

UNIVERSIDADE FEDERAL DE SÃO CARLOS – UFSCAR

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS – CCET

DEPARTAMENTO DE ESTATÍSTICA – DES

ANDERSON LUIZ ARA-SOUZA

REDES BAYESIANAS:
UMA INTRODUÇÃO APLICADA A CREDIT SCORING

ORIENTADOR: PROF. DR. FRANCISCO LOUZADA NETO

SÃO CARLOS
2010

ANDERSON LUIZ ARA-SOUZA

REDES BAYESIANAS:

UMA INTRODUÇÃO APLICADA A CREDIT SCORING

*Relatório apresentado à comissão julgadora do concurso de
Iniciação Científica do 19º Simpósio Nacional de
Probabilidade e Estatística (SINAPE) - realizado na Estância
de São Pedro-SP, durante 26 a 30 de 2010.*

ORIENTADOR: PROF. DR. FRANCISCO LOUZADA NETO

SÃO CARLOS
2010

“A verdadeira medida de um homem não é como ele se comporta em
momentos de conforto e conveniência, mas como ele se
mantém em tempos de controvérsia e desafio”

-Martin Luther King

AGRADECIMENTOS

À minha família, principalmente meus pais, Carmen e Valdeci, por todo esforço, compreensão e apoio para dar continuidade aos meus estudos. À minha avó Aparecida, por estar sempre ao meu lado interessada e solidária aos meus avanços. À minha irmã Crystiane pela tolerância e horas de distração.

Aos meus verdadeiros amigos, mesmo não havendo a necessidade de citar seus nomes, pois quem me conhece, os conhece. Agradeço pelo intenso apoio e longas conversas sobre os muitos variados assuntos, que contribuíram para nos fortalecer em que somos hoje.

À professora Teresa Cristina Martins Dias, pela amizade e por ter acreditado em meu potencial desde meu ingresso no Bach. em Estatística.

Ao meu orientador Francisco Louzada Neto pela amizade, oportunidades e toda a experiência que tem me passado nesses anos de trabalho.

A todos os docentes e funcionários do Departamento de Estatística da UFSCar, pela formação e estrutura disponível.

RESUMO

A Inteligência Artificial, segundo Tafner *et al.* (1995), é uma área catalisadora do desejo humano de reproduzir inteligência em mecanismos não-biológicos; constitui-se em um conjunto de técnicas de programação para resolver problemas. Ela procura imitar as formas de resolução de problemas do mesmo modo que o homem o faz.

Além disso, a Inteligência Artificial é uma área interdisciplinar dentro das Ciências Exatas, sendo explorada com maior intensidade pelas áreas de Ciências da Computação e Engenharia Elétrica, com o auxílio de metodologia e pensamento estatístico.

Segundo Mitchel, citado por Rezende (2004, p.93), entre os métodos estatísticos, destaca-se o aprendizado Bayesiano, que utiliza um modelo probabilístico baseado no conhecimento prévio do problema, o qual é combinado com exemplos de treinamento para determinar a probabilidade final de uma hipótese.

Assim, as Redes Bayesianas são uma abordagem interpretativa e analítica para o raciocínio probabilista e tem sido utilizada recentemente em diversas áreas como, por exemplo, estimação de risco operacional, diagnóstico médico, *credit scoring*, projeto de jogos computacionais, imputação de dados, entre outras.

Desta forma, a técnica de Redes Bayesianas é um método de modelagem e de decisão, sendo alternativo às técnicas comumente utilizadas como, por exemplo, Regressão Logística e Análise Discriminante.

Esse trabalho de conclusão de curso, desenvolvido de julho a dezembro de 2008, tem a finalidade de apresentar de forma clara e objetiva fundamentos básicos da técnica de Redes Bayesianas, realizar sua aplicação em exemplos, alguns deles envolvendo dados reais, e apresentar os resultados de um estudo de simulação que compara a capacidade preditiva das Redes Bayesianas com um procedimento padrão, comumente utilizado na modelagem de dados com respostas dicotômicas.

Palavras Chaves: Inteligência Artificial Probabilística, Redes Bayesianas, Aprendizado Bayesiano, *Naive Bayes*, *Credit Scoring*.

ÍNDICE

1. INTRODUÇÃO.....	08
1.1. INTELIGÊNCIA ARTIFICIAL.....	09
1.2. <i>CREDIT SCORING</i>	12
1.3. PROBABILIDADES.....	12
1.3.1. Thomas Bayes.....	13
1.3.2. Conceitos probabilistas.....	13
1.3.2.1. Probabilidade e suas propriedades.....	13
1.3.2.2. Probabilidade condicional.....	16
1.3.2.3. Independência probabilística.....	17
1.3.2.4. Teorema de Bayes.....	18
1.3.2.5. As distribuições Multinomial e Dirichlet.....	19
1.3.3. As Redes Bayesianas são Bayesianas?	20
1.4. COMENTÁRIOS FINAIS.....	21
2. REDES BAYESIANAS.....	22
2.1. ESTRUTURA.....	23
2.1.1. Elementos Básicos.....	23
2.1.2. Estruturas de Teoria de Grafos.....	23
2.1.3. Hierarquia entre nós.....	25
2.1.4. Formalização Estatística da Estrutura.....	25
2.1.5. Tabela de probabilidade condicional.....	26
2.1.6. Exemplo Básico de uma Rede Bayesiana.....	26
2.2. EVIDÊNCIA.....	28
2.3. PROPRIEDADES MARKOVIANAS.....	29
2.4. A PROPRIEDADE DE D-SEPARAÇÃO.....	31
2.5. EQUIVALÊNCIA DE MARKOV.....	32
2.6. MÉTODO GERAL PARA A CONSTRUÇÃO DE UMA REDE BAYESIANA.....	33
2.7. COMENTÁRIOS FINAIS.....	34

3. INFERÊNCIA EM REDES BAYESIANAS.....	35
3.1. ALGORITMOS DE INFERÊNCIA EXATA.....	36
3.1.1 Passagem de mensagem.....	36
3.1.1.1. Passagem de mensagem para árvore simples.....	40
3.1.1.2. Passagem de mensagem para poli-árvore.....	45
3.1.2. Método de formação de agrupamentos.....	48
3.2. ALGORITMOS DE INFERÊNCIA APROXIMADA.....	50
3.2.1. Amostragem por Rejeição.....	51
3.2.2. Ponderação de Probabilidade.....	53
3.2.3. Gibbs Sampling	55
3.3. COMENTÁRIOS FINAIS.....	55
4. APRENDIZAGEM EM REDES BAYESIANAS.....	56
4.1. ESTIMAÇÃO DE PARÂMETROS.....	56
4.1.1. Estimação via Máxima verossimilhança.....	57
4.1.2 Estimação Bayesiana.....	60
4.2. ESTIMAÇÃO DE ESTRUTURA.....	62
4.2.1 Algoritmo PC.....	63
4.2.2 Algoritmo K2.....	63
4.2.3 Estimação Híbrida utilizando o <i>Software R</i>	64
4.3. COMENTÁRIOS FINAIS.....	68
5. <i>SOFTWARE HUGIN</i>	69
5.1. PRIMEIROS PASSOS.....	70
5.1.1. Instalação.....	70
5.1.2. Acessando Arquivos.....	70
5.1.3. Compilação.....	71
5.1.4. Salvando os arquivos.....	71
5.2. CONSTRUINDO UMA REDE BAYESIANA.....	71
5.3. INFERÊNCIA EM UMA REDE BAYESIANA.....	74
5.4. ESTIMAÇÃO DA REDE ATRAVÉS DOS DADOS.....	76
5.5. COMENTÁRIOS FINAIS.....	83

6. REDES BAYESIANAS PARA CLASSIFICAÇÃO.....	84
6.1. ESTRUTURAS DE CLASSIFICAÇÃO.....	84
6.2. REDE BAYESIANAS SIMPLES.....	84
6.3. REDE BAYESIANAS SIMPLES COM K-DEPENDÊNCIA.....	86
6.3.1. INFORMAÇÃO MÚTUA.....	87
6.4. CURVA ROC.....	88
6.3. REGRESSÃO LOGÍSTICA.....	88
6.4. REGRESSÃO LOGÍSTICA VERSUS <i>NAIVE</i> BAYES.....	89
6.5. COMENTÁRIOS FINAIS.....	92
7. CONSIDERAÇÕES FINAIS.....	94
REFERÊNCIAS BIBLIOGRÁFICAS.....	95

1. INTRODUÇÃO

As técnicas Bayesianas, além de serem amplamente utilizadas dentro da área de Estatística, são também utilizadas em outros enredos. Um deles é o da Inteligência Artificial, uma área que pode ser denominada Inteligência Artificial Probabilística.

Os métodos e teorias Bayesianos para Inteligência Artificial Probabilística, freqüentemente referidos como aprendizado Bayesiano, vêm sendo aplicados em diversos contextos, desde a área da saúde como áreas de entretenimento (ver por exemplo, Korb e Nicholson, 2004). Um caso particular de aplicação do aprendizado Bayesiano é para a área de *credit scoring*, área importante onde métodos precisos de classificação são requeridos.

Porém, existe uma grande escassez de textos para tais procedimentos na literatura especializada nacional, principalmente na área da Estatística, visto que esse assunto é abordado, geralmente, dentro do cenário da Ciência da Computação.

Mesmo assim, a Inteligência Artificial pode ser considerada uma área interdisciplinar dentro das Ciências Exatas e, muitas vezes, é utilizada com o auxílio de metodologia e pensamento estatístico. Essa contextualização é um assunto ainda não explorado em cursos tradicionais de Graduação em Estatística. Quando tal investigação existe, ocorre apenas em nível de Pós-Graduação.

Desta forma, o presente trabalho tem entre seus principais objetivos apresentar uma abordagem introdutória e didática da técnica de Redes Bayesianas, a fim de desmistificar a técnica de nome geralmente conhecido, mas pouco investigado na área da Estatística. Toda a exemplificação da técnica será apresentada no decorrer do texto e no enredo de *credit scoring*. Tais exemplos são construídos baseados em um conjunto de dados reais.

O presente capítulo expõe contextualizações importantes para a Inteligência Artificial, *credit scoring* e teoria básica de probabilidades. O Capítulo 2 apresenta conceitos básicos da técnica de Redes Bayesianas. No Capítulo 3, métodos comuns de inferência em redes Bayesianas são apresentados. Idéias básicas sobre estimação de estruturas e de probabilidades são exibidas no Capítulo 4. O Capítulo 5 apresenta um *software* importante para área e sua compatibilidade com o *Software R*. Subseqüentemente, o Capítulo 6 apresenta a estrutura *Naive Bayes* em Redes Bayesianas utilizada para classificação, bem como sua comparação com a técnica de Regressão Logística. Por fim, o Capítulo 7 exhibe comentários finais sobre o trabalho.

1.1. INTELIGÊNCIA ARTIFICIAL

A inteligência sempre esteve em destaque entre as ciências desenvolvidas pelo ser humano, sendo um intenso objeto de estudo. Podemos notar esse fato, por exemplo, através do estudo da inteligência por filósofos Gregos como Platão, Aristóteles e Sócrates desde 460 a.C.

Assim, com a evolução das ferramentas tecnológicas, o homem desenvolveu máquinas para realizar serviços comuns a ele, sendo a mais extraordinária, até hoje, o computador. A partir daí, muitos processos tecnológicos foram focados para desenvolvê-lo ainda mais, a fim de transformá-lo em uma máquina totalmente pensante e capaz de realizar qualquer tarefa de forma inteligente. Todo esse processo criou a ciência conhecida atualmente como Inteligência Artificial.

De uma forma mais geral, segundo Rich (1988), a Inteligência Artificial (I.A.) pode ser entendida como o estudo de como fazer os computadores realizarem tarefas em que, no momento, as pessoas são melhores. Além disso, segundo Tafner *et al.* (1995), é uma área catalisadora do desejo humano de reproduzir inteligência em mecanismos não-biológicos, constitui-se em um conjunto de técnicas de programação para resolver problemas. Ela procura imitar as formas de resolução de problemas do mesmo modo que o homem o faz. Ainda, para Russel e Norvig (2004) as definições de Inteligência Artificial podem ser divididas em quatro grandes grupos, sendo estes os grupos de sistemas que pensam como seres humanos, sistemas que pensam racionalmente, sistemas que atuam como seres humanos e sistemas que atuam racionalmente. Neste contexto, a analogia entre racional e humano é dado pela perfeição de um sistema racional, ou seja, o sistema racional não é suscetível aos erros comuns realizados por seres humanos.

Como podemos notar, existe uma grande dificuldade de definir precisamente o que é I.A. devido ao atual estado da arte da ciência, que envolve inclusive questões filosóficas de interpretação do pensamento humano.

Além disso, o “inteligente” se remete ao próprio processo de pensamento humano, o qual se refere a raciocinar baseado no conhecimento que possui sobre determinado problema, ignorando se tal processo é perfeito ou não.

Desta forma, um sistema de I.A. é baseado em quantidades de conhecimento a respeito do domínio de um problema em estudo, ou seja, informação sobre o escopo do problema. Uma forma tradicional de quantificar tal conhecimento é a partir de uma base de dados.

A partir dos anos 80, técnicas específicas surgiram motivadas pela grande aplicação da área em domínios especializados como a diagnose médica, análise química, reconhecimento de padrões etc., sendo realizadas hoje rotineiramente. Desta forma, problemas realizados por apenas um pequeno contingente de especialistas, como a avaliação de crédito ou a diagnose de doenças, podem ser realizados por sistemas inteligentes comumente denominados na literatura especializada de sistemas especialistas.

A lista seguinte contém um resumo de alguns problemas que se enquadram dentro deste contexto (Rich, 1988):

- Jogos
- Prova de teoremas
- Resolução de problemas gerais
- Percepção
 - Visão
 - Fala
- Compreensão da linguagem natural
- Resolução de problemas especializados
 - Matemática simbólica
 - Diagnose
 - Análise química
 - Projeto de engenharia

Assim, para cada uma dessas tarefas seleciona-se um conjunto de informações que representam o escopo do problema, geralmente baseado no conhecimento empírico ou de um especialista, buscando especificar o máximo possível os critérios para o sucesso do sistema.

Através desta breve introdução sobre I. A. e sistemas especialistas, notamos que existe grande relação entre os métodos estatísticos e alguns objetivos específicos de Inteligência Artificial. De uma forma mais precisa, esse relacionamento ocorre no âmbito da motivação para a análise, uma vez que ambas as áreas buscam informações refletidas em variáveis e em um conjunto de observações representativas do problema, visando resolver problemas específicos como o de classificação e dependência entre variáveis em diversos enredos.

Desta forma, métodos estatísticos são também utilizados na grande área de Inteligência Artificial. Estes vêm a contribuir e concorrer com os métodos mais comuns, mais precisamente para a resolução de sistemas especialistas. Um exemplo disso é a aplicabilidade

para os mesmos fins das técnicas de Redes Neurais, uma metodologia originariamente desenvolvida dentro da área de I.A. e de Regressão Logística, dentro da área da Estatística.

Além disso, confirmando o relacionamento comum entre as duas áreas, podemos realizar uma breve consulta à literatura especializada e encontraremos facilmente assuntos como *Data Mining*, Modelagem e Aprendizado sendo direcionados para ambas.

Entre os métodos estatísticos utilizados em I.A., destaca-se o aprendizado Bayesiano, que utiliza um modelo probabilístico baseado no conhecimento prévio do problema, o qual é combinado com exemplos de treinamento para determinar a probabilidade final de uma hipótese (Mitchel citado por Rezende, 2004, p.93).

Neste contexto, a técnica de Redes Bayesianas é uma poderosa ferramenta de aprendizado Bayesiano que utiliza raciocínio probabilista e é aplicada no contexto de I.A..

O raciocínio probabilista, bem como o raciocínio lógico, é uma das abordagens principais utilizadas para modelar sistemas que agem racionalmente dentro da área de Inteligência Artificial. Marques e Dutra (1999) consideram que o raciocínio lógico retira suas conclusões sobre a base ponderada do conhecimento prévio a respeito do problema. O raciocínio probabilista, por sua vez, é adequado para situações onde não se conhece previamente todo o escopo do problema, ou seja, para representar inferências prováveis, porém incertas como, por exemplo, no contexto: “Muitas vezes bons pagadores possuem um saldo devedor menor”.

Em I.A., as probabilidades não estão presentes apenas no âmbito deste tipo de raciocínio, mas também estão presentes em outros contextos como, por exemplo, aquele da probabilidade lógica. Segundo Haenni (2005), o raciocínio probabilista e a probabilidade lógica podem ser interpretados como casos extremamente opostos dentro do universo de estudo do raciocínio, chamado de argumentação probabilística.

Em referência ao tipo de raciocínio de embasamento, podemos utilizar o “Inteligência Artificial Probabilística” para categorizar esta subárea do conhecimento em I. A..

De uma forma geral, a técnica de Redes Bayesianas é uma abordagem interpretativa e analítica para a Inteligência Artificial Probabilística e tem sido utilizada recentemente em diversas áreas como, por exemplo, estimação de risco operacional, diagnóstico médico, *credit scoring*, projeto de jogos computacionais, imputação de dados, entre outras. Desta forma, pode ser utilizada para fins de decisão, classificação, análise de dependência entre variáveis, sendo alternativa às técnicas comumente utilizadas como, por exemplo, Regressão Logística e Análise Discriminante.

1.2. CREDIT SCORING

A necessidade de análise de crédito nasceu nos primórdios do comércio conjuntamente com a concessão de empréstimos de dinheiro ou com a autorização de compras a pagar futuramente, pois, desde aquela época, quando um comerciante oferece demasiado crédito à pessoa errada, este corre o risco de perder dinheiro e ter futuros problemas financeiros. Com o passar dos anos, os comerciantes começaram a levantar informações sobre os solicitantes de crédito e catalogá-los para decidir se emprestariam ou não determinada quantia em dinheiro.

Com o desenvolvimento da ciência em análise de dados refletida em métodos precisos, hoje *credit scoring* é um método de avaliação de risco de crédito para aplicação de empréstimos (Mester, 1997). Baseado em métodos estatísticos para análise de dados, tal método produz um *score* para cada cliente, quantificando o risco deste cliente ser bom ou mau pagador, a fim de minimizar as perdas ou ganhos de uma empresa, geralmente financeira.

Por ter como objetivo final a classificação binária de uma determinada característica, na área de *credit scoring* são aplicadas diversos métodos de tratamento de dados, dentre elas temos modelos probito, logito, regressão por árvores, análise discriminante, redes neurais e, também, Redes Bayesianas, entre outras.

Neste trabalho, além de introduzir a teoria básica de Redes Bayesianas, temos por objetivo expor exemplos de aplicações em *credit scoring* para as manipulações mais importantes da técnica. Especificamente, iremos expor a aplicação do algoritmo de *Naive Bayes*, associado a uma estrutura particular de Redes Bayesianas, para a classificação de indivíduos como maus pagadores. Além disso, iremos visualizar o relacionamento, dentro de uma estrutura estimada de Redes Bayesianas, das variáveis de um conjunto de dados reais de *credit scoring*.

1.3. PROBABILIDADES

O cálculo das probabilidades teve origem em estudos de jogos de azar na Idade Média. Assim, em 1654, o desenvolvimento desta ciência deu-se devido a uma série de cartas trocadas entre dois matemáticos e pensadores notáveis, Blaise Pascal (1623-1662) e Pierre de Fermat (1601-1665), sobre problemas com apostas em jogo composto por moedas e dados.

Desde então, a teoria de probabilidades foi amplamente estudada, como pelo também renomado Thomas Bayes, e sendo hoje utilizada em diversos procedimentos das Ciências Exatas.

Nesta seção, introduzimos uma breve história sobre Thomas Bayes e conceitos fundamentais em probabilidade que são necessários para o entendimento da teoria de Redes Bayesianas.

1.3.1. Thomas Bayes

Nascido em Londres no ano de 1702 e falecido em Kent, a 58 km de Londres, em 1761, o inglês Thomas Bayes (Figura 1.1) foi matemático e reverendo da igreja presbiteriana e imortalizado por formular um importante teorema de probabilidade, o qual intitula seu nome e deu origem, anos depois, a um novo ramo da ciência estatística denominada Estatística Bayesiana.

Sua família possuía o alinhamento não conformista – título dado a europeus não-anglicanos ou que prezam a liberdade religiosa – e, antes de seu nascimento, havia feito fortuna no setor da cutelaria, arte de fabricar instrumentos cortantes, um ramo importante em Sheffield, cidade de origem do avô de Thomas Bayes, Richard Bayes.



Figura 1.1. Única Ilustração conhecida de Thomas Bayes

Desta forma, o pai de Thomas, Joshua Bayes (1638–1703), foi enviado à Academia de Frankland (Inglaterra), onde estudou filosofia e religião. Em 1694, Joshua partiu para Londres, onde havia o primeiro grupo de presbiterianos ingleses. Lá foi ordenado ministro e,

em poucos anos, tornou-se um homem respeitado. Durante esse período, Joshua Bayes casou-se com Anne Carpenter e juntos tiveram 7 filhos. Assim, Thomas Bayes, seguindo os passos do pai e por ser o filho mais velho, tornou-se ministro.

Thomas Bayes estudou teologia na Universidade de Edimburgo (Escócia) e em 1731 assumiu a paróquia de Tunbridge Wells, em Kent. Historicamente, publicou apenas dois livros em vida, o primeiro chamado de *Benevolência divina* (1731) e o segundo chamado de *Uma Introdução a doutrina dos fluxions*, no qual ele defendia Isaac Newton contra a crítica de George Berkley, conhecido filósofo irlandês da época. Após sua morte, outro trabalho de sua autoria foi revelado, *Ensaio buscando resolver um problema na doutrina das probabilidades*, no qual havia a formulação do Teorema de Bayes. Era creditado que esse artigo fornecia uma prova da existência de Deus.

Para maiores detalhes sobre a vida de Thomas Bayes consultar Bellhouse (2004), uma completa biografia realizada em comemoração ao seu 300º aniversário de nascimento.

1.3.2. Conceitos probabilistas

As Redes Bayesianas são ferramentas que utilizam o raciocínio probabilista, ou seja, toda sua metodologia é baseada em probabilidades, especialmente a probabilidade condicional. Para melhor exposição da teoria de Redes Bayesianas, uma breve revisão da teoria de probabilidades será apresentada abaixo.

1.3.2.1. Probabilidade e suas propriedades

Em poucas palavras, a probabilidade pode ser introduzida, segundo Costa Neto e Cymbalista (2006), como sendo o número que mede a maior ou menor possibilidade de ocorrência de diversos eventos.

Porém, o conceito de probabilidade é, historicamente, cenário de ampla discussão e tem sido definido de diferentes maneiras, sendo que algumas são as definições de probabilidade freqüentista, clássica e subjetiva.

Hoje em dia, a definição axiomática, dada por Komolgorov em 1933, é comumente adotada e considera que a probabilidade é uma função definida numa classe \mathfrak{S} de eventos de Ω que satisfaz as seguintes condições:

$$(a) P(A) \geq 0 \text{ para todo } A \in \mathfrak{S}; \quad (1.1)$$

(b) Se $(A_n)_{n \geq 1}$ é uma seqüência de eventos de \mathfrak{S} , que são mutuamente exclusivos, então:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n); \quad (1.2)$$

$$(c) P(\Omega) = 1. \quad (1.3)$$

onde A é um evento no espaço \mathfrak{S} e Ω é um conjunto de eventos de interesse denominado espaço amostral.

A definição acima origina as seguintes propriedades:

$$(d) P(\emptyset) = 0 \quad (1.4)$$

$$(e) \text{ Se } E, F, \dots, K \text{ são eventos mutuamente exclusivos} \\ P(E \cup F \cup \dots \cup K) = P(E) + P(F) + \dots + P(K); \quad (1.5)$$

$$(f) P(\overline{E}) = 1 - P(E) \quad (1.6)$$

$$(g) P(E \cup F) = P(E) + P(F) - P(E \cap F) \quad (1.7)$$

entre outras.

Desta forma, uma forma objetiva de atribuição de probabilidade ao evento F é dada por (1.8).

$$P(F) = \frac{\#F}{\#\Omega}, \quad (1.8)$$

onde $\#F$ é número de resultados favoráveis ao evento F e $\#\Omega$ é o número de resultados totais, ou seja, o número de resultados no espaço amostral Ω .

