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 2016

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
  - *Phù hợp để giải quyết bài toán thực tế bằng Học máy*!
- **Chú ý năm nay:**
  - Không làm lọc thư rác
  - Không sử dụng giải thuật Naïve Bayes

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 học máy**. Mạng nơ-ron nhân tạo, hoặc Máy vectơ hỗ trợ, …
- **Tập dữ liệu**. Một tập các ví dụ; mỗi ví dụ bao gồm biểu diễn của một trang Web và nhãn lớp (thể loại)

Dự đoán lượng tải apps

- **Mô tả bài toán**. Hệ thống cần dự đoán số lượng tải (downloads) cho một app khi nó được đưa lên Google Play Store (https://play.google.com). Số lượng tải là các số nguyên. Dự đoán được đưa ra dựa vào các mô tả của app đó.
- **Đầu vào**. Các mô tả về một app mới dưới dạng texts.
- **Đầu ra**. Số lượng downloads
- **Phương pháp học máy**. SVM (hoặc một phương pháp khác)
- **Tập dữ liệu**. Một tập các apps với mô tả dạng văn bản; mỗi app đã biết số lượng downloads.

Dự đoán đánh giá người dùng apps

- **Mô tả bài toán**. Hệ thống cần dự đoán đánh giá của người dùng cho một mobile app khi nó được đưa lên App Store. Các đánh giá sẽ thuộc 5 mức khác nhau {1*, 2*, 3*, 4*, 5*}. Dự đoán đc đưa ra dựa vào mô tả của app đó.
- **Đầu vào**. Các mô tả về một app mới dưới dạng texts.
- **Đầu ra**. Đánh giá thuộc {1*, 2*, 3*, 4*, 5*}
- **Phương pháp học máy**. SVM (hoặc một phương pháp khác)
- **Tập dữ liệu**. Một tập các apps với mô tả dạng văn bản; mỗi app đã được đánh giá chất lượng.

Dự đoán tiện nghi hotels

- **Mô tả bài toán**. Hệ thống cần dự đoán mức độ tiện nghi và khả năng người dùng yêu thích một hotel khi hotel mới/chuẩn bị ra đời. Các đánh giá sẽ thuộc 5 mức khác nhau {1*, 2*, 3*, 4*, 5*}. Dự đoán đc đưa ra dựa vào mô tả về hotel đó.
- **Đầu vào**. Các mô tả về một hotel mới, bao gồm vị trí địa lý, các dịch vụ cung cấp, mô tả phòng ốc, …
- **Đầu ra**. Đánh giá thuộc {1*, 2*, 3*, 4*, 5*}
- **Phương pháp học máy**. kNN (hoặc phương pháp khác)
- **Tập dữ liệu**. Một tập các hotels với mô tả dạng văn bản; mỗi hotel đã được đánh giá chất lượng.
- Tham khảo các hotels tại đây: www.booking.com

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 máy**. Học dựa trên láng giềng gần nhất (k-NN)
- **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

So sánh thử nghiệm các phương pháp

- **Mô tả vấn đề**. Một bài toán thực tế phù hợp để giải quyết bằng học máy (vd: phân loại văn bản)
- **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 phù hợp
  - Đối với mỗi phương pháp học máy đã chọn, cài đặt một hệ thống học máy tương ứng để giải quyết bài toán, hoặc sử dụng thư viện đã có.
  - So sánh hiệu năng của các hệ thống học máy này đối với cùng một (hoặc một số) tập dữ liệu đã chọn, với nhiều tiêu chí khác nhau (accuracy, time, …).
  - Ví dụ, sinh viên có thể so sánh về hiệu năng giữa phương pháp phân loại SVM và phương pháp rừng ngẫu nhiên.