TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MÔ HÌNH ĐỒ THỊ CÓ XÁC SUẤT**

**TÌM HIỂU VỀ THƯ VIỆN PGMPY VÀ ỨNG DỤNG**

*Người hướng dẫn*: **TS NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **ĐÀO VIỆT DUY – MSSV: 51603070**

**NGUYỄN DUY HÀN LÂM – MSSV: 196005004**

Khoá  **: 2019**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MÔ HÌNH ĐỒ THỊ CÓ XÁC SUẤT**

**TÌM HIỂU VỀ THƯ VIỆN PGMPY VÀ ỨNG DỤNG**

*Người hướng dẫn*: **TS NGUYỄN CHÍ THIỆN**

*Người thực hiện*: **ĐÀO VIỆT DUY – MSSV: 51603070**

**NGUYỄN DUY HÀN LÂM – MSSV: 196005004**

Khoá  **: 2019**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Nguyễn Chí Thiện;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Đào Việt Duy*

*Nguyễn Duy Hàn Lâm*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Tài liệu này xin trình bày về thư viện pgmpy của Python trong nghiên cứu về lĩnh vực mô hình đồ thị có xác suất và xây dựng ứng dụng dựa trên thư viện này.

MỤC LỤC

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN ii](#_Toc57499431)

[TÓM TẮT iii](#_Toc57499432)

[MỤC LỤC 1](#_Toc57499433)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 2](#_Toc57499434)

[CHƯƠNG 1: MÔ HÌNH ĐỒ THỊ XÁC SUẤT 5](#_Toc57499435)

[1.1 Giới thiệu 5](#_Toc57499436)

[1.1.1 Đồ thị xác suất 5](#_Toc57499437)

[1.1.2 Mô hình đồ thị vô hướng 8](#_Toc57499438)

[1.1.3 Suy luận và học (Inference and learning) 9](#_Toc57499439)

[1.2 Create, define model 10](#_Toc57499440)

[1.3 Suy luận thống kê 13](#_Toc57499441)

[1.4 Learning 17](#_Toc57499442)

[1.4.1 Parameter 17](#_Toc57499443)

[1.4.2 Structure 20](#_Toc57499444)

[1.5 Tóm tắt chương 1 23](#_Toc57499445)

[CHƯƠNG 2: ỨNG DỤNG PHÂN TÍCH NHIỄM BỆNH 24](#_Toc57499446)

[2.1 Bayesian 24](#_Toc57499447)

[2.1.1 Giới thiệu 24](#_Toc57499448)

[2.1.2 Học 26](#_Toc57499449)

[2.1.3 suy luận xác suất 27](#_Toc57499450)

[2.2 Phân tích đề tài 30](#_Toc57499451)

[2.2 Xây dựng chương trình 31](#_Toc57499452)

[2.4 Kết quả học mô hình 35](#_Toc57499453)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1.1. Mô hình đồ thị biểu thị mối tương quan giữa R, W1, W2 5](#_Toc57499380)

[Hình 1.2. Ví dụ về mô hình đồ thị 6](#_Toc57499381)

[Hình 1.3. Ví dụ về mô hình đồ thị 7](#_Toc57499382)

[Bảng 1.1. Bảng biểu thị xác suất giữa W1 và R 7](#_Toc57499383)

[Bảng 1.2. Bảng biểu thị xác suất giữa W2 và R 7](#_Toc57499384)

[Hình 1.4. Ví dụ về đồ thị vô hướng 8](#_Toc57499385)

[Hình 1.5. Mô hình đồ thị biểu diễn các mối tương quan của mô hình 1 [1] 11](#_Toc57499386)

[Hình 1.6. Tạo mô hình 1 bằng pgmpy [1] 11](#_Toc57499387)

[Hình 1.7. Khởi tạo các giá trị xác suất của mô hình 1 [1] 11](#_Toc57499388)

[Hình 1.8. Thêm các giá trị xác suất vừa khởi tạo vào mô hình 1 [1] 12](#_Toc57499389)

[Hình 1.9. Xuất các mối quan hệ độc lập của mô hình 1 [1] 13](#_Toc57499390)

[Hình 1.10. Mô hình “asia” [2] 14](#_Toc57499391)

[Hình 1.11. Thực hiện lấy mô hình “asia” [3] 14](#_Toc57499392)

[Hình 1.12. In các giá trị đỉnh, cạnh và xác suất quan hệ của mô hình “asia” [3] 15](#_Toc57499393)

[Hình 1.13. Thực hiện loại bỏ các biến của mô hình “asia” [3] 15](#_Toc57499394)

[Hình 1.14. Tiến hành in các giá trị xác suất quan hệ mong muốn [3] 15](#_Toc57499395)

[Hình 1.15. Các bảng thống kê các giá trị xác suất quan hệ mong muốn [3] 16](#_Toc57499396)

[Hình 1.16. Kiểm tra xem các thuộc tính có bị ảnh hưởng bởi các thuộc tính khác hay không [3] 16](#_Toc57499397)

[Hình 1.17. Kết quả sau khi chạy câu lệnh hình 1.16 [3] 17](#_Toc57499398)

[Hình 1.18. Mô hình “alarm” [4] 18](#_Toc57499399)

[Hình 1.19. Lấy và tạo mô hình “alarm” [5] 18](#_Toc57499400)

[Hình 1.20. Tạo mô hình “alarm” bằng các cặp cạnh đã được định nghĩa [5] 19](#_Toc57499401)

[Hình 1.21. Tính toán các giá trị xác suất cho mô hình bằng MaximumLikelihoodEstimator [5] 19](#_Toc57499402)

[Hình 1.22. Tính toán các giá trị xác suất cho mô hình bằng BayesianEstimator [5] 20](#_Toc57499403)

[Hình 1.23. Chuẩn bị các thư viện cho việc học cấu trúc [6] 21](#_Toc57499404)

[Hình 1.24. Lấy dữ liệu mô hình alarm [6] 21](#_Toc57499405)

[Hình 1.25. Xây dựng mô hình bằng thuật toán PC biến thể stable [6] 21](#_Toc57499406)

[Hình 1.26. Xây dựng mô hình bằng thuật toán PC biến thể orig [6] 21](#_Toc57499407)

[Hình 1.27. Xây dựng hàm score để đánh giá thuật toán [6] 22](#_Toc57499408)

[Hình 1.28. Kết quả đánh giá thuật toán PC biến thể “stable” [6] 22](#_Toc57499409)

[Hình 1.29. Kết quả đánh giá thuật toán PC biến thể “orig” [6] 22](#_Toc57499410)

[Hình 1.30. Tiến hành xây dựng học cấu trúc cho mô hình bằng Hill-climb [6] 22](#_Toc57499411)

[Hình 2.1. Minh họa mô hình mạng Bayesian 25](#_Toc57499412)

[Hình 2.2. Minh họa inference suy luận xác suất Bayesian 28](#_Toc57499413)

[Hình 2.3. Minh họa inference suy luận xác suất Bayesian 29](#_Toc57499414)

[Hình 2.4. Minh họa inference suy luận xác suất Bayesian 30](#_Toc57499415)

[Hình 2.5. Chuẩn bị thư viện cho chương trình 31](#_Toc57499416)

[Hình 2.6. Tiến hành nạp dữ liệu vào chương trình 32](#_Toc57499417)

[Hình 2.7. Lọc ra các giá trị riêng biệt cho từng loại bệnh, symptom 32](#_Toc57499418)

[Hình 2.8. Xây dựng hàm tạo các cặp cạnh cho mô hình 33](#_Toc57499419)

[Hình 2.9. Tiến hành xây dựng mô hình 33](#_Toc57499420)

[Hình 2.10. Tiến hành thống kê symptom và disease 33](#_Toc57499421)

[Hình 2.11. Xây dựng hàm chuyển đổi sang dataframe 34](#_Toc57499422)

[Hình 2.12. Tính xác suất cho Symptom\_1 bằng MaximumLikelihooddEstimator 34](#_Toc57499423)

[Hình 2.13. Tính xác suất cho Disease bằng MaximumLikelihooddEstimator 34](#_Toc57499424)

[Hình 2.14. Tính xác suất cho Disease bằng BayesianEstimator 34](#_Toc57499425)

[Hình 2.15. Xuất kết quả tính xác suất của Disease bằng MaximumLikelihooddEstimator ra file CSV 34](#_Toc57499426)

[Hình 2.16. Xuất kết quả tính xác suất của Disease bằng BayesianEstimator ra file CSV 35](#_Toc57499427)

[Hình 2.17. Thống kê các triệu chứng cho các bệnh khác nhau 35](#_Toc57499428)

[Hình 2.18. Thống kê số lần xuất hiện các triệu chứng của Symptom\_1 36](#_Toc57499429)

[Hình 2.19. Bảng thống kê xác suất giữa symptom và disease 37](#_Toc57499430)

