Redes Bayesianas

André Ricardo Gonçalves
andreric [at] dca.fee.unicamp.br
www.dca.fee.unicamp.br/~andreric

Sumário

1 Redes Bayesianas

Em muitos problemas reais não há informações completas sobre o ambiente, seja por falha na coleta dos dados, imprecisão do aparelho de coleta ou até mesmo sendo a informação de impossível obtenção. Nestes casos técnicas que trabalham com o raciocínio probabilístico podem ser interessantes.

Métodos de raciocínio probabilístico podem trabalhar bem em ambientes onde existem informações parciais (incompletas) ou informações aproximadas (não exatas), ou seja, tais métodos podem ser aplicados sobre incertezas. Em ambientes de incerteza é possível utilizar-se de ferramentas como a Teoria da Probabilidade com enfoque *Bayesiano*, que considera a probabilidade como o grau de certeza da ocorrência de um evento.

Estes modelos ainda podem ser estendidos a casos onde um banco de exemplos está disponível, e até mesmo onde há falta de informação nos bancos de exemplos, nestes casos os modelos estimarão tais informações, por meio de um processo de *imputação*.

1.1 Cálculo de Probabilidades

A Probabilidade é um campo da matemática que estuda e analisa a ocorrência de fenômenos aleatórios. Fenômenos aleatórios são experimentos repetidos sob as mesmas condições produzem resultados que não se pode prever com certeza (MORGADO et al., 2001). Outros conceitos importantes dentro da probabilidade são definidos a seguir.

Definição 1.1.1 *Espaço amostral é o conjunto de todos os resultados possíveis de um experimento aleatório.*

Definição 1.1.2 *Evento é qualquer subconjunto do espaço amostral.*

Meyer (2000) apresenta uma definição formal do conceito de probabilidade.

Definição 1.1.3 *Dado um experimento ϵ e S o espaço amostral associado a ϵ. A cada evento A associaremos um número real representado por $P(A)$, denominado de probabilidade de A e que satisfaça as seguintes propriedades:*

1. $0 \leq P(A) \leq 1$.

2. $P(S) = 1$.

3. *Se A e B forem mutuamente exclusivos , então* $P(A \vee B) = P(A) + P(B)$

Da propriedade (1) é possível identificar que os valores das probabilidades estarão no intervalo [0,1]. Pela propriedade (2) concluímos que a soma de todos os eventos do espaço amostral é igual a 1, e a propriedade (3) diz que, sendo dois eventos mutuamente exclusivos, ou seja, se um está presente então o outro estará ausente, a união das probabilidades é igual à soma das mesmas isoladas. Esta probabilidade é também chamada de **probabilidade incondicional**, pois não depende de nenhuma condição anterior.

1.1.1 Probabilidade Condicional e Independência Condicional

Ao contrário da probabilidade incondicional, a probabilidade condicional depende de uma condição anterior. Representada por $P(B|A)$, a probabilidade condicional pode ser interpretada como: "A probabilidade da ocorrência do evento B, dada a ocorrência do evento A". Se calcularmos $P(B|A)$, estaremos essencialmente calculando $P(B)$ em relação ao espaço amostral reduzido de A (MEYER, 2000).

A definição formal da probabilidade condicional, como observa (HAZZAN; IEZZI, 2004), utiliza-se do conceito de freqüência relativa. Seja um experimento repetido *n* vezes e seja n_A, n_B e $n_{A \wedge B}$, o número de vezes que ocorreram os eventos A, B e A $\wedge$ B. Sendo assim o termo $n_{A \wedge B}/n_A$ representa a freqüência relativa de B condicionada a ocorrência do evento A.

A partir disso é possível afirmar que

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} \tag{1.1}$$

desde que $P(A) > 0$. Sendo assim existem duas maneiras de calcular a probabilidade condicionada $P(B|A)$ (MEYER, 2000):

1. Diretamente, considerando a probabilidade de B em relação ao espaço amostral reduzido de A;

2. Aplicando a definição acima, onde $P(A \wedge B)$ e $P(A)$ são calculados em relação ao espaço amostral original.

Uma importante conseqüênciada probabilidade condicional é Teorema da multiplicação (HAZZAN; IEZZI, 2004):

Teorema 1.1.1 *(Teorema da Multiplicação) A probabilidade de dois eventos ocorrerem simultaneamente é o produto da probabilidade de um deles pela probabilidade do outro dado o primeiro.*

O teorema da multiplicação é representado pela Eq. (1.2)

$$P(A \wedge B) = P(B|A) \cdot P(A) \tag{1.2}$$

Outro conceito importante é a *independência de eventos*. Um evento A independe de B se:

$$P(A|B) = P(A) \tag{1.3}$$

ou seja, A independe de B se a ocorrência de B não afeta a probabilidade de A. Observando o evento B é possível concluir que B também independe de A, pois