Para melhor entendimento dos termos probabilísticos, considere os itens 1, 2, 3 e 4 da Figura 1.2, os quais exibem uma visualização freqüente na literatura da teoria de

probabilidades baseada na diagramação de Eüller-Venn para os eventos e o seu espaço amostral.

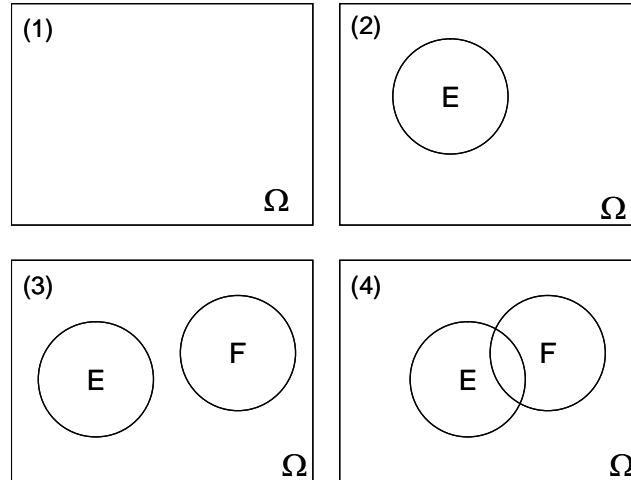


Figura 1.2. Diagramas de Eüller-Venn

Na Figura 1.2, o item (1) exhibe todo o espaço amostral Ω , o item (2) exhibe o evento E sob o espaço amostral, o item (3) exhibe os eventos E e F sendo mutuamente exclusivos, ou seja, $P(E \cap F)=0$ e, finalmente, o item (4) exhibe os eventos E e F como não exclusivos.

1.3.2.2. Probabilidade condicional

A probabilidade condicional trata do fato de que muitas vezes temos conhecimento que um determinado evento já ocorreu, desta forma surge o interesse de calcular a probabilidade de outro evento de interesse e possivelmente relacionado ao evento anterior.

Denotamos como $P(E | F)$ à probabilidade de ocorrência do evento E, sabendo que o evento F ocorreu, ou simplesmente, a probabilidade de E dado F.

Desta forma, temos (1.9).

$$P(E | F) = \frac{P(E \cap F)}{P(F)} \quad (1.9)$$

Analogamente,

$$P(E \cap F) = P(E | F)P(F) \text{ ou } P(E \cap F) = P(F | E)P(E) \quad (1.10)$$

Assim temos também, generalizando (1.10) e considerando a notação $P(E \cap F) = P(E, F)$,

$$P(E_1, E_2, \dots, E_n) = P(E_1)P(E_2 | E_1)P(E_3 | E_1, E_2) \cdots P(E_n | E_1, E_2, \dots, E_{n-1}) \quad (1.11)$$

Além disso, considerando $E_1 \cdots E_n$ eventos exclusivos e exaustivos, ou seja, eventos que não possuem intersecção e sua união é igual ao espaço amostral Ω , temos para um evento F,

$$P(F) = \sum_{k=1}^n P(F | E_k)P(E_k) \quad (1.12)$$

A propriedade (1.12) é comumente denominada de fórmula de probabilidades totais. Note que esta permite calcular a probabilidade de um evento F quando se conhece as probabilidades de um conjunto de eventos distintos, sendo que sua união forma o espaço amostral.

1.3.2.3. Independência probabilística

Assim como a probabilidade condicional, a dependência probabilística é uma das propriedades fundamentais utilizadas na teoria de Redes Bayesianas.

Basicamente, podemos considerar que os eventos E e F são independentes quando existe a relação (1.13).

$$P(E | F) = P(E) \text{ ou } P(F | E) = P(F) \quad (1.13)$$

A relação (1.13) vem de outra propriedade básica de independência probabilística entre dois eventos, apontada em (1.14).

$$P(E, F) = P(E)P(F) \quad (1.14)$$

Para chegarmos à relação (1.13) basta aplicar (1.14) em (1.9).

Note que a relação de independência não está ligada à relação de exclusividade de eventos.

1.3.2.4. Teorema de Bayes

Como anteriormente, considere o evento F e $E_1 \cdots E_n$ eventos exclusivos e exaustivos, ou seja, que não possuem intersecção e que a união deles forma o espaço amostral. Assim, o Teorema de Bayes é definido em (1.15)

$$P(E_j | F) = \frac{P(E_j)P(F | E_j)}{\sum_{i=1}^n P(E_i)P(F | E_i)} \quad (1.15)$$

O teorema de Bayes é uma junção do teorema de probabilidade condicional e da fórmula de probabilidades totais. Assim, $P(E_j)$ pode ser denominada como probabilidade a priori, $P(F | E_j)$ como verossimilhança e $P(E_j | F)$ como probabilidade a posteriori, ou seja, a probabilidade posterior à observação do evento F . Além disso, o denominador é a decomposição de $P(F)$, ou seja, pode ser considerado como constante normalizadora; desta forma, (1.15) pode ser reescrito na forma (1.16).

$$P(E_j | F) \propto P(E_j)P(F | E_j) \quad (1.16)$$

sendo \propto indicador de proporcionalidade. Em outros termos, podemos dizer que a probabilidade a posteriori é proporcional à probabilidade a priori multiplicada pela verossimilhança.

1.3.2.5. As distribuições Multinomial e Dirichlet

Estas duas distribuições, aqui introduzidas, são amplamente utilizadas no contexto de Redes Bayesianas quando métodos de estimação bayesiana são requeridos.

Considere uma variável aleatória X discreta que represente um experimento com r possíveis resultados, sendo que cada tipo de resultado possui uma probabilidade específica

$P(X = x_r) = p_r$ e $\sum_{i=1}^r p_i = 1$. Além disso, o experimento é repetido de forma independente N

vezes, de forma que a variável X_i seja o número de vezes que o resultado x_i está presente na amostra com $i=1, \dots, r$. Temos que a variável X segue distribuição Multinomial, sendo sua função densidade de probabilidade expressa pela fórmula (1.17).

$$P(X | N, p) = \frac{N!}{x_1! x_2! \dots x_r!} p_1^{x_1} p_2^{x_2} \dots p_r^{x_r} \quad (1.17)$$

sendo $\sum_{i=1}^r x_i = N$.

Considerando o termo $\frac{N!}{x_1! x_2! \dots x_r!}$ como normalizador, temos

$$P(X | N, p) \propto p_1^{x_1} p_2^{x_2} \dots p_r^{x_r} \quad (1.18)$$

Além disso, temos que para um vetor $p=(p_1, p_2, \dots, p_r)$ de valores desconhecidos com $\sum_{i=1}^r p_i = 1$, podemos assumir que p segue distribuição Dirichlet com parâmetros $\alpha = (\alpha_1, \dots, \alpha_r)$ com $\alpha_i > 1$, $E(p_i) = \alpha_i/\alpha_0$ e função densidade de probabilidade expressa pela fórmula (1.19).

$$P(p | \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_r)} p_1^{\alpha_1-1} p_2^{\alpha_2-1} \dots p_r^{\alpha_r-1} \quad (1.19)$$

Da mesma forma, podemos considerar o termo $\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_r)}$ como normalizador, assim temos (1.20).

$$P(p|\alpha) \propto p_1^{\alpha_1-1} p_2^{\alpha_2-1} \dots p_r^{\alpha_r-1} \quad (1.20)$$

Assumindo como priori $P(p|\alpha)$ e como verossimilhança $P(X|N, p)$, temos que a posteriori $P(p|X)$ é dada pela expressão (1.21) a qual tem distribuição Dirichlet com parâmetros $\alpha = (\alpha_1 + x_1, \dots, \alpha_r + x_r)$, ou seja $E(p_i) = \frac{\alpha_i + x_i}{\alpha_0 + N}$, sendo $\alpha_0 = \sum_{i=1}^r \alpha_i$.

$$P(p|X) \propto p_1^{\alpha_1+x_1-1} p_2^{\alpha_2+x_2-1} \dots p_r^{\alpha_r+x_r-1} \quad (1.21)$$

Notamos que neste caso a posteriori possui o mesmo tipo de distribuição que a priori, assim dizemos que a família Dirichlet é conjugada para amostras com distribuição Multinomial.

1.3.3. As Redes Bayesianas são Bayesianas?

Existe uma grande discussão na literatura sobre se as Redes Bayesianas são realmente Bayesianas ou não. Alega-se que esse termo seja uma nomenclatura inadequada. Korb e Nicholson (2004) evidenciam a pronúncia formal do Professor Geoff Webb, especialista em *Data Mining* da universidade australiana de Monash, que declarou dois pontos de vista:

1. A técnica de Redes Bayesianas pode ser considerada um método de *Data Mining* que utiliza métodos não-Bayesianos.
2. As Redes Bayesianas são um método para representar probabilidades, que podem ser interpretadas de forma Bayesiana ou não.

Deste modo, notamos que atualmente essa discussão pode gerar bastante polêmica entre os especialistas da área. Porém, temos que o objetivo fundamental da técnica é realizar inferência e estimativas com base em condicionamentos de informações, o que gera uma ponte de ligação sólida com a filosofia Bayesiana.

Ainda assim, como mostramos neste trabalho, os métodos de estimação dentro da teoria de Redes Bayesianas podem ser realizados por métodos Bayesianos ou não-Bayesianos.

1.4. COMENTÁRIOS FINAIS

Neste capítulo, apresentamos uma conceituação básica sobre Inteligência Artificial e sua subdivisão denominada Inteligência Artificial Probabilística, sendo que tal termo provém do tipo de raciocínio utilizado, ou seja, o raciocínio probabilista. Além disso, introduzimos a necessidade de análise para a área de *credit scoring*

A respeito da teoria de probabilidade, exibimos importantes propriedades probabilistas que serão utilizadas ao decorrer do trabalho, sendo as mais importantes a propriedade de dependência, o Teorema de Bayes e o relacionamento entre as distribuições de probabilidade Multinomial e Dirichlet.

Assim, estamos prontos para introduzir os conceitos básicos de Redes Bayesianas.

2. REDES BAYESIANAS

As Redes Bayesianas, também conhecidas como Redes casuais, Rede de crença e Gráficos de dependência probabilística, surgiram na década de 80 e têm sido aplicadas em uma grande variedade de atividades do mundo real (Bobbio *et al.*, 2001). Algumas aplicações atuais se estendem às áreas como finanças (Chang *et al.*, 2000), saúde (Abicalaff, Amaral e Dias, 2004) (Korb e Nicholson, 2004), desenvolvimento de jogos (Vieira Filho e Albuquerque, 2007), entre outras.

Ainda, as Redes Bayesianas vêm sendo bastante utilizadas em áreas financeiras para a estimação de risco operacional e *credit scoring* (ex: Sistema BayesCredit, um sistema criado por Nykredit, uma das principais empresas no mercado dinamarquês de financiamento imobiliário) e possui vários programas específicos disponíveis como, por exemplo, os *softwares* Netica (www.norsys.com) e Hugin (www.hugin.com).

Segundo Neapolitan (2004), a técnica de Redes Bayesianas surgiu no contexto no qual há um grande número de variáveis e o objetivo de verificar qual a influência probabilística não direta de uma variável para as demais.

Assim, a teoria de Redes Bayesianas combina princípios de Teoria de grafos, teoria de probabilidades, Ciência da Computação e Estatística (Ben-Gal, 2007).

Além disso, as Redes Bayesianas podem ser consideradas como uma representação visual e informativa da tabela de probabilidade conjunta de todas as variáveis que envolvem o domínio do problema.

Desta forma, na literatura especializada, uma terminologia específica é utilizada para definir tipos de variáveis, dependências probabilísticas e outras propriedades das Redes Bayesianas. Neste trabalho, optamos por simplificar tal terminologia, quando possível, aproximando-a de termos utilizados na modelagem estatística de dados.

Esse capítulo tem como objetivo introduzir conceitos básicos da teoria de Redes Bayesianas, que envolvem os tipos de estruturas de teoria de grafos, noções de evidência, propriedade markoviana, equivalência, noção de independência, definição básica para construção e ordem das variáveis, bem como exibir breves exemplos.

2.1. ESTRUTURA

Nesta seção serão introduzidos conceitos elementares dentro da estrutura gráfica de uma Rede Bayesiana, em sua maioria um conjunto de nomenclaturas originadas através das relações visualmente perceptíveis da estrutura gráfica.

2.1.1. Elementos Básicos

As Redes Bayesianas são uma representação gráfica de variáveis e suas relações para um problema específico, 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. Esses elementos podem ser visualizados na Figura 2.1.



Figura 2.1. Elementos básicos da Teoria de Grafos

2.1.2. Estruturas de Teoria de Grafos

Existem diversos tipos de aplicações da Teoria de Grafos na literatura. Maiores detalhes podem ser encontrados em Feofiloff *et al.*(2007).

Além disso, existem diversos tipos de estruturas básicas dentro da Teoria de Grafos. Para uma visualização geral, tais estruturas são exibidas na Figura 2.2.

A teoria de Redes Bayesianas é construída considerando grafos direcionados, conectados e acíclicos, frequentemente referenciados pela sigla DAG (*directed acyclic graph*).

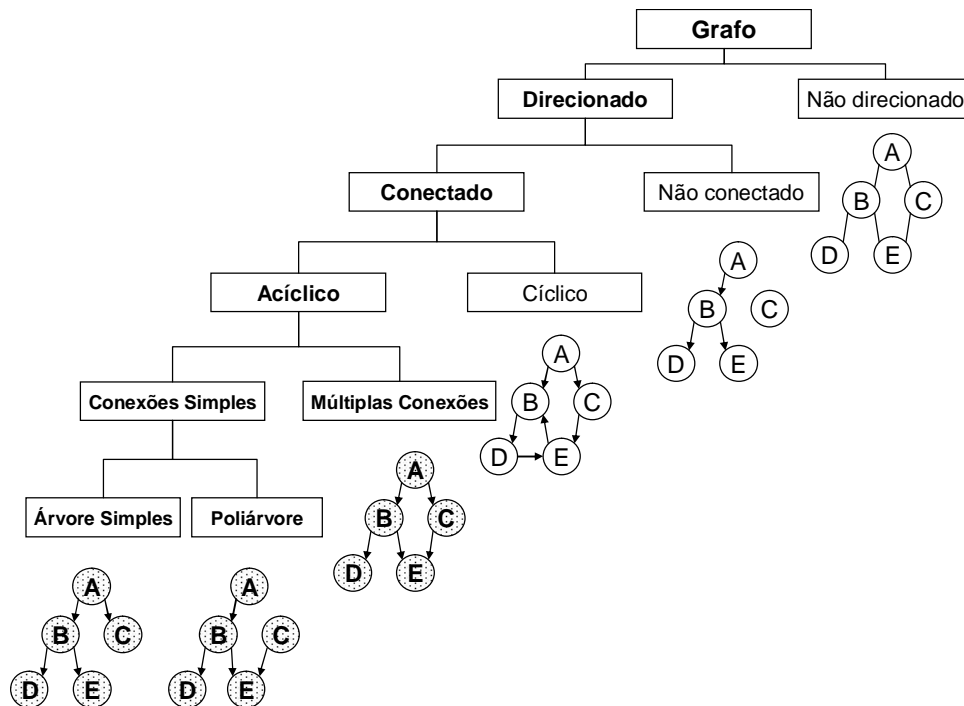


Figura 2.2. Estruturas básicas existentes dentro da Teoria de grafos

O termo “direcionado” faz referência à presença de direção dos arcos, o termo “conectado” é utilizado para designar que todos os nós estão conectados na rede e, por fim, o termo “acíclico” se refere à propriedade de não retorno para um nó após seguida a direção dos arcos.

Desta forma, através da Figura 2.2, notamos que as Redes Bayesianas envolvem apenas alguns tipos de estruturas básicas: a estrutura de conexões simples, que engloba as estruturas de árvore simples e poliárvore, e a estrutura de múltiplas conexões.

Para as estruturas de conexões simples é dada a regra geral de que existe apenas um caminho que liga uma variável a outra, independente da direção dos arcos. Analogamente, para as estruturas de múltiplas conexões há mais de um possível caminho que liga uma variável a outra, independente da direção dos arcos.

A subdivisão das estruturas de conexão simples se dá pelo número de nós que originam a rede, ou seja, nós que não possuem nenhum arco chegando, apenas arcos partindo.

Assim, como notamos na Figura 2.2, as estruturas de árvores simples possuem apenas uma variável que origina a rede (variável A) e as estruturas de poliárvore possuem duas (ou mais) variáveis que originam a rede (variáveis A e C). Estas variáveis geralmente possuem um nome específico o qual será apresentado no próximo item.

2.1.3. Hierarquia entre nós

Dentro da terminologia de Redes Bayesianas, outros termos também são comuns e utilizados para considerar a hierarquia de nós dentro da rede, o que é o caso dos termos pai e filho. Esses termos referem-se à relação de dependência direta entre dois nós por meio do arco que os conecta, o nó de onde o arco parte é designado nó pai, o nó de onde o arco chega com sua ponta é designado nó filho. Assim, considerando a estrutura de simples conexões da Figura 2.2, o nó A é pai do nó B, sendo o nó B filho do nó A. Analogamente, o nó B é pai dos nós C e D, sendo os mesmos filhos do nó B.

Além disso, um nó que não possui filhos é chamado de folha e um nó que origina a rede, ou seja, que não possui pais, é chamado de raiz.

Os nós antecedentes a um determinado nó A, ou seja, o(s) pai(s) e seus respectivos pais e assim por diante, são denominados como ancestrais de A. Da mesma forma, os nós derivados de determinado nó A, ou seja, o(s) filho(s) e seus respectivos filhos e assim por diante, são denominados como descendentes de A, analogamente a uma estrutura de genealogia.

2.1.4. Formalização Estatística da Estrutura

Como dito anteriormente, em Redes Bayesianas cada variável aleatória do estudo é representada por um nó. Por esse motivo, iremos substituir o termo “nó” pelo termo “variável”, ou seja, ao nos referirmos ao nó A, iremos representá-lo pelo termo variável A. Estendendo esse conceito para a hierarquia de nós, temos que a variável A é pai da variável B.

Os valores das variáveis podem ser de qualquer tipo de escala, contínua ou discreta. Porém, neste trabalho, a maior parte do texto irá tratar de variáveis discretas, pois, segundo Korb e Nicholson (2004), a tecnologia de redes Bayesianas é primeiramente direcionada ao tratamento de variáveis discretas, como por exemplo, para a confecção de algoritmos de inferência. Além disso, as variáveis contínuas podem ser facilmente transformadas em variáveis discretas através de simples categorizações. Analogamente, as técnicas

desenvolvidas até o presente momento para variáveis contínuas não são ainda tão desenvolvidas. Por exemplo, é condição básica a uma variável discreta não possuir variáveis-pai contínuas.

Assim, de uma forma geral para este trabalho, uma Rede Bayesiana é definida pelo trio (ξ, θ, X) , onde ξ é uma estrutura DAG e θ é um conjunto de parâmetros específicos de distribuições de probabilidades condicionais envolvendo um conjunto X de variáveis aleatórias discretas.

2.1.5. Tabela de probabilidade condicional

Outro elemento importante dentro da estrutura de Redes Bayesianas é a tabela de probabilidade condicional (CPT). Trata-se da exibição dos parâmetros de probabilidade condicional da variável sendo condicionada a seu(s) pai(s).

Por exemplo, dado o conjunto de três variáveis A , B e C , todas dicotômicas assumindo valores binários, onde A e B são pais da variável C , temos a Tabela 2.1.

Tabela 2.1. Tabela de Probabilidade Condicional $P(C|A,B)$

C	A	B	P(C A,B)
1	1	1	θ_1
1	1	0	θ_2
1	0	1	θ_3
1	0	0	θ_4
0	1	1	θ_5
0	1	0	θ_6
0	0	1	θ_7
0	0	0	θ_8

Com base nas definições acima, podemos exibir um exemplo de Rede Bayesiana.

2.1.6. Exemplo Básico de uma Rede Bayesiana

Considere uma Rede Bayesiana dada sua estrutura já conhecida e relacionando seguintes variáveis binárias:

- Sexo { M, F };
- Idade { <20 anos, >=20 anos };
- Créditos Anteriores { 1, >1 };
- *Credit Rating* { Bom , Ruim }.

Assim, a rede é representada pela Figura 2.3.

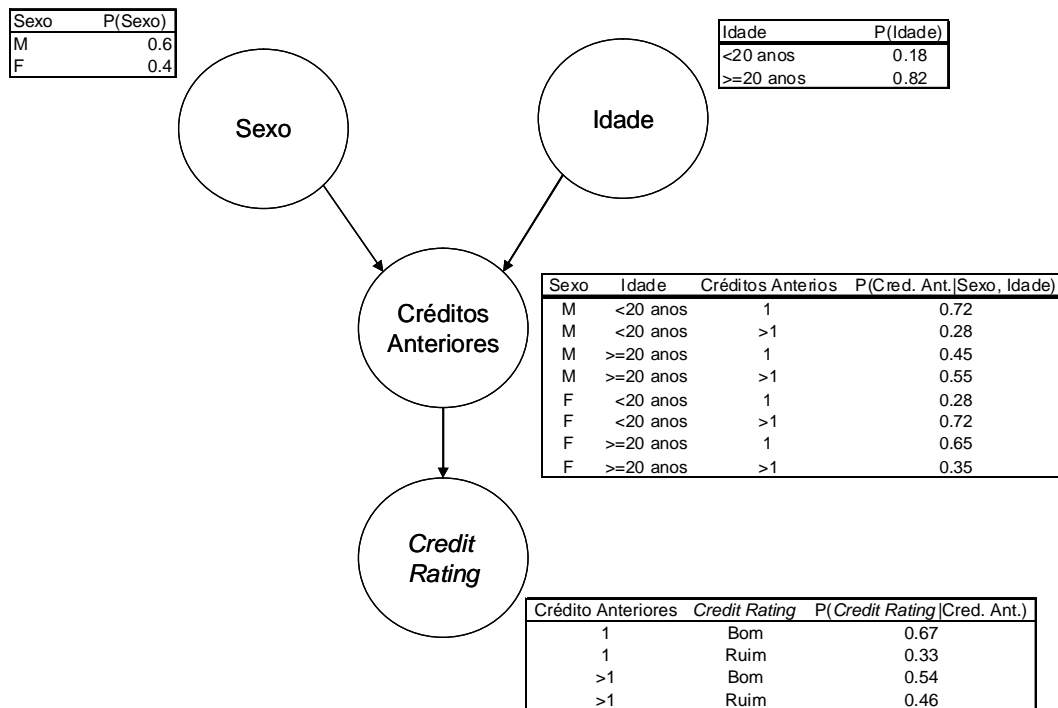


Figura 2.3 – Exemplo de Rede Bayesiana para dados de *Credit Scoring*

Considerando o exemplo da Figura 2.3 temos que as variáveis Sexo, Idade, Créditos Anteriores e *Credit Rating* são representadas por seu respectivo nó na rede, sendo Sexo e Idade variáveis-pai da variável Créditos Anteriores e Créditos Anteriores pai da variável *Credit Rating*. Ainda realizando uma análise hierárquica, as variáveis Sexo e Idade são classificadas na rede como variáveis-raiz e *Credit Rating* como folha.

Além disso, notamos que Sexo e Idade influenciam diretamente a variável Créditos Anteriores, que por sua vez influencia probabilisticamente de uma forma direta a variável *Credit Rating*.

Interpretando os relacionamentos, se o cliente é do sexo masculino, ou não, isso influencia na probabilidade do cliente ter um, ou mais, créditos anteriores realizados na instituição. Se o cliente é menor de 20 anos, ou não, também influencia a probabilidade do cliente ter um ou mais créditos anteriores realizados na instituição. Assim, a probabilidade do cliente ter, ou não, realizado requisição de créditos anteriormente na instituição financeira influencia a probabilidade dele ser classificado como um bom pagador ou mau pagador.

Para cada uma das variáveis e seus cruzamentos condicionais, temos uma tabela de probabilidade condicional (CPT) explicando numericamente a chance da cada categoria – evento – ocorrer dado premissas anteriores.