CHƯƠNG 1: MÔ HÌNH ĐỒ THỊ XÁC SUẤT

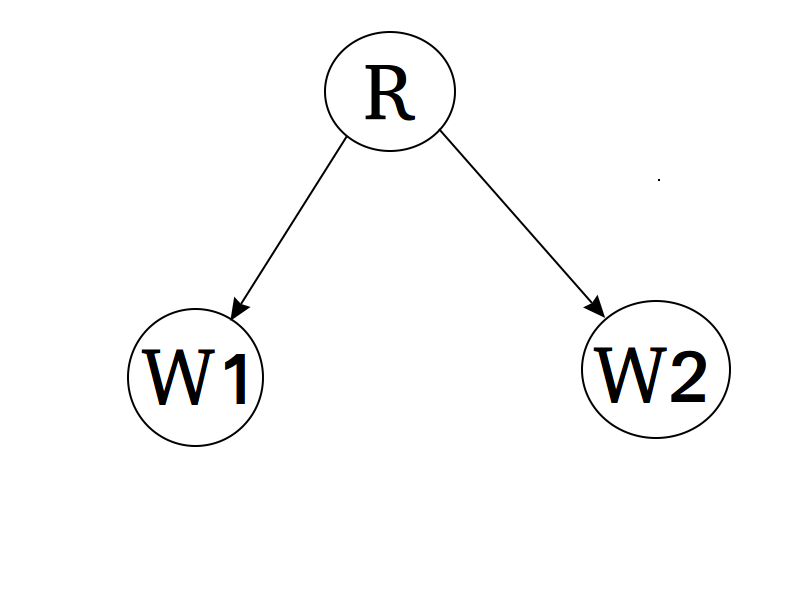
* 1. Giới thiệu

Khi đưa vào huấn luyện mạng Neural cho các mô hình lớn, bộ dữ liệu đầu vào thường rất lớn và quan hệ giữa các biến trong bộ dữ liệu rất phức tạp. Đến một lúc nào đó, dữ liệu không thể chỉ đơn thuần xuất ra bảng để phân tích vì có quá nhiều loại. Với mô hình đồ thị xác suất, ta có thể hiển thị thông tin dưới dạng gói gọn, xúc tích hơn trong kiểu đồ thị, thể hiện thông tin quan hệ các biến ngẫu nhiên phụ thuộc vào các biến khác.

1.1.1 Đồ thị xác suất

Mọi mô hình đồ thị xác suất (Probabilistic Graphical Model – PGM) bao gồm hai loại thành phần: các nút, hay gọi là biến ngẫu nhiên và các cạnh phụ thuộc thống kê giữa các nút với nhau. Đôi khi tất cả các biến đó có thể có sẵn để quan sát, trong khi đôi khi chỉ có tập hợp con của chúng.

Ví dụ ta có ba biến nhị phân (có/không) biểu thị sự hiện diện của mưa (R), nước trên cửa sổ (W1) và nước trên đường (W2). Nếu không có bất kỳ thông tin bổ sung nào, chúng ta biết rằng có mối quan hệ nhân quả giữa mưa và hai biến số khác:



1. Mô hình đồ thị biểu thị mối tương quan giữa R, W1, W2

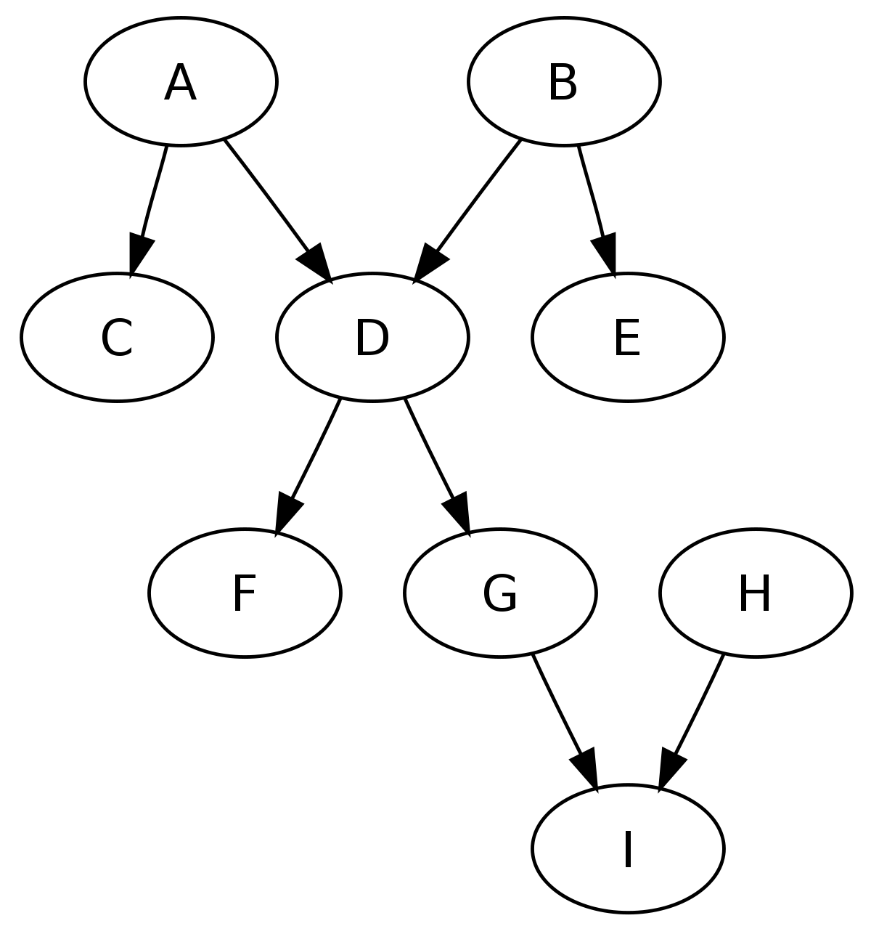
Chỉ có mưa mới có thể làm ướt cửa sổ và đường xá, nhưng ngược lại thì không. Ngoài ra, không có chu kỳ. Đây là mô hình đồ thị có hướng không có chu kỳ (Directed acyclic graph – DAG), loại PGM phổ biến nhất.

Nếu chúng ta quan sát thấy rằng khi R = Yes, chúng ta biết rằng xác suất W1 = Yes và W2 = Yes cao. Khi R = No, xác suất W1 = Yes và W2 = Yes sẽ rất thấp.

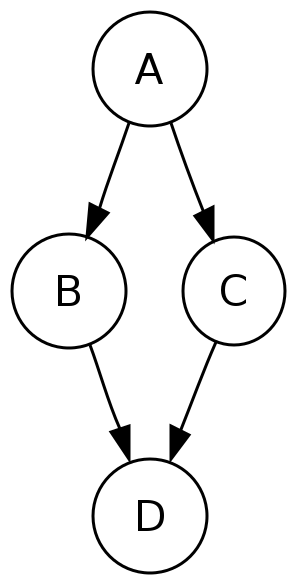
Khi chúng ta biết giá trị của R, xác suất cửa sổ ướt và đường ướt trở nên độc lập. Chúng có thể cao hoặc thấp, nhưng những sự kiện này không ảnh hưởng lẫn nhau nữa.

Ngoài ra, đây là Singly Connected DAG (SC-DAG), nghĩa là chỉ có một đường đi duy nhất, vô hướng giữa bất kỳ cặp nút nào.

Một số ví dụ khác:



1. Ví dụ về mô hình đồ thị



1. Ví dụ về mô hình đồ thị

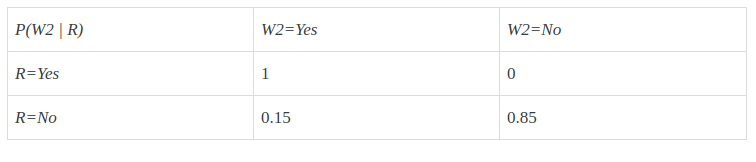
Ta có thể biểu diễn mô hình (hình 1.1) dưới dạng bảng xác suất có điều kiện (CPT), trong đó mỗi bảng biểu thị xác suất có điều kiện của việc quan sát một biến cho trước một biến khác.

Xác suất cửa sổ bị ướt khi mưa:

|  |  |  |
| --- | --- | --- |
| P(W1| R) | W1 = Yes | W1 = No |
| R = Yes | 0.9 | 0.1 |
| R = No | 0.03 | 0.97 |

1. Bảng biểu thị xác suất giữa W1 và R

Xác suất đường bị ướt khi mưa:



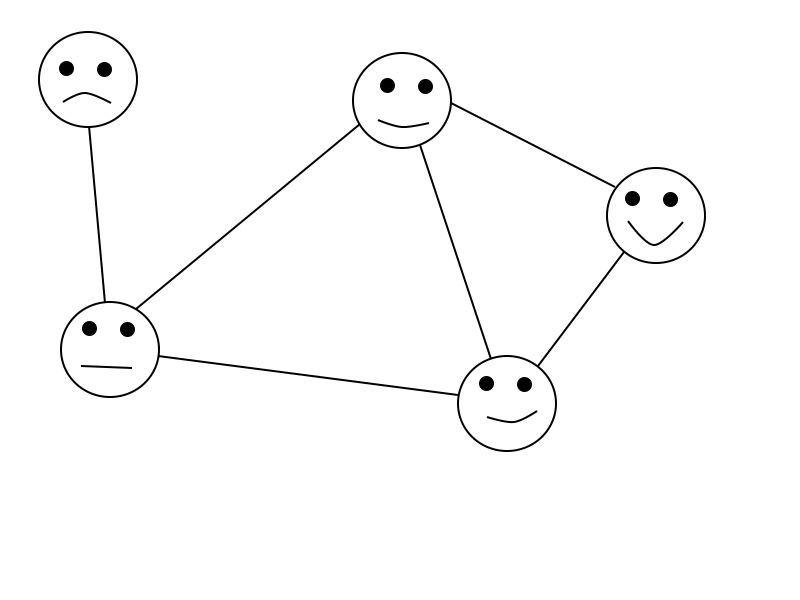
1. Bảng biểu thị xác suất giữa W2 và R

Bayes hay Bayesian network (BN) là những trường hợp đặc biệt của DAG. Chúng có thể được kết nối đơn lẻ hoặc nhiều lần và sử dụng logic Bayes để giải thích các xác suất. Nói tóm lại, điều đó có nghĩa là họ bắt đầu với một số giả định trước về sự phụ thuộc và sử dụng quy tắc Bayes để cập nhật niềm tin của họ về dữ liệu. Tương tự, hầu hết các mô hình DAG thực sự là Bayesian.

