ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Final Project

Data Science

Đánh giá điểm khi xem phim

GVHD: Trần Trung Kiên
SV: Tôn Thất Tâm Định - 1512112
SV: Ngô Văn Hùng - 1412214
SV: Trương Ngọc Khải - 1412245

TP. HỒ CHÍ MINH, THÁNG 12/2018

Mục lục

Báo cáo này trình bày lại quá trình thực hiện đồ án cuối kì của nhóm.

1 Lý do chọn đề tài

Lựa chọn một bộ phim để xem luôn là một câu hỏi khó với mỗi người. Hiện tại có quá nhiều sự lựa chọn với ngành phát triển phim ảnh cực kì mạnh. Điều này dẫn đến mong muốn lựa chọn một bộ phim ưng ý với một vài đánh giá review từ phim.

Với thực trạng như vậy có rất nhiều trang review phim được sinh ra, nhưng đôi khi chúng không chuẩn xác hoặc với một số người chúng không chính xác về mặt nào đó.

Thế nên, với tinh thần đam mê phim ảnh nhóm chúng em sẽ phát triển đồ án môn học với đề tài này: Dự đoán điểm phim.

2 Một số kiến thức nền tảng về machine learning

2.1 Supervised learning

2.1.1 Classification

Bài toán phân lớp có thể được hiểu là với một tập dữ liệu đầu vào ta cần phải phân nhóm ...

2.1.2 Regression

Bài toán dự đoán được hiểu nôm na là với một tập data set đầu vào gồm 2 thành phần X, Y. Ta sẽ biểu diễn mối liên hệ giữa X và Y dưới dạng một hàm số và dùng hàm số này để dự đoán cho dữ liệu mới. Một số thuật toán regression phổ biến như linear regression, ridge regression,...

2.2 Unsupervised learning

Bài toán unsupervised learning là bài toán học không giám sát, tức là dữ liệu đầu vào chưa có label, nên nhiệm vụ của chúng ta là phải tìm một hàm số để phân tách dữ liệu đã cho thành các cụm rồi từ đó tiến hành label.

2.3 Một số thành tố cơ bản của một thuật toán machine learning

Model = Representation + Evaluation + Optimization

2.3.1 Representation

Trong supervised learning, một model phải được biểu diễn bởi một hàm xác suất. Tập các phân lớp được gọi là không gian giả thuyết của model. Việc chọn lựa một model tương ứng vơi việc chọn lựa không gian giả thuyết mà từ đó có thể học được trên không gian đó.

2.3.2 Evaluation

Khi đã có được giả thuyết rồi, công việc của ta là phải đưa ra một hàm số(risk function) để đánh giá độ chính xác của giả thuyết đó.

2.3.3 Optimization

Sau khi đã có thuật toán train(training algorithm hay còn gọi là learning algorithm) tương ứng với giả thuyết được chọn thì tức là ta đã có được hàm số biểu diễn cho classifier. Công việc bây giờ của ta là đi tìm các điểm cực trị của hàm để optimize kết quả học.

2.3.4 Overfiting

Vấn đề overfitting hay còn gọi là rotation estimation, là vấn đề xảy ra khi hàm số mà ta học được chỉ biểu diễn được đúng trên dữ liệu đã được trani(hay còn gọi là học vẹt) mà không thể dự đoán chính xác khi cho vào một bộ dữ liệu mới.

2.3.5 Model selection

Khi đã có một tập các hàm số cho trước, vấn đề của chúng ta bây giờ là chọn hàm số nào để có thể biểu diễn chính xác nhất dự đoán.

3 Khảo sát các phương pháp

Đây là bài toán thuộc dạng bài dự đoán giá trị(regression). Trước đây đã có một số bài toán tương tự thuộc dạng này được tổ chức trên kaggle như: https://www.kaggle.com/surya635/house-price-prediction, https://www.kaggle.com/c/house-prices-advanced-regression-techniques, ... Phương hướng giải quyết chung của những bài toán thuộc dạng này là chọn feature phù hợp để mô hình hóa bài toán(feature engineering).