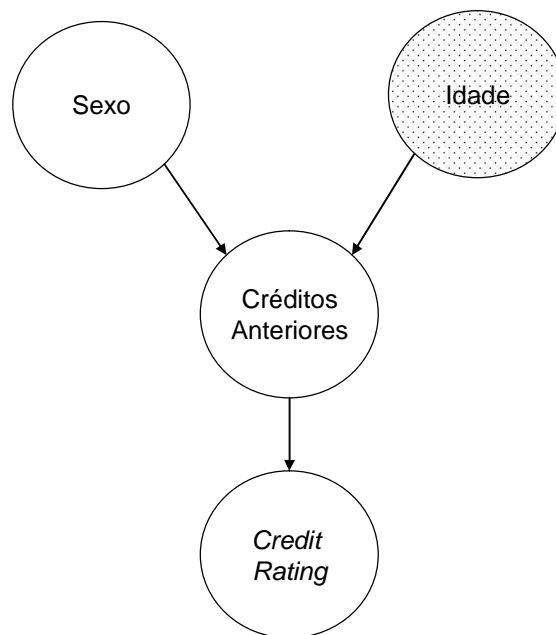


Figura 2.4. Rede Bayesiana tendo como evidencia a variável Idade (Idade <20)

2.2. EVIDÊNCIA

Dada a estrutura gráfica DAG, outra definição é importante para a teoria de Redes Bayesianas. Esta é denominada como evidência e refere-se ao fato de uma variável ser

indicada pelo usuário da rede, ou seja, uma variável aleatória com valor conhecido e acoplado à Rede Bayesiana com estrutura já conhecida. Basicamente, podemos definir uma evidência com uma observação.

Considere o exemplo da Figura 2.3. Desta forma, observamos que um novo cliente possui a idade de 18 anos; assim, na rede, indicamos a variável Idade para a categoria respectiva, ou seja, definimos Idade <20 anos. A variável idade é classificada como uma evidência para a rede. A Figura 2.4 exibe uma demonstração visual para Idade <20 anos.

As evidências são úteis quando existe o objetivo de realização de inferência probabilística para a rede em estudo. Este procedimento será visto com mais detalhes posteriormente.

2.3. PROPRIEDADES MARKOVIANAS

Assim como em alguns tipos de processos estocásticos, a dinâmica de uma Rede Bayesiana é controlada pela propriedade de Markov, a qual rege que não existem dependências diretas entre as variáveis de uma Rede Bayesiana que não estão explícitas através da apresentação orientada dos arcos, ou seja, cada variável possui dependência direta apenas de sua(s) variável(eis) pai.

A partir de todas as propriedades acima, temos que uma Rede Bayesiana é um par (ξ, θ) definido sobre um conjunto de variáveis aleatórias $X = \{X_1, X_2, \dots, X_K\}$, onde cada X_i corresponde a uma variável da rede, satisfazendo a propriedade de Markov:

$$P[X_i | X_j, \text{pais}(X_i)] = P[X_i | \text{pais}(X_i)] \quad (2.1)$$

Além disso, consideremos a distribuição de probabilidade conjunta de uma Rede Bayesiana com k variáveis e a propriedade (2.1), temos que em uma Rede Bayesiana (ξ, θ) , definida sobre um conjunto de variáveis aleatórias $X = \{X_1, X_2, \dots, X_K\}$, a probabilidade conjunta de toda a rede é dada através da expressão (2.2).

$$P(X_1 = x_1, \dots, X_K = x_k) = \prod_{i=1}^k P[X_i = x_i | \text{pais}(X_i)] \quad (2.2)$$

Ou seja, as propriedades probabilísticas estão intimamente ligadas com o condicionamento da variável com seu (s) pai (s) respectivo (s). Note que (2.2) é resultado direto do desenvolvimento do Teorema de Bayes visto na seção 1.3.2.4., dada a propriedade (2.1).

Para o exemplo da Figura 2.3, a variável Sexo e Idade são independentes, pois não existe nenhum arco relacionando-as. Além disso, *Credit Rating* é diretamente independente de Sexo e Idade, a variável *Credit Rating* depende apenas diretamente da variável Créditos Anteriores, a qual é sua variável-pai.

Uma Rede Bayesiana na qual cada dependência probabilística entre as variáveis é dada por um único arco é chamada de Rede perfeita (Korb e Nicholson, 2004).

Outro conceito muito utilizado na teoria de Redes Bayesianas é a cobertura de Markov, que consiste no conjunto formado pelas variáveis-pai, variáveis-filhos e pais dos filhos de uma determinada variável. Como exemplo, temos que a cobertura de Markov para a variável Idade da Figura 2.4 envolve a variável Créditos Anteriores (variável-filho da variável Idade) e a variável Sexo (variável-pai de uma variável-filho da variável Idade), note que a variável Idade não possui variáveis-pai, se estas existissem seriam consideradas na cobertura de Markov. Outro exemplo de cobertura de Markov pode ser visualizado na Figura 2.5 que exibe a cobertura de Markov para a variável A.

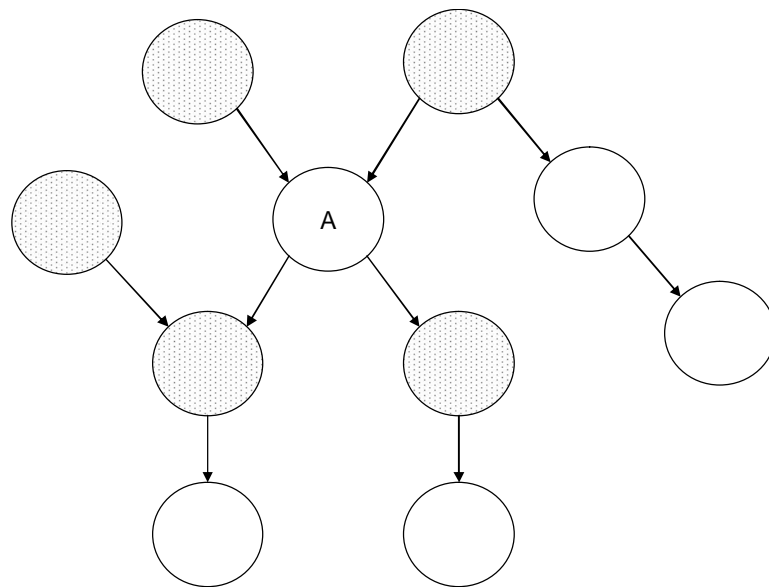


Figura 2.5. Cobertura de Markov de A representada pelas variáveis-nó cinza

2.4. A PROPRIEDADE DE D-SEPARAÇÃO

Através das propriedades markovianas, notamos que uma variável é independente de outra se não existe nenhum arco conectando-as. Porém, é possível definir independência quando existe entre as variáveis analisadas um grupo específico de variáveis, podendo ser um grupo de evidências, por exemplo.

Neste caso, surge a conceito de d-separação. Para defini-la consideremos alguns tipos de conexões dadas em Neopolitan (2004). Seja X , Z e Y variáveis de uma Rede Bayesiana (ξ, V) , definimos alguns tipos de conexão:

1. Se $X \rightarrow Z \rightarrow Y$, temos um relacionamento *head-to-tail*;
2. Se $X \leftarrow Z \rightarrow Y$, temos um relacionamento *tail-to-tail*;
3. Se $X \rightarrow Z \leftarrow Y$, temos um relacionamento *head-to-head*.

Além disso, podemos definir $A \subset V$, sendo X e $Y \in V - A$. Desta forma, para os casos 1 e 2, se consideramos que $Z \in A$, a variável Z bloqueará o caminho entre X e Y . Para o caso 3, se consideramos que Z e seus descendentes $\notin A$, a variável Z bloqueará o caminho entre X e Y . Se o caminho entre duas variáveis, ou conjunto de variáveis, é bloqueado, dizemos que essas variáveis, ou conjuntos, são d-separados.

A Figura 2.6, retirada de Marques e Dutra (1999), ilustra os três casos de d-separação, onde os conjuntos U e W são d-separados.

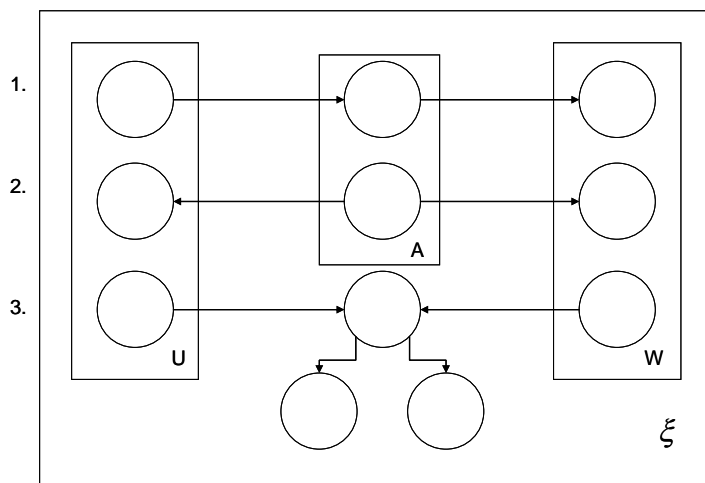


Figura 2.6. Tipos de d-separação, U e W d-separados

Maiores detalhes sobre d-separação são dados em Neapolitan (2004).

2.5. EQUIVALÊNCIA DE MARKOV

Existem inúmeras estruturas possíveis no enredo de Redes Bayesianas. Porém, podemos construir para cada conjunto de variáveis um grupo de estruturas extremamente semelhantes, chamadas de equivalentes de Markov.

Segundo Neapolitan (2004), dois grafos são equivalentes quando mantêm as mesmas independências condicionais. Ou seja, dois grafos são considerados equivalentes quando conservam as mesmas ligações de arcos entre as variáveis independente da direção, com exceção às ligações *head-to-head*, ou seja, quando uma variável-filho possui mais que uma variável-pai.

Assim, considere o exemplo da Figura 2.7.

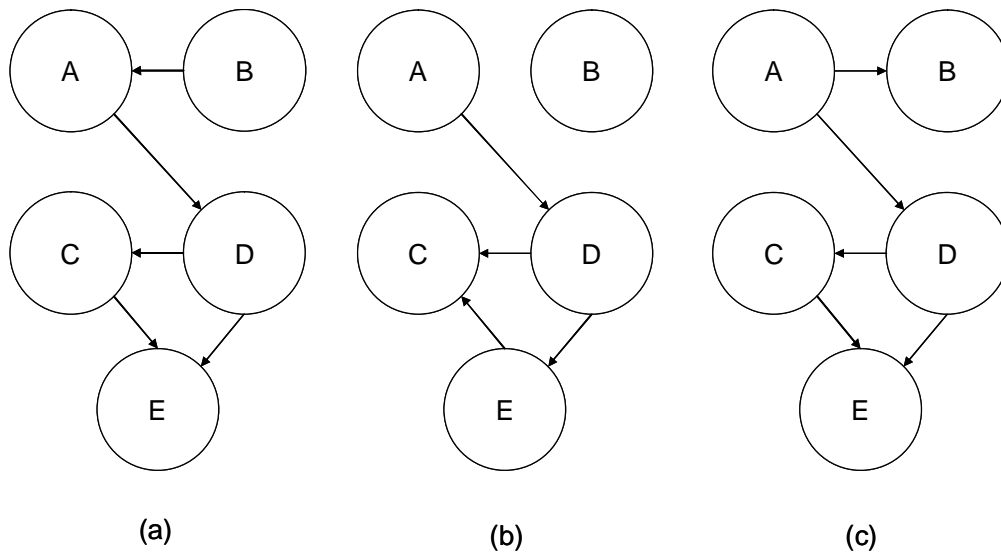


Figura 2.7. Exemplo de identificação de redes Bayesianas Markov equivalentes

Analisando a Figura 2.7, notamos que a estrutura (a) não é equivalente a (b), pois além de não preservar a conexão *head-to-head* $C \rightarrow E \leftarrow D$, a estrutura (b) não mantém a conexão entre as variáveis A e B. Esses mesmos motivos fazem (b) não equivalente à estrutura (c).

Comparando a estrutura (a) com (c), notamos que existe apenas diferença entre a direção de ligação entre as variáveis A e B, ou seja, (a) e (c) são equivalentes. Dizemos que (a) e (c) pertencem à mesma classe de equivalência markoviana.

2.6. MÉTODO GERAL PARA A CONSTRUÇÃO DE UMA REDE BAYESIANA

A construção de uma Rede Bayesiana não é trivial, além de existir vários métodos para a estimação de estruturas de rede através do conjunto de dados, os métodos podem ser influenciados por fatores como a ordem e escolha das variáveis que compõem o problema. Esse problema proporciona atualmente intensas pesquisas buscando um método ótimo para estimação de estruturas DAG para domínios de problemas práticos.

Porém, de uma forma geral, Pearl (1988) criou um algoritmo baseando-se nas propriedades 2.1 e 2.2, no qual, dado um conjunto de variáveis discretas ordenadas, constrói uma Rede Bayesiana única, adicionando às variáveis a rede em sua ordem e acrescentando arcos para a formação da estrutura. Assim, cada variável é conectada às variáveis antigas da rede, o que garante que a estrutura seja sempre acíclica.

O algoritmo de Pearl é dado abaixo.

1. Escolha um conjunto de variáveis X_i que em suposição descreva o problema;
2. Escolha uma ordem para as variáveis;
3. Para todas as variáveis em ordem, faça:
 - a. Escolha a variável X e adicione-a na rede;
 - b. Determine os pais da variável X dentre os nós que já estão na rede, que satisfaça $P[X_i | X_j, pais(X_i)] = P[X_i | pais(X_i)]$.
 - c. Construa a tabela de probabilidade condicional (CPT) para X .

Para uma Rede Bayesiana ser adequada, ela deve ser perfeita, ou seja, todos arcos devem expressar corretamente as dependências entre as variáveis.

Desta forma, é fácil notar que para a construção de uma Rede Bayesiana devemos escolher uma ordem correta para as variáveis, pois diferentes ordens podem gerar Redes Bayesianas diferentes. Desta forma, Korb e Nicholson (2004) sugerem que primeiramente consideremos as variáveis possíveis a serem raízes e suas variáveis independentes, a seguir as demais variáveis.

Outros métodos de construção de Redes Bayesianas serão apresentados no decorrer do trabalho.

2.7. COMENTÁRIOS FINAIS

Neste capítulo, foram apresentados conceitos básicos sobre a técnica de Redes Bayesianas, sendo estes de suma importância para o entendimento geral do método. Alguns dos conceitos mais importantes englobam a propriedade de d-separação, base para diversos tipos de cálculos, e a propriedade de cobertura de Markov utilizada em algoritmos para estimação de probabilidades condicionais.

Além disso, introduzimos a idéia básica para a criação de uma estrutura de Redes Bayesianas. Porém, a construção geral de uma estrutura não é trivial, além de existir vários métodos para este mesmo objetivo.

Neste contexto, nos próximos dois capítulos exibimos como inferências e estimações podem ser realizadas.

3. INFERÊNCIA EM REDES BAYESIANAS

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.

Porém, como visto anteriormente, existem diversos tipos de estruturas de Redes Bayesianas, assim as inferências probabilistas podem ser realizadas de diferentes formas assumindo tipos de estruturas específicas e com diferentes quantidades de variáveis.

Entretanto, existem fatores que dificultam o processo de inferência, como a complexidade da estrutura em análise e o número de variáveis. Ou seja, para poucas variáveis e uma estrutura gráfica dirigida acíclica (DAG) não complexa, a propagação das probabilidades pode ser facilmente calculada. À medida que existe um grande número de variáveis e estruturas mais complexas, existe grande dificuldade no cálculo das probabilidades da rede.

Marques e Dutra (1999) consideram que existe maior dificuldade quando o problema é modelado de uma forma complexa, ou seja, quando a estrutura considerada para a inferência possui relações complexas entre as variáveis. Assim, em algumas situações, uma rede com apenas uma dezena de variáveis pode necessitar de um tempo computacional muito grande para término da inferência (às vezes inviável), enquanto uma rede contendo milhares de variáveis e uma estrutura simplista pode levar apenas alguns instantes de processamento.

Buscando contornar tal problema, utilizamos algoritmos de inferência probabilística que facilitem os cálculos, sendo essa uma área de pesquisa em Redes Bayesianas.

Deste modo, alguns algoritmos são utilizados de forma particular para tipos gerais de estruturas, basicamente as estruturas citadas na seção 2.1.2.

Considerando as estruturas mais comuns, existem dois tipos de algoritmos para realizar inferência probabilista, estes são chamados de algoritmos exatos e algoritmos aproximados. Os algoritmos exatos produzem resultados mais satisfatórios, porém exigem um grande esforço computacional na presença de um alto número de variáveis na rede. Os algoritmos aproximados são construídos através de métodos de simulação, tendo uma precisão inferior, porém maior velocidade de processamento que os algoritmos exatos.

Neste capítulo, serão expostos os mais comuns tipos de algoritmos exatos e aproximados, bem como suas utilizações em diferentes tipos de estruturas de Redes Bayesianas.

3.1. ALGORITMOS DE INFERÊNCIA EXATA

Os algoritmos de inferência exata, de uma forma geral, são mais complexos computacionalmente. Na prática, são indicados quando existe um número de variáveis inferior a trinta e seis (Korb e Nicholson, 2004).

Alguns dos algoritmos exatos freqüentemente encontrados na literatura são:

- Passagem de mensagens;
- Método de formação de agrupamentos.

Esses algoritmos serão apresentados nesta seção.

3.1.1 Passagem de mensagem

Este algoritmo é utilizado para um tipo de estrutura de Redes Bayesianas chamada de árvores, tanto para árvores simples quanto para poli-árvores.

Tais estruturas são definidas na seção 2.1.2, porém, como uma breve revisão, temos que o termo “árvore simples” é utilizado para definir Redes Bayesianas com estrutura de grafos DAG de simples conexões que possuem apenas uma variável como raiz. Além disso, cada variável da rede possui precisamente um pai. O termo “poli-árvore” é utilizado para definir Redes Bayesianas com a mesma estrutura de árvore dada na definição de árvore simples, porém possui duas ou mais variáveis como raiz.

Um exemplo de árvore simples e outro de poli-árvore podem ser visualizados na Figura 3.1 e 3.2, respectivamente.

Notamos através da Figura 3.1 que a estrutura da rede apresenta apenas uma variável-raiz, a variável X. Além disso, todas as variáveis possuem apenas uma variável-pai. Para a Figura 3.2, notamos que a estrutura apresenta duas variáveis-raiz, neste caso existe pelo menos uma variável que possui pelo menos duas variáveis-pai.

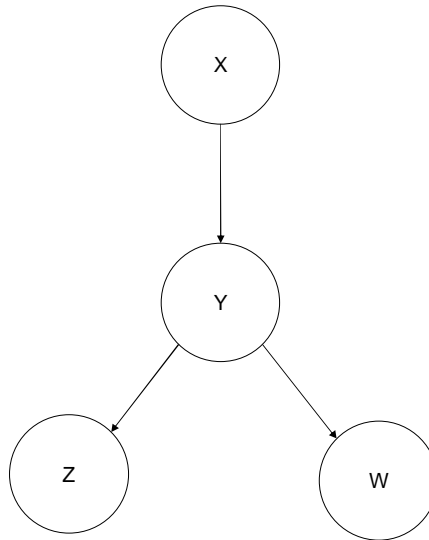


Figura 3.1. Rede Bayesiana com estrutura árvore simples

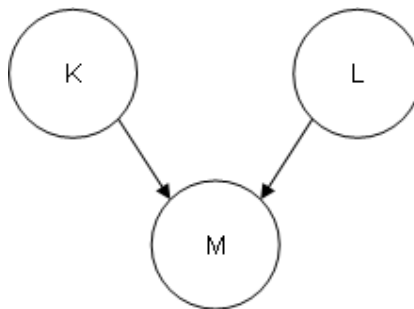


Figura 3.2. Rede Bayesiana com estrutura poli-árvore

O algoritmo de passagem de mensagem de Pearl baseia-se na repetição do teorema de Bayes considerando a estrutura de independência entre determinados conjuntos de variáveis presentes em uma Rede Bayesiana, proporcionada por propriedades markovianas como visto anteriormente. Assim, o teorema de Bayes é aplicado para variáveis, dadas as evidências vizinhas; tais variáveis passam para seus novos vizinhos, assim por diante até envolver toda a rede.

Para sua formalização, realizamos uma breve exposição probabilista de passagens importantes do algoritmo dentro da teoria de Redes Bayesianas. Algumas passagens são extraídas de uma demonstração completa encontrada em Neapolitan (2004).

Consideremos E um conjunto contendo todas as variáveis instanciadas, evidências da rede. Assim, tendo uma variável X como referência, podemos dividir esse conjunto em $E = \{E^+, E^-\}$, sendo que E^+ representa o conjunto de variáveis de evidências e ancestrais de X , e E^- representa o conjunto de variáveis-evidência descendentes de X . A visualização gráfica desta estrutura é dada pela Figura 3.2.

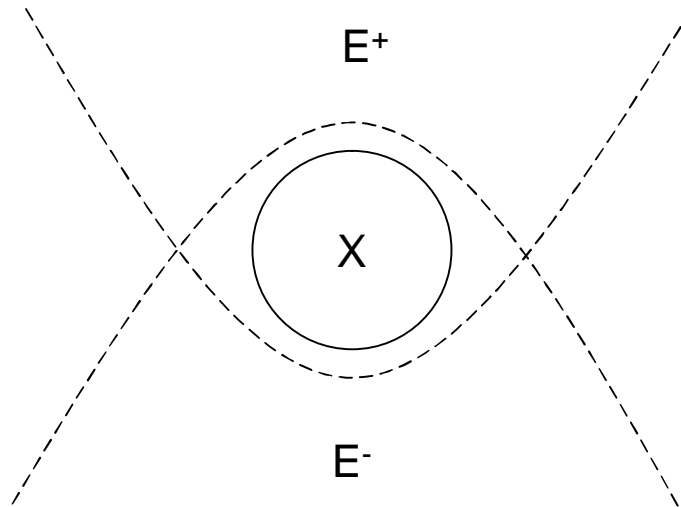


Figura 3.2. Dimensões do conjunto E de uma Rede Bayesiana

Estamos interessados em calcular a probabilidade de $P(X | E)$, ou seja, dado o conjunto de variáveis instanciadas, temos o objetivo de calcular a probabilidade de X .

Assim, seguem as seguintes passagens matemáticas.

$$\begin{aligned}
P(X | E) &= P(X | E^+, E^-) \\
&= P(X | E^+, E^-) = \frac{P(E^+, E^- | X)P(X)}{P(E^+, E^-)} \\
&= \frac{P(E^+ | X)P(E^- | X)P(X)}{P(E^+, E^-)} \\
&= \frac{P(X | E^+)P(E^+)P(E^- | X)P(X)}{P(X)P(E^+, E^-)} = \alpha P(X | E^+)P(E^- | X)
\end{aligned}$$

Para o desenvolvimento realizado utilizamos na segunda linha o Teorema de Bayes e na terceira linha a propriedade de d-separação vista na seção 2.4. Em sua finalização, consideramos α como constante, pois não depende de X , ou seja, $\alpha = \frac{P(E^-)}{P(E^-, E^+)}$.

Através deste desenvolvimento e considerando $P(X | E^+) = \pi(x)$ e $P(E^- | X) = \lambda(x)$, chegamos à igualdade (3.1).

$$P(X | E) = \alpha \lambda(x) \pi(x) \tag{3.1}$$

Korb e Nicholson (2007) interpretam os termos $\pi(x)$ e $\lambda(x)$, respectivamente, como verossimilhança e priori de X . Mas também consideram que os termos representam mensagens probabilísticas, ou seja, mensagens enviadas de uma variável para outra para a realização da inferência em toda a rede.

O algoritmo passagem de mensagem, como dito anteriormente, é aplicado para dois tipos de estrutura, a estrutura de árvore simples e a estrutura de poli-árvore. Desta forma, apresentamos os dois procedimentos respectivamente.

Todo o desenvolvimento destes algoritmos pode ser encontrado detalhadamente em Neapolitan (2004).

3.1.1.1. Passagem de mensagem para árvore simples

Generalizando esse procedimento para todas as variáveis da rede e adotando alternativas de fácil demonstração probabilística para as variáveis-raiz, evidências e variáveis-folha, temos o Algoritmo 3.1.

Algoritmo 3.1. Considere (G, θ) como uma Rede Bayesiana com a estrutura DAG de árvore simples e $G = \{V, \xi\}$, sendo V o conjunto de variáveis, E o conjunto de evidências e ξ a estrutura destas variáveis em grafos, sendo $E \subset V$. Para cada variável X são definidos mensagens λ , valores λ , mensagens π e valores π .