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(B) \cdot P(A|B)}{P(A)} = \frac{P(B) \cdot P(A)}{P(A)} = P(B) \tag{1.4}$$

utilizando o teorema da multiplicação é possível identificar ainda

$$P(A \wedge B) = P(A) \cdot P(B|A) = P(A) \cdot P(B) \tag{1.5}$$

A partir disso é possível definir formalmente a independência de dois eventos:

Definição 1.1.4 *Dois eventos A e B são chamados de independentes se:*

$$P(A \wedge B) = P(A) \cdot P(B)$$

Outro conceito importante é a independência condicional, uma extensão da independência entre dois eventos. A independência condicional pode ser definida como:

Definição 1.1.5 *Um evento X é condicionalmente independente de Y dado Z se a distribuição de probabilidade que rege Z é independente de Y dado o valor de Z, que pode ser representado*

$$P(X|Y \wedge Z) = P(X|Z)$$

1.1.2 Teorema de Bayes

Considere uma partição de um espaço amostral S um conjunto de eventos A_1,A_2,A_3, ..., A_n, os eventos A_i são mutuamente exclusivos e sua união é S. Agora dado outro evento B com probabilidade $P(B) > 0$ então:

$$B = S \wedge B = (A_1 \vee A_2 \vee .... \vee A_n) \wedge B$$

onde $A_i \wedge B$ são mutuamente exclusivos. Conseqüentemente a probabilidade da ocorrência de B é dada por:

$$P(B) = P(A_1 \wedge B) + P(A_2 \wedge B) + ... + P(A_n \wedge B) = \sum_i P(A_i \wedge B)$$

Utilizando-se do teorema da multiplicação 1.1.1, temos que:

$$P(B) = \Sigma_i P(A_i \wedge B) = \Sigma_i P(B|A_i) \cdot P(A_i) \tag{1.6}$$

Além do mais é possível notar que

$$P(A_i \wedge B) = P(B|A_i) \cdot P(A_i) = P(A_i) \cdot P(B)$$

resolvendo em ordem a $P(A_i|B)$, chega-se o Teorema de Bayes (PAULINO; TURKMAN; MURTEIRA, 2003)

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)} \tag{1.7}$$

A definição formal do Teorema de Bayes apresentada por (LIPSCHUTZ, 1993) é mostrada pelo teorema 1.1.2.

Teorema 1.1.2 *Suponha A_1,A_2,A_3, ..., A_n ser uma partição de S e B, um evento qualquer. Então para qualquer i*

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)} \tag{1.8}$$

Uma interpretação do teorema de Bayes consiste em considerar os eventos A_i como "causas" do evento B, sendo atribuído probabilidades deste evento atuar na ocorrência de B. Esta probabilidade é calculada antes da realização do experimento, sendo designada como a probabilidade *a priori* de A_i. Após a realização do experimento, é conhecido que o evento B ocorreu, então a probabilidade *a priori* é revista por meio da fórmula de Bayes e então passa a atribuir aos eventos A_i, *i*= 1,2,...n as probabilidades *a posteriori* $P(A_i|B)$, *i = 1,2,...n* (CRAMÉR, 1955) (PAULINO; TURKMAN; MURTEIRA, 2003).

Como observado por (PAULINO; TURKMAN; MURTEIRA, 2003) o Teorema de Bayes é para muitos, um dos poucos resultados da matemática que se propõe a caracterizar a aprendizagem com a experiência, ou seja, a modificação de atitude inicial em relação as "causas" depois de ter a informação adicional de que certo acontecimento ou acontecimentos se realizaram.

1.1.3 Variáveis aleatórias e Distribuição de Probabilidade Conjunta

De acordo com (MEYER, 2000) uma variável aleatória é uma função que associa a cada elemento um valor real. O conjunto de valores que uma variável aleatória X

pode assumir é chamado de espaço de X. Uma variável aleatória é dita ser discreta se o espaço é finito e contável (NEAPOLITAN, 2003).

De acordo com (CHARNIAK, 1991), a distribuição de probabilidade conjunta (*joint probability distribution*) de um conjunto de variáveis aleatórias X_1, X_2, .., X_n é definida como $P(X_1 \wedge X_2 \wedge .. \wedge X_n)$, para todos os valores de X_1, X_2, .., X_n. A distribuição conjunta de um grupo de variáveis aleatórias fornece toda a informação sobre a distribuição.

A distribuição de probabilidade pode ser representada em uma tabela, como mostra o exemplo abaixo.

Exemplo 1.1.1 *Para um conjunto de variáveis aleatórias binárias {a,b} a distribuição de probabilidade conjunta pode ser representada como mostra a tabela 1.*

	a	$\neg$ *a*
b	*0.04*	*0.06*
$\neg$ *b*	*0.01*	*0.89*

Tabela 1: Distribuição de probabilidade conjunta de duas variáveis binárias

Para *n* variáveis booleanas a distribuição conjunta terá 2^n valores. De qualquer forma a soma de toda a distribuição conjunta é igual a 1, pois a probabilidade de todas as possíveis respostas deve ser 1 (CHARNIAK, 1991).