Mô hình Markov ẩn (Hidden Markov Model – HMM) và Mô hình không gian-trạng thái Gaussian tuyến tính (Gaussian State Space – GSS) cũng là những loại SC-DAG đặc biệt.

1.1.2 Mô hình đồ thị vô hướng

Còn được gọi là Trường ngẫu nhiên Markov (Markov Random Field – MRF). Giống như các mô hình có hướng, chúng thể hiện sự phụ thuộc có điều kiện giữa các biến ngẫu nhiên, nhưng không bao hàm quan hệ nhân quả có thứ tự. Ví dụ một mô hình đơn giản: Điểm hạnh phúc giữa mọi người.



1. Ví dụ về đồ thị vô hướng

Mức độ hạnh phúc của mỗi người trong ví dụ này ảnh hưởng đến điểm số của những người khác. Khi chúng ta quan sát một trong số họ, nó cũng cung cấp một số thông tin về hạnh phúc của những người còn lại, vì tất cả chúng ta đều được kết nối với nhau.

Trong số các loại phổ biến khác của mô hình đồ thị vô hướng là Trường ngẫu nhiên có điều kiện (Conditional Random Fields) và Restricted Boltzmann Machines.

Nhân tiện, DAG cung cấp quy trình suy luận đơn giản hơn và dễ giải thích hơn, trong khi các mô hình vô hướng có nhiều tiềm năng hơn để mô tả các mối quan hệ phức tạp.

Factor Graph cũng rất giống với MRF. Hơn nữa, các mô hình nhân tố truyền thống có thể được mở rộng để bao gồm cả mô hình Được hướng và Không có hướng. Trên thực tế, có rất nhiều nghiên cứu tương tự được thực hiện với mục tiêu tìm ra các kết nối sâu hơn giữa PGM, NN và các công cụ, thuật toán tính xác suất khác.

1.1.3 Học và học (Inference and learning)

Nếu đã quan sát một số biến, phân phối nào sẽ có các biến khác? Phương pháp phổ biến là sử dụng quy tắc Bayes, P (A cho trước B = b) = P (A và B = b) / P (B = b). Ý tưởng là: để có được xác suất của sự kiện A cho rằng B đã nhận giá trị b, bạn cần phải lấy xác suất cả A và B = b đã xảy ra và mở rộng quy mô nó bằng xác suất B = b, vì bạn đã biết rằng nó đã xảy ra.

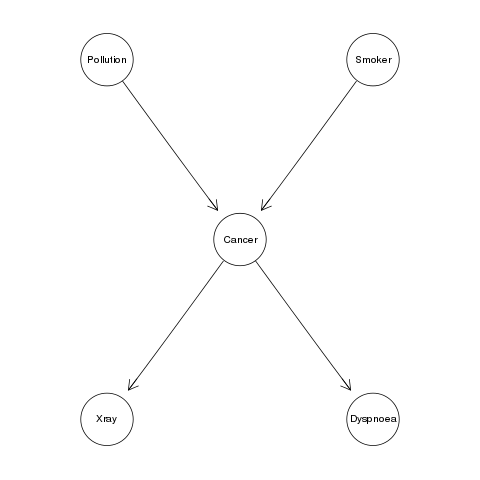
Tương tự như vậy, bằng cách tính toán các xác suất chung như A và B và các phân phối biên như B = b từ dữ liệu đào tạo, bạn có thể suy ra CPT. Khi một số biến bị ẩn, bạn có thể ước tính phân phối của chúng từ dữ liệu quan sát bằng phương pháp Tối đa hóa kỳ vọng (EM) hoặc Markov Chain Monte Carlo (MCMC).

Hơn nữa, các phép tính suy luận trong PGM có thể được thực hiện cục bộ. Thông tin về ảnh hưởng mà các nút gửi cho nhau thường được gọi là thông điệp. Chúng thường được chia theo loại Pi (mang xác suất trước) và Lambda (mang xác suất có thể xảy ra). Thuật toán phổ biến để trao đổi các thông điệp đó là Truyền bá niềm tin (BP), đây cũng là sự tổng quát của thuật toán Chuyển tiếp-Quay lại cho HMM và thuật toán Làm mịn Kalman cho GSS. Trong trường hợp đồ thị có chu kỳ hoặc vô hướng, nó được gọi là Loopy BP, vì những thông báo đó có thể đi theo các vòng gần nhau.

1.2 Create, define model

Trong thư viện pgmpy, cấu trúc mô hình không phụ thuộc tham số hóa( CPDs). Vì thế, ta có thể định hình mô hình trước, sau đó định hình tham số đầu vào. Như vậy, tham số có thể thay đổi tùy chỉnh tùy ý mà không ảnh hưởng đến mô hình đã được định hình ban đầu.

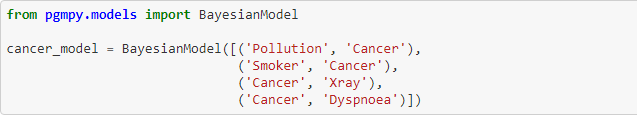
Để tiện cho việc giải thích, ta sẽ gói gọn lại trong việc thông qua ví dụ để giải thích cách sử dụng thư viện để định nghĩa mô hình gồm 5 trạng thái: Pollution, Smoker, Cancer, Xray, Dysponea (ta gọi tắt là mô hình 1).



1. Mô hình đồ thị biểu diễn các mối tương quan của mô hình 1 [1]

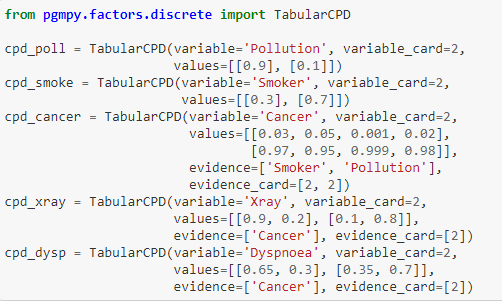
Giả định tạo một mô hình có hướng Bayesian từ đầu như trong ảnh trên.

Mô hình Bayesian được pgmpy cung cấp, BayesianModel, có thể được tạo bằng cách nhập vào danh sách các cạnh.



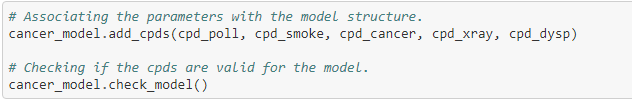
1. Tạo mô hình 1 bằng pgmpy [1]

Mỗi node trong mô hình đều có CPD tương ứng, trong trường hợp ảnh trên có 5 node Pollution, Smoke, Cancer, Xray, Dypnoea. Pgmpy cũng cung cấp lớp định nghĩa CPD là TabularCPD.



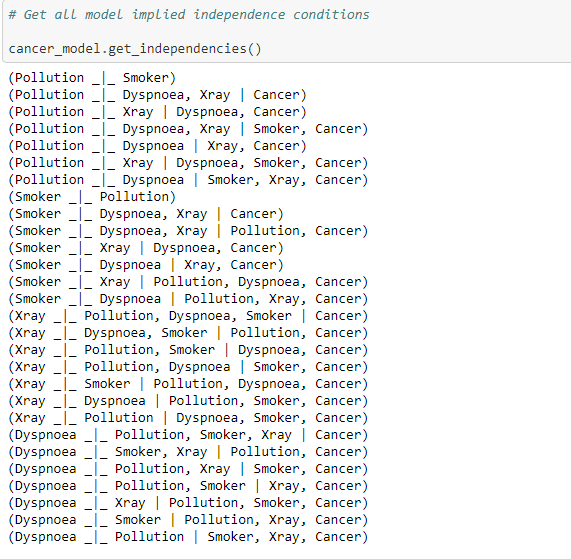
1. Khởi tạo các giá trị xác suất của mô hình 1 [1]

Sau khi đã định nghĩa mô hình và tham số đầu vào, giờ ta có thể them tham số vào mô hình thông qua hàm add\_cpd. Có thể dùng hàm check\_method để kiểm tra CPD có được định nghĩa phù hợp cho mô hình không.