1. Mensagens λ

- a. Para cada Y variável-filho de X , para todos os valores de x ,

$$\lambda_Y(x) \equiv \sum_y P(y|x) \lambda(y) \quad (3.2)$$

2. Valores λ

- a. Se $X \in E$ e X possui o valor \hat{x} , ou seja, caso X seja instanciada,

$$\lambda(\hat{x}) \equiv 1 \text{ e } \lambda(x) \equiv 0 \quad (3.3)$$

- b. Se $X \notin E$ e X é uma folha, para todos os valores de x ,

$$\lambda(x) \equiv 1 \quad (3.4)$$

- c. Se $X \notin E$ e X não é uma folha, para todos os valores de x ,

$$\lambda(x) \equiv \prod_F \lambda_i(x) \quad (3.5)$$

onde F é indicador para todas as variáveis-filho da variável X .

3. Mensagens π

a. Se W é pai da variável X, então para todos os valores de x ,

$$\pi_x(w) \equiv \pi(w) \prod_F \lambda_i(w) \quad (3.6)$$

onde F indica para todas as variáveis-filho da variável W, exceto a variável X.

4. Valores π

a. Se $X \in E$ e X possui o valor \hat{x} , ou seja, caso X seja instanciada,

$$\pi(\hat{x}) \equiv 1 \text{ e } \pi(x) \equiv 0 \quad (3.7)$$

b. Se $X \notin E$ e X é uma raiz, para todos os valores de x ,

$$\pi(x) \equiv P(x) \quad (3.8)$$

c. Se $X \notin E$ e X não é uma raiz e W é pai de X, para todos os valores de x ,

$$\pi(x) \equiv \sum_w P(x | w) \pi_x(w) \quad (3.9)$$

5. Dadas as definições acima, para cada variável X cada um de seus respectivos valores x , chegamos à expressão (3.1).

Para ilustrar a aplicabilidade do algoritmo 3.1, seus cálculos são realizados através de um exemplo.

Exemplo 3.1: Considere um conjunto de dados constituído de 3 variáveis dicotômicas e 24 observações referentes à *credit scoring*, sendo as variáveis:

- Sexo { Masculino, Feminino };
- Créditos Anteriores { Um, Diferente de um };
- *Credit Rating* { Bom, Ruim }.

Para este problema considere a possível estrutura de Rede Bayesiana exibida na Figura 3.3.

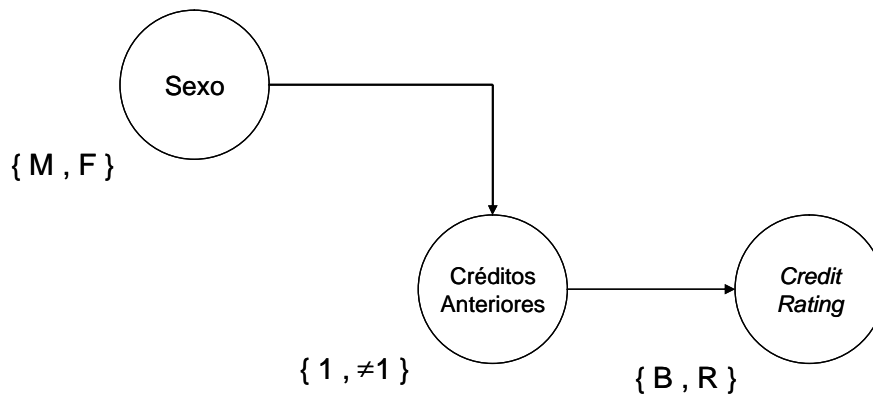


Figura 3.3. Possível Rede Bayesiana para dados aplicados à *credit scoring*

Realizamos a atualização de probabilidades considerando a variável Créditos Anteriores conhecida, ou seja, a variável Créditos Anteriores como evidência, sabemos apenas que um cliente possui somente um crédito anterior na instituição financeira. Desta forma, temos a estrutura apresentada na Figura 3.4.

Assim, aplicando as definições do Algoritmo 3.1, temos a seguinte resolução.

Primeiramente, devemos considerar a localização de cada variável na rede, ou seja, S é uma variável-raiz, CA é uma variável-evidência e CR é uma variável-folha. Desta forma, podemos estabelecer mensagens para cada valor de cada variável.

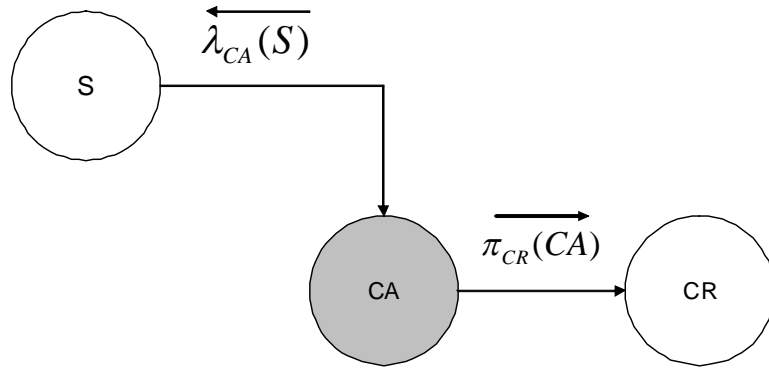


Figura 3.4. Mensagens para o Algoritmo de passagem de mensagem para dados de *credit scoring*

– Mensagens λ

Aplicando a propriedade 2.a do Teorema 3.1, temos:

$$\lambda(CA = 1) = 1$$

$$\lambda(CA \neq 1) = 0$$

Aplicando a propriedade 2.b do Teorema 3.1, temos:

$$\lambda(CR = Bom) = 1$$

$$\lambda(CR = Ruim) = 1$$

Aplicando as propriedades 2.c e 1.a do Teorema 3.1, temos:

$$\begin{aligned} \lambda(S = M) &= \lambda_{CA}(S = M) \\ &= P(CA = 1 | S = M)\lambda(CA = 1) + P(CA \neq 1 | S = M)\lambda(CA \neq 1) \\ &= 0.60 \times 1 + 0.40 \times 0 \\ &= 0.60 \end{aligned}$$

$$\begin{aligned} \lambda(S = F) &= \lambda_{CA}(S = F) \\ &= P(CA = 1 | S = F)\lambda(CA = 1) + P(CA \neq 1 | S = F)\lambda(CA \neq 1) \\ &= 0.58 \times 1 + 0.42 \times 0 \\ &= 0.58 \end{aligned}$$

– Mensagens π

Aplicando a propriedade 4.a do Teorema 3.1, temos:

$$\pi(CA = 1) = 1$$

$$\pi(CA \neq 1) = 0$$

Aplicando a propriedade 4.b do Teorema 3.1, temos:

$$\pi(S = M) = P(S = M)$$

$$= 0.79$$

$$\pi(S = F) = P(S = F)$$

$$= 0.21$$

Aplicando a propriedade 3 do Teorema 3.1, temos:

$$\pi_{CR}(CA = 1) = \pi(CA = 1)$$

$$\pi_{CR}(CA \Rightarrow 1) = \pi(CA \Rightarrow 1)$$

Note que CA possui apenas CR de filho na rede.

Aplicando a propriedade 4.c e os resultados da propriedade 3, temos:

$$\pi(CR = Bom)$$

$$= P(CR = Bom | CA = 1)\pi_{CR}(CA = 1) + P(CR = Bom | CA \neq 1)\pi_{CR}(CA \neq 1)$$

$$= 0.71 \times 1 + 0.60 \times 0$$

$$= 0.71$$

$$\pi(CR = Ruim)$$

$$= P(CR = Ruim | CA = 1)\pi_{CR}(CA = 1) + P(CR = Ruim | CA \neq 1)\pi_{CR}(CA \neq 1)$$

$$= 0.29 \times 1 + 0.40 \times 0$$

$$= 0.29$$

Desta forma, aplicamos a propriedade 5 para cada um dos valores de cada variável da rede.

$$\begin{aligned}
P(S = M \mid CA = 1) \\
&= \alpha \lambda(S = M) \pi(S = M) \\
&= \alpha 0.58 \times 0.79 \\
&= \alpha 0.4740
\end{aligned}$$

$$\begin{aligned}
P(S = F \mid CA = 1) \\
&= \alpha \lambda(S = F) \pi(S = F) \\
&= \alpha 0.58 \times 0.21 \\
&= \alpha 0.1218
\end{aligned}$$

Normalizando os resultados, temos:

$$P(S = M \mid CA = 1) = \frac{\alpha 0.4740}{\alpha 0.4740 + \alpha 0.1218} = 0.796$$

$$P(S = F \mid CA = 1) = \frac{\alpha 0.1218}{\alpha 0.4740 + \alpha 0.1218} = 0.204$$

Aplicando a definição (1) para a variável *Credit Rating*, temos:

$$\begin{aligned}
P(CR = Bom \mid CA = 1) &= 0.71 \\
P(CR = Ruim \mid CA = 1) &= 0.29
\end{aligned}$$

Notamos que a probabilidade $P(CR = Bom \mid CA)$ é exatamente igual à probabilidade obtida via tabela de probabilidade condicional (CPT), pois CR é filho direto de CA. Ou seja, o algoritmo é necessário quando a variável-evidência não é pai direto de uma determinada variável, porém é capaz de calcular a probabilidade de qualquer variável dada, uma ou mais variáveis-evidência na rede.

3.1.1.2. Passagem de mensagem para poli-árvore

Este algoritmo é extremamente semelhante ao procedimento de inferência para árvore simples, porém é embasado na existência de duas ou mais variáveis-raiz e considera que ao menos uma variável da rede possui pelo menos duas variáveis-pai.

Algoritmo 3.2. Considere (G, θ) como uma Rede Bayesiana com a estrutura DAG de poli-
 árvore e $G = \{V, \xi\}$, sendo V o conjunto de variáveis e ξ a estrutura destas variáveis em
 grafos e sendo $E \subset V$. Para cada variável X são definidos mensagens λ , valores λ ,
 mensagens π e valores π .

1. Mensagens λ

a. Para cada Y variável-filho de X , para todos os valores de x ,

$$\lambda_Y(x) \equiv \sum_y \left[\sum_i \left(P(y | x, w_1, w_2, \dots, w_k) \prod_k \pi_Y(w_i) \right) \right] \lambda(y) \quad (3.10)$$

onde F é indicador para todas as variáveis-filho da variável X .

2. Valores λ

a. Se $X \in E$ e X possui o valor \hat{x} , ou seja, caso X seja instanciada,

$$\lambda(\hat{x}) \equiv 1 \text{ e } \lambda(x) \equiv 0 \quad (3.11)$$

b. Se $X \notin E$ e X é uma folha, para todos os valores de x ,

$$\lambda(x) \equiv 1 \quad (3.12)$$

c. Se $X \notin E$ e X não é uma folha, para todos os valores de x ,

$$\lambda(x) \equiv \prod_F \lambda_i(x) \quad (3.13)$$

onde F é indicador para todas as variáveis-filho da variável X .

3. Mensagens π

a. Se W é pai da variável X , então para todos os valores de x ,

$$\pi_X(w) \equiv \pi(w) \prod_F \lambda_i(w) \quad (3.14)$$

onde F indica para todas as variáveis-filho da variável W , exceto a variável X .

4. Valores π

a. Se $X \in E$ e X possui o valor \hat{x} , ou seja, caso X seja instanciada,

$$\pi(\hat{x}) \equiv 1 \text{ e } \pi(x) \equiv 0 \quad (3.15)$$

b. Se $X \notin E$ e X é uma raiz, para todos os valores de x ,

$$\pi(x) \equiv P(x) \quad (3.16)$$

c. Se $X \notin E$ e X não é uma raiz e W_1, \dots, W_j são variáveis-pai de X , para todos os valores de x ,

$$\pi(x) \equiv \sum_w \left(P(x | w_1, w_i, \dots, w_j) \prod_w \pi_x(w_i) \right) \quad (3.17)$$

5. Dadas as definições acima, para cada variável X cada um de seus respectivos valores x , chegamos à expressão (3.1).

Notamos que a maior diferença entre os dois algoritmos é a generalização para a possibilidade da presença de mais de uma variável-pai.

Segundo Korb e Nicholson (2004), as notações utilizadas através letras gregas λ e π podem parecer confusas a primeira vista, mas, em suma, podemos definir as mensagens dentro da estrutura inteligente da modelagem como:

- π : notação de mensagem probabilística utilizada na direção do arco, de pai para filho, ou seja, $\pi_{filho}(pai)$;
- λ : notação de mensagem probabilística utilizada contra a direção do arco, de filho para pai, ou seja, $\lambda_{pai}(filho)$.

3.1.2. Método de formação de agrupamentos

Para a realização do cálculo de propagação de probabilidades em estruturas de múltiplas conexões um tempo muito extenso é necessário. Por esse motivo, um dos métodos comumente utilizados para a redução de esforço computacional é o de formação de agrupamentos, também conhecidos como algoritmos de formação de árvores.

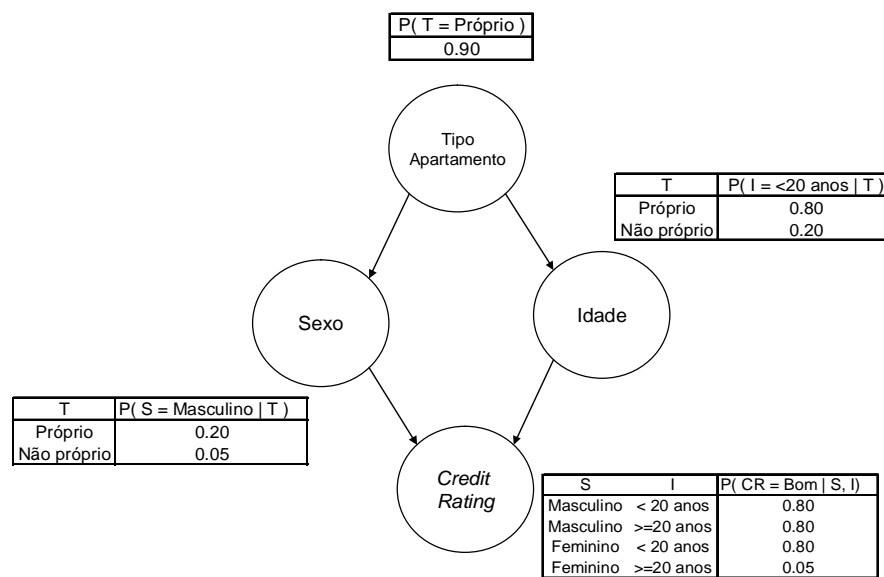


Figura 3.5. Estrutura de Múltiplas conexões para dados de *credit scoring*

A idéia principal deste método é manipular uma Rede Bayesiana de estrutura de múltiplas conexões a fim de transformá-la em uma estrutura do tipo árvore (poli ou simples). A partir daí, aplicamos o algoritmo de passagem de mensagem, ou outro algoritmo desejado para este tipo de estrutura.

Exemplo 3.2: Para ilustrar o procedimento, consideramos algumas variáveis utilizadas em exemplos anteriores, acrescentando à variável tipo de moradia. Deste modo, temos as letras S, I, T e CR representando respectivamente as variáveis:

- Sexo { Masculino, Feminino };
- Idade { <20 anos; ≥ 20 anos };
- Tipo de Moradia { Própria, Não própria };
- *Credit Rating* { Bom, Ruim }.

Para tais variáveis supomos a Rede Bayesiana com estrutura de múltiplas conexões exibida na Figura 3.4. Em critério de simplificação, algumas categorias foram ocultadas, porém são facilmente verificadas.

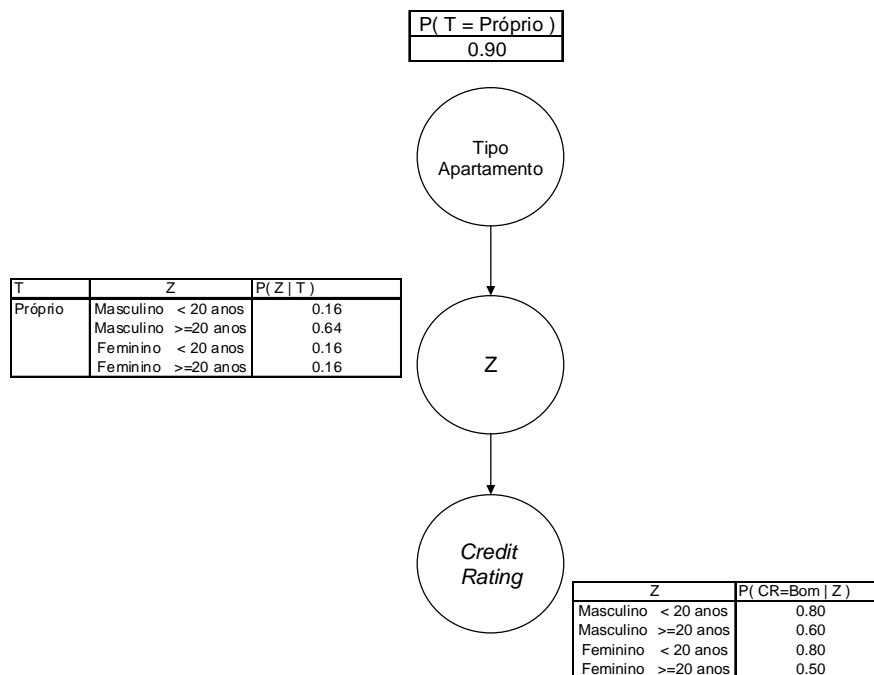


Figura 3.6. Estrutura após o *clustering* para dados de *credit scoring*

Assim, aplicamos um simples método de formação de agrupamentos, mais conhecido como *clustering*, ou seja, unimos as categorias das variáveis Sexo e Idade, transformando-as em uma única variável cruzada. Ou seja, criamos uma variável nova, sendo $Z = \{(\text{Masculino}, <20 \text{ anos}), (\text{Feminino}, <20 \text{ anos}), (\text{Masculino}, \geq 20 \text{ anos}), (\text{Feminino}, \geq 20 \text{ anos})\}$. A junção é apresentada na Figura 3.6.

Desta forma, a estrutura se transforma de múltiplas conexões para árvores simples e o método de passagem de mensagem pode ser aplicado.

Esse método é uma ferramenta de fácil entendimento e de aplicação, porém seu processo pode ser computacionalmente lento, devido ao tempo extenso para cálculo da propagação. Além disso, a utilização deste método pode gerar gigantescas tabelas de CPT, pois estas aumentam quando existe uma alta complexidade de ligação entre as variáveis. Mesmo assim, algoritmos do tipo de junção de árvore são amplamente utilizados em ferramentas comerciais em Redes Bayesianas.

3.2. ALGORITMOS DE INFERÊNCIA APROXIMADA

Para contornar o problema de dependências complexas que ocasionam um grande intervalo de tempo para a resolução de inferências exatas, os métodos de inferência aproximada são uma poderosa alternativa. Tais algoritmos baseiam-se em processos de simulação dentro da estrutura da rede, onde a cada passo uma mesma estatística é observada. Neste caso, todas elas são relacionadas à probabilidade posteriori da variável consulta dado um conjunto de evidências.

De uma forma geral, estes métodos são uma forma de extrair amostras de uma distribuição difícil de amostrar, com base em uma distribuição fácil de amostrar. Ou seja, extrair amostras de uma estrutura da rede em estudo dado evidências, para inferir para a distribuição condicional de uma variável específica, dado um conjunto de evidências.

Alguns dos algoritmos exatos freqüentemente encontrados na literatura e citados por Russel e Norvig (2004) são:

- Amostragem por Rejeição;
- Ponderação de Probabilidade;
- Gibbs Sampling.

Estes algoritmos serão apresentados subseqüentemente nesta seção.

3.2.1. Amostragem por Rejeição

A idéia básica deste algoritmo é percorrer toda a rede, dada a estrutura de arcos e CPTs, gerando N amostras e verificando se aquele tipo de amostra é consistente, ou seja, se ela contém a evidência dada. Caso isso não ocorra, tal amostra é rejeitada.

Desta forma, a probabilidade condicional para uma variável em questão X dada uma evidência e é estimada pela contagem do número de ocorrências em cada uma de suas categorias, dividida pelo número de amostras consistências.

Genericamente:

$$P(X | e) = \frac{P(X, e)}{P(e)} \cong \frac{\text{Contagem das categorias de X}}{\text{Número de amostras consistentes}} \quad (3.18)$$

O algoritmo para este tipo de amostragem é dado abaixo:

Algoritmo 3.3:

Indique:

X	Variável consulta;
E=e	Variáveis instanciadas (evidência);
(ξ, θ)	Estrutura da Rede Bayesiana;
N	Número de repetições a serem realizadas;
C[x _j]	Contador para cada categoria j da variável consulta.

Para i=1 até N

x_j= amostre com base em RB;

Se x_j é consistente com e então

C[x_j]= C[x_j]+1;

Retorne $\frac{C[x_j]}{\sum_j C[x_j]}$

Exemplo 3.3: Considerando a Rede Bayesiana da Figura 3.5, vamos supor que desejamos estimar $P(Idade | Sexo = M)$. Então, geramos uma quantidade fixa de amostras, neste caso igual a 100. Desta forma, verificamos para cada amostra se esta possui ou não a evidência desejada, caso não possua, o algoritmo rejeita a amostra. Por exemplo, a amostra $\langle T=Próprio, Sexo=M, Idade=<20anos, Credito=B \rangle$ não é rejeitada, pois se enquadra com a evidência da rede (sexo = masculino).

Realizando a simulação das 100 amostras, apenas 18 são consistentes com a evidência, dentre elas, 3 com idade maior que 20 anos e 15 com a idade menor ou igual há 18 anos. Assim, de (3.18) temos:

$$P(Idade | Sexo = M) = \frac{\langle 15,3 \rangle}{18} = (0.833, 0.166)$$

Ou seja,

$$P(Idade \Rightarrow 20 | Sexo = M) = 0.166$$

$$P(Idade \leq 20 | Sexo = M) = 0.833$$

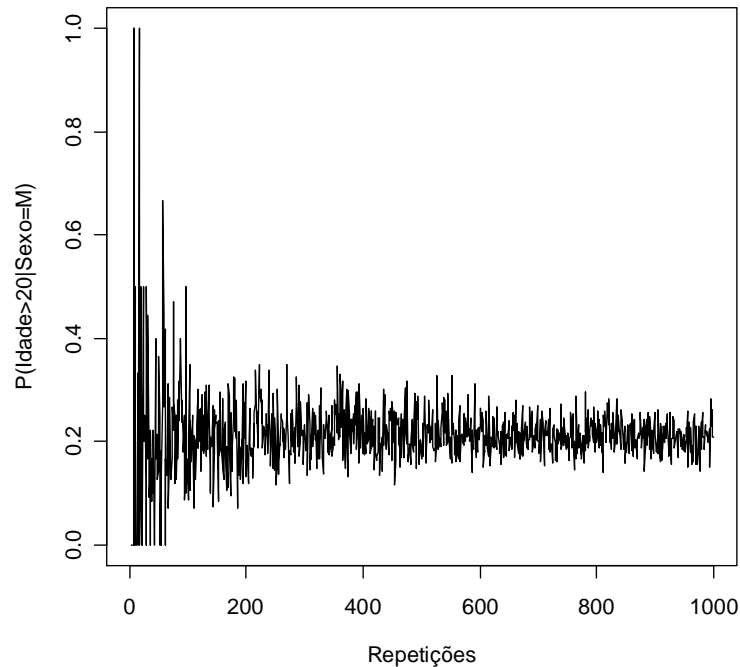


Figura 3.7. Convergência de probabilidade em amostragem por rejeição

Quanto maior o número de repetições, maior será a precisão de estimação. Desta forma, construímos o Gráfico da Figura 3.7, realizando para esse exemplo repetições de 1 a 1000 e verificando a probabilidade $P(Idade \Rightarrow 20 | Sexo = M)$.

Notamos através da Figura 3.7 que a probabilidade em foco converge para 0.21, sendo que existe instabilidade para repetições de tamanho 0 a 200.

Esse algoritmo torna-se preciso com o aumento do número de repetições, porém seu maior problema é o descarte de um grande número de amostras geradas, note que anteriormente apenas 18 amostras foram utilizadas em um total de 100 repetições.

Segundo Russel e Norvig (2004), o número de amostras consistentes é ainda menor quando aumentamos o número de variáveis-evidência.