1.2 Inferência Bayesiana

O processo de obtenção da probabilidade *a posteriori* a partir da probabilidade *a priori* é chamado de Inferência Bayesiana (NEAPOLITAN, 2003).

As inferências Bayesianas sobre uma variável aleatória Y, são baseadas em probabilidades subjetivas ou credibilidades *a posteriori* associadas aos valores do espaço de Y e condicionadas pelo valor particular de um evento X (PAULINO; TURKMAN; MURTEIRA, 2003). Probabilidades subjetivas diferentemente das probabilidades relativas não podem ser obtidas por simples repetição de um experimento, ela é a medida do nível de "confiança" que se tem sobre a verdade de uma determinada proposição. Por exemplo, a probabilidade de uma pessoa ter uma doença A não pode ser obtida como em um experimento de lançamento de dados.

Neapolitan (2003) apresenta as etapas realizadas no processo de modelagem de uma situação a fim de obter informações adicionais sobre ela e para isso utiliza-se da inferência bayesiana:

1. Identificação das variáveis aleatórias do modelo, que representaram as características ou causas e efeitos dentro da situação;

2. Determinação do conjunto mutuamente exclusivo de valores para cada uma das variáveis. Esses valores podem ser obtidos considerando os diferentes estados que a característica pode estar;

3. Decidir as probabilidades de uma variável aleatória ter seu valor, ou seja, calcular a distribuição das probabilidades, o que nem sempre pode ser obtido diretamente;

4. Utilizando dos relacionamentos entre variáveis, identificando as dependências e posteriormente calculando as probabilidades condicionais é possível a obtenção da distribuição das probabilidades.

Neapolitan (2003) observa ainda que a especificação das variáveis e seus valores devem ser precisos o suficiente para satisfazer os requerimentos da situação modelada. Com a situação modelada e com as probabilidades calculadas é possível inferir qualquer indagação sobre a situação.

1.3 Redes Bayesianas

A aplicação da inferência bayesiana sobre um número pequeno de variáveis relacionadas é um processo relativamente simples. Mas em situações reais onde um grande número de variáveis e estados é encontrado a inferência pode não ser trivial.

Uma rede Bayesiana, também chamada de rede de crença, rede probabilística ou rede causal, pode ser vista como um modelo que utiliza teoria dos grafos, condições de Markov e distribuição de probabilidades para representar uma situação, suas variáveis e estados e a partir disto realizar inferências.

Quando uma situação possui um grande número de características (variáveis) surgem alguns problemas, como relatado por (NEAPOLITAN, 2003), considerando que a distribuição de probabilidade conjunta não é prontamente acessível o número exponencial de cálculos necessário na aplicação do teorema de Bayes 1.1.2 torna a inferência impraticável.

Mitchell (1997) define que as redes Bayesianas descrevem a distribuição de probabilidade sobre um conjunto de variáveis. Já (MARQUES; DUTRA, 2008) afirma que matematicamente uma rede bayesiana é uma representação compacta de uma tabela de probabilidades conjunta do universo do problema e que pelo ponto de vista de um especialista esta técnica constitui em um modelo gráfico que representa de forma simples as relações de causalidade das variáveis de um sistema.

Em redes Bayesianas a representação das variáveis e relações é feita utilizando Teoria dos Grafos. As variáveis são os nós e os arcos identificam as relações entre as variáveis, formando um grafo dirigido e sem ciclos, **DAG!** (**DAG!**), como mostra a figura 1. Neste exemplo a variável Z é condicionada as variáveis X e Y.

Uma Rede Bayesiana consiste do seguinte (MARQUES; DUTRA, 2008):

- Um conjunto de variáveis e um conjunto de arcos ligando as variáveis;
- Cada variável possui um número limitado de estados mutuamente exclusivos;
- As variáveis e arcos formam um grafo dirigido e sem ciclos **DAG!**;
- Para cada variável A que possui como pais B_1, ..., B_n existe uma tabela de probabilidade condicional (TPC) $P(A|B_1 \wedge ... \wedge B_m)$.

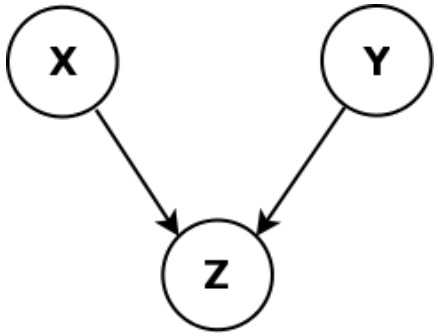

Figura 1: Grafo construído a partir de variáveis e suas relações

Caso a variável A for não possua um pai, a tabela de probabilidade é reduzida a probabilidade incondicional $P(A)$.