1. Thêm các giá trị xác suất vừa khởi tạo vào mô hình 1 [1]

Sau khi định nghĩa mô hình, ta có thể kiểm định các quan hệ giữa các biến với nhau.



1. Xuất các mối quan hệ độc lập của mô hình 1 [1]

1.3 Suy luận thống kê

Trong tiểu mục này ta sẽ thực hiện các ví dụ về xuất các giá trị suy luận thống kê bằng pgmpy.

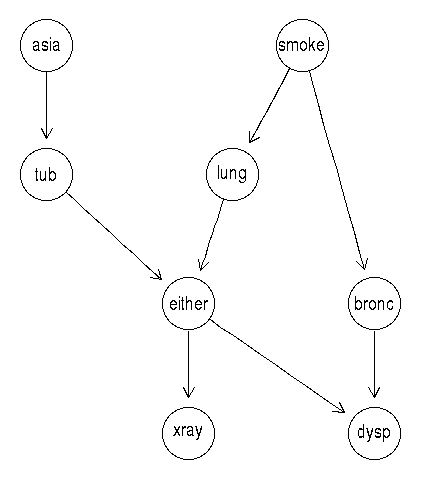
Suy diễn thống kê là quá trình sử dụng dữ liệu phân tích để suy ra các thuộc tính của phân phối xác suất cơ bản. Phân tích thống kê suy diễn đưa ra một tập hợp các đặc tính cần quan tâm.

Trong học máy, một tập các giá trị suy diễn chỉ có thể được xuất từ mô hình khi một tập dữ liệu lớn khác đã đào tạo và huấn luyện cho mô hình.

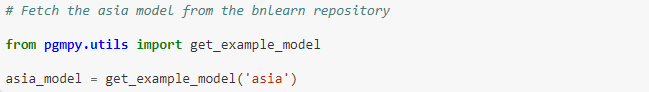
Khác với suy diễn thống kê, suy diễn thống kê sẽ sử dụng chính dữ liệu hiện tại để sinh ra tập các giá trị cần được quan sát, quan tâm.

Phần này ta sẽ thực hiện trên mô hình mẫu có tên là “asia”.

Mô hình “asia” có dạng như sau:

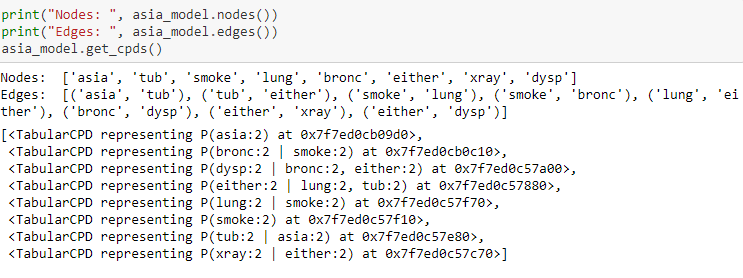


1. Mô hình “asia” [2]



1. Thực hiện lấy mô hình “asia” [3]

Tiếp theo ta sẽ thực hiện việc in các đỉnh (node), cạnh (edge) và các xác suất quan hệ của mô hình “asia”.



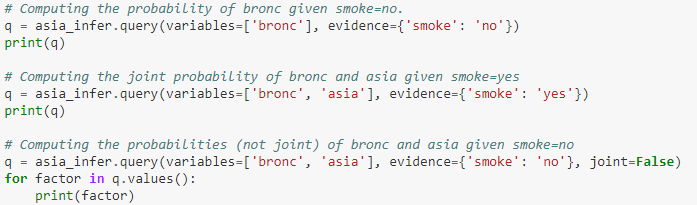
1. In các giá trị đỉnh, cạnh và xác suất quan hệ của mô hình “asia” [3]

Tiếp theo ta sẽ tiến hành trả về đồ thị quy nạp được hình thành bằng cách loại bỏ các biến của mô hình “asia”.



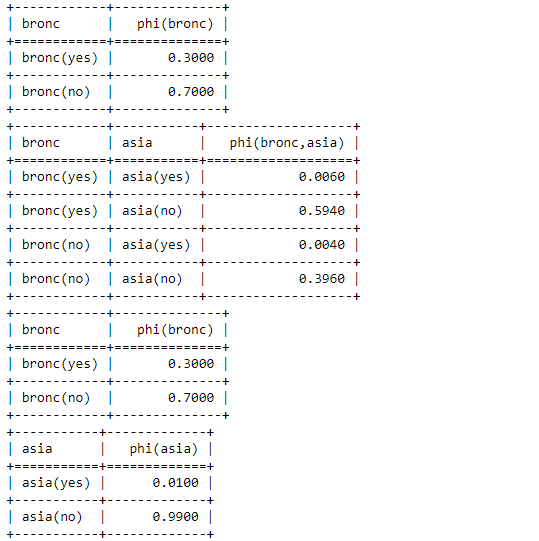
1. Thực hiện loại bỏ các biến của mô hình “asia” [3]

Tiếp theo ta sẽ tiến hành tính toán và in các giá trị xác suất quan hệ mà ta quan tâm.



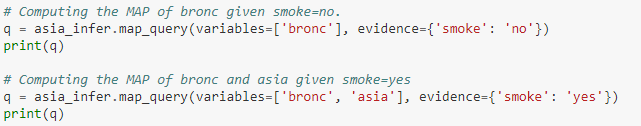
1. Tiến hành in các giá trị xác suất quan hệ mong muốn [3]

Sau khi chạy các câu lệnh ở hình 1.14, ta thu được kết quả:



1. Các bảng thống kê các giá trị xác suất quan hệ mong muốn [3]

Để xem 1 thuộc tính nào đó có bị ảnh hưởng bởi 1 thuộc tính khác không, ta sử dụng hàm map\_query.



1. Kiểm tra xem các thuộc tính có bị ảnh hưởng bởi các thuộc tính khác hay không [3]

Sau khi chạy câu lệnh ở hình 1.16, ta thu được kết quả:



1. Kết quả sau khi chạy câu lệnh hình 1.16 [3]

Như vậy ở hình 1.17 ta thu được kết quả đó là thuộc tính “bronc” có bị ảnh hưởng bởi thuộc tính “smoke” và thuộc tính “asia” không bị ảnh hưởng bởi thuộc tính “smoke”.

1.4 Learning

1.4.1 Parameter

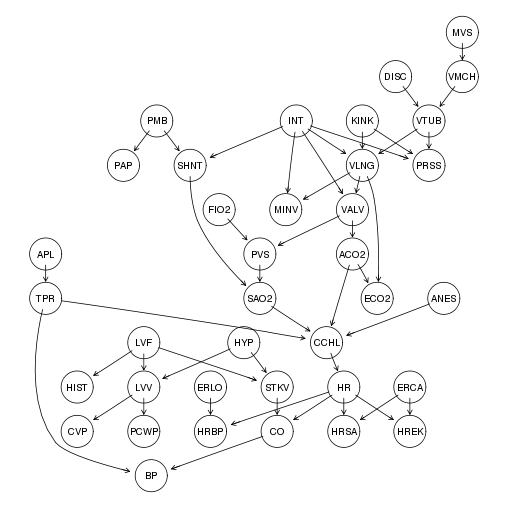
Pgmpy có hai cách chính để học tham số:

\_ Ước tính MaximumLikelihood.

\_ Ước tính Bayesian.

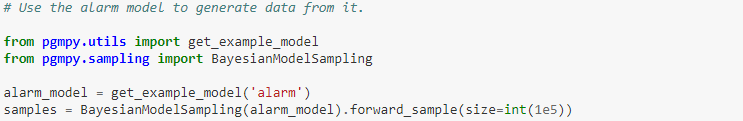
Để minh họa cho tiểu mục này, ta sẽ sử dụng mô hình và tham số có sẵn là alarm và sẽ học tham số mô hình từ dữ liệu đã có sẵn.

Mô hình “alarm” có dạng như sau:



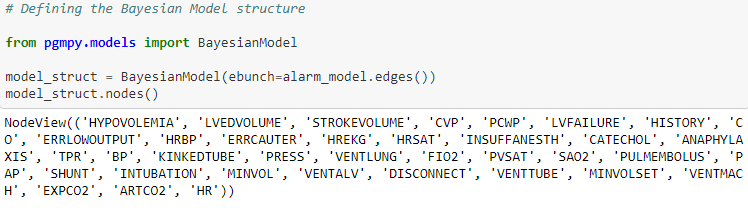
1. Mô hình “alarm” [4]

Bước đầu, ta lấy dữ liệu được pgmpy cung cấp sẵn.