4 Phương án giải quyết của của nhóm

Để thực hiện bài toán trên, chúng ta cần phải thu thập dữ liệu các bộ phim từ các trang đánh giá trên internet. Thu thập từ nhiều nguồn: imdb.com, the-numbers.com/movie ...

Sau khi đã có dữ liệu, ta sẽ huấn luyện một mô hình máy học dạng regression như linear regression hoặc ridge regression dựa trên tập dữ liệu mà chúng ta thu thập được. Với output thu được sẽ là điểm phim được máy đánh giá.

Với output thu được ta có thể trả lời được những câu hỏi mà người dùng cần

5 Tóm tắt quá trình thực hiện

5.1 Thu thập dữ liệu

Viết một chương trình thu thập dữ liệu từ trang : the-numbers.com/movie

Chương trình sử dụng thư viện crapy của python.

5.2 Xử lý dữ liệu

5.3 Huấn luyện và đưa ra hàm dự đoán

6 Phương hướng giải quyết

Viết crawler thu thập dữ liệu từ IMDB.com. Sau khi thu thập được dữ liệu chúng ta huấn luyện một mô hình máy học dựa trên input là những dữ liệu của các bộ phim và xuất ra kết quả mong

muốn sẽ là đánh giá điểm của bộ phim đó. Cuối cùng chúng ta sẽ cho ra một mô hình có khả năng dự đoán một bộ phim bất kì và điểm số này sẽ giải quyết vấn đề của bạn là có muốn xem bộ phim này hay không?

7 Phát biểu bài toán

Thu thập dữ liệu về các phim cũng như điểm đánh giá phim trên website : IMDB (internet movie database) - là điểm đánh giá một bộ phim từ rất nhiều người dùng, cũng như người xem. Vì vậy nó sẽ có độ khách quan cao hơn so với số ít lượng chuyên gia đánh giá phim.

8 Phương pháp giải quyết

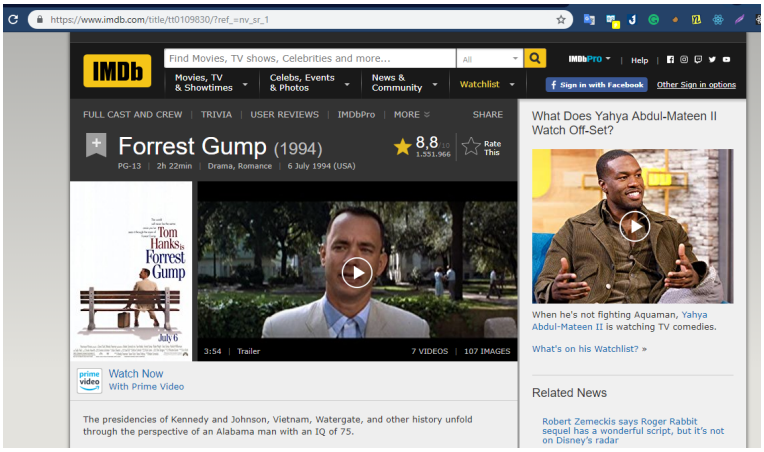
8.1 Thu thập dữ liệu:

Sử dụng thư viện crapy của python thua thập 5000 bộ phim từ trang :

http://www.the-numbers.com/movie/budgets/all

	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	12/18/2009	Avatar	$425,000,000	$760,507,625	$2,783,918,982
2	12/18/2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,058,662,225
3	5/24/2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$963,420,425
4	11/6/2015	Spectre	$300,000,000	$200,074,175	$879,620,923
5	7/20/2012	The Dark Knight Rises	$275,000,000	$448,139,099	$1,084,439,099
6	7/2/2013	The Lone Ranger	$275,000,000	$89,302,115	$260,002,115
7	3/9/2012	John Carter	$275,000,000	$73,058,679	$282,778,100
8	11/24/2010	Tangled	$260,000,000	$200,821,936	$586,581,936
9	5/4/2007	Spider-Man 3	$258,000,000	$336,530,303	$890,875,303
10	5/1/2015	Avengers: Age of Ultron	$250,000,000	$459,005,868	$1,404,705,868
11	5/6/2016	Captain America: Civil War	$250,000,000	$408,084,349	$1,151,684,349
12	3/25/2016	Batman v Superman: Dawn of Justice	$250,000,000	$330,360,194	$868,160,194

Từ tựa đề phim tiến hành tìm kiếm dữ liệu trên website.