Uma rede Bayesiana é a representação correta de um domínio caso a **condição de Markov** seja satisfeita. A condição de Markov é definida por (NEAPOLITAN, 2003) como:

Definição 1.3.1 *(Condição de Markov) Suponha a distribuição de probabilidade conjunta das variáveis aleatórias em um conjunto de nós V em um* **DAG!** *$\mathbb{G} = (V, E)$. Então dizemos que (G, P) satisfazem a condição de Markov se cada variável $X \in V$, X é condicionalmente independente dos nós não descendentes dados seus pais.*

A condição de Markov afirma que as variáveis não-descendentes não fornecem informações adicionais sobre a variável em questão.

De acordo com (PEARL, 1988), considerando F_X e Pa_X o conjunto de filhos e dos pais do nó X respectivamente, e ainda Pa_{F_x} como o conjunto dos pais dos descendentes diretos de X. O conjunto de nós formados pela união destes três conjuntos é denominado de *Markov Blanket*. Os nós pertencentes ao *Markov Blanket* são os únicos nós da rede necessários para predizer o comportamento do nó.

De acordo com (MARQUES; DUTRA, 2008), uma vez definida topologia da rede (distribuição dos nós e os relacionamentos entre as variáveis), é preciso determinar as probabilidades dos nós que participam em dependências diretas e utilizar estas para computar as demais probabilidades desejadas.

O exemplo abaixo, extraído de (RUSSELL; NORVIG, 1995), mostra as etapas de identificação das características (variáveis), seus conjunto de valores e a construção topológica da rede (mapa causal).

Exemplo 1.3.1 *"Um novo alarme contra assaltos é instalado, mesmo sendo muito confiável na detecção de assaltos ele pode disparar caso ocorra um terremoto. Os dois vizinhos João e Maria se disponibilizaram a telefonar caso o alarme dispare. João sempre liga quando ouve o alarme, entretanto algumas vezes ele confunde o alarme com o telefone e também liga nestes*

casos. Já a Maria gosta de ouvir música alta e às vezes não houve o alarme disparar, não ligando nestes casos.“

A modelagem do domínio pode ser representada como segue:

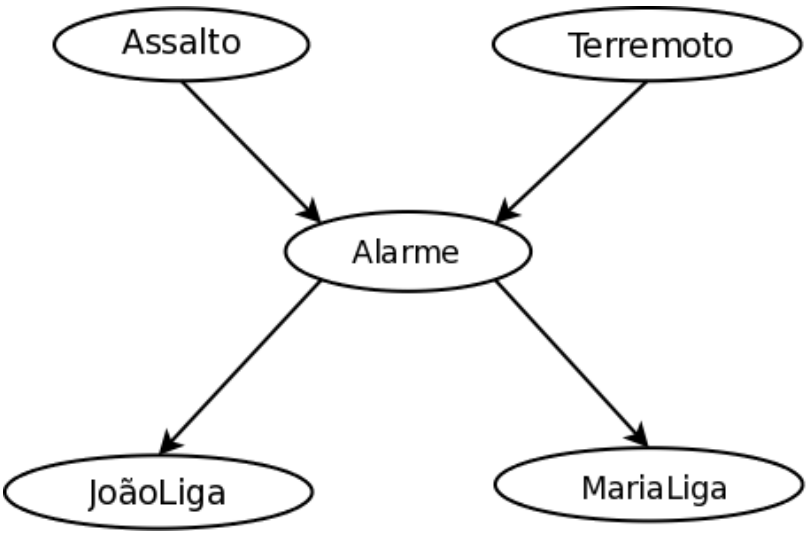

Figura 2: Representação de uma Rede Bayesiana do domínio

É possível notar que as condições da Maria estar ouvindo música e do telefone estar tocando, consequentemente confundindo João, não estão sendo expressas na representação. Essas condições estão implícitas, associados à incerteza relacionada pelos arcos Alarme → JoãoLig e Alarme → MariaLig, pois calcular estas probabilidades seria muito dispendioso ou até impossível. Sendo assim o sistema pode manipular um grande número de probabilidades, mesmo de forma aproximada (MARQUES; DUTRA, 2008).

Após a definição da topologia da rede é necessário calcular a tabela de probabilidade condicional, a qual expressará as probabilidades condicionais de cada variável (nó) dado seus pais (predecessores imediatos). A tabela 1.3 mostra a tabela da variável representada na rede pelo nó *Alarme*, dado seus pais *Assalto* e *Terremoto*.

Assalto	Terremoto	$P(Alarme\|Assalto, Terremoto)$	
		V	F
V	V	0.95	0.05
V	F	0.95	0.05
F	V	0.29	0.71
F	F	0.001	0.999

Tabela 2: Tabela de probabilidade condicional do nó *Alarme*

Os nós que não possuem pai *Assalto* e *Terremoto*, as probabilidades incondicionais são atribuídas por um especialista ou de modo freqüêncista, utilizando a freqüência relativa da ocorrência destes eventos. Para isso um banco de exemplos satisfatoriamente grande deve ser considerado, a fim de obter valores fidedignos da proporção.