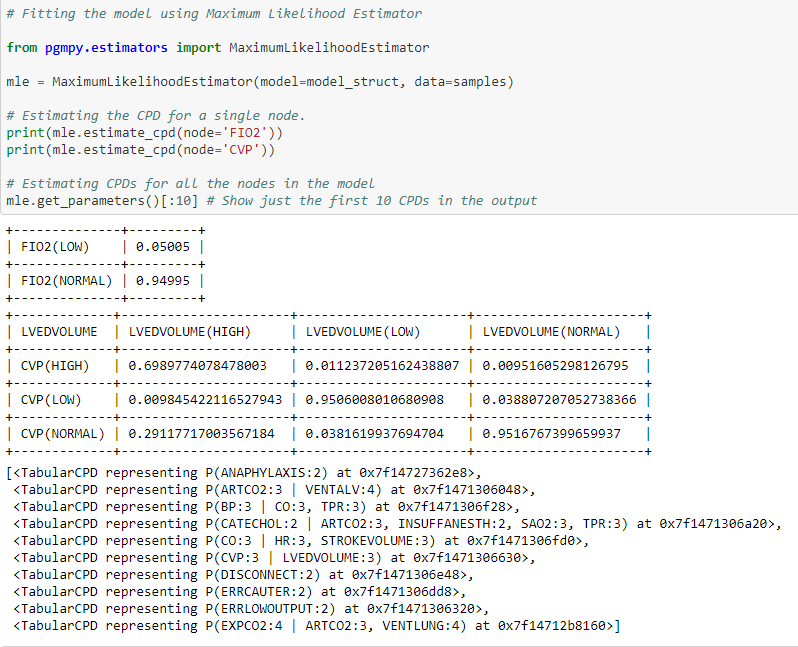
1. Lấy và tạo mô hình “alarm” [5]

Tiếp đến, ta có thể tạo mô hình “alarm” các dữ liệu các cặp cạnh đã được định nghĩa sẵn.

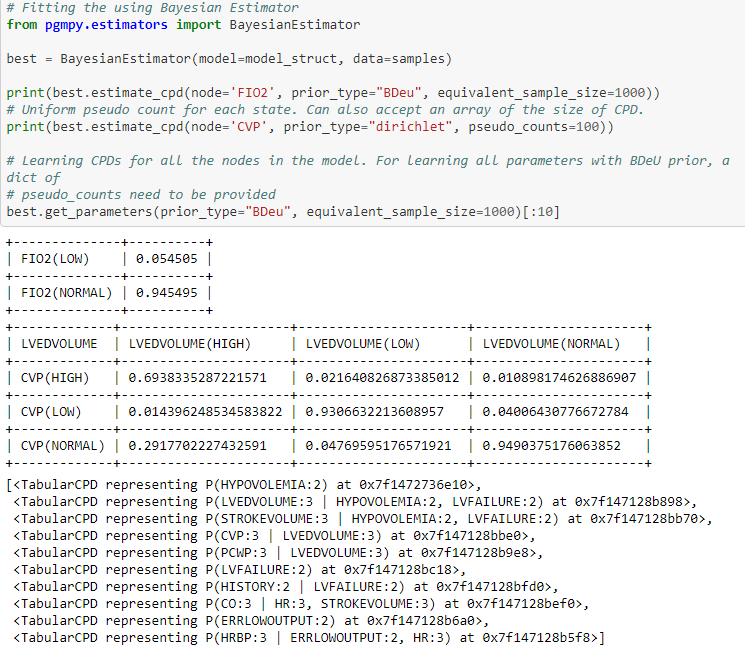


1. Tạo mô hình “alarm” bằng các cặp cạnh đã được định nghĩa [5]

Với mô hình và dữ liệu có được, ta có thể tiến hành học các tham số bằng MaximumLikelihoodEstimator hay BayesianEstimator.



1. Tính toán các giá trị xác suất cho mô hình bằng MaximumLikelihoodEstimator [5]



1. Tính toán các giá trị xác suất cho mô hình bằng BayesianEstimator [5]

1.4.2 Structure

Ở thời điểm bài bài báo cáo này, pgmpy đã có sẵn 3 thuật toán học cấu trúc:

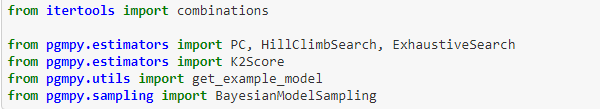
* Thuật toán dựa trên ràng buộc (Prototypical Constraint-based – PC) với các biến thể ổn định và song song.
* Thuật toán tìm kiếm leo đồi (Hill-Climb Search).
* Thuật toán tìm kiếm vét cạn (Exhaustive search).

Với thuật toán PC các phép thử có điều kiện có thể thực hiện trên pgmpy gồm:

* Kiểm định Pearsonr (Pearsonr test).
* Kiểm thử Chi bình phương (Chi-square test).

Để minh họa cho học cấu trúc, thay vì phải định hình lại cấu trúc hay tham số đầu vào như tiểu mục tạo và định hình mô hình. Ta sẽ sử dụng mô hình và dữ liệu có sẵn của pgmpy là alarm.

Đầu tiên nhập các thư viện cần có và chuẩn bị mô hình, dữ liệu.



1. Chuẩn bị các thư viện cho việc học cấu trúc [6]

Lấy dữ liệu của mô hình alarm.



1. Lấy dữ liệu mô hình alarm [6]

Sau khi đã chuẩn bị mô hình và tài nguyên. Ta có thể học cấu trúc mô hình bằng thuật toán PC với biến thể ổn định (stable) hay orig.

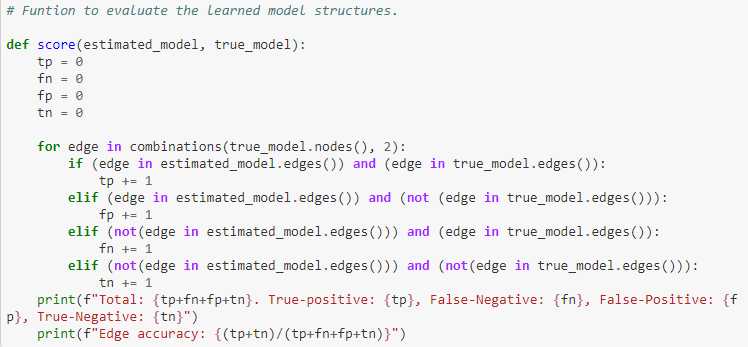


1. Xây dựng mô hình bằng thuật toán PC biến thể stable [6]



1. Xây dựng mô hình bằng thuật toán PC biến thể orig [6]

Tiếp theo ta sẽ xây dựng hàm score để đánh giá mức độ chính xác của quá trình học cấu trúc.



1. Xây dựng hàm score để đánh giá thuật toán [6]

Cuối cùng ta thu được kết quả đánh giá của việc xây dựng mô hình bằng thuật toán PC với biến thể “stable”:



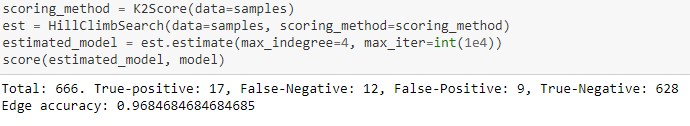
1. Kết quả đánh giá thuật toán PC biến thể “stable” [6]

Và kết quả đánh giá của việc xây dựng mô hình bằng thuật toán PC với biến thể “orig”:



1. Kết quả đánh giá thuật toán PC biến thể “orig” [6]

Tiếp theo ta sẽ thực hiện thử việc xây dựng cấu trúc mô hình bằng thuật toán leo đồi và tiến hành xuất kết quả đánh giá:



1. Tiến hành xây dựng học cấu trúc cho mô hình bằng Hill-climb [6]

1.6 Đánh giá mô hình học cấu trúc

1.5.1 Chi-square test cho PC

1.5.2 Pearsonr test cho PC

1.5.3 K2 score cho Hill-climb search và Exhaustive search

1.5.4 BDeu score cho Hill-climb search và Exhaustive search

1.5.5 Bic score Hill-climb search và Exhaustive search

1.7 Tóm tắt chương 1

Như vậy, trong chương 1, ta đã nắm một số khái niệm khái quát về mô hình đồ thị xác suất. Bao gồm các yếu tố cơ bản của đồ thị, đồ thị có hướng, đồ thị vô hướng, ta cũng tìm hiểu về các phép tính toán suy luận để tìm ra xác suất giữa quan hệ các biến ngẫu nhiên, và học tham số, học cấu trúc của mô hình. Ta cũng đã tìm hiểu về thư viện PGMPY dùng để tạo mô hình, vẽ đồ thị xác suất. Thông qua các minh họa, ta đã tìm hiểu về cách ứng dụng thư viện pgmpy tạo, định hình mô hình từ đầu, định hình các tham số đầu vào để có thể dễ dàng thay đổi tham số khi cần mà không ảnh hưởng mô hình. Ta cũng tìm hiểu cách suy luận bằng phương pháp loại bỏ biến của thư viện pgmpy. Xây dựng mô hình để học cấu trúc từ dữ liệu đạ có, học tham số trên dữ liệu.

Sau khi đã tìm hiểu về các chức năng cơ bản của thư viện pgmpy cùng lý thuyết cơ bản của mô hình đồ thị xác suất, ta sẽ sang bước tiếp theo trong bài báo cáo này là xây dựng ứng dụng minh họa.

