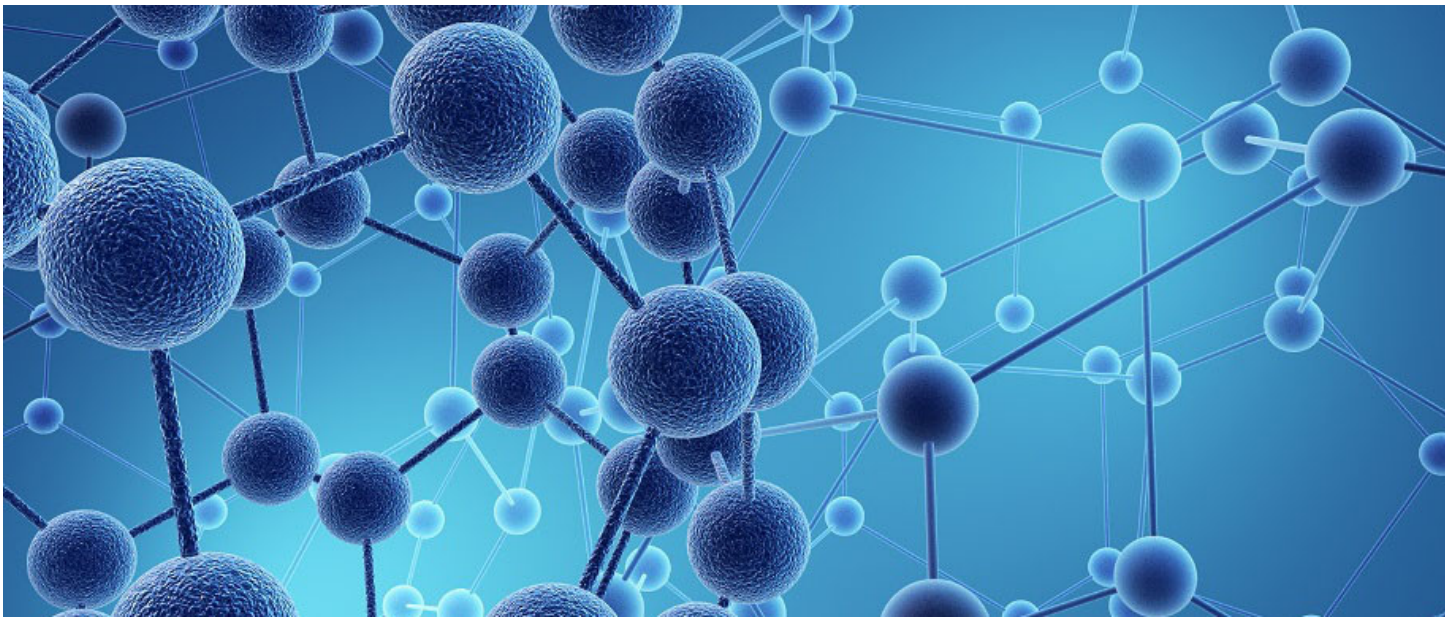
Data Science - Inferência em Redes Bayesianas Aplicado a Análise de Fatores de Risco em Trombose Coronária | LinkedIn



Data Science - Inferência em Redes Bayesianas Aplicado a Análise de Fatores de Risco em Trombose Coronária

Publicado em 2 de maio de 2017 Editar artigo | Visualizar estatísticas

Robson Fernandes
Cientista de Dados | Analista de Inteligência Artificial | Professor Universitário
4 artigos

282 0 2

Continuando nossos estudos sobre Data Science. Neste artigo abordarei uma introdução as Redes Bayesianas e Inferência em Redes Bayesianas aplicado a análise de fatores de risco em trombose coronária.

Ao longo deste artigo, iremos analisar nosso conjunto de dados, etapas de criação, aprendizagem da estrutura, treinamento e inferência a partir de Redes Bayesianas.

Os estudos aqui realizados tomam como base o excelente livro "Bayesian Networks in R - with Applications in Systems Biology" escrito pelos autores, Ph.D.:

- *Radhakrishnan Nagarajan*
- *Marco Scutari*
- *Sophie Lèbre*

Para facilitar o desenvolvimento do algoritmo em R, será utilizado o pacote "bnlearn", desenvolvimento pelo **Ph.D. Marco Scutari**, autor do livro.

Introdução

As **redes bayesianas** foram desenvolvidas no início dos anos 1980 para facilitar a tarefa de predição e “abdução” em sistemas de inteligência artificial.