Com as tabelas de probabilidade condicional de cada nó calculada, é possível obter a distribuição de probabilidade conjunta e conseqüentemente inferir qualquer evidência sobre o domínio. A figura 3 mostra as tabelas probabilidade condicional de cada nó da Rede Bayesiana da figura 2.

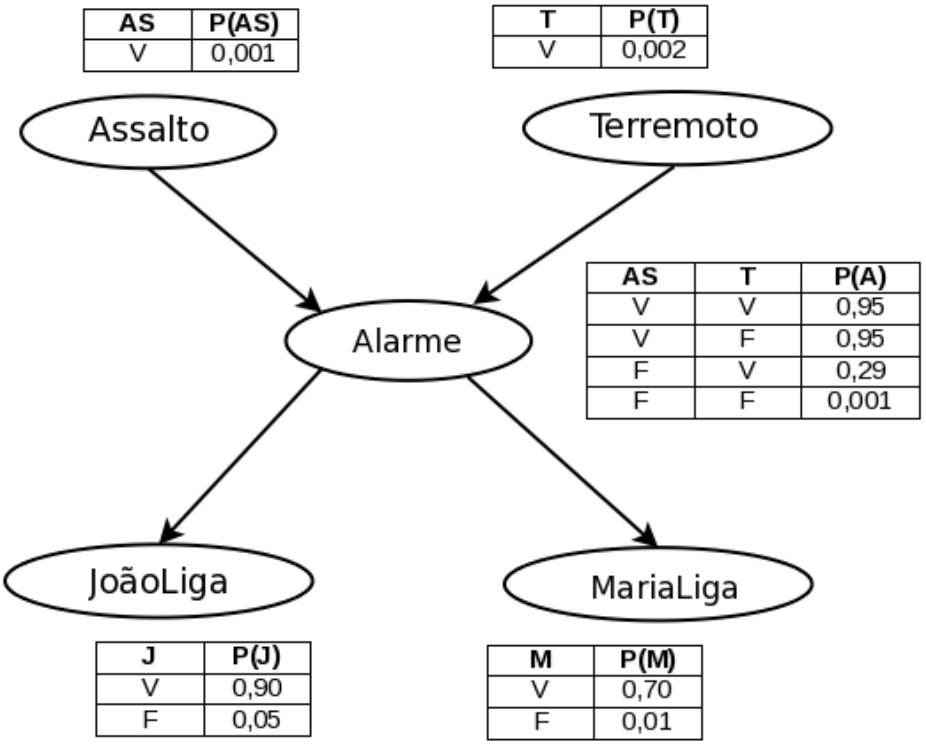

Figura 3: Representação de uma Rede Bayesiana do domínio

1.3.1 Cálculo da distribuição de probabilidade conjunta

Com as tabelas de probabilidade condicional calculadas podemos obter a distribuição de probabilidade conjunta de todo o domínio.

Sendo X_i um nó da rede e *pa(X_i)* representando os pais de X_i. Dessa maneira $X_1, X_2, ..., X_n$ identifica todos os nós do domínio e denotaremos por $P(X_1, X_2, ..., X_n)$ como a distribuição de probabilidade conjunta da rede.

O teorema a seguir define o cálculo da distribuição de probabilidade conjunta de todos os nós, como sendo o produto da probabilidade condicional de todos os nós dados seus pais.

Teorema 1.3.1 *Se uma rede bayesiana satisfaz a condição de Markov, então sua distribuição de probabilidade conjunta é igual ao produto das probabilidades condicionais de todos os nós dado os valores de seus pais.*

A prova do teorema 1.3.1 pode ser encontrada em (NEAPOLITAN, 2003). De uma maneira matemática podemos definir a distribuição de probabilidade conjunta como

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i)) \tag{1.9}$$

Com isso podemos concluir que as tabelas de probabilidade condicional constituem uma representação distribuída da tabela de probabilidade conjunta do domínio em questão.

Do exemplo 1.3.1, poderíamos querer obter a probabilidade do alarme disparar sem que um assalto e nem um terremoto tenha ocorrido, além de ambos, João e Maria, ligarem. Podemos representar esta indagação por:

$$\begin{aligned} &P(A \wedge \neg AS \wedge \neg T \wedge J \wedge M) \\ &\quad = P(J|A) \text{ x } P(M|A) \text{ x } P(A|\neg AS \wedge \neg T) \text{ x } P(\neg AS) \text{ x } P(\neg T) \\ &\quad = 0.9 \text{ x } 0.70 \text{ x } 0.001 \text{ x } 0.999 \text{ x } 0.998 = 0.00062 \end{aligned}$$

Como observado por (RUSSELL; NORVIG, 1995), o processo geral para construção de uma rede Bayesiana é dado pelo algoritmo 1..