CHƯƠNG 2: ỨNG DỤNG PHÂN TÍCH NHIỄM BỆNH

* 1. Bayesian

2.1.1 Giới thiệu

Mạng Bayes (BN) được xác định bởi:

một cấu trúc mạng, một đồ thị xoay chiều có hướng G, trong đó mỗi nút vi ∈ V tương ứng với một biến ngẫu nhiên Xi;

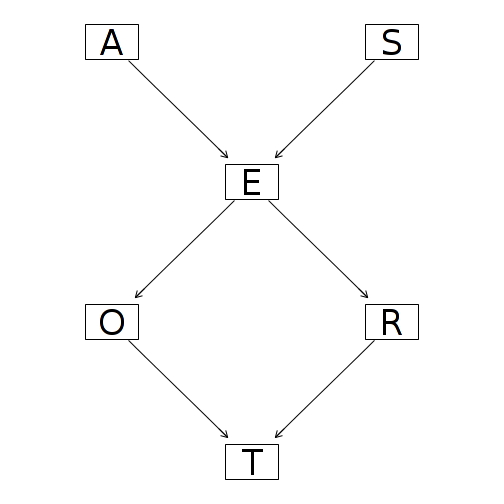
phân phối xác suất toàn cục X (với tham số Θ), có thể được tính theo nhân tử thành các phân phối xác suất cục bộ nhỏ hơn theo các cung có trong biểu đồ.

Vai trò chính của cấu trúc mạng là thể hiện các mối quan hệ độc lập có điều kiện giữa các biến trong mô hình thông qua sự phân tách bằng đồ thị, do đó xác định nhân tử của phân phối toàn cục:

phân hủy thành các phân phối cục bộ

Mỗi phân phối cục bộ có tập tham số riêng ΘXi; và ⋃ ΘXi nhỏ hơn nhiều so với Θ vì nhiều tham số được cố định bởi thực tế là các biến chúng thuộc về độc lập.

Vì vậy, thành phần đầu tiên là một đồ thị xoay chiều có hướng như thế này:



1. Minh họa mô hình mạng Bayesian

Thành phần thứ hai của BN là phân phối xác suất P (X). Sự lựa chọn phải được sao cho BN:

* có thể được học một cách hiệu quả từ dữ liệu;
* linh hoạt (nó có thể mã hóa nhiều loại hiện tượng hợp lý);
* dễ dàng truy vấn để thực hiện suy luận.

Ba lựa chọn phổ biến nhất trong tài liệu (cho đến nay), là:

* BN rời rạc, trong đó X và Xi ∣ ΠXi; và ΘXi là xác suất có điều kiện
* Gaussian BNs (GBN), trong đó X là chuẩn đa biến và Xi ∣ ΠXi là chuẩn đơn biến được xác định bởi mô hình hồi quy tuyến tính
* BN Gaussian tuyến tính có điều kiện (CLGBN), trong đó X là hỗn hợp của các chuẩn tắc đa biến và Xi ∣ ΠXi là đa thức, chuẩn đơn biến hoặc hỗn hợp của các chuẩn.
  + Xi rời rạc chỉ được phép có phần tử cấp trên rời rạc (ký hiệu là ΔXi), được giả định tuân theo phân phối đa thức được tham số hóa với bảng xác suất có điều kiện;
  + Xi liên tục được phép có cả phần tử cấp trên rời rạc và liên tục (ký hiệu là ΓXi, ΔXi ∪ ΓXi = ΠXi, và phân phối cục bộ của chúng là

có thể được viết dưới dạng hỗn hợp các hồi quy tuyến tính

chống lại các phần tử cấp trên liên tục với một thành phần cho mỗi cấu hình δXi của phần tử cấp trên rời rạc. Nếu Xi không có phần tử cấp trên rời rạc, thì hỗn hợp sẽ chuyển thành một hồi quy tuyến tính duy nhất.

2.1.2 Học

Lựa chọn mô hình và ước tính BN được gọi chung là học tập và thường được thực hiện như một quy trình gồm hai bước:

* học cấu trúc, học cấu trúc mạng từ dữ liệu;
* học tham số, học các phân phối cục bộ được ngụ ý bởi cấu trúc đã học ở bước trước.

Quy trình làm việc này là Bayesian; đã cho một tập dữ liệu D và nếu chúng ta biểu thị các tham số của phân phối toàn cục là X với Θ, chúng ta có:

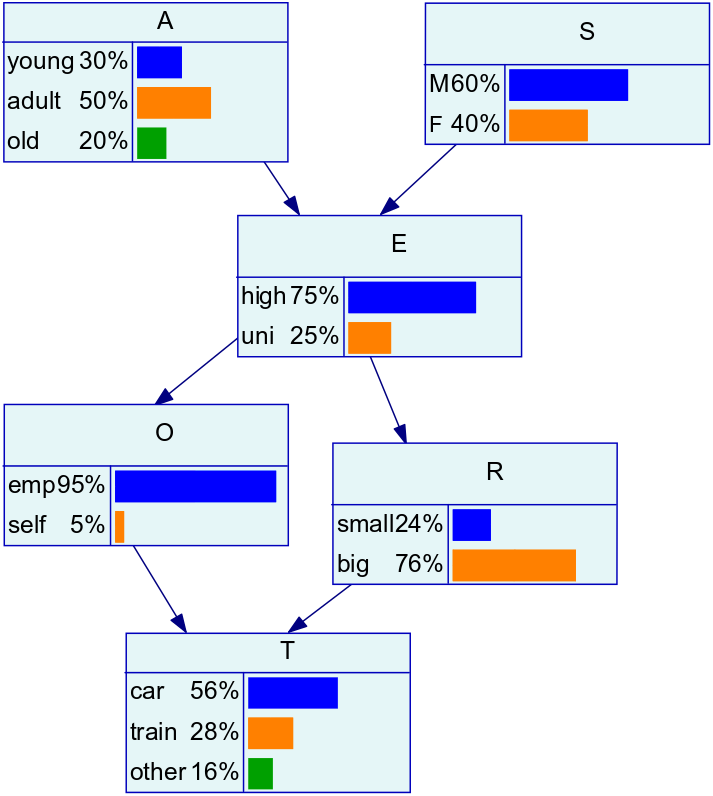
Với là kết quả học được, là kết quả học cấu trúc và là kết quả học tham số.

2.1.3 suy luận xác suất

Suy luận trên BN thường bao gồm các truy vấn xác suất có điều kiện (CP) hoặc tối đa một truy vấn sau thử nghiệm (MAP). Ý tưởng chung của truy vấn CP là:

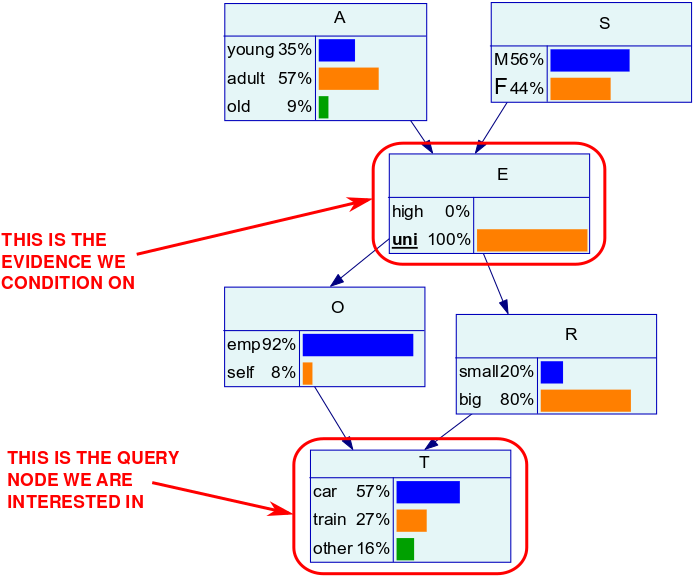
* Ta có một số bằng chứng, đó là chúng ta biết giá trị của một số biến và chúng ta sửa các nút cho phù hợp
* chúng ta muốn xem xét xác suất của một số sự kiện liên quan đến (một tập hợp con của) các biến khác có điều kiện dựa trên bằng chứng chúng ta có.

Giả sử chúng ta bắt đầu từ một BN rời rạc với DAG được sử dụng ở trên.



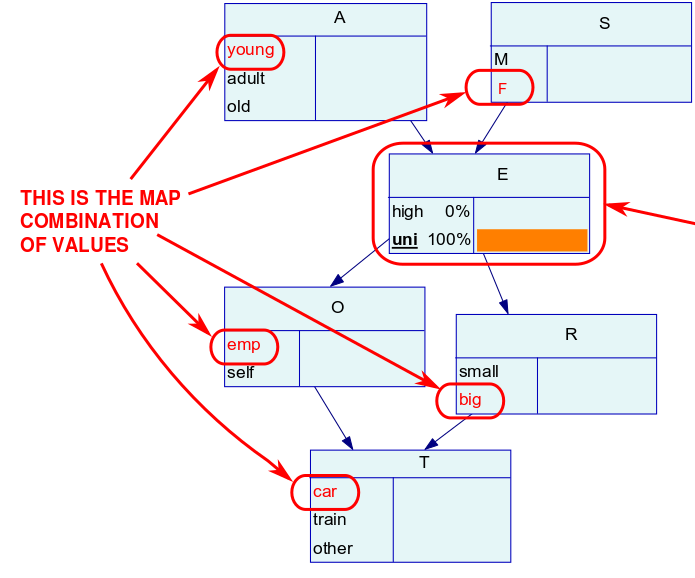
1. Minh họa inference suy luận xác suất Bayesian

Về mặt đồ họa, một truy vấn CP trông như thế này:



1. Minh họa inference suy luận xác suất Bayesian

Mặt khác, mục tiêu của truy vấn MAP là tìm tổ hợp các giá trị cho (một tập hợp con của) các biến trong mạng có xác suất cao nhất được đưa ra một số bằng chứng. Nếu bằng chứng là một cá thể mới được quan sát một phần, thì việc thực hiện truy vấn MAP giống như một bài tập dự đoán cổ điển.