3.2.2. Ponderação de Probabilidade

O algoritmo de ponderação de probabilidade evita o descarte de amostras que ocorre para a amostragem por rejeição. Desta forma, gera apenas amostras que possuem o valor instanciado da variável-evidência, ou seja, fixa valores para as evidências e amostra as demais variáveis. Cada evento associado à variável em estudo possui um peso específico, medida pelo produto das probabilidades condicionais para cada variável-evidência, dadas suas respectivas variáveis-pai.

Os procedimentos deste algoritmo são dados abaixo:

Algoritmo 3.4:

Indique:

X	Variável consulta;
E=e	Variáveis instanciadas (evidência);
(ξ, θ)	Estrutura da Rede Bayesiana;
N	Número de repetições a serem realizadas;
W[x _j]	Contador de pesos cada categoria j da variável consulta.

Para i=1 até N

X_j, w = Amostrar {

Se X_j é uma evidência

W=w P(E_j|pais(E_j))

```

        Senão  $X_j = \text{amostra aleatória } P(X_j | \text{pais}(X_j))$ 
    }
     $W[x_j] = W[x_j] + w$ 

    Retorne  $\frac{W[x_j]}{\sum_j W[x_j]}$ 

```

Exemplo 3.4: Considerando novamente a Rede Bayesiana da Figura 3.4, aplicamos esse algoritmo com 100 repetições, sendo suas probabilidades estimadas:

$$P(\text{Idade} \Rightarrow 20 \mid \text{Sexo} = M) = 0.258$$

$$P(\text{Idade} \leq 20 \mid \text{Sexo} = M) = 0.742$$

Esse algoritmo possui rápida convergência, sendo esses valores de probabilidade mais precisos do que os gerados pelo algoritmo de rejeição. Para ilustrar a convergência do algoritmo de Ponderação de Probabilidade, observamos a Figura 3.8.

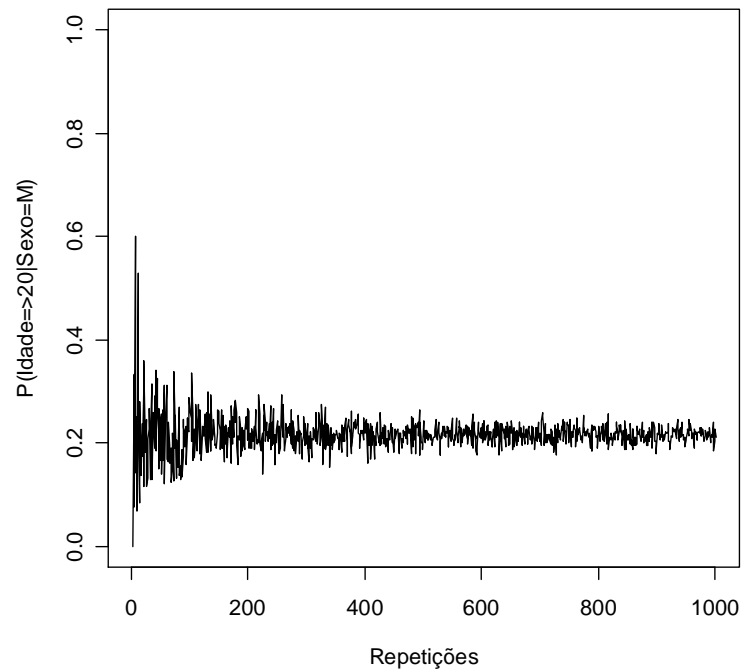


Figura 3.8. Convergência de probabilidade em ponderação de probabilidade

Segundo Russel e Norvig (2004), esse algoritmo é muito eficiente, porém pode sofrer degradação quando o número de evidências aumenta.

3.2.3. Gibbs Sampling

Outro tipo comum de algoritmo para inferência é o Gibbs Sampling, ou Amostrador de Gibbs, sendo este um método particular de MCMC. A idéia básica deste algoritmo é gerar cada evento a partir do estado atual da rede, ou seja, podemos imaginar que a rede se encontra em um determinado estado e cada variável possui um valor específico. Assim, o próximo estado é gerado por uma passagem aleatória dependente do estado atual. Para esse tipo de procedimento, as variáveis de evidências são consideradas fixas.

Conceitos importantes como probabilidade de transição e cobertura de Markov são importantes para a aplicação desse algoritmo.

Mesmo assim, tal algoritmo não é freqüentemente implementado em *softwares* básicos de Redes Bayesianas, devido a sua complexidade de manipulação das probabilidades para cada estado da rede. Um *software* indicado na literatura para realizar esse tipo de inferência é o *software* WinBUGS (www.mrc-bsu.cam.ac.uk/bugs/winbugs/).

Maiores detalhes e uma aplicação mais profunda sobre o Amostrador de Gibbs em Redes Bayesianas podem ser encontradas em Russel e Norvig (2004).

3.3. COMENTÁRIOS FINAIS

Observamos que o termo “inferência”, neste contexto, está intimamente ligado à verificação de probabilidades condicionais após a observação de determinadas informações, denominadas evidências. Para isso, utilizamos métodos específicos para diferentes tipos de estrutura.

Com isso, notamos que, dependendo do tipo de algoritmo utilizado, teremos esforços computacionais diferenciados de processamento, relacionados ao alto grau de complexidade do algoritmo e da estrutura em questão.

Assim, algoritmos são freqüentemente utilizados para o cumprimento de objetivos em Redes Bayesianas, um deles é o de estimação de estrutura e de probabilidades condicionais. Estes métodos serão apresentados no capítulo seguinte.

4. APRENDIZAGEM EM REDES BAYESIANAS

O termo “aprendizado” é muito comum no contexto de Inteligência Artificial e denota a assimilação de experiência que gera a capacidade de um agente ou sistema obter sucesso em determinada tarefa.

Desta forma, o aprendizado estatístico está intimamente ligado ao processo de aprendizagem quando existem incerteza e variabilidade. Para isso, através de um conjunto de dados, utilizamos o processo de estimação e validação do sistema em estudo, sendo aplicada qualquer técnica estatística que se enquadre ao domínio do problema.

Assim, devido à dificuldade da construção de uma Rede Bayesiana unicamente consultando um especialista, existe o interesse de se estimar todos os elementos da rede, estes sendo sua estrutura, e as probabilidades condicionais de cada CPT, também chamadas de parâmetros ou elementos numéricos.

Até o presente momento, assumimos que as estruturas e as probabilidades condicionais já estavam definidas. Porém, a partir de agora temos o interesse de estimar a rede por completo.

Neste capítulo, exibimos de uma forma rápida métodos para estimação conhecidos na literatura. Assim, apresentamos métodos específicos para ambos objetivos, a estimação de parâmetros e a estimação de estrutura.

4.1. ESTIMAÇÃO DE PARÂMETROS

Neste caso, estamos interessados em estimar as probabilidades condicionais para cada variável-nó da rede. Estes procedimentos podem ser realizados para conjuntos de dados completos e incompletos, sendo aqui apresentado apenas o método de estimação para dados completos.

Porém, um procedimento utilizado quando a base de dados é incompleta é o algoritmo EM. Basicamente, se alguma variável possui uma falta de informação, também conhecido como *missing*, este algoritmo utiliza os casos observados para estimar os valores faltantes. Esse método é utilizado pelo *software* Hugin, também para dados completos assumindo o conjunto de *missing* como vazio.

Além disso, a estimação pode ser realizada utilizando estimadores de máxima verossimilhança e estimadores bayesianos, tais abordagens serão dadas nas Seções 4.1.1 e 4.1.2, respectivamente.

4.1.1. Estimação via Máxima verossimilhança

Esse processo de estimação é extremamente simples, não consideram nenhum tipo de conhecimento a priori, sendo suas estimativas baseadas em frequências relativas e contagens através da base de dados.

Para esta abordagem, considere que cada variável-nó X_i possua r_i estados possíveis, sendo indicados por $x_i^1, x_i^2, x_i^3, \dots, x_i^{r_i}$, dado o j -ésimo pai pa_i e estrutura ξ conhecida. Assim temos (4.1).

$$P(X_i = x_i^k \mid pa_i^j, \xi) = \theta_{ijk} = \frac{P(X_i = x_i^k, pa_i^j)}{P(pa_i^j)} \cong \frac{f(x_i^k, pa_i^j)}{f(pa_i^j)} \quad (4.1)$$

onde $f(\cdot)$ denota frequência relativa.

Note que nenhuma suposição a priori foi dada sobre qualquer um dos elementos em análise. Porém, a forma mais clara de exibir tal pensamento é através de um exemplo.

Exemplo 4.1. Considere um conjunto de dados constituído de 3 variáveis dicotômicas e 24 observações referentes à *credit scoring*, sendo as variáveis:

- Sexo { Masculino, Feminino };
- Créditos Anteriores { Um, Diferente de um };
- *Credit Rating* { Bom, Ruim }.

Assim, o conjunto de dados é exposto na Tabela 4.1. Para este problema considere a possível estrutura de Rede Bayesiana exibida na Figura 4.1.

Através da Figura 4.1, notamos que existe apenas uma variável-raiz e todas as demais variáveis possuem somente uma variável-pai.

Tabela 4.1. Conjunto de dados referentes à *credit scoring*.

<i>Credit Rating</i>	<i>Sexo</i>	<i>Créditos Anteriores</i>
Ruim	Masculino	Diferente de um
Bom	Masculino	Um
Ruim	Feminino	Diferente de um
Bom	Feminino	Diferente de um
Bom	Masculino	Um
Bom	Masculino	Um
Ruim	Masculino	Um
Bom	Masculino	Diferente de um
Bom	Masculino	Diferente de um
Ruim	Masculino	Diferente de um
Ruim	Masculino	Um
Ruim	Feminino	Um
Ruim	Masculino	Diferente de um
Bom	Masculino	Diferente de um
Bom	Feminino	Um
Bom	Masculino	Um
Bom	Masculino	Um
Ruim	Feminino	Um
Bom	Masculino	Um
Bom	Masculino	Um
Bom	Masculino	Um
Bom	Masculino	Diferente de um
Bom	Masculino	Diferente de um
Bom	Masculino	Um

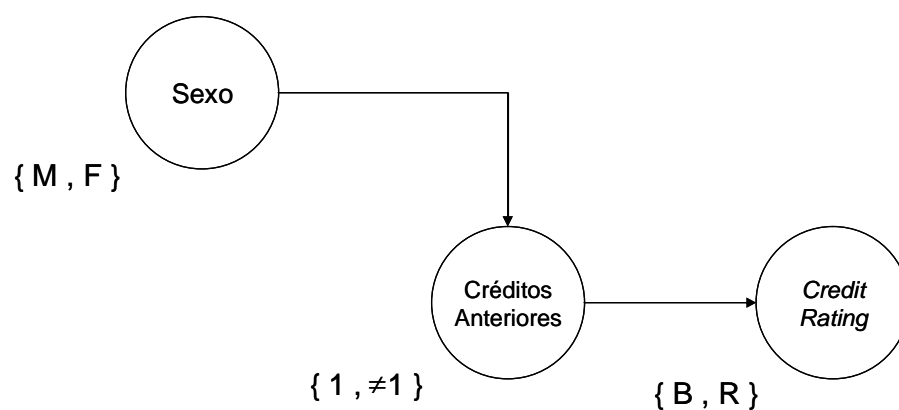


Figura 4.1. Possível Rede Bayesiana para dados aplicados a *credit scoring*

Para facilitar os cálculos, a variável *Sexo* será representada pela letra *S*, a variável *Créditos Anteriores* pela sigla *CA*, e a variável *Credit Rating* pela sigla *CR*.

Desta forma, levando em consideração a estrutura de relacionamento apresentada, necessitamos dos cálculos das probabilidades $P(S)$, $P(CA|S)$ e $P(CR|CA)$.

Assim, $P(S)$ é estimada facilmente através da frequência relativa calculada via Tabela 3.1. Para o cálculo das probabilidades $P(CA|S)$ e $P(CR|CA)$, partimos de tabelas de distribuição conjunta obtidas das tabelas cruzadas entre as variáveis de interesse. As probabilidades conjuntas $P(CA,S)$ e $P(CR,CA)$ são estimadas através das Tabelas 4.2 e 4.3, respectivamente.

Tabela 4.2. Probabilidade conjunta $P(CA,S)$

		S		Total
		F	M	
CA	1	0.13	0.46	0.58
	$\neq 1$	0.08	0.33	0.42
Total		0.21	0.79	1.00

Tabela 4.3. Probabilidade conjunta $P(CR,S)$

		CA		Total
		1	$\neq 1$	
CR	Ruim	0.17	0.17	0.33
	Bom	0.42	0.25	0.67
Total		0.58	0.42	1.00

Note que em cada tabela as células referentes ao total são as probabilidades marginais de cada categoria, ou seja, para a Tabela 4.2 a probabilidade marginal da variável CA. Fixando CA em 1, a probabilidade marginal é dada por $P(CA = 1) = 0.58$.

Assim, através do Teorema de Bayes visto na seção 1.3.2.4, no qual, por exemplo, $P(CR|CA) = P(CR,CA)/P(CA)$, realizamos o cálculo de cada célula de probabilidade conjunta dividida por sua respectiva célula de probabilidade marginal.

As probabilidades condicionais $P(CA|S)$ e $P(CR|CA)$ são estimadas através das Tabelas de probabilidade condicionais (CPT) 4.4 e 4.5, respectivamente.

Tabela 4.4. Probabilidade condicional $P(CA | S)$

		S	
		F	M
CA	1	0.60	0.58
	$\neq 1$	0.40	0.42

Tabela 4.5. Probabilidade condicional $P(CR | CA)$

		CA	
		1	$\neq 1$
CR	Ruim	0.29	0.40
	Bom	0.71	0.60

Deste modo, a Rede Bayesiana pode ser expressa pela Figura 4.2.

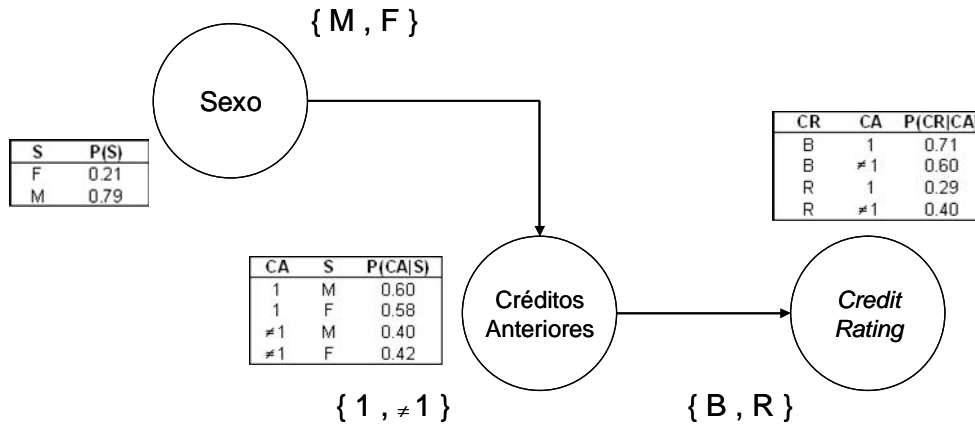


Figura 4.2. Possível Rede Bayesiana com CPT para dados de *credit scoring*

4.1.2 Estimação Bayesiana

Considere θ o parâmetro numérico da rede, sendo instância da variável Θ e a Rede Bayesiana em estudo com estrutura ξ conhecida. Desta forma, temos que X representa todo o conjunto de dados fornecido. Assim, nosso problema se resume em estimar a $P(\Theta | X, \xi)$.

Desta forma temos as seguintes suposições:

- X é uma amostra aleatória de dados completos com distribuição multinomial.
- θ_{ij} representa todas as probabilidade da variável X_i em qualquer um de seus estados, dados seus pais no j -ésimo estado e sendo independentes dada a amostra X .

$$P(\theta | X, \xi) = \prod_i \prod_j P(\theta_{ij} | X, \xi)$$

Dadas as suposições anteriores, podemos assumir que $\theta_{ij} | \xi \sim \text{Dirichlet}(\theta_{ij} | a_{ij1}, \dots, \alpha_{ijr_i})$ e temos que $\theta_{ij} | X, \xi \sim \text{Dirichlet}(\theta_{ij} | a_{ij1} + N_{ij1}, \dots, \alpha_{ijr_i} + N_{ijr_i})$, como visto na seção 1.3.2.4. Sendo N_{ijk} a frequência na amostra X com que a variável X_i assume o k -ésimo estado, condicionada ao j -ésimo estado dos pais.

Assim, tomando a média da distribuição de $\theta_{ij} | X, \xi$ como estimador de θ_{ijk} temos que (4.2).

$$P(X_i = x_i^k | pa_i^j, \xi) = \frac{(1 + N_{ijk})}{(r_i + N_{ij})} \quad (4.2)$$

Para maiores detalhes da formulação deste procedimento para o aprendizado dos parâmetros de uma rede Bayesiana, consultar Luna (2004).

Exemplo 4.2: Para aplicação desta técnica, considere o conjunto de dados de exemplo 4.1, mais especificamente a Tabela 4.3. Assim, podemos construir a nova tabela a seguir.

Tabela 4.6. Frequência CR, CA

		CA		Total
		1	$\neq 1$	
CR	Ruim	4	4	16
	Bom	10	6	8
Total		14	10	24

Assim, podemos realizar os cálculos a partir de (4.2).

$$P(CR = Ruim | CA = 1) = \frac{(1 + N_{CR,1,Ruim})}{(r_{CA} + N_{CR,1})} = \frac{(1 + 4)}{(2 + 14)} = 0.312$$

Tabela 4.7. Probabilidade condicional $P(CR | CA)$

		CA	
		1	$\neq 1$
CR	Ruim	0.312	0.417
	Bom	0.688	0.583

Note que os valores da Tabela 4.7 são bastante similares aos encontrados na Tabela 4.5.

4.2. ESTIMAÇÃO DE ESTRUTURA

Para este tipo de estimação estamos interessados na busca da melhor estrutura de Redes Bayesianas para um determinado conjunto de dados, ou seja, a melhor disposição de dependências e independências entre as variáveis que explique melhor o problema em estudo.

Segundo Hrushchka (2003), a estimação de estrutura de uma Rede Bayesiana, também conhecida na literatura como aprendizado de estrutura, pode ser dividida em duas partes: a primeira baseada em uma busca heurística e a segunda baseada no conceito de independência condicional dos atributos da rede. Assim, algoritmos são requeridos para ambos os tipos de estimação.

Os algoritmos de busca heurística pesquisam a melhor estrutura com base na busca de uma pontuação adequada, assim, começam com uma rede sem arcos e, gradativamente, adicionam arcos ligando variável a variável, analisando um determinado *score* em cada passagem. Por fim, indica como sendo a melhor estrutura aquela com o máximo *score* obtido. Uma desvantagem desse tipo de algoritmo é que ele depende diretamente da ordenação inicial das variáveis.

Os algoritmos baseados no conceito de independência condicional utilizam a propriedade de d-separação (Seção 2.4), o que diminui significativamente o esforço computacional.

Porém, Magalhães (2007) aponta a aplicação de métodos híbridos como uma terceira alternativa para estimação de estrutura, os quais se utilizam de uma composição dos algoritmos de busca por pontuação e dos baseados em propriedades de d-separação.

Bottcher e Dethlefsen (2003) sugerem um algoritmo híbrido implementado no pacote *deal* do *Software R*.

A abordagem aqui apresentada é meramente informativa, pois existem várias abordagens sobre algoritmos de aprendizado de estrutura, sendo que a área está em constante desenvolvimento (Russell e Norvig, 2004). Assim, nesta seção, apresentamos de forma sucinta o algoritmo K2, que busca maximizar a métrica de determinada função, também apresentamos o algoritmo PC, implementado no *Software Hugin*, como sendo um algoritmo baseado em propriedades de d-separação. Além disso, apresentamos como realizar o aprendizado híbrido utilizando o *software R*.

4.2.1 Algoritmo PC

O algoritmo PC foi proposto por Spirtes, Glymour e Scheines (1991), levando assim no nome as iniciais de seus principais criadores, Peter Spirtes e Clark Glymour. A idéia básica do algoritmo é realizar testes estatísticos para determinar grupos de variáveis independentes, utilizando o critério de d-separação. Geralmente, o teste utilizado é o teste estatístico de Qui-quadrado, sendo calculado através do conjunto de dados (Abellan et al, 2006). Assim, os testes são realizados a cada par de variáveis da rede. Tal processo considera que se o teste é significativo, as variáveis se encontram conectas e, assim, estabelecem a orientação dos arcos, através do critério de d-separação.

4.2.2 Algoritmo K2

O algoritmo K2 é considerado um dos mais importantes dentre todos os algoritmos que se referenciam a busca de pontuação para estimação de estrutura. Assim, sua idéia base é, partindo de uma ordenação das variáveis, a fim de tornar a estrutura acíclica, pesquisar entre os $2^{n(n-1)/2}$ tipos de configurações de estruturas de rede e verificar qual entre elas maximiza a função *score* dada por (3.3) (Hrushchka, 1997).

$$P(\xi | X) = c \prod_{i=1}^m \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk} ! \quad (4.3)$$

Onde, X é a base de dados com n observações, ξ representa a dimensão de estrutura, m é o número de variáveis, r_i é a quantidade total de possíveis valores que a variável X_i ($i = 1, \dots, m$) pode assumir. O termo q_i está relacionado às possíveis configurações dos pais. O valor de N_{ijk} representa a quantidade total de observações em X onde a variável X_i está no k -ésimo estado e os seus pais apresentam a j -ésima configuração. A constante c é a constante de proporcionalidade. Já N_{ij} é o número total de observações em X onde se tem X_i com qualquer um de seus possíveis valores e com a j -ésima configuração.

4.2.3. Estimação Híbrida utilizando o *Software R*

O tipo de estimação realizada por Bottcher e Dethlefsen (2003) está relacionado a um método de busca particular, denominada pesquisa ávida com reinícios aleatórios (Heckerman *et al.*, 1995). Esse método perturba a Rede Bayesiana através de inserções, remoções ou troca de sentidos de arcos, comparando os *scores* de duas diferentes estruturas de rede. Note que o método ainda é essencialmente influenciado por busca de pontuação.

Um possível método para realizar tal comparação é baseado na utilização da razão posteriori, dada em (4.4).

$$\frac{P(\xi | X)}{P(\xi^* | X)} = \frac{P(\xi)}{P(\xi^*)} \times \frac{P(X | \xi)}{P(X | \xi^*)} \quad (4.4)$$

Esse método de aprendizado também é aplicado na existência de variáveis contínuas na rede. Porém, não há metodologia desenvolvida para Redes Bayesianas exclusivamente contínuas, sendo denominadas mistas as redes que possuem ambos os tipos de variáveis.

Para a abordagem de Redes Bayesianas mistas, os autores consideram uma restrição para capacitar métodos exatos locais. Deste modo, não é permitido que variáveis discretas possuam variáveis-pai contínuas.

Todo esse processo de estimação para estrutura de rede e o método de estimação Bayesiana de parâmetros é baseado na distribuição Dirichlet e estão disponíveis no pacote *deal* do *Software R*.

Não há conhecimento de nenhum pacote ou algoritmo implementado em R para realização de inferências probabilísticas em Redes Bayesianas. Devido a esse fato o pacote *deal* possui compatibilidade com o *Software* Hugin, permitindo a exportação da rede após o término da estimação. Este procedimento é atualmente poderoso, pois o *Software* Hugin não é capaz de realizar qualquer tipo de estimação na presença de variáveis contínuas.

Assim, os principais comandos do pacote *deal* são apresentados na Tabela 4.2.

Tabela 4.2. Funções importantes para ajuste de Rede Bayesiana no *Software* R

Nome da Função	Descrição
<code>network()</code>	Transforma um conjunto de dados em um objeto da classe <code>network</code> .
<code>jointprior()</code>	Calcula a probabilidade conjunta, dado um objeto de classe <code>network</code> .
<code>learn()</code>	Realiza a estimação das probabilidades.
<code>getnetwork()</code>	Capta a rede sem arcos, parte integrante do começo da estimação da estrutura.
<code>heuristic()</code>	Realiza a pesquisa ávida com reinícios aleatórios.
<code>savenet()</code>	Salva a rede para ser utilizada pelo <i>Software</i> Hugin.