As Redes Bayesianas, também conhecidas como redes de opinião, redes causais e gráficos de dependência probabilística, são modelos gráficos para raciocínio (conclusões) baseados em incerteza, onde os nós representam as variáveis (discretas ou contínuas), e os arcos representam conexões diretas entre eles.

Tal representação é comumente chamada de grafo, sendo este um elemento fundamental da rede.

O estudo dos grafos é realizado pelo ramo da matemática denominado Teoria de Grafos e diz respeito ao estudo das relações de seus elementos, os quais são comumente chamados de nós e arcos. Os nós são elementos principais os quais representam as variáveis aleatórias consideradas no problema e são representados por círculos. Os arcos são setas que representam a relação de direta dependência entre um nó e outro, ou seja, representa a dependência probabilística direta entre duas variáveis.

"A principal vantagem de raciocínio probabilístico sobre raciocínio lógico é fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará" - Russel

Conjunto de Dados

O conjunto de dados possui prováveis fatores de risco para trombose coronária, compreendendo dados de 1841 homens.

O conjunto de dados *coronary* contém 6 variáveis:

- **Smoking**(*Tabagismo*): um fator de dois níveis com níveis "no" e "yes".
- **M. Work**(*Trabalho mental extenuante*): um fator de dois níveis com níveis "no" e "yes".
- **P. Work**(*Trabalho físico extenuante*): um fator de dois níveis com níveis "no" e "yes".
- **Pressure**(*Pressão arterial sistólica*): um fator de dois níveis com níveis <140 e >140.
- **Proteins**(*Proporção de beta e alfa lipoproteínas*): um fator de dois níveis com níveis <3 e >3.
- **Family**(*Anamnese familiar de doença coronária*): um fator de dois níveis com níveis "neg" e "pos".

Fonte: Reinis Z, Pokorny J, Basika V, Tiserova J, Gorican K, Horakova D, Stuchlikova E, Havranek T, Hrabovsky F (1981). "Prognostic Significance of the Risk Profile in the Prevention of Coronary Heart Disease". Bratisl Lek Listy, 76, 137-150. Published on Bratislava Medical Journal, in Czech.

Estrutura de Aprendizagem

Existem vários algoritmos de aprendizagem de estruturas em Redes Bayesianas e alguns estão disponíveis no pacote **bnlearn.** Neste artigo vamos utilizar o algoritmo Hill Climbing.

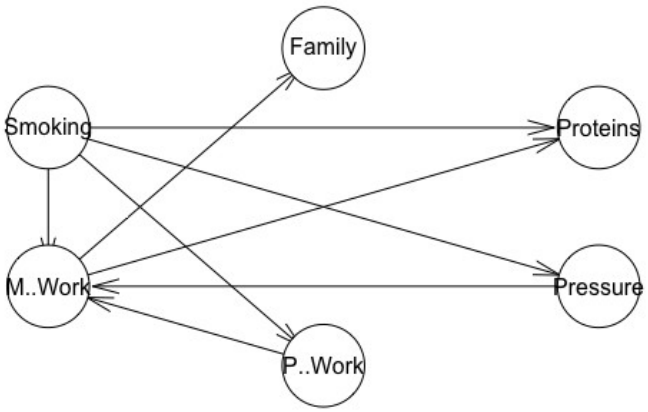
O código abaixo irá coletar a base de dados Coronary disponível dentro do pacote **bnlearn** e fará o processo de aprendizado da estrutura da Rede Bayesiana utilizando o algoritmo Hill Climbing, posteriormente, fará a plotagem gráfica da Rede.

```
require("bnlearn")

#Coleta da Base de Dados Coronary
coronaryDataFrame <- data.frame(coronary)

#Aprendizagem da rede bayesiana usando algoritmo Hill-Climbing (HC)
res <- hc(coronaryDataFrame)
#Plot da Rede
plot(res)
```

Na figura abaixo temos a representação gráfica da Rede Bayesiana gerada a partir do código em R. Podemos observar que o algoritmo descobriu as dependências condicionais entre as variáveis do conjunto de dados.

