Chương 3: Các bộ dữ liệu

Xuyên suốt cuốn sách này, tất cả cách mô hình và phương pháp được áp dụng cho các bộ dữ liệu thực tế đều có sẵn trên mạng. Ta sẽ dùng các tập dữ liệu khác nhau cho các tác vụ khác nhau: Phân loại, hồi quy, và phân loại văn bản.

0.1 Thuê xe đạp (Hồi quy)

Tập dữ liệu sau thống kê số lượng xe đạp được thuê mỗi ngày từ công ty thuê xe đạp Capital-Bikeshare ở Washington D.C, Mỹ, cùng với thông tin thời tiết và các mùa. Dữ liệu đã được cấp phép sử dụng tự do bởi Capital-Bikeshare. Nhóm tác giả Fanaee-T and Gama (2013) sau đó đã thêm thông tin thời tiết và mùa tương ứng. Mục tiêu cho bộ dữ liệu này là dự đoán có bao nhiêu chiếc xe được thuê dựa trên thông tin cho trước về thời tiết và mùa. Bộ dữ liệu có thể truy cập tải miễn phí tại đây: UCI Machine Learning Repository.

Các đặc trưng (features) mới đã được thêm vào bộ dữ liệu nêu trên và không phải tất cả các đặc trưng có sẵn sẽ được dùng trong các ví dụ trong sách này. Danh sách các đặc trưng được dùng là:

- Số lượng xe đạp được thuê bởi cả khách hàng thông thường (casual) và thành viên (registered). Số lượng này sẽ là đầu ra của mô hình hồi quy.
- Thông tin về mùa, xuân, hạ, thu, hoặc đông.
- Ngày nghỉ lễ hoặc không nghỉ lễ.
- Năm 2011 hoặc 2012.
- Số lượng ngày tính từ 1/1/2011 (ngày đầu tiên trong bộ dữ liệu). Đặc trưng này được cân nhắc để theo dõi xu hướng dữ liệu theo thời gian.

- Ngày làm việc hay cuối tuần.
- Tình hình thời tiết ngày đó, ví dụ:
  - Trời quang (clear), ít mây (few clouds), hơi mây (partly cloudy), hoặc nhiều mây (cloudy).
  - Sương (mist), sương + mây, sương + mây vụn (broken clouds), sương + ít mây.
  - Tuyết nhẹ, mưa nhẹ + giông (thunderstorm) + mây rải rác (scattered clouds), mưa nhẹ + mây rải rác.
  - Mưa nặng (heavy rain) + mưa tuyết nhẹ (ice pellets) + giông + sương, tuyết + sương.
- Nhiệt độ theo độ Celsius.
- Độ ẩm tương đối theo phần trăm (0% - 100%).
- Tốc độ gió theo km/giờ.

Bộ dữ liệu trong sách đã được xử lý tương đối (slightly processing). Bạn đọc có thể tìm mã R trong trang Github cùng với tệp final RData file.

0.2 Bình luận rác trên Youtube (Phân loại văn bản)

Dữ liệu được dùng là tập dữ liệu về phân loại bình luận rác (Alberto, Lochter, và Almeida (2015)).

Các bình luận này được thu thập thông qua Youtube API từ 5 trong số 10 video được xem nhiều nhất trên YouTube vào nửa đầu năm 2015. Cả 5 đều là video âm nhạc. Một trong số đó là ”Gangnam Style“ của PSY. Các nghệ sĩ khác là Katy Perry, LMFAO, Eminem và Shakira.

Nhìn qua một số bình luận, ta thấy các bình luận được gắn nhãn thủ công là ”rác“ hoặc ”có ý nghĩa". Bình luận ”rác“ được mã hóa bằng ”1“ và ”có ý nghĩa“ bằng ”0".

Bạn cũng có thể truy cập YouTube và xem phần bình luận. Nhưng làm ơn đừng bị cuốn vào địa ngục YouTube và cuối cùng xem video những con

Bảng 1:

CONTENT	CLASS
Huh, anyway check out this you[tube] channel: kobyoshi02	1
Hey guys check out my new channel and our first vid THIS IS US THE MONKEYS!!! I'm the monkey in the white shirt, please leave a like comment and please subscribe!!!!	1
just for test I have to say murdev.com	1
me shaking my sexy ass on my channel enjoy ^_^	1
watch?v=vtaRGgvGtWQ Check this out .	1
Hey, check out my new website!! This site is about kids stuff. kidsmediausa . com	1
Subscribe to my channel	1
i turned it on mute as soon is i came on i just wanted to check the views...	0
You should check my channel for Funny VIDEOS!!	1
and u should.d check my channel and tell me what I should do next!	1

khỉ ăn cắp và uống cocktail từ khách du lịch trên bãi biển. Trình phát hiện bình luận rác Google Spam có lẽ cũng đã thay đổi rất nhiều kể từ năm 2015.

Xem qua video phá vỡ kỷ lục số lượt xem "Gangnam Style" tại đây.

Nếu bạn muốn thao tác với dữ liệu, bạn có thể dùng file RData cùng với tập lệnh R với một số hàm tiện lợi trong Github của cuốn sách.

0.3 Các nguy cơ gây ung thư cổ tử cung (Phân loại)

Bộ dữ liệu ung thư cổ tử cung bao gồm các chỉ số và nguy cơ để dự đoán liệu một người phụ nữ sẽ bị ung thư cổ tử cung hay không. Các tính năng bao gồm thống kê dân số (như tuổi), lối sống, và lịch sử y tế. Dữ liệu có thể được tải xuống từ kho lưu trữ của UCI Machine Learning và được mô tả bởi Fernandes, Cardoso và Fernandes (2017).

Một số đặc trưng (features) được sử dụng trong các ví dụ của sách là:

- Tuổi hiện tại
- Số lượng bạn tình

- Tuổi lần đầu quan hệ tình dục
- Số lần mang thai
- Có dùng thuốc lá hay không?
- Số năm đã dùng thuốc lá
- Có dùng thuốc tránh thai ảnh hưởng nội tiết tố?
- Số năm đã dùng thuốc tránh thai ảnh hưởng nội tiết tố
- Có dùng vòng tránh thai hay không?
- Số năm đã dùng vòng tránh thai
- Bệnh nhân có từng mắc bệnh lây qua đường tình dục hay không?
- Số lượng bệnh truyền nhiễm qua đường tình dục đã được chẩn đoán
- Thời gian kể từ khi chẩn đoán bệnh (STD) đầu tiên
- Thời gian kể từ lần chẩn đoán bệnh (STD) cuối cùng
- Kết quả sinh thiết "khoẻ mạnh" hay "ung thư" (mục tiêu/kết quả)

Kết quả sinh thiết đóng vai trò như là tiêu chuẩn vàng để chẩn đoán ung thư cổ tử cung. Đối với các ví dụ trong cuốn sách này, kết quả sinh thiết đã được sử dụng làm nhãn. Các giá trị bị thiếu trong mỗi cột được thay thế bởi giá trị xuất hiện nhiều nhất, đây là một giải pháp chưa tốt, vì giá trị có thể tương quan với xác suất mà giá trị bị thiếu. Rất có thể sẽ co sai lệch bởi vì các câu hỏi mang tính riêng tư rất cao. Nhưng đây không phải là một cuốn sách về việc xử lí việc thiếu dữ liệu, nên giải phá này được tạm chấp nhận.

Để mô phỏng lại các ví dụ của cuốn sách với bộ dữ liệu này, hãy tìm tập lệnh R tiền xử lý (preprocessing R-script) và tệp final RData trong Github của cuốn sách.