Algoritmo 1: Algoritmo para construção de uma Rede Bayesiana

```
1 begin
2     Escolher um conjunto de variáveis relevantes X_i que descrevam o domínio;
3     Escolher uma ordem para as variáveis;
4     while Existir variáveis do
5         Selecione uma variável X_i e adicione um nó na rede;
6         Determine os nós pais pa(X_i), dentre os nós que já estejam na rede, de modo
          que a condição de Markov seja satisfeita;
7         Determine a tabela de probabilidade condicional para X_i;
8     end
9 end
```

A condição de que os novos nós devem ser conectados aos nós antigos, garantem que o grafo seja sempre acíclico.

1.3.2 Inferência em redes Bayesianas

Com a rede Bayesiana definida, pode-se extrair conhecimento nela representado através de um processo de inferência. De acordo com (HRUSCHKA JR., 2003) existem vários métodos para realização de inferência, dentre os métodos tradicionais destacam-se o de propagação em poliárvores (PEARL, 1988) e o de eliminação de variáveis (COZMAN, 2000) com suas variações.

Como destacado por (RUSSELL; NORVIG, 1995), inferências podem ser realizadas sobre 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.

O autor supracitado ainda afirma que as redes Bayesianas, podem ser utilizadas para outros fins, como:

- Tomar decisões baseadas em probabilidades;
- Decidir quais evidências adicionais devem ser observadas, a fim de obter total conhecimento do domínio;
- Realizar uma *análise sensitiva* para entender quais aspectos do modelo tem maior impacto sobre determinadas variáveis;
- Explicar os resultados de uma inferência probabilística ao usuário.

1.3.3 Aprendizagem Bayesiana

A aprendizagem Bayesiana pode ser visto como uma forma de obter a representação interna da rede que define um dado domínio de modo a facilitar a extração do conhecimento.

Dentro do processo de aprendizagem é necessário calcular as distribuições de probabilidade (parâmetros numéricos) e identificar a estrutura da rede, ou seja, identificar variáveis e as relações de interdependência dadas pelos arcos (HRUSCHKA JR., 2003).

O processo de aprendizagem é dividido em duas partes: aprendizagem da estrutura (e relações entre as variáveis); e a aprendizagem dos parâmetros numéricos (distribuição de probabilidade).

Ambas as partes, estrutura e parâmetros, podem ser aprendidas por meio de um especialista e indutivamente.

Por aprendizagem com especialista entende-se que o conhecimento será transmitido por meio de um especialista, que será responsável por definir e/ou supervisionar a construção da rede baseando-se em seu conhecimento. Já a aprendizagem indutiva utiliza-se de um banco de dados de exemplos, e partindo deste a rede é construída automaticamente.

Diversos algoritmos foram propostos na literatura de redes Bayesianas, com objetivo de encontrar a estrutura que represente fielmente o domínio modelado; e algoritmos que determinam as distribuições de probabilidade, considerando aprendizagem indutiva.

De acordo com (HRUSCHKA JR., 2003), o processo de obtenção dos parâmetros numéricos é geralmente mais simples do que a construção da estrutura da rede.

A aprendizagem dos parâmetros numéricos, considerando que a rede já está estruturada, pode ser estimados através das freqüências relativas, caso exista uma quantidade de dados significativa de uma amostra aleatória.

Para a aprendizagem de estrutura, (HRUSCHKA JR., 2003) observa que existem várias metodologias na literatura sendo que cada uma aplica-se melhor em um tipo de aplicação. Por serem bastante específicas não é possível definir qual é a melhor.

Dentre os métodos existentes destacam-se:

- Métodos de Verossimilhança Máxima;
- Métodos de Teste de Hipóteses;
- Métodos de Verossimilhança Extendidos;
- Métodos *"Minimum Information Complexity"*;
- Métodos *"Resampling"*;
- Métodos Bayesianos, destacando o clássico algoritmo $K2$ (COOPER; HERSKOVITS, 1992).

1.4 Classificador Naive Bayes

Uma rede bayesiana pode ser modelada como um classificador, calculando a probabilidade de $P(C|V)$, onde C representa a classe analisada e V o conjunto de variáveis que descrevem os padrões.

O classificador mais importante dentre os classificadores Bayesianos é o *Naive Bayes*, descrito em (DUDA; HART, 1973). É um modelo simples que se destaca pelos sucessos obtidos na aplicação em diversos problemas, mesmo comparado à classificadores mais complexos (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

Este modelo descreve um caso particular de uma rede Bayesiana, o qual considera que as variáveis do domínio são condicionalmente independentes, ou seja, uma característica não é relacionada com a outra. Em decorrência desta restrição utiliza-se o termo *"naive"*. A figura 4 mostra a estrutura da rede *Naive Bayes*, considerando sua restrição.