1. Minh họa inference suy luận xác suất Bayesian

2.2 Phân tích đề tài

Sau khi đã tìm hiểu về kiến thức mô hình đồ thị xác suất, ta thực hiện một ứng dụng nhỏ để áp dụng thư viện vào thực tiễn. Dùng mô hình Bayesian để thu dự đoán khả năng nhiễm dựa trên các triệu chứng của người bệnh.

Dữ liệu được thực nghiệm lấy từ nguồn: <https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=dataset.csv>.

Dữ liệu chứa 41 loại bệnh. Mỗi người bị bệnh sẽ đi kèm với 3 triệu chứng (symptom) liên quan (Symptom\_1, Symptom\_2, Symptom\_3). Symptom\_1 có 34 loại triệu chứng khác nhau. Symptomm\_2 có 48 loại triệu chứng khác nhau. Symptom\_3 có 54 triệu chứng liên quan.

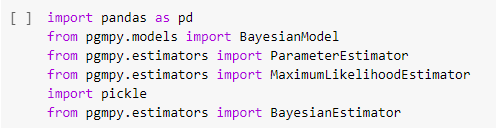
Ta sẽ sử dụng pgmpy và bayesian network để xây dựng bộ giá trị xác suất của mối quan hệ giữa 3 symptom với loại bệnh.

Từ đó ta sẽ sử dụng bộ giá trị này để biết được xác suất người đó sẽ mắc từng loại bệnh là bao nhiêu khi người đó có 3 triệu chứng tương ứng. Từ đó ta sẽ chọn ra giá trị xác suất cao tương ứng với khả năng cao người đó sẽ bị bệnh đó khi mắc 3 triệu chứng tương ứng.

* 1. Xây dựng chương trình

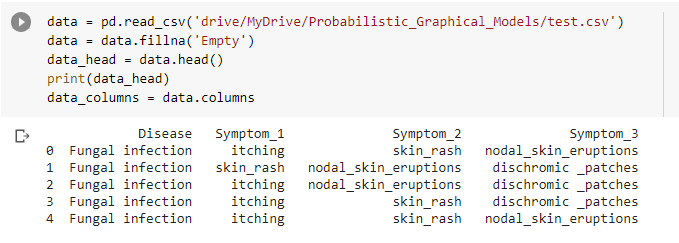
Ta sẽ tiến hành xây dựng chương trình cho ứng dụng.

Đầu tiên ta sẽ chuẩn bị các thư viện.



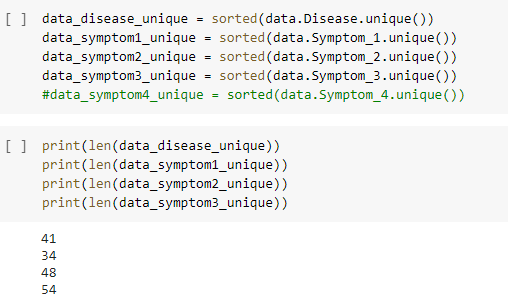
1. Chuẩn bị thư viện cho chương trình

Tiếp theo ta sẽ nạp dữ liệu vào chương trình. Đồng thời ta sẽ thay các giá trị rỗng bằng “Empty”.



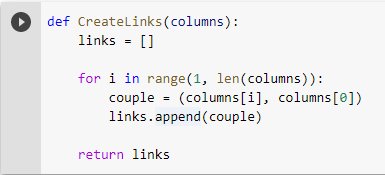
1. Tiến hành nạp dữ liệu vào chương trình

Ta sẽ tiến hành lọc ra các giá trị riêng biệt cho từng loại bệnh, symptom trong dữ liệu.



1. Lọc ra các giá trị riêng biệt cho từng loại bệnh, symptom

Tiếp theo ta sẽ xây dựng hàm tạo các cặp cạnh cho mô hình.



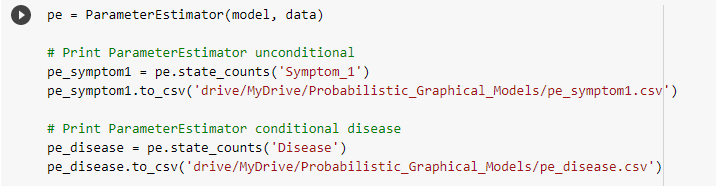
1. Xây dựng hàm tạo các cặp cạnh cho mô hình

Từ các cặp cạnh này ta sẽ tiến hành xây dựng mô hình bayesian network.



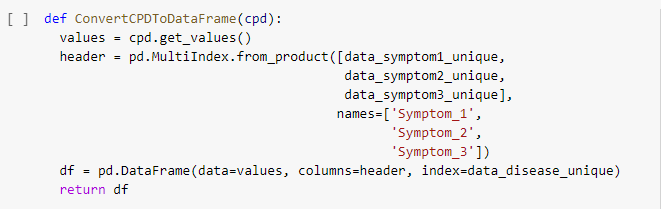
1. Tiến hành xây dựng mô hình

Tiếp theo ta sẽ tiến hành thống kê cho từng loại symptom, và số lần xuất hiện bệnh ứng với các symptom.



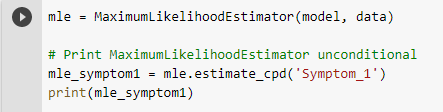
1. Tiến hành thống kê symptom và disease

Do dữ liệu quá lớn và IDE không hỗ trợ việc in kết quả khi dữ liệu lớn nên ta sẽ tiến hành xây dựng hàm để chuyển đổi từ loại của pgmpy sang dataframe của pandas để xuất ra csv để ta có thể xem được.



1. Xây dựng hàm chuyển đổi sang dataframe

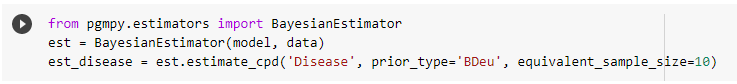
Cuối cùng ta sẽ tiến hành tính toán các giá trị xác suất bằng MaximumLikelihooddEstimator và BayesianEstimator.



1. Tính xác suất cho Symptom\_1 bằng MaximumLikelihooddEstimator

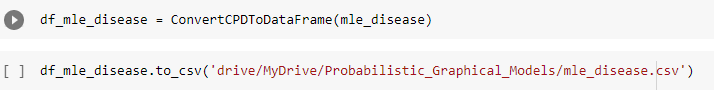


1. Tính xác suất cho Disease bằng MaximumLikelihooddEstimator

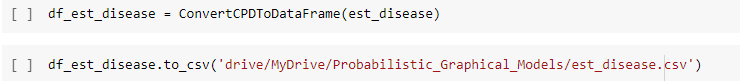


1. Tính xác suất cho Disease bằng BayesianEstimator

Cuối cùng ta sẽ xuất kết quả ra file csv.