Para ilustrar o procedimento de ajuste de uma Rede Bayesiana, exibimos um exemplo.

Exemplo 4.3. Para realizar um breve exemplo de estimação de estrutura e de probabilidades via o *Software* R, consideramos um conjunto de dados reais em *credit scoring*, composto por 499 observações e 5 variáveis, sendo a variável de interesse Y (*Credit Rating*). Uma extensão desse conjunto de dados será utilizada para exibir como os procedimentos de estimação podem ser realizados através do *Software* Hugin.

Assim, temos as seguintes variáveis:

Y = *Credit Rating*

- Mau
- Bom

X1 = Saldo da conta corrente

- Sem movimento
- Sem balanço
- \$300
- $\leq \$300$

X2 = Duração do Crédito em meses

- <21
- ≥ 21

X4 = Valor da Poupança em Dólares

- Sem economias
- <140
- 140-700
- 700-1400
- >1400

X3 = Valor do Crédito em Dólares

- $< 4\,828.15$
- $\geq 4\,828.15$

A estrutura estimada da Rede Bayesiana pode ser visualizada pela Figura 4.3. Além disso, salvamos essa estrutura em um arquivo chamado *rede.net* a fim de utilizá-lo no próximo capítulo, onde será apresentado o *Software* Hugin.

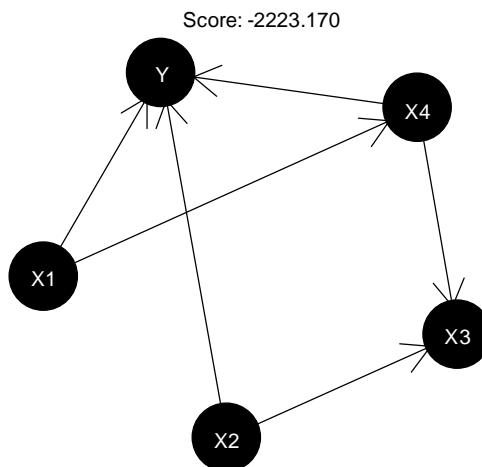


Figura 4.3. Estrutura de Rede Bayesiana estimada através do *Software* R

Através da Figura 4.3, notamos que a variável Y é influenciada probabilisticamente pelas variáveis X1, X2 e X4. Ou seja, a probabilidade de um indivíduo ser classificado como bom ou mau pagador está relacionada diretamente com seu saldo na conta corrente, a duração do crédito em meses e o valor da poupança.

Todo o processo de estimação é realizado através do *Software R* e seus códigos de programação são apresentados abaixo.

```
dados=read.csv("C:\\...\\CreditScoring.csv",
sep=';',header=T) # Abrindo o conjunto de dados

require(deal) # Abrindo o pacote deal

# Transformado o conjunto de dados para o tipo Frame
x=data.frame(dados)

# Transformando x para a classe network
fit      <- network(x)

# Calculando as probabilidades conjuntas
fit.prior <- jointprior(fit)

# Calculando as probabilidades conjuntas
fit      <- getnetwork(learn(fit,x,fit.prior))

# Realizando a busca heurística. Os atributos restart e degree
# estão associados aos números de reinício e a força de
# perturbação. O atributo removecycles quando TRUE, remove
# redes que são cíclicas da estimação, retornando apenas as
# acíclicas.

hisc      <- heuristic(fit,x,fit.prior,restart=2,degree=5,
trace=F,removecycles=T)
```

```
#Exibe a Rede estimada, visualizada através da Figura 3.3
plot(hisc$nw)

# Salva a rede na pasta Meus documentos, se o sistema
# operacional for Windows.
savenet(hisc$nw,file("rede.net"))
```

4.3. COMENTÁRIOS FINAIS

Para os métodos de estimação de estrutura e probabilidades condicionais, exibimos dois tipos de procedimentos para a estimação de probabilidades e três tipos de métodos para a estimação de estrutura, sendo estes os métodos mais comuns utilizados na literatura especializada.

Além disso, apresentamos como um procedimento de estimação de Redes Bayesianas pode ser realizado utilizando o *Software R*. Assim, temos um importante relacionamento entre este *software* e o *Software Hugin*, que será apresentado no capítulo seguinte.

5. SOFTWARE HUGIN

O *Software* Hugin (<http://www.hugin.com>) foi construído pela empresa Hugin Expert S/A fundada em 1989 e localizada em Aalborg, Dinamarca. Assim, foi implementado como instrumento de análise para metodologias voltadas à área da saúde, mais especificamente para a diagnose muscular. Hoje em dia, tornou-se um sistema comercial altamente conhecido e direcionado a soluções utilizando a tecnologia de Redes Bayesianas.

De uma forma geral, existem vários tipos de versões e licenças para este *software*, cada uma direcionada a uma finalidade diferente. Por exemplo, as versões Hugin Developer, Explorer e OTM são voltadas apenas para o uso comercial, as versões Hugin Researcher, Classroom e Educational são voltadas apenas uso acadêmico.

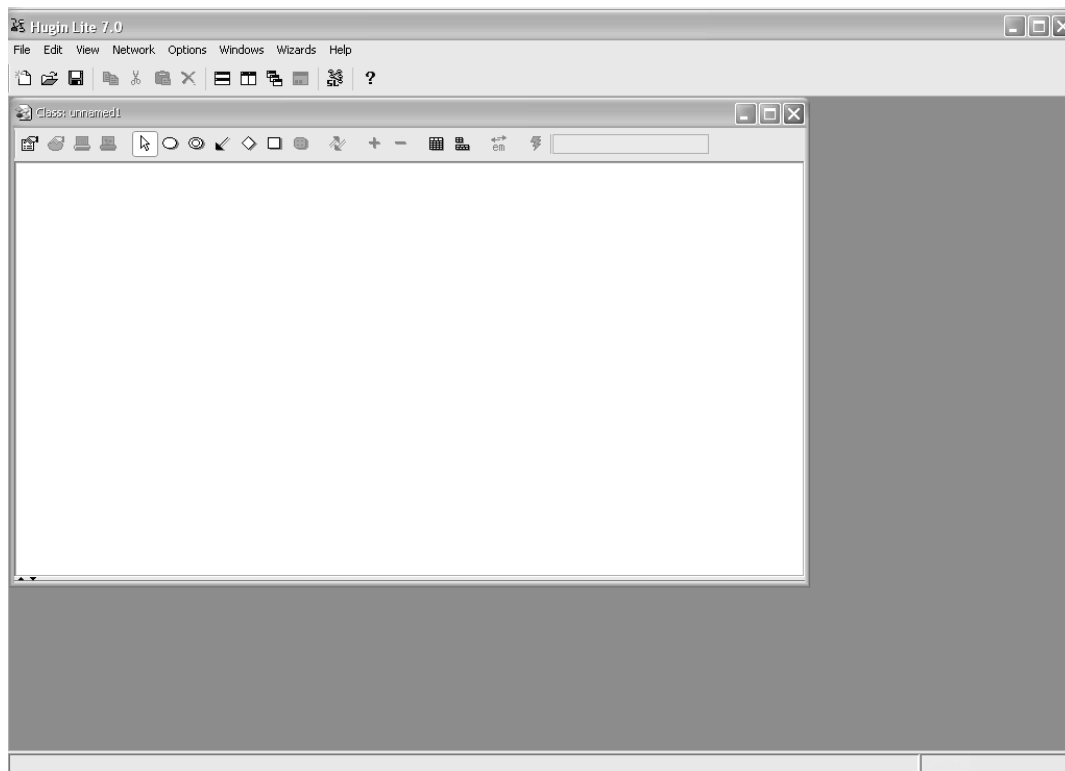


Figura 5.1. Interface Gráfica do *Software* Hugin Lite

Mesmo assim, a empresa fornece uma versão demonstrativa e para uso pessoal do *software*, tal versão é conhecida como Hugin Lite, sendo uma edição limitada das licenças Hugin Developer e Researcher.

Entre outras limitações, o Hugin Lite fornece a manipulação de um conjunto de dados com apenas 500 observações e com até 50 categorias. O download desta versão pode ser realizado gratuitamente no site da empresa e está disponível para os sistemas operacionais Windows (95/98/NT4/2000/XP), Solaris Sparc, Solaris x86, Linux e Mac OS X 10.4, 10.5.

A página principal do *Software* Hugin Lite pode ser visualizada na Figura 5.1.


Nesta seção, apresentamos como as técnicas básicas em Redes Bayesianas podem ser realizadas no Hugin Lite, além de proporcionar um rápido entendimento das funções básicas do *software*.

5.1. PRIMEIROS PASSOS


5.1.1. Instalação



Para realizar o download do arquivo HuginLiteR70.msi através do site, um breve cadastro deve ser preenchido, no qual o *e-mail* é de extrema importância, nele será enviado o código de utilização do *software*. Após realizar o download do arquivo, um duplo clique inicializará o procedimento de instalação, avance e forneça o código que foi enviado para seu e-mail quando requisitado, continue avançando e aguarde a instalação do programa. Assim, para todas as versões do Windows, ele pode ser acessado em *Iniciar > Programas > Hugin Lite 7.0 > Hugin Lite 7.0*.

5.1.2. Acessando Arquivos


As Redes Bayesianas são armazenadas em arquivos no formato específico *.net*. Além disso, o Hugin Lite possui uma pasta contendo alguns exemplos de Redes Bayesianas, esta pode ser acessada através do ícone  ou com um duplo clique em *File > Open* que se encontra no menu principal. Para qualquer versão do Windows, a pasta de exemplos está localizada no endereço *Arquivos de programas\Hugin Expert\Hugin Lite 7.0\Samples*.

5.1.3. Compilação


Após a Rede Bayesiana ter sido aberta e antes mesmo de probabilidades de propagações serem definidas ou evidências instanciadas, devemos acionar *Switch Run* , ou *Network > Run*, para iniciar a compilação da rede.

Após isso, uma nova janela se abrirá ao lado esquerdo da tela, mostrando todas as variáveis-nó com suas respectivas probabilidades, clicar no botão *Expand note list*  exibe todas as probabilidades a posteriori, dada uma evidência atual. Note que o programa é iniciado sem nenhuma evidência, ou seja, o conjunto de evidências é vazio. O botão *Collapse node list*  oculta todas as probabilidades a posteriori.

5.1.4. Salvando os arquivos

Quando uma rede está aberta, o salvamento é efetuado de maneira simples. Selecione *Save As* no menu *File* ou pressione o botão *Save the select network* , encontre o melhor local para o salvamento, digite o nome do arquivo e pressione *Save*.



5.2. CONSTRUINDO UMA REDE BAYESIANA


Uma Rede Bayesiana pode ser criada e/ou editada facilmente através do *Software Hugin*. Para abrir o módulo de edição, basta clicar no botão *Switch to edit mode* .


Os procedimentos para edição são similares aos de construção de redes, aqui o termo “construção” é entendido como transferir uma rede já existente para o *software*.

Exemplo 5.1: Considere a estrutura apresentada no capítulo de inferência probabilística, sendo novamente exibido na Figura 5.2. Vamos construir tal Rede Bayesiana no *Software Hugin*.

O módulo de criação e edição de uma Rede Bayesiana é, geralmente, inicializado conjuntamente com o *software*. Tal módulo é apresentado na Figura 5.3.

Para adicionar uma variável-nó discreta basta clicar em *Discrete change tool* , para adicionar uma variável-nó contínua basta clicar em *Continuous change tool* . Adicionamos

quatro variáveis-nó discretas como é indicado na Figura 5.2, para isso basta clicar em  e clicar novamente dentro da janela principal do módulo. Caso for necessário, a variável-nó pode ser arrastada para um lugar desejado, para isso basta segurar o clique e arrastar com o mouse.

Para adicionar os arcos indicando a dependência direta entre as variáveis, basta clicar no botão *Link tool* , clicar segurando o botão e arrastar até a variável-filho desejada. Ou seja, Tipo de imóvel para Idade e Sexo, e estas para *credit rating*, como indicado na Figura 5.2.

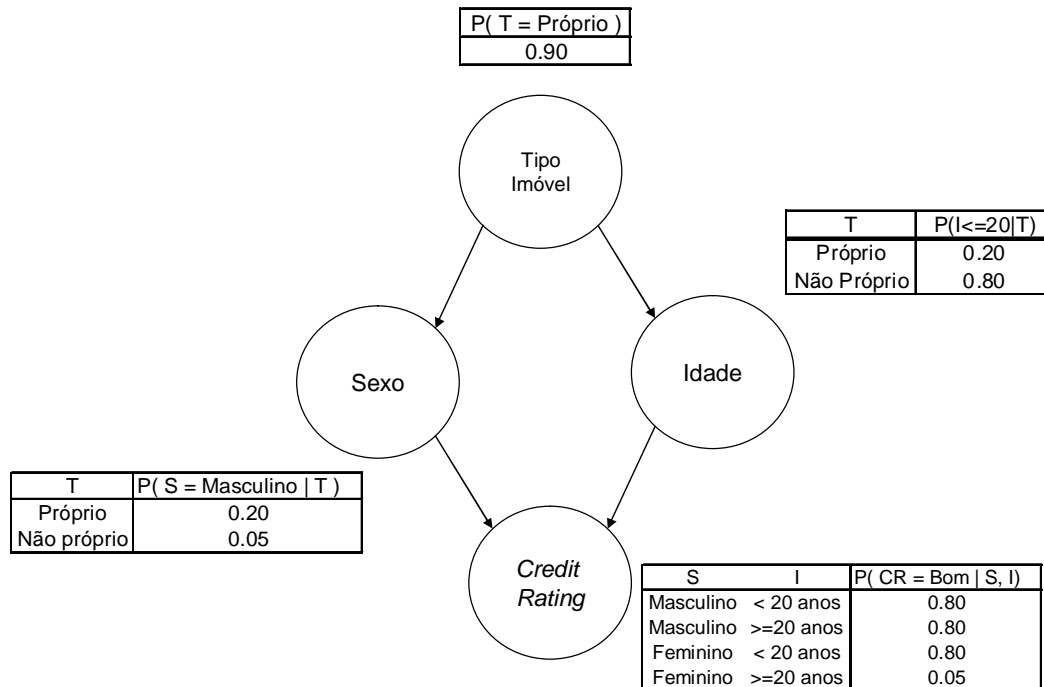


Figura 5.2. Rede Bayesiana a ser transferida para o *software* Hugin

Desta forma, até o presente momento devemos ter encontrado a estrutura indicada acima.

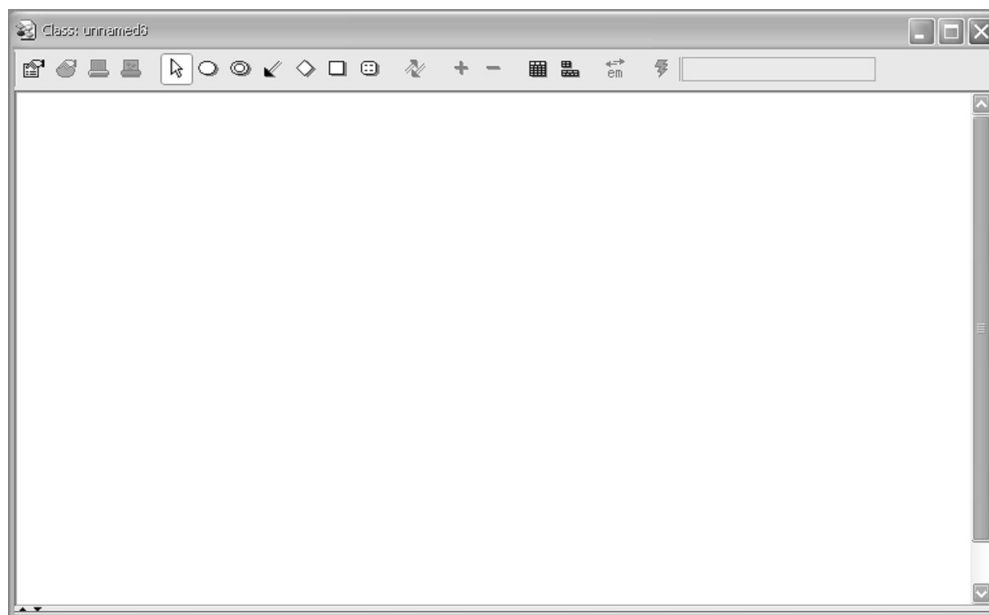


Figura 5.3. Módulo de Edição e construção de uma Rede Bayesiana no Hugin

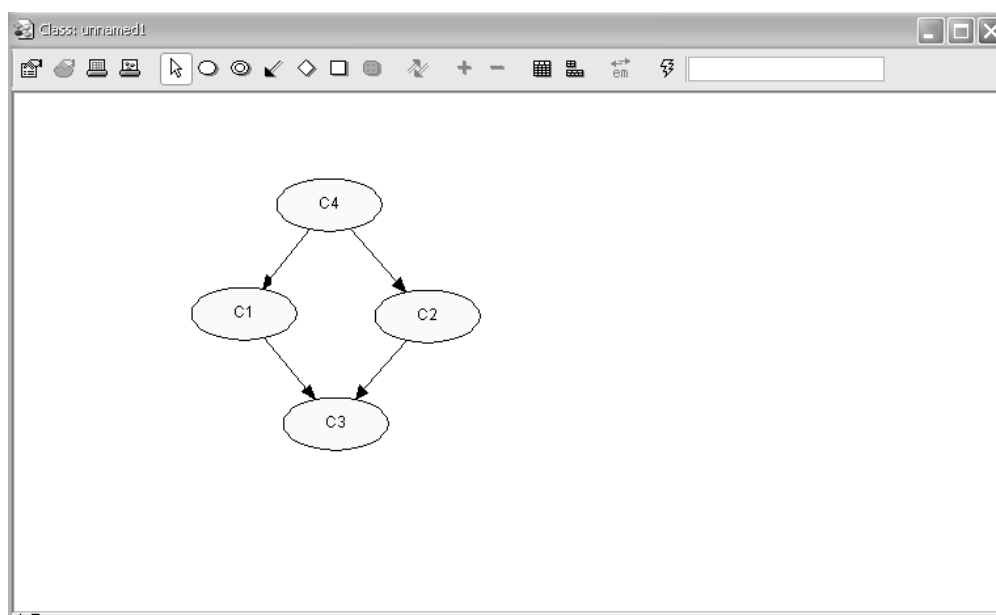
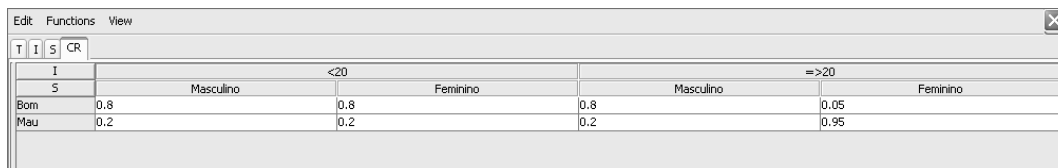


Figura 5.4. Módulo de Edição com a construção parcial da Rede da Figura 5.2

Observando a Figura 5.4, notamos que ainda é necessário estabelecer os nomes para as variáveis-nó, bem como seus possíveis estados e suas probabilidades.

Deste modo, para alterar os estados e nome de cada variável-nó, clicamos duas vezes sobre cada variável e alteramos o campo *name* da aba *node* e os estados são alterados através do botão *rename* da aba *States*.

Para inserir as probabilidades de cada CPT associada a cada variável-nó, clicamos com o botão direito sobre a variável respectiva e selecionamos *Open Tables*. A partir daí, digitamos todos os valores dados pela estrutura da Figura 5.2.



		<20		=>20	
		Masculino	Feminino	Masculino	Feminino
S	Bom	0.8	0.8	0.8	0.05
	Mau	0.2	0.2	0.2	0.95

Figura 5.5. Tabelas de probabilidade condicional no Hugin para a rede da Figura 5.2


A Figura 5.5 exibe a tabela de probabilidade condicional para a variável *Credit Rating*, note que as demais tabelas da rede se encontram nas abas vizinhas.


Assim, a rede está pronta para outros procedimentos.

5.3. INFERÊNCIA EM UMA REDE BAYESIANA

A partir de uma Rede Bayesiana com estrutura já definida e com suas tabelas de probabilidades condicionais já construídas, podemos realizar os procedimentos de inferência utilizando o *software* Hugin.

Essa habilidade é uma das mais importantes deste *software*, pois ele efetua a inferência probabilística de forma rápida e fácil. Para maiores informações sobre os tipos de algoritmos para inferência probabilística implementados no Hugin consultar Jensen *et al.*(1994).

O processo de propagação de probabilidades é procedente da compilação da rede. Desta forma, para realizar esse procedimento, basta abrir uma Rede Bayesiana e acionar o botão *Switch Run* , ou *Network > Run*.

Assim, acionar o botão *Expand note list*  para exibir todas as probabilidades dado um conjunto de evidencias. Como dito anteriormente, neste momento o conjunto de evidências é vazio.

Deste modo, basta clicar duas vezes sobre a categoria de qualquer uma das variáveis que se deseja evidenciar, ou seja, informar tal observação para a rede. Automaticamente, o *software* já realizará a propagação de probabilidades para as demais variáveis-nó, exibindo a probabilidade a posteriori de todas as categorias dado o novo conjunto de evidências.

Exemplo 5.2: Considere a Rede Bayesiana construída no Exemplo 5.1, para ela observamos que um cliente é do sexo masculino e possui 38 anos de idade. Assim, estas categorias são instanciadas, tornando-se evidências. Note que a Figura 5.6 exibe a rede sem a propagação da evidência, já a Figura 5.7 exibe a figura com a propagação para os nós Tipo de Apartamento (T) e *Credit Rating* (TR).

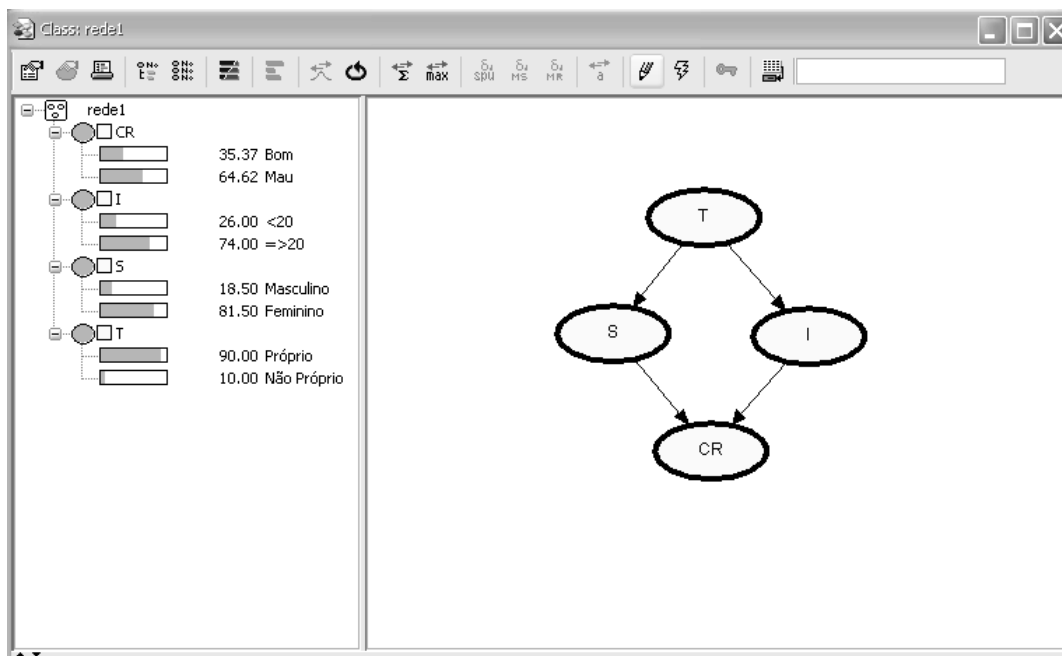


Figura 5.6. Rede Bayesiana sem propagação de probabilidades

Note que para a rede já com as probabilidades propagadas, a probabilidade de o cliente ser bom pagador, dado que ele é do sexo masculino e possui idade maior ou igual a 20 anos, cresce de 0.3537 para 0.80. Da mesma forma, a probabilidade de o cliente possuir imóvel próprio, dado que ele é do sexo masculino e maior que 20 anos, cresce de 0.90 para 0.9931.