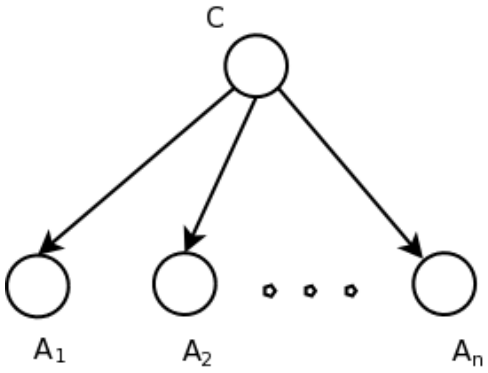

Figura 4: Estrutura de uma rede *Naive Bayes*

A classificação é então feita aplicando o teorema de Bayes para calcular a probabilidade de C dado uma particular instância de A_1, A_2, ..., A_n e então predizendo a classe com a maior probabilidade *a posteriori* (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). De outra maneira:

$$classificador(A_1, A_2, ..., A_n) = \arg\max_c P(c) \prod_i P(A_i|c) \quad (1.10)$$

O processo de aprendizagem do *Naive Bayes* é feito de maneira indutiva, apresentando um conjunto de dados de treinamento e calculando a probabilidade condicional de cada atributo A_i, dado a classe C (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

O algoritmo 2, baseado em (CARVALHO,), identifica as etapas do treinamento do *Naive Bayes*.

Algoritmo 2: Algoritmo de aprendizagem do *Naive Bayes*

```
Input: Exemplos para treinamento
begin
    for cada classe C_j do
        Obtenha probabilidade incondicional P(C_j);
        for cada atributo A_i de um exemplo do
            Obtenha a probabilidade estimada P(A_i|C_j);
        end
    end
end
```

Com a rede treinada é possível realizar a classificação de novos padrões, utilizando a definição 1.10.

A probabilidade incondicional das classes C_j pode ser obtida por meio do conhecimento de um especialista ou atribuindo probabilidades iguais para todas as classes.

Vários outros algoritmos foram propostos como melhorias do *Naive Bayes*, como o *Tree Augmented Naive Bayes (TAN)* apresentado por (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997), o *BN Augmented Naive Bayes (BAN)*, o *General Bayesian Network (GBN)* (CHENG; GREINER, 1999), entre outros.

1.5 Dificuldades na aplicação

Mitchell (1997) identifica algumas dificuldades práticas na aplicação de redes *Bayesianas*, como a necessidade de conhecimento inicial de muitas probabilidades incondicionais. Quando estas probabilidades não são conhecidas, elas são muitas vezes estimadas, com base em conhecimento de especialistas, dados disponíveis previamente e hipóteses sobre a forma das distribuições de probabilidades.

O autor supracitado ainda observa outro empecilho, o significativo custo computacional necessário para determinar a hipótese Bayesiana ótima em casos mais gerais, porém em casos mais restritos o custo pode ser reduzido.

1.6 Aplicações

Diversas aplicações em várias áreas do conhecimento obtiveram ótimos resultados comparados à outras técnicas. Dentre as áreas aplicadas destacam-se, diagnóstico médico ((HECKERMAN, 1990), (LONG; FRASER; NAIMI, 1997)), aprendizagem de mapas (BASYE; VITTER, 1997), interpretação de linguagem (GOLDMAN, 1990), visão (LEVITT; AGOSTA; BINFORD, 1990) entre outros.

Além das aplicações acima descritas, uma rede Bayesiana pode ser utilizada como um classificador, como o *Naive Bayes*. Este classificador vem sendo utilizado em várias áreas, mesmo sendo um modelo simples, ele tem obtido sucesso comparado à outros classificadores mais sofisticados. Áreas essas como classificação textual ((PENG; SCHUURMANS, 2003), (MCCALLUM; NIGAM, 1998)), filtros *anti-spam* (ANDROUTSOPOULOS et al., 2000) (uma aplicação particular das classificações textuais), identificação de genes (bioinformática) (YOUSEF et al., 2006), entre outros.

1.7 Conclusão

As redes Bayesianas utilizam dos conceitos de mapas causais, para modelar domínios. Mapas causais estes que descrevem as variáveis (nós) e as relações de causa e efeito entre elas, na forma de um grafo acíclico. A intensidade das relações é dada pelas tabelas de probabilidade condicional de cada variável, que quantifica as probabilidades de ocorrência de um evento dado seus pais.

O cálculo das probabilidades é obtido com a aplicação do teorema de Bayes, a partir das probabilidades *a priori*, adquiridas com o auxílio de um especialista ou através de um banco de dados.

Com isso podemos concluir que uma rede Bayesiana que represente corretamente um domínio, pode ser considerada um método bastante atrativo para armazenamento e extração de conhecimento. E ainda podemos destacar o não menos relevante método de classificação *Naive Bayes*, o qual foi provado por inúmeros trabalhos que mesmo possuindo fortes restrições, é incrivelmente eficiente.