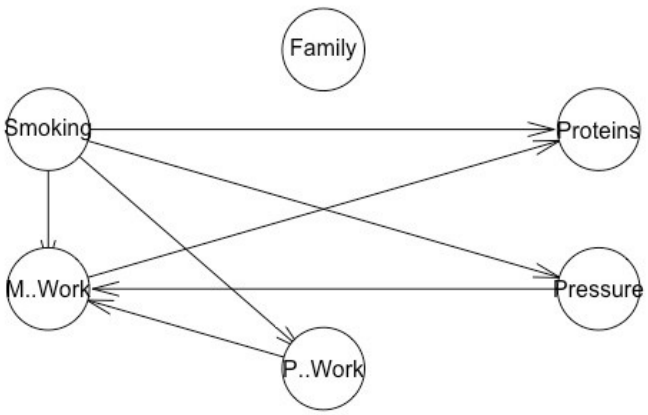


A causalidade entre alguns nós é intuitiva. Por exemplo, o nó $Pressure$*(Pressão) arterial sistólica,* é influenciado pelo nó $Smoking$*(Tabagismo).*

No entanto, algumas relações extraídas do conjunto de dados não parecem estar corretas. Por exemplo, não faz muito sentido o nó $Family$ (*Anamnese familiar de doença coronária), ser influenciado pelo nó* $M.\ Work$ (*Trabalho mental extenuante*).

Neste caso, vamos modificar a estrutura gerado removendo o link entre os nós $M.\ Work$ e $Family$.

```
#Remover o Link entre nós "M..Work", "Family"
res <- drop.arc(res, "M..Work", "Family")
#Plot da Rede
plot(res)
```



Treinamento

Depois de aprender a estrutura da Rede Bayesiana, precisamos descobrir as tabelas de probabilidade condicional (CPTs) em cada nó. A função *bn.fit* executa o algoritmo EM (Expectation–Maximization) para aprender CPTs para diferentes nós no gráfico acima.

Abaixo temos o código em R para realizar o ajuste da Rede Bayesiana utilizando a função bn.fit.

```
16 #Rede Bayesiana Ajustada
17 bnAjustado <- bn.fit(res, data = coronaryDataFrame)
```

Vamos analisar a tabela de probabilidade condicional gerada para o nó *Pressure* (*Pressão).* Para isto, execute o código abaixo.

```
19 #Tabela de Probabilidade Condicional - Pressure
20 print(bnAjustado$Pressure)
```

Como saída temos a Tabela de Probabilidade Condicional abaixo. Observem que o nó *Pressure* (*Pressão)* é condicionada a variável *Smoking(Tabagismo),* conforme a representação gráfica da Rede Bayesiana.

```
  Parameters of node Pressure (multinomial distribution)

Conditional probability table:

        Smoking
Pressure        no       yes
    <140 0.5359001 0.6125000
    >140 0.4640999 0.3875000
```

Inferência

Uma vez que realizados todos os processos para montagem, treinamento e ajustes na Rede Bayesiana, podemos agora inferir a partir da rede, ou seja, extrair conhecimento da rede.

No contexto de Redes Bayesianas, o termo “inferência”, também conhecido como atualização de crença (*belief updating*), é comumente utilizado para referenciar a atualização de probabilidades por toda a estrutura da rede dada um conjunto de evidências. Ou seja, segundo Korb e Nicholson (2004), trata-se de um mecanismo para cálculo da

distribuição *posteriori* de probabilidade para um conjunto de variáveis, dado um conjunto de evidências, ou seja, variáveis aleatórias com valores instanciados.

O processo de inferência podem ser realizado sobre as Redes Bayesianas, em quatro maneiras distintas:

1. **Diagnósticos**: partindo dos efeitos para as causas;
2. **Causa**: partindo das causas para os efeitos;
3. **Intercausal**: entre causas de um efeito comum;
4. **Mistas**: combinação de dois ou mais tipos descritos acima.

Vamos analisar na prática o seguinte questionamento, para extrair conhecimento (inferir) a partir da nossa Rede Bayesiana.

Qual é a chance de que um não fumante com pressão maior que 140 ter um nível de Proteínas inferior a 3

Para isto, vamos considerar as evidências e extrair a probabilidade do evento ocorrer, conforme o código em R abaixo, utilizando a função *cpquery*.

```
22 #Inferência em Redes Bayesianas
23 cpquery(bnAjustado,
24         event = (Proteins=="<3"),
25         evidence = ( Smoking=="no" & Pressure==">140" ) )
```

Ao executar o código teremos que

A chance de que um não fumante com pressão maior que 140 ter um nível de Proteínas inferior a 3 é de aproximadamente 62%

Download

Link para download do algoritmo em R.