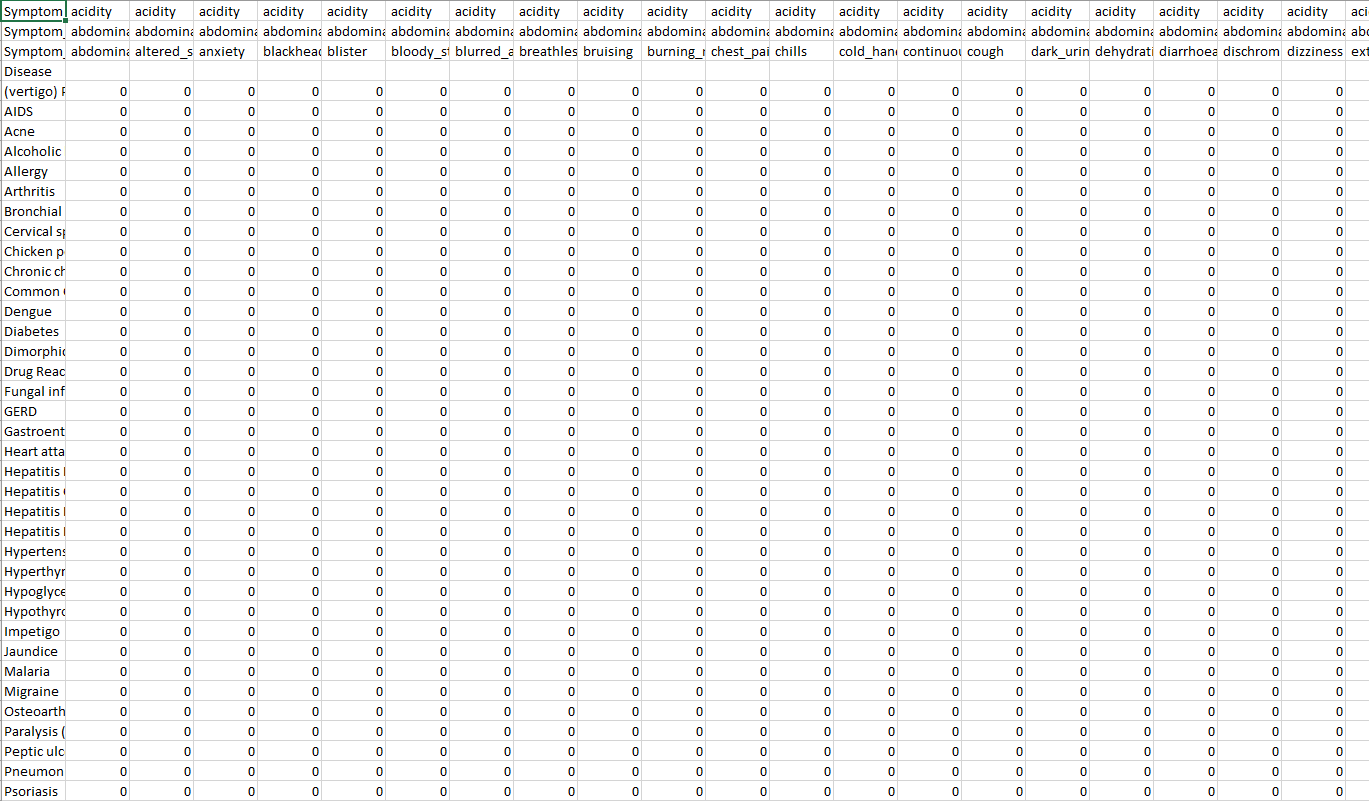
1. Xuất kết quả tính xác suất của Disease bằng MaximumLikelihooddEstimator ra file CSV



1. Xuất kết quả tính xác suất của Disease bằng BayesianEstimator ra file CSV

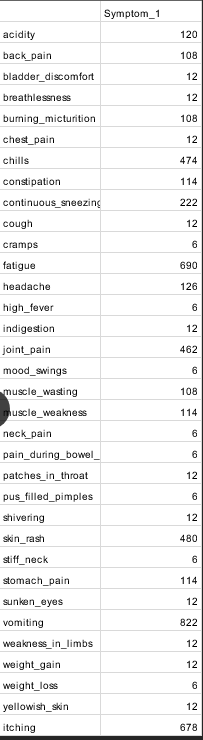
2.4 Kết quả học mô hình

Ta có một số bản khác nhau. Bảng thứ nhất thống kê tất các bệnh khác nhau. Với từng bệnh ta có một nhóm 3 các triệu chứng khác nhau



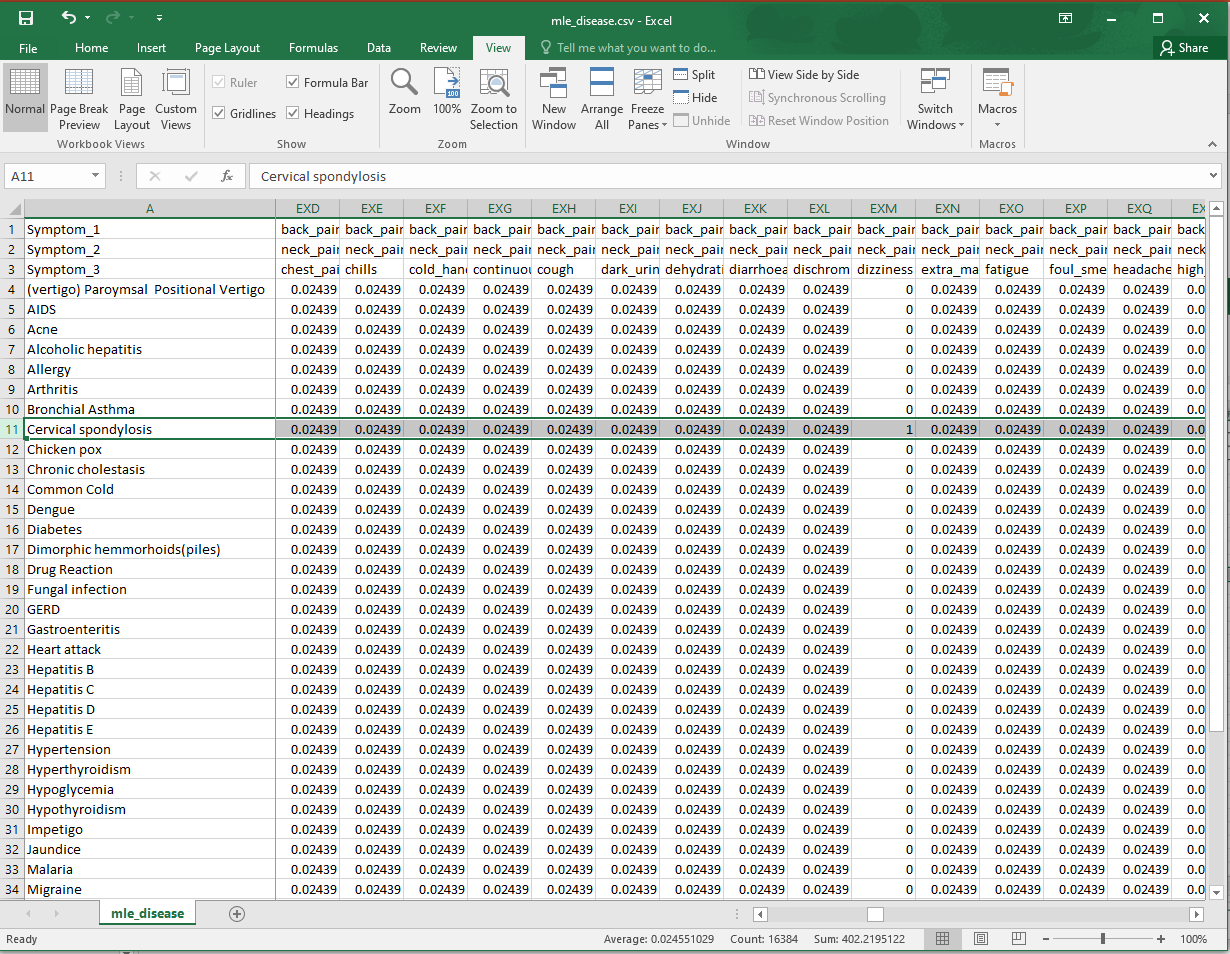
1. Thống kê các triệu chứng cho các bệnh khác nhau

Bảng tiếp theo ghi nhận có bao nhiêu triệu chứng đã được liệt kê trong nhóm các triệu chứng có thể xảy ra trong từng dịch bệnh.



1. Thống kê số lần xuất hiện các triệu chứng của Symptom\_1

Tiếp theo ta sẽ cho chương trình tính các giá trị xác suất quan hệ giữa 3 symptom với từng loại bệnh.



1. Bảng thống kê xác suất giữa symptom và disease

Từ hình 2.7 ta có thể thấy rằng ứng với 1 người có các triệu chứng là “back\_pain”, “neck\_pain”, “dizziness” sẽ có khả năng mắc bệnh “AIDS” là 0. Tuy nhiên khả năng mắc bệnh “Cervical spondylosis” là 1. Từ đó ta có thể suy luận rằng khi 1 người bị có các triệu chứng “back\_pain”, “neck\_pain”, “dizziness” sẽ có khả năng cao mắc bệnh “Cervical spondylosis”.

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

**Tiếng Anh**

1. <https://github.com/pgmpy/pgmpy/blob/dev/examples/Creating%20a%20Discrete%20Bayesian%20Network.ipynb>
2. [https://www.bnlearn.com/bnrepository/discrete-small.html#asia](https://www.bnlearn.com/bnrepository/discrete-small.html%23asia)
3. <https://github.com/pgmpy/pgmpy/blob/dev/examples/Inference%20in%20Discrete%20Bayesian%20Networks.ipynb>
4. [https://www.bnlearn.com/bnrepository/discrete-medium.html#alarm](https://www.bnlearn.com/bnrepository/discrete-medium.html%23alarm)
5. <https://github.com/pgmpy/pgmpy/blob/dev/examples/Learning%20Parameters%20in%20Discrete%20Bayesian%20Networks.ipynb>
6. <https://github.com/pgmpy/pgmpy/blob/dev/examples/Structure%20Learning%20in%20Bayesian%20Networks.ipynb>
7. <https://medium.com/@Dezhic/understanding-probabilistic-graphical-models-658b6fa40184>
8. <https://medium.com/@neerajsharma_28983/intuitive-guide-to-probability-graphical-models-be81150da7a>
9. <https://blog.statsbot.co/probabilistic-graphical-models-tutorial-and-solutions-e4f1d72af189>
10. <https://blog.statsbot.co/probabilistic-graphical-models-tutorial-d855ba0107d1>