Referências

ANDROUTSOPOULOS, I. et al. An evaluation of naive bayesian anti-spam filtering. In: *Workshop on Machine Learning in the New Information Age*. [s.n.], 2000. p. 9–17. Disponível em: <http://arxiv.org/abs/cs.CL/0006013>.

BASYE, T. D. K.; VITTER, J. S. Coping with uncertainty in map learning. *Machine Learning*, Springer Netherlands, v. 29, n. 1, October 1997.

CARVALHO, F. *Aprendizagem Bayesiana*. Apresentação de Slides. Acessado em: 05 de Outubro de 2008. Disponível em: <http://www.cin.ufpe.br/~compint/aulas-IAS/kdd-011-/Bayes.ppt>.

CHARNIAK, E. Bayesian networks without tears: making bayesian networks more accessible to the probabilistically unsophisticated. *AI Mag.*, American Association for Artificial Intelligence, Menlo Park, CA, USA, v. 12, n. 4, p. 50–63, 1991. ISSN 0738-4602. Disponível em: <http://portal.acm.org/citation.cfm?id=122716>.

CHENG, J.; GREINER, R. Comparing bayesian network classifiers. In: . Morgan Kaufmann Publishers, 1999. p. 101–108. Disponível em: <http://citeseer.ist.psu.edu/115216.html>.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, v. 09, n. 4, p. 309–347, October 1992. Disponível em: <http://www.springerlink.com/content/t2k011n123r16831/fulltext.pdf>.

COZMAN, F. G. Generalizing variable elimination in bayesian networks. In: *In Workshop on Probabilistic Reasoning in Artificial Intelligence*. [S.l.: s.n.], 2000. p. 27–32.

CRAMÉR, H. *Elementos da Teoria da Probabilidade e algumas de suas aplicações*. São Paulo: Mestre Jou, 1955.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis*. [S.l.]: John Wiley Sons Inc, 1973. Hardcover.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. *Machine Learning*, v. 29, n. 2-3, p. 131–163, 1997. Disponível em: <http://citeseer.ist.psu.edu-/friedman97bayesian.html>.

GOLDMAN, R. *Probabilistic Approach to Language Understanding*. [S.l.], 1990.

HAZZAN, S.; IEZZI, G. *Fundamentos de Matemática Elementar vol. 5*. [S.l.]: Atual, 2004.

HECKERMAN, D. E. *Probabilistic Similarity Networks*. [S.l.], 1990.

HRUSCHKA JR., E. R. *Imputação Bayesiana no contexto da Mineração de Dados*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, Outubro 2003. Disponível em: <http://www.coc.ufrj.br/teses/doutorado/inter/2003/teses-/HRUSCHKA%20JUNIOR_ER_03_t_D_int.pdf>.

LEVITT, T. S.; AGOSTA, J. M.; BINFORD, T. O. Model-based influence diagrams for machine vision. In: *UAI '89: Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*. Amsterdam, The Netherlands: North-Holland Publishing Co., 1990. p. 371–388. ISBN 0-444-88738-5.

LIPSCHUTZ, S. *Probabilidade*. 4. ed. São Paulo: Makron Books, 1993.

LONG, W. J.; FRASER, H. S. F.; NAIMI, S. Reasoning requirements for diagnosis of heart disease. *Artificial Intelligence in Medicine*, v. 10, n. 1, p. 5–24, 1997. Disponível em: <http://citeseer.ist.psu.edu/william97reasoning.html>.

MARQUES, R. L.; DUTRA, I. *Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações*. Rio de Janeiro: [s.n.], 2008. Disponível em: <www.cos.ufrj.br/˜ines-/courses/cos740/leila/cos740/Bayesianas.pdf>.

MCCALLUM, A.; NIGAM, K. *A comparison of event models for Naive Bayes text classification*. 1998. Disponível em: <http://citeseer.ist.psu.edu/489994.html>.

MEYER, P. L. *Probabilidade: Aplicações à Estatística*. 2. ed. [S.l.]: LTC, 2000.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997.

MORGADO, A. C. et al. *Análise Combinatória e Probabilidade*. Rio de Janeiro: SBM, 2001.

NEAPOLITAN, R. E. *Learning Bayesian Networks*. [S.l.]: Prentice Hall, 2003.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. [S.l.]: Morgan Kaufmann, 1988. Paperback.

PENG, F.; SCHUURMANS, D. *Combining Naive Bayes and n-Gram Language Models for Text Classification*. 2003. Disponível em: <http://citeseer.ist.psu.edu/572782.html>.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: a modern approach*. New Jersey: Prentice Hall, 1995.

YOUSEF, M. et al. Combining multi-species genomic data for microrna identification using a naive bayes classifier machine learning for identification of microrna genes. *Bioinformatics*, The Wistar Institute, Philadelphia, PA 19104, USA., March 2006. ISSN 1367-4803. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/16543277>.