Ou seja, para esse conjunto de dados, se as estimativas de probabilidades fossem corretas poderíamos dizer que um indivíduo de sexo masculino, com idade maior que 20 anos e toda a estrutura de dependência entre as variáveis, possui um risco menor de ser mau

pagador. Observe que não foi necessária a coleta de informação sobre o tipo de imóvel do cliente.

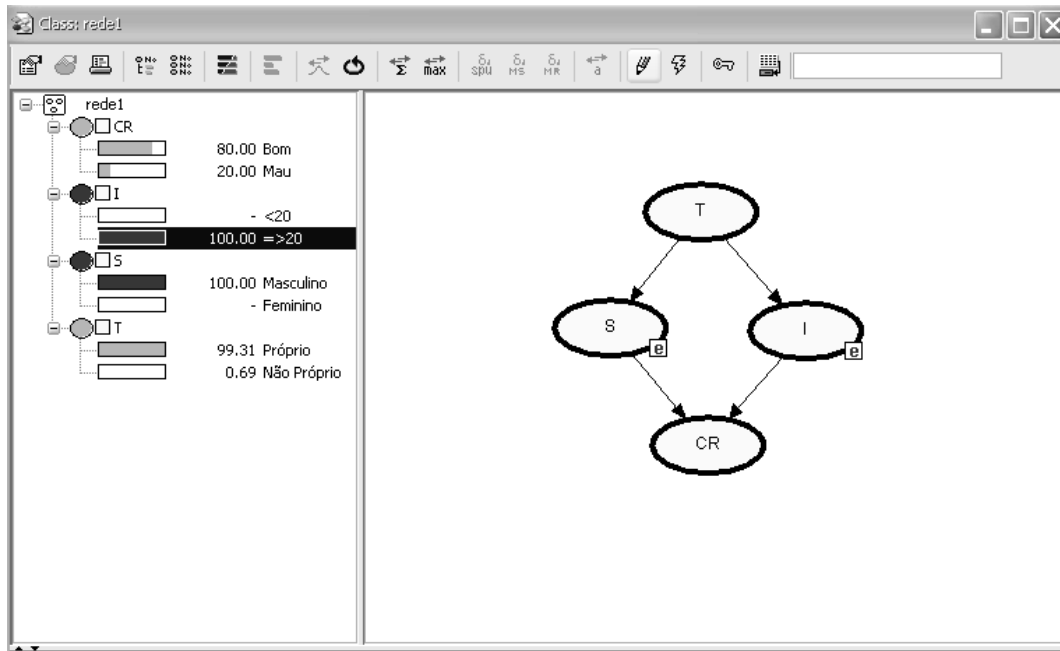


Figura 5.7. Rede Bayesiana com propagação de probabilidades

Note também que os valores que o *software* apresenta estão em porcentagem e não em valor de probabilidade de 0 a 1.

5.4. ESTIMAÇÃO DA REDE ATRAVÉS DOS DADOS

O *Software* Hugin proporciona a facilidade de através de um conjunto de dados em formato *.dat* realizar a estimação da estrutura de rede, bem como estimar as probabilidades em cada variável-nó.

Para os procedimentos de estimação o *software* possui duas possibilidades de algoritmos a serem aplicados: o PC e o NPC. O primeiro, como visto anteriormente, baseia-se na construção de conjuntos condicionalmente dependentes ou independentes realizando testes estatísticos. O segundo estima a rede através de critérios de dependência ou independência estabelecidos pelo usuário.

Uma maneira rápida de iniciar os procedimentos de estimação basta acessar o *Wizards* > *Learning Wizard ...* através do menu principal.

Neste trabalho, iremos apenas demonstrar a estimação da rede através do algoritmo PC. Após o exemplo 5.3, exibiremos como o arquivo *redes.net*, estimado pelo *Software R*, pode ser utilizado no Hugin.

Exemplo 5.3: Devido às restrições do Hugin Lite, consideramos um conjunto de dados reais com 499 observações, 28 categorias divididas em 8 covariáveis, as quais dizem respeito à variável-resposta *Credit Rating*. Esse conjunto de dados é salvo no formato *.dat* e é exibido na Figura 5.8. Abaixo segue o detalhamento das variáveis.

X1 = Saldo da conta corrente

- Sem movimento
- Sem balanço
- \$300
- <= \$300

X5= Tempo de Emprego em anos

- <1
- 1-5
- 5-8
- >8

X2 = Duração do Crédito em meses

- <21
- >=21

X6 = Parcela do Rendimento em %

- < 15
- 15-25
- 25-35
- >35

X3 = Valor do Crédito em Dólares

- < 4 828.15
- >= 4 828.15

X7 = Idade em anos

- <34
- >=34

X4 = Valor da Poupança em Dólares

- Sem economias
- <140
- 140-700
- 700-1400
- >1400

X8 = Tipo de imóvel

- Próprio
- Alugado
- Cedida

Y = *Credit Rating*

- Mau
- Bom

Desta forma, podemos clicar em *Wizards > Learning Wizard ...* para dar início a estimação da rede.

```
Y, X1, X2, X3, X4, X5, X6, X7, X8
"bad", "no running account", ">=21", "B", "no savings", "5-8 years", "25-35", "<34", "rented"
"good", "no balance", ">=21", "A", ">1400", "1-5 years", "25-35", ">=34", "rented"
"bad", ">$300", ">=21", "A", "no savings", "unemployed", "< 15", "<34", "rented"
"good", "no running account", ">=21", "A", "no savings", "> 8 years", "25-35", ">=34", "owned"
"good", ">$300", ">=21", "B", "no savings", "5-8 years", "< 15", "<34", "rented"
"good", "no balance", "<21", "B", "<140", "5-8 years", "15- 25", ">=34", "rented"
"bad", "no running account", ">=21", "B", "no savings", "<1 year", "25-35", "<34", "rented"
"good", "no balance", "<21", "B", ">1400", "> 8 years", "< 15", ">=34", "rented"
"good", ">$300", "<21", "B", "no savings", "1-5 years", "< 15", "<34", "owned"
"bad", "no balance", ">=21", "B", "140-700", "1-5 years", "< 15", "<34", "rented"
"bad", "no balance", ">=21", "A", "no savings", "5-8 years", "15- 25", ">=34", "rented"
"bad", "no running account", "<21", "B", "no savings", "unemployed", "< 15", "<34", "rented"
:
```

Figura 5.8. Rede Bayesiana com propagação de probabilidades

Começando o procedimento, a primeira janela é indicada na Figura 5.9 e faz referência à importação do arquivo contendo o conjunto de dados. Assim, basta acionar o botão *Browser*, procurar o arquivo e seguir para a próxima tela, através do botão *Next*. Note que existe a opção para que o programa leia ou não o nome das variáveis na primeira linha do arquivo de dados. Como indicado na Figura 5.8, na qual podemos observar os nomes das variáveis, devemos manter essa opção.

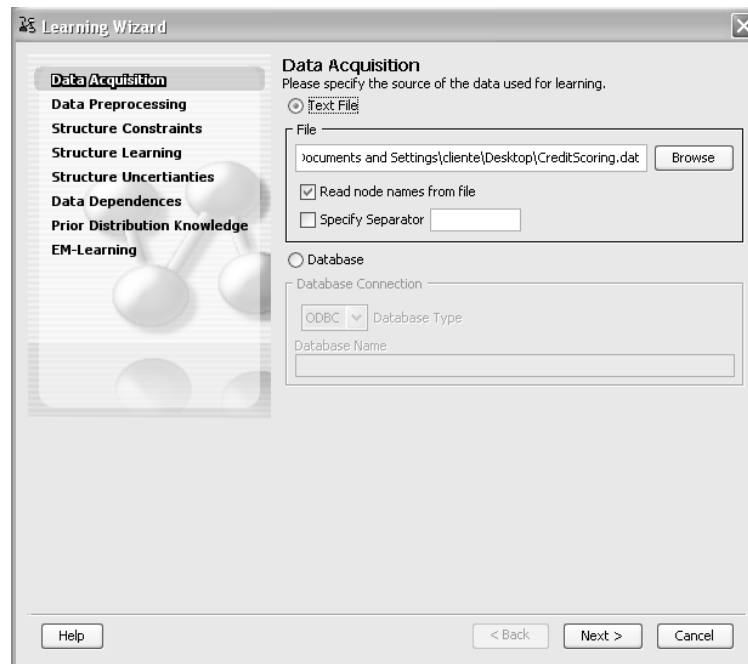


Figura 5.9. Tela inicial do *Learning Wizard*

A próxima tela fornece a possibilidade para usuário editar o conjunto de dados, como remover variáveis através dos botões *Include All* e *Exclude All*, bem como alterar o nome das variáveis através do botão *Label Variables*. Além disso, essa tela fornece uma análise de frequência de cada variável através do botão *Analyze*. Neste exemplo, nenhuma alteração é realizada e prosseguimos através da tecla *Next*.

A terceira janela disponibiliza a construção de contrastes e relações importantes a qual o usuário acredita através de sua experiência como profissional. Neste exemplo, nenhuma alteração é realizada e prosseguimos através da tecla *Next*.

A janela seguinte pode ser visualizada na Figura 5. 10 e solicita o tipo de algoritmo a ser utilizado, bem como o nível de significância para os testes estatísticos. Para o exemplo, selecionamos o nível de significância de 0.05 e o algoritmo PC.

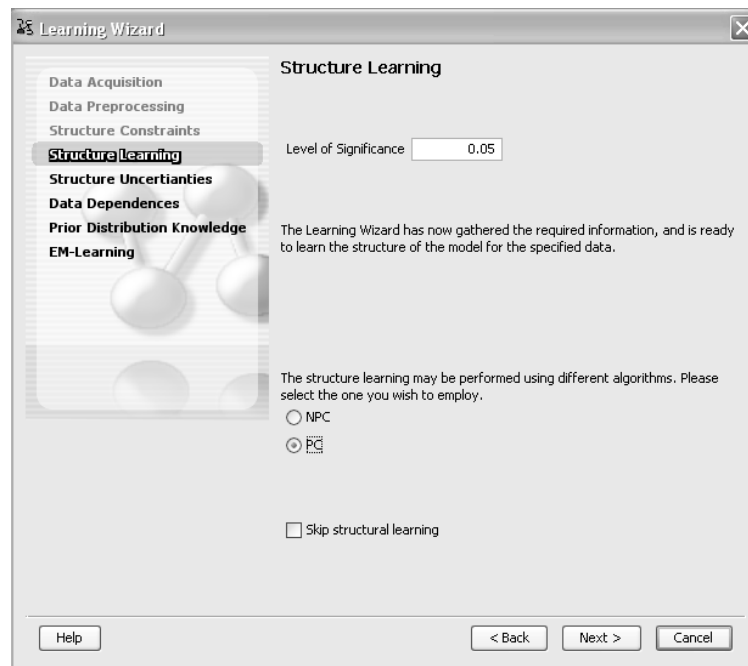


Figura 5.10. Janela do procedimento para selecionar o algoritmo a ser utilizado

A janela seguinte exibe a possibilidade de o usuário inserir manualmente os valores das CPTs, porém através deste procedimento a estimação de probabilidades condicionais pode ser dada através da base de dados na janela seguinte. Assim, nenhuma alteração é realizada e prosseguimos através da tecla *Next*.

A última tela é exibida na Figura 5.11 e solicita para o usuário informar um limiar de convergência para a realização da estimação das probabilidades através do algoritmo de estimação EM, ou seja, um erro máximo permitido para verificar a convergência do algoritmo. Novamente, nenhuma alteração é realizada e terminamos o procedimento através da tecla *Finish*.

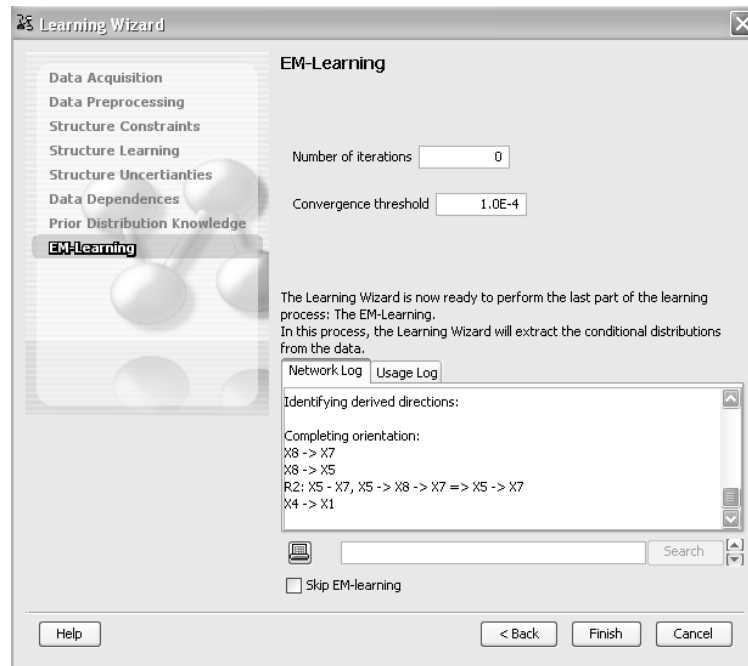



Figura 5.11. Finalização através da estimação das probabilidades via algoritmo EM.

Na Figura 5.12 é exibida a Rede Bayesiana estimada através dos dados. Note que os relacionamentos à primeira vista são confusos, porém podem ser organizados da forma indicada pela Figura 5.13. O procedimento de organização é totalmente manual, porém o *software* disponibiliza um botão no menu principal, através de *Network > Layout the nodes* . Porém, o processo manual é mais flexível, basta clicar e arrastar cada variável-nó para o local desejado.

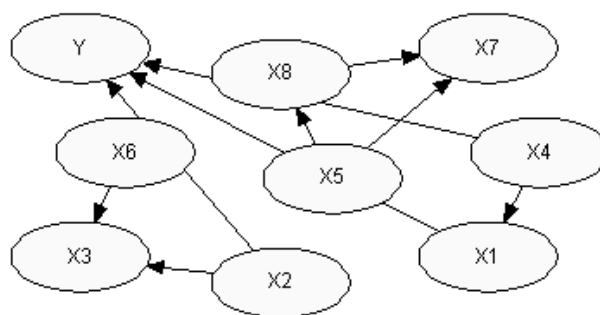


Figura 5.12. Rede Bayesiana estimada pelo *software* Hugin

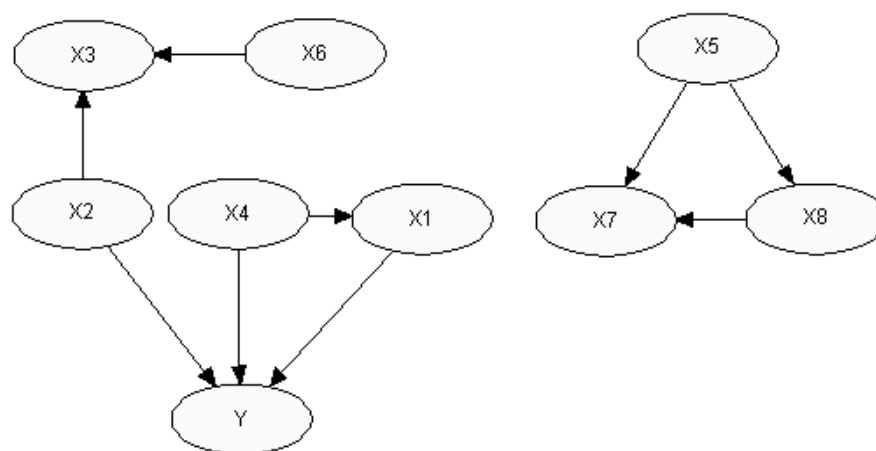



Figura 5.13. Rede Bayesiana estimada pelo Hugin após organização

Através da Rede Bayesiana indicada na Figura 5.13, podemos tirar inúmeras conclusões. Porém, notamos claramente que as variáveis são divididas em dois grupos diferentes, um contendo apenas as variáveis X5, X7 e X8 e o outro contendo as demais variáveis. Ou seja, existe uma relação intrínseca entre o tempo no emprego atual, idade e tipo de imóvel, a qual não se relaciona de forma alguma ao fato de o cliente ser dado como bom ou mau pagador.

Notamos também que a variável Y é influenciada diretamente apenas pelas variáveis X1, X2 e X4. Ou seja, o cliente ser classificado como bom ou mau pagador está diretamente relacionado com seu saldo da conta corrente, a duração do crédito e ao valor da poupança do cliente.

Para realizar os procedimentos de inferência basta acionar o botão *Switch Run*  e proceder da forma como foi vista anteriormente.

No Capítulo 4, exibimos métodos de estimação para a estrutura e para as probabilidades das CPTs de uma Rede Bayesiana. Agora exibimos como o arquivo salvo anteriormente, *rede.net*, pode ser utilizado para realização da propagação de probabilidades.

Exemplo 5.4: Para abrir o arquivo acione *File > Open...* ou dê um duplo clique sobre o ícone do arquivo. Assim, uma vez que ele se encontre no *Software Hugin*, podemos organizar a rede da forma indicada na Figura 5.13.

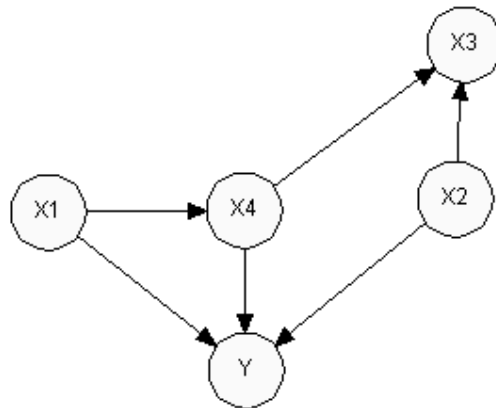



Figura 5.13. Rede Bayesiana estimada pelo *Software R* e editada no *Software Hugin*

Desta forma, podemos realizar os procedimentos vistos anteriormente para a realização de inferências, ou seja, acionando *Switch Run*  e selecionando as categorias de evidência.

Assim, selecionamos como evidência a variável *X1*, sendo instanciada para a categoria sem movimentação, isto é, o cliente não possui movimentação em sua conta corrente. A Figura 5.14 exibe as mudanças de probabilidades para esse caso.

Note que a probabilidade de ter mais que US\$ 1400 em sua conta, dado que o cliente não possui movimentação em conta corrente, tem um decréscimo de 0.1595 para 0.0661 e a probabilidade de ser classificado como mau pagador cresce de 0.3862 para 0.4661.

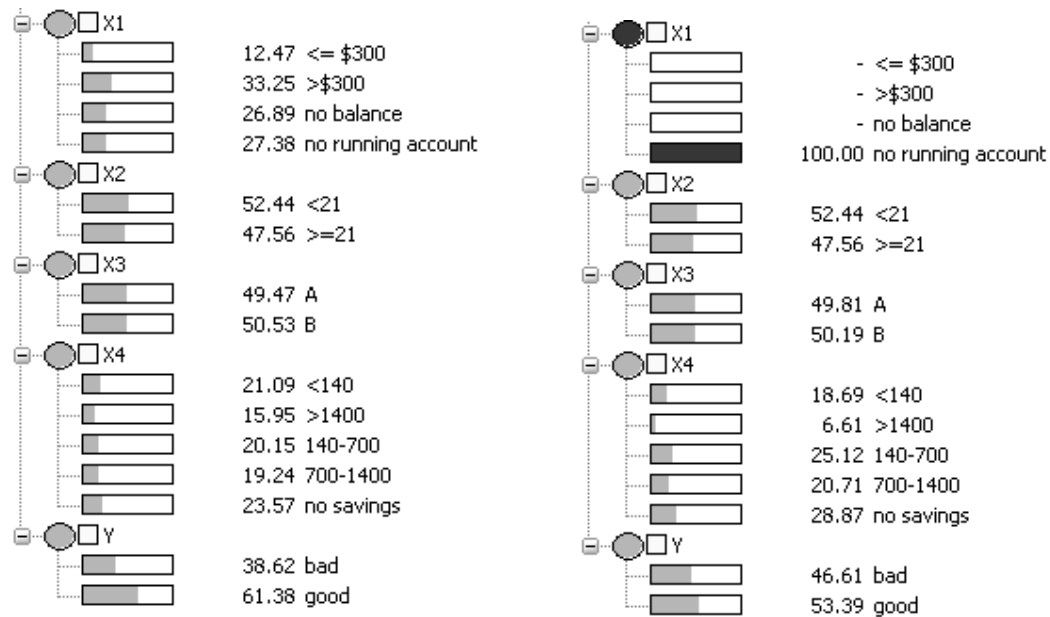


Figura 5.14. Inferência probabilística realizada através do *Software Hugin*.

5.5. COMENTÁRIOS FINAIS

Neste capítulo, apresentamos o *Software Hugin*, sendo este um aplicativo fundamental na área de Redes Bayesianas, bem como os procedimentos passo a passo de como uma Rede Bayesiana pode ser tratada utilizando este *software*.

Assim, englobamos todo o processo de definições básicas da técnica Redes Bayesianas e seus métodos. Desta forma, apresentamos no próximo capítulo uma abordagem específica da técnica para análises de classificação.

6. REDES BAYESIANAS PARA CLASSIFICAÇÃO

Neste capítulo apresentamos uma estrutura geral de Redes Bayesianas utilizada para realizar análise de classificação, dado um grupo de variáveis-categoria e uma variável de interesse a ser classificada. Além disso, exibimos um estudo de simulação estocástica para uma estrutura conhecida e particular desta estrutura geral, o classificador de *Naive Bayes*.

6.1. ESTRUTURAS DE CLASSIFICAÇÃO

Os métodos de classificação são motivos de intensa pesquisa na área da Estatística. Neste contexto, vários métodos tradicionalmente conhecidos são utilizados como, por exemplo, Regressão Logística, árvores de decisão, redes neurais, análise discriminante etc. (Friedman *et al.*, 1997).

Da mesma forma, Redes Bayesianas com estruturas particulares podem ser aplicadas a problemas de classificação e, neste contexto, também são conhecidas como classificadores bayesianos.

Neste trabalho, consideramos a estrutura de Rede Bayesiana Simples, popularmente conhecida como classificador de *Naive Bayes*, e a estrutura de Redes Bayesiana Simples com K-Dependência, também conhecida como classificador bayesiano com k-dependência. (KDB) (Sahami, 1996).

6.2. REDE BAYESIANAS SIMPLES

A construção de uma Rede Bayesiana Simples, conhecida também como classificador de *Naive Bayes*, está baseada no cálculo da distribuição de probabilidade a posteriori $P(Y/X)$, onde $Y=(y_1, y_2, \dots, y_k)$ é a variável aleatória a ser classificada apresentando k categorias e $X=(X_1, X_2, \dots, X_p)$ é um conjunto de p variáveis explicativas discretas.

Para o cálculo da probabilidade condicional $P(Y/X)$ este método assume independência probabilística entre as variáveis explicativas, facilitando a aplicação do método computacionalmente.

Desta forma, $P(Y/X)$ é dada por (6.1)

$$P(Y = y_k | x_1, \dots, x_p) \propto P(Y = y_k) \prod_{i=1}^p P(x_i | Y = y_k) \quad (6.1)$$

Desta forma, o método baseia-se em calcular a probabilidade de uma respectiva observação pertencer a cada uma das categorias e classifica a observação na categoria mais plausível. Se a classificação em foco for binária, podemos utilizar a curva ROC para inferir sobre a classificação.

A Figura 6.1 exibe o caso geral de uma Rede Bayesiana Simples.

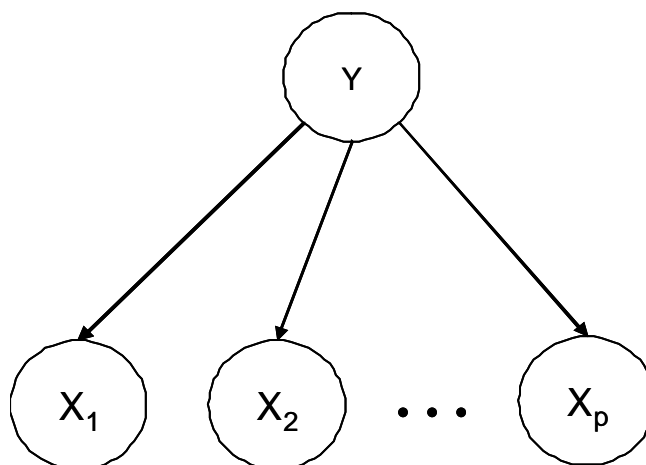


Figura 6.1 – Rede Bayesiana Simples

Através da Figura 6.1, notamos que todas as variáveis explicativas X_i possuem apenas Y como variável-pai, ou seja, Y é a única variável-raiz, a qual origina a rede.