8.2 Tiền sử lý dữ liệu

Ta sẽ br nhưng cột(feature) như language, content$_r aiting, country, num_u ser_f or_r eviews, ..., num_v oted_u sersØbtØutinh$

Sau đó, ta sẽ chia dữ liệu thành 3 phần:
Train gồm 3250 dòng.
Validation gồm 1083 dòng.
Test gồm 1086 dòng.
Mục đích của phần này nhằm chống overfitting.

Kế đến, ta sẽ tiến hành xử lí những cột có giá trị bị thiếu: Cột color thiếu 27 dòng, facenumber$_i n_p osterthiu21dng, budgetthiu198dng...$ Œivinhngctint64, $taschuynnsangfloat64(float64schoØchnhxc$

Chúng ta sẽ sử dụng mean, mode của tập huấn luyện để điền giá trị thiếu cho các tập validation và test. Vì tập huấn luyện thì chúng ta sẽ có đủ lớn để tính toán các giá trị này, còn đối với các dữ liệu mang tính chất kiểm thử mô hình, thường sẽ là các dòng đơn, chúng ta sẽ không thể nào tính toán được mean và mode của chúng.

Sau khi đã điền giá trị thiếu, ta sẽ chuyển cột object sang dạng one-hot, ở đây cột Color chỉ có 2 giá trị là Color / Black and White.

Thể loại của phim cũng sẽ được rất nhiều người quan tâm khi quyết định xem phim nên cột genres cũng ảnh hưởng rất nhiều đến quá trình huấn luyện. Vì thế nên tiến hành xử lí dữ liệu text cho cột genres (một bộ phim có thể có nhiều genres như Romace | Comedy) nên sẽ sử dụng phương pháp Bag of Words để chuyển những giá trị của dòng thành vector với các giá trị là 0 và 1. Ngoài ra, chúng ta cũng cần chuẩn hoá giá trị của các cột để cột có mean là 0 và độ lệch chuẩn (std) là 1. Cuối cùng thêm cột x0 và chuẩn bị huấn luyện.[1]

8.3 Máy học:

Ở đây ta sẽ sử dụng các mô hình Linear Regression và Ridge Regression. Bây giờ ta sẽ phát biểu bài toán lại để có thể thấy rõ mô hình hồi qui của ta.

Tập train có dữ liệu đầu vào sẽ là một ma trận có kích thước 27 x 1085. Trong đó 26 cột đầu tiên là các feature extract được từ 1085 bộ phim, cột cuối cùng là điểm phim do người dùng bình chọn.

Đặt ma trận X[1085 x 26] là input, ma trận Y[1085 x 1]. Ta giả sử có tồn tại một hàm số f(X) Y, và ta sẽ tiến hành đi tìm hàm số này dựa trên các thông số đã có.

Đầu tiên, ta giả sử giữa Y và X tồn tại một quan hệ tuyến tính tức là Y = aF(X) + b. Khi đó ta sẽ đi tìm a, b thỏa mãn phương trình này. Hàm linear regression trong sklearn đã giúp ta làm việc này một cách đơn giản. Sau khi đã có được hàm f rồi ta sẽ tiến hành đánh giá độ chính xác của hàm dựa trên tập test(được lưu trong file test.csv) với độ đo hàm lỗi là Mean Square Error(bình phương tối tiểu).

Mặt khác, ta sẽ đặt giả thuyết rằng hàm số f(X) Y là một hàm số bậc hai, và ta sử dụng mô hình Ridge Regression. Tiến hành tương tự như trên và đo độ lỗi cũng là hàm Mean Square Error.

9 Kết quả thực nghiệm

Sau khi chạy thực nghiệm ta có kết quả được lưu ghi vào bảng sau:

10 Kết luận

Tài liệu

[1] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469–472, 1985.