**. Với một tập các trang Web, hệ thống cần phải gán (phân loại) mỗi trang Web vào một trong số các thể loại (vd: “Kinh doanh”, “Thể thao”, “Công nghệ”, …)
- **Đầu vào**. Biểu diễn nội dung của một trang Web (vd: một vectơ các tần xuất xuất hiện của các từ khóa)
- **Đầu ra**. Thể loại phù hợp của trang Web đó
- **Phương pháp**. Phân lớp Naïve Bayes (Naïve Bayes classification), hoặc Mạng nơ-ron nhân tạo (Artificial neural network)
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn của một trang Web và nhãn lớp (thể loại)

Phân nhóm kết quả học tập của sv

- **Mô tả bài toán**. Hệ thống cần phân (gom) nhóm các sinh viên dựa trên một tập các yếu tố xác định trước (vd: kết quả trung bình học kỳ, giới tính, quê quán, số lượng các học phần đăng ký cho học kỳ, tỷ lệ tham dự các buổi học trên lớp,…)
- **Đầu vào**. Một vectơ các giá trị thuộc tính mô tả sinh viên
- **Đầu ra**. Nhóm mà sinh viên đó thuộc vào
- **Phương pháp**. Phân cụm K-means
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ là một vectơ các giá trị thuộc tính biểu diễn cho một sinh viên

Dự đoán mức độ rủi ro của hồ sơ vay tín dụng

- **Mô tả vấn đề**. Với một tập các hồ sơ xin vay tín dụng (tài chính), hệ thống cần phải dự đoán (phân loại) mức độ rủi ro của mỗi hồ sơ xin vay – để quyết định chấp nhận hay từ chối yêu cầu
- **Đầu vào**. Biểu diễn của một hồ sơ xin vay tín dụng (vd: một vectơ các giá trị thuộc tính)
- **Đầu ra**. Dự đoán (phân loại) mức độ rủi ro (vd: "thấp" – chấp nhận; "cao" – từ chối)
- **Phương pháp**. Phân lớp bằng cây quyết định (Decision tree classification), hoặc Phân lớp Naive Bayes (Naïve Bayes classification)
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn của một hồ sơ xin vay tài chính và nhãn lớp (mức độ rủi ro) tương ứng

Gợi ý các trang Web

- **Mô tả vấn đề**. Với một tập các trang Web mà một người dùng đã xem, hệ thống cần phải xác định (dự đoán) những trang Web nào (chưa được xem) mà người dùng đó thích xem. Ý tưởng (giả sử): hai người dùng xem 2 tập tương tự các trang Web, thì sẽ có xu hướng thích xem cùng các trang Web trong tương lai
- **Đầu vào**. Danh mục các trang Web mà người dùng đã xem (mỗi trang Web được xác định bởi định danh (ID), chứ không quan tâm đến nội dung)
- **Đầu ra**. Một tập (nhỏ, có chọn lọc) các trang Web chưa xem được gợi ý đến cho anh ta
- **Phương pháp**. Học dựa trên láng giềng gần nhất (Nearest neighbor learning), hoặc Lọc cộng tác (Collaborative filtering)
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ bao gồm định danh của một người dùng và danh sách (IDs) các trang Web mà người dùng đó đã xem

Phân tích giỏ hàng (Shopping basket analysis)

- **Mô tả vấn đề**. Dựa trên lịch sử mua hàng (bao gồm một tập các giao dịch mua hàng) của khách hàng, hệ thống cần xác định thói quen (các mẫu – patterns) mua hàng của khách hàng thể hiện dưới dạng các luật *NẾU mua các mặt hàng A,B,C THÌ cũng mua các mặt hàng X,Y*
- **Đầu vào**. Một tập các mặt hàng mà một người dùng đã lựa chọn (mua) trong phiên mua sắm hiện tại của anh ta
- **Đầu ra**. Một tập các mặt hàng mà anh ta sẽ quan tâm (có khả năng cao là sẽ mua)
- **Phương pháp**. Phát hiện các luật kết hợp (association rule mining) bằng giải thuật Apriori
- **Tập dữ liệu**. Một tập các giao dịch (transactions) mua hàng của khách hàng, trong đó mỗi giao dịch là một tập các mặt hàng mà một khách hàng đã mua trong phiên mua sắm của anh ta

So sánh thử nghiệm các f.f. HM/KPDL

- **Mô tả vấn đề**. Một bài toán thực tế phù hợp để giải quyết bằng học máy/khai phá dữ liệu (vd: một trong các bài toán vừa đề cập ở trên)
- **Tập dữ liệu**. Một tập dữ liệu phù hợp đối với bài toán được giải quyết
- **Nhiệm vụ:**
  - Lựa chọn một số (2-3) phương pháp học máy/khai phá dữ liệu phù hợp
  - Đối với mỗi phương pháp đã chọn, cài đặt một hệ thống tương ứng để giải quyết bài toán
  - So sánh hiệu năng của các hệ thống này đối với cùng một (hoặc một số) tập dữ liệu đã chọn
  - Ví dụ, sinh viên có thể so sánh về hiệu năng giữa phương pháp phân lớp Naive Bayes và phương pháp phân lớp bằng cây quyết định trong việc dự đoán (đánh giá) mức độ rủi ro của các hồ sơ xin vay tài chính