Porém, na maioria das vezes a suposição de independência entre as variáveis explicativas não condiz com a realidade, ou seja, o método não leva em conta a possível relação de dependência probabilística entre as variáveis explicativas.

Assim, outras estruturas de Redes Bayesianas devem ser utilizadas, uma possível alternativa é apresentada a seguir.

6.3. REDE BAYESIANAS SIMPLES COM K-DEPENDÊNCIA

Este método, ao contrário do anterior, considera possíveis relações de dependência entre as variáveis explicativas. Desta forma, uma Rede Bayesiana Simples com k -Dependência trata-se de uma Rede Bayesiana Simples a qual permite em sua estrutura que cada variável explicativa X_i possua no máximo k variáveis-pai. Em outras palavras $\text{pais}(X_i) = \{ Y, Z \}$ onde Z é um conjunto de no máximo k variáveis explicativas.

Assim, temos que uma Rede Bayesiana Simples (*Naive-Bayes*) é caso particular de uma Rede Bayesiana com k -dependência, sendo $k=0$. Note também que k pode variar de 0 a $1-p$, onde p é o número de variáveis explicativas consideradas.

A Figura 6.2 exibe o caso geral de uma Rede Bayesiana Simples com k -dependência com $k=1$.

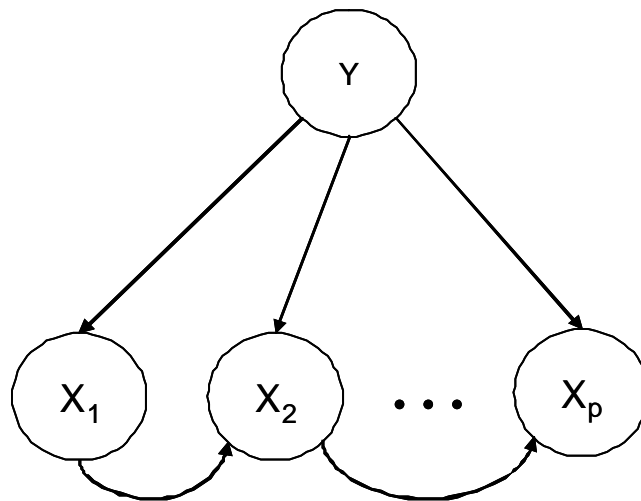


Figura 6.2 – Rede Bayesiana Simples com k -dependência ($k=1$)

Para realizar o ajuste de tal estrutura através de um conjunto de dados, Sahami (1996) propõe o seguinte algoritmo:

1. Para cada variável X_i , calcule a medida de informação mútua $I(X_i, Y)$;
2. Para cada par de variáveis explicativas, calcule a medida de informação mútua condicional $I(X_i, X_j | Y)$;

3. Defina S como a lista de variáveis explicativas utilizadas, inicialmente considere S como vazio;
4. Inicie a Rede Bayesiana com a variável de classificação Y ;
5. Repita até a lista S conter todas as variáveis explicativas:
 - a. Selecione a variável explicativa X_{max} que ainda não está contida em S e que possua a maior medida $I(X_{max}, Y)$;
 - b. Adicione à rede a variável X_{max} ;
 - c. Adicione um arco de Y para X_{max} ;
 - d. Adicione $m = \min(|S|, k)$ arcos partindo das m X_j variáveis explicativas com o maior valor $I(X_{max}, X_j | Y)$;
 - e. Adicione X_{max} à lista S ;
6. Calcule as tabelas de probabilidades condicionais considerando a estrutura construída.

Note que o algoritmo necessita do cálculo da medida de informação mútua entre as variáveis X_i e Y , $I(X_i, Y)$, a qual definimos subseqüentemente.

6.3.1. Informação Mútua

Foi desenvolvida em um ramo da teoria da probabilidade e da matemática estatística que lida com problemas relacionados à comunicação denominada Teoria da Informação e introduzida por Shannon (1948).

A Informação mútua está relacionada de forma muito próxima com testes estatísticos como o teste Chi-square. E é definida por:

$$I(X, Y) = H(X, Y) - H(X) - H(Y) \quad (6.2)$$

Onde $H(X, Y) = -\sum_x \sum_y p(x, y) \log(p(x, y))$ é a entropia conjunta das variáveis X e Y e

$H(X) = -\sum_x p(x) \log(p(x))$ é a entropia da variável aleatória X .

Analogamente $I(X_1, X_2 | Y) = E(I(X_1, X_2) | Y)$.

A medida de informação mútua expressa a quantidade de informação que X compartilha com Y . Ou seja, quando X e Y são independentes temos que $I(X, Y) = 0$

De uma forma geral, quando temos o interesse em decidir entre duas categorias estamos realizando uma classificação dicotômica, a discriminação destas categorias pode ser realizada através da utilização da curva ROC.

6.4. CURVA ROC

A curva ROC (curva Característica Operativa do Receptor), introduzida em 1993 por Zweig e Campbell, pode ser definida, geometricamente, como um gráfico em que para a abscissa temos a medida de 1-especificidade e para a ordenada temos a medida de sensibilidade, sendo esse plano designado unitário, pois cada eixo possui tamanho 1. A sensibilidade é responsável pela proporção de indivíduos com a característica do modelo, a especificidade é responsável pela proporção de indivíduos sem a característica de interesse que é identificada corretamente pelo modelo.

Assim, a curva ROC é construída variando o ponto de corte de classificação e através da amplitude dos *scores*, para ambos os casos temos os *scores* como probabilidades. Um exemplo de curva ROC é exibido na Figura 6.3.

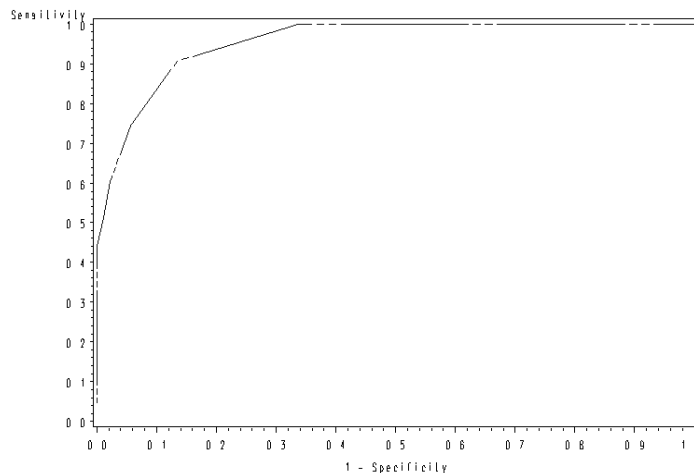


Figura 6.3 – Exemplo de Curva ROC

Uma curva ROC obtida ao longo da diagonal principal corresponde a uma classificação obtida sem a utilização de qualquer ferramenta preditiva, ou seja, sem a presença de modelos. Conseqüentemente, a curva ROC deve ser interpretada de forma que, quanto mais a curva estiver distante da diagonal principal, melhor o desempenho do modelo associado a ela.

Para definir o melhor ponto de corte, temos que escolher o ponto que maximize conjuntamente a sensibilidade e a especificidade da classificação. Sendo assim escolhemos o ponto mais próximo do eixo superior esquerdo do gráfico, ou seja, temos que o melhor ponto de corte é o que possui menor distância euclidiana do ponto (0,1).

6.5. REGRESSÃO LOGÍSTICA

O modelo de Regressão Logística, também conhecido por modelo logístico, é, em geral, um método de classificação bastante solidificado e utilizado para tratar problemas relacionados à classificação dicotômica em várias áreas do conhecimento.

Este modelo estabelece uma relação entre a probabilidade de ocorrência dos resultados de uma variável-resposta dicotômica, que normalmente é representada pelos termos “sucesso” e “fracasso” ou “bom” e “mau”, e variáveis explicativas, sendo estas categóricas ou contínuas.

Este modelo é representado por (Hosmer e Lemeshow, 1989).

Dessa forma, considerando Y como a variável de classificação, sendo a categoria de sucesso igual a 1, e X_i como variável explicativa, a probabilidade de sucesso para a variável de classificação é dada pelo modelo indicado em 6.3.

$$P(Y = 1) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}}. \quad (6.3)$$

Assim, através da probabilidade expressa por $P(Y=1)$, decidimos se classificamos ou não um indivíduo como possuidor da característica em estudo. Geralmente, um ponto de corte é especificado para tal decisão. Um método frequentemente utilizado para estimar este ponto é através da curva ROC.

6.6. REGRESSÃO LOGÍSTICA VERSUS NAIVE BAYES

Como podemos observar, aqui são apresentados dois métodos de classificação, o método de *Naive Bayes*, utilizado especialmente em aplicações computacionais de aprendizado para classificação (Ex: *Plugin ThuderBayes* utilizado para o Gerenciador de e-

mails da Firefox), e o método de Regressão Logística, método solidificado e utilizado em diversos enredos.

Desta forma, realizamos uma avaliação comparativa entre os métodos, utilizando um método exaustivo de amostragem na qual retiramos K amostras de tamanho n e verificamos a mesma estatística para cada uma delas, com o objetivo de estudar as distribuições destas estatísticas para as K amostras. Utilizamos 399 replicações (K=399), sendo esse número utilizado por Hall (1986) para construção de intervalos de confiança utilizando a técnica *Bootstrap* (Efron, 1982).

Deste modo, através de uma base de dados artificiais, analisamos o rendimento de ambas as formas de classificação. Assim, geramos 4 tipos de populações com o mesmo tamanho, possuindo uma variável dicotômica para classificação e com diferentes quantidades de observações com a característica de interesse. Ou seja, cada população composta por 50%, 25%, 10% e 1% de observações com a característica de interesse, respectivamente. Além disso, fixamos os tamanhos de amostra de 100, 300, 1000 e 10000 observações.

A base de dados artificiais foi gerada seguindo o contexto de dados de *credit scoring*, no qual consideramos a característica de interesse como maus pagadores e, segundo Breiman (1998), a distribuição dos bons pagadores segue uma normal p-variada, com média (0,..., 0) e matriz de covariância $4I_p$, onde I_p é matriz identidade de ordem p. E a distribuição dos maus pagadores segue uma normal p-variada, com média $\left(\frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}} \right)$ e matriz de covariância I_p , sendo p o número de covariáveis em análise. Neste caso, utilizamos 10 covariáveis (p=10).

Porém, para atender os critérios de ambas as técnicas, especialmente para a técnica de *Naive Bayes*, categorizamos todas as covariáveis considerando como categoria verdadeiro se a variável é maior ou igual a 0 e como categoria falso se menor que zero. Além disso, consideramos apenas amostras que, para as observações da categoria de classificação, possuíam pelo menos uma observação para cada categoria das covariáveis. A fim de a técnica ser aplicada ao procedimento de *Naive Bayes*.

As Tabelas 6.1 e 6.2 exibem para ambas as técnicas as estatísticas de TTA (taxa total de acerto), TAB (taxa de acerto dos bons pagadores) e TAM (taxa de acerto dos maus pagadores), bem como os respectivos intervalos de confiança considerando um nível de significância de 5%.

Tabela 6.1 – Avaliação das técnicas de Regressão Logística e *Naive Bayes*, n=100 e 300

		n=100		n=300	
Configuração	Medidas	Técnica Estatística para Classificação			
		Naive-Bayes	Reg. Log.	Naive-Bayes	Reg. Log.
1	TTA	0.722 (0.650; 0.810)	0.729 (0.640; 0.810)	0.681 (0.630; 0.730)	0.683 (0.633; 0.730)
	TAB	0.727 (0.592; 0.852)	0.733 (0.612; 0.844)	0.685 (0.606; 0.771)	0.687 (0.612; 0.762)
	TAM	0.718 (0.596; 0.836)	0.727 (0.596; 0.845)	0.679 (0.603; 0.760)	0.680 (0.606; 0.756)
2	TTA	0.739 (0.640; 0.850)	0.747 (0.630; 0.850)	0.686 (0.620; 0.763)	0.690 (0.613; 0.764)
	TAB	0.896 (0.811; 0.966)	0.900 (0.812; 0.960)	0.873 (0.826; 0.916)	0.873 (0.824; 0.917)
	TAM	0.493 (0.349; 0.655)	0.503 (0.341; 0.656)	0.424 (0.329; 0.518)	0.429 (0.325; 0.531)
3	TTA	0.782 (0.660; 0.921)	0.800 (0.660; 0.930)	0.717 (0.607; 0.804)	0.717 (0.613; 0.817)
	TAB	0.972 (0.931; 1.000)	0.974 (0.938; 1.000)	0.960 (0.930; 0.984)	0.961 (0.933; 0.984)
	TAM	0.316 (0.170; 0.534)	0.343 (0.185; 0.600)	0.227 (0.143; 0.329)	0.229 (0.143; 0.324)
4	TTA	---	---	0.787 (0.607; 0.954)	0.796 (0.620; 0.947)
	TAB	---	---	0.996 (0.989; 1.000)	0.996 (0.989; 1.000)
	TAM	---	---	0.079 (0.032; 0.200)	0.083 (0.031; 0.200)

Tabela 6.2 – Avaliação das técnicas de Regressão Logística e *Naive Bayes*, n=1000 e 10000

		n=1 000		n=10 000	
Configuração	Medidas	Técnica Estatística para Classificação			
		Naive-Bayes	Reg. Log.	Naive-Bayes	Reg. Log.
1	TTA	0.661 (0.634; 0.689)	0.662 (0.634; 0.689)	0.666 (0.662;0.670)	0.653 (0.647;0.659)
	TAB	0.662 (0.617; 0.704)	0.663 (0.619; 0.707)	0.677 (0.664;0.686)	0.733 (0.722;0.742)
	TAM	0.661 (0.619; 0.704)	0.662 (0.621; 0.704)	0.656 (0.650;0.664)	0.571 (0.562;0.582)
2	TTA	0.664 (0.619; 0.708)	0.664 (0.622; 0.708)	0.653 (0.646;0.668)	0.648 (0.637;0.663)
	TAB	0.857 (0.827; 0.886)	0.858 (0.828; 0.885)	0.858 (0.853;0.863)	0.856 (0.849;0.862)
	TAM	0.399 (0.348; 0.452)	0.399 (0.351; 0.452)	0.390 (0.382;0.400)	0.384 (0.375;0.395)
3	TTA	0.678(0.598; 0.749)	0.679 (0.604; 0.759)	0.639 (0.619;0.664)	0.640 (0.620;0.667)
	TAB	0.952(0.936; 0.966)	0.952 (0.936; 0.966)	0.947 (0.943;0.949)	0.947 (0.943;0.950)
	TAM	0.195(0.156; 0.245)	0.196 (0.155; 0.240)	0.171 (0.164;0.178)	0.171 (0.166;0.179)
4	TTA	0.751(0.597; 0.910)	0.753 (0.604; 0.912)	0.675 (0.586;0.765)	0.674 (0.587;0.757)
	TAB	0.997(0.993; 0.999)	0.997 (0.993; 0.999)	0.995 (0.994;0.997)	0.995 (0.994;0.997)
	TAM	0.035(0.017; 0.073)	0.035 (0.018; 0.075)	0.021 (0.016;0.028)	0.021 (0.016;0.027)

Assim, temos para as respectivas configurações:

- Configuração 1: 50% Bons pagadores e 50% de Maus pagadores;
- Configuração 2: 75% Bons pagadores e 25% de Maus pagadores;
- Configuração 3: 90% Bons pagadores e 10% de Maus pagadores;
- Configuração 4: 99% Bons pagadores e 1% de Maus pagadores.

Note que para a configuração 4 não é possível realizar o processo de simulação com tamanho de amostra 100, pois existe apenas uma observação sendo classificada como mau pagador, em média. Sendo assim, nem todas as covariáveis possuirão uma observação para cada uma de suas categorias.

Observando ambas as Tabelas 6.1 e 6.2, notamos que existe uma grande queda na assertividade de ambos os modelos quando o desbalanceamento cresce, ainda mais quando o número de bons pagadores é grande, gerando uma alta proporção de acerto para TAB.

Além disso, notamos que existe uma grande aproximação dos resultados para ambas as técnicas, as estatísticas de desempenho são aproximadamente iguais.

Mesmo assim, existe um leve ganho da Regressão Logística com relação ao algoritmo de *Naive Bayes*, variando em média de 1% a 2%, para todos os tipos de configuração realizados, de uma forma geral. Porém, esse fato não acontece para $n=10000$, especialmente para os conjuntos de dados balanceados ou com baixo desbalanceamento (configuração 1 e 2), note que o rendimento do algoritmo *Naive Bayes* é superior para todas as estatísticas, sendo o maior e mais importante encontrado para a TAM da configuração 1, com um ganho médio de 8.5%. Ainda assim, quando existe o aumento do desbalanceamento do conjunto de dados essa diferença tende a desaparecer.

Existe também uma leve diferença entre as estatísticas de TAM em ambos os métodos para $n=100$.

Assim, podemos considerar que, com a utilização de variáveis-categoria para realizar a modelagem, ambas as técnicas possuem o grau de precisão extremamente semelhante. Mesmo, a Regressão Logística tendo um sólido embasamento e sendo uma técnica bem estruturada e dotada de outros resultados auxiliares importantes, como a análise de impacto de variáveis através da interpretação dos coeficientes do modelo e razão de chances. Analogamente, a técnica de *Naive Bayes* assume uma grosseira suposição teórica e é pouco explorada pela área da Estatística, onde seu estudo é mais focalizado dentro da Ciência da Computação.

Estes resultados também são indicativos que a Regressão Logística possui um leve em ganho em comparação a *Naive Bayes* quando é utilizada para amostras pequenas (aproximadamente 100), independente do balanceamento da amostra. Da mesma forma, a técnica de *Naive Bayes* possui um leve ganho em comparação a Regressão Logística quando é utilizada para grandes amostras (aproximadamente 10 000) balanceadas ou com baixo desbalanceamento.

6.7. COMENTÁRIOS FINAIS

Neste capítulo, notamos que as Redes Bayesianas podem ser utilizadas com objetivos de classificação, sendo a estrutura de *Naive Bayes* particular a esse procedimento.

Além disso, realizamos um exaustivo procedimento de simulação para avaliar o rendimento da estrutura de *Naive Bayes* comparada à técnica de Regressão Logística. Desta forma, notamos que estas são extremamente semelhantes, mas se diferenciam levemente em alguns tipos de estrutura de dados.

Com isso, abordamos todos os tópicos inicialmente almejados em nossa proposta, sendo que o próximo capítulo finaliza este trabalho com considerações finais sobre a técnica de Redes Bayesianas.

7. CONSIDERAÇÕES FINAIS

Neste trabalho foram expostas idéias básicas, breves definições e implementações da técnica de Redes Bayesianas, a fim de contribuir à falta de textos científicos introdutórios sobre o assunto e de fácil entendimento para alunos de graduação. Além disso, construímos todo o embasamento de exemplos no enredo de *credit scoring*, área de grande aplicação para a técnica, na qual as Redes Bayesianas auxiliam a verificar, de forma gráfica, a dinâmica das variáveis financeiras e, também, prever a probabilidade de um cliente ser classificado como bom ou mau pagador.

Apresentamos, também, um importante relacionamento entre o *Software Hugin* e o *Software R*, sendo o primeiro aplicativo utilizado mundialmente para a realização de todo o ajuste de uma Rede Bayesiana.

Por fim, apresentamos o método de *Naive Bayes*, um método estruturado em Redes Bayesianas e utilizado para classificação, bem como sua comparação com a técnica de Regressão Logística. Verificamos que ambas as técnicas possuem rendimentos extremamente semelhantes, mesmo com a técnica de *Naive Bayes* possuindo embasamentos e formulações teóricas mais grosseiras que a técnica de Regressão Logística.

Notamos também que as técnicas de Redes Bayesianas estão em atual progresso, abrangendo diversos tipos de pesquisa, como o desenvolvimento de algoritmos para aprendizado de estrutura, algoritmos para aprendizado de probabilidades condicionais e técnicas de classificação. Porém, sendo pouco exploradas pela comunidade estatística quando comparadas aos demais assuntos da área.

Além disso, através do presente estudo, possíveis trabalhos podem ser realizados para verificar se existe um ganho significativo entre a técnica de *Naive Bayes* e Regressão Logística. Em caso afirmativo, qual deve ser o grau de desbalanceamento, tamanho de amostra e número de covariáveis a serem considerados.

REFERÊNCIAS BIBLIOGRÁFICAS

ABELLAN J.; GOMEZ-OLMEDO M.; MORAL. S. **Some variations on the PC algorithm.** In Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM'06), pages 1-8, 2006.

ABICALAFFE, C.; AMARAL, V. F.; DIAS, J. S.. **Aplicação da Rede Bayesiana na Prevenção da Gestão de Alto Risco.** In: Congresso Brasileiro de Informática Médica, Ribeirão Preto. Anais do Congresso Brasileiro de Informática Médica, v. 1. p. 1-1, 2004.

BELLHOUSE, D. R..**The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth.** Statistical Science. Volume 19, N. 1, 3-43, 2004.

BEN-GAL, I.. **Bayesian Networks.** Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons, 2007.

BOBBIO, A.; PORTINALE, L.; MINICHINO, M.; CIANCAMERLA, E.. **Improving the Analysis of Dependable Systems by Mapping Fault Trees into Bayesian Networks.** Reliability Engineering & System Safety, Vol. 71, p.249-260, 2001.

BØTTCHER, S. G.; DETHLEFSEN, C.. **Learning Bayesian Networks with R.** Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, March 20-22, 2003.

BREIMAN, L. **Arcing classifiers.** The Annals of Statistics, N. 26, p. 801-849, 1998.

CHANG, K. C.; FUNG, R; LUCAS, A.; OLIVER R.; SHIKALOFF, N. **Bayesian networks applied to credit scoring.** IMA Journal of Mathematics Applied in Business and Industry. London: Oxford University Press, N. 11, p. 1-18, 2000.

COSTA NETO, P. L. O. ; CYMBALISTA, M. . **Probabilidades.** 2ª. ed. São Paulo: Edgard Blücher, 2006.

EFRON, B.. **The jackknife, the bootstrap, and other resampling plans**. Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38 , 1982.

FEOFILLOFF, P. **Uma introdução sucinta à teoria dos grafos**. São Paulo: Universidade de São Paulo, 2007. Disponível em <<http://www.ime.usp.br/pf/teoriadosgrafos/>>. Acesso em 17 de outubro de 2008.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. **Bayesian network classifiers**. Machine Learning, 29(2-3):131–163, 1997.

HAENNI, R.. **Towards a Unifying Theory of Logical and Probabilistic Reasoning**. 4th International Symposium on Imprecise Probabilities and Their Applications, Pittsburgh, Pennsylvania, 2005.

HALL, P.. **On the number of bootstrap simulations required to construct a confidence interval**. Annals of Statistics, 14,1453-1462, 1986.

HECKERMAN, D.; GEIGER D.; CHICKERING D.. **Learning Bayesian networks: The combination of knowledge and statistical data**. Machine Learning, 1995.

HRUSCHKA, E. R.. **Propagação de Evidências em Redes Bayesianas: Diagnóstico sobre Doenças Pulmonares**. Tese (Mestrado em Ciência da Computação) – Universidade de Brasília, Brasília- DF, 1997.

JENSEN F.; JENSEN F. V.; DITTMER S. L.. **From influence diagrams to junction trees**. In R. L. de Mantaras and D. Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 367-373, Seattle, Washington, July 29-31, 1994.

KORB, K. B.; NICHOLSON, A. E.. **Bayesian artificial intelligence**. London: Chapman & Hall/CRC Press UK, 2004.

LUNA, J. E. O.. **Algoritmos EM para Aprendizagem de Redes Bayesianas a partir de Dados Incompletos**. Tese (Mestrado em Ciência da Computação) – Universidade Federal do Mato Grosso do Sul, Campo Grande - MS, 2004.

MAGALHÃES, I. B.. **Avaliação de redes Bayesianas para imputação de variáveis qualitativas e quantitativas**. Tese (Doutorado em Engenharia) - POLI-USP, São Paulo, 2007.

MARQUES, R. L.; DUTRA, I.. **Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações**. Maio de 1999. Disponível em: <<http://www.cos.ufrj.br/~ines/courses/cos740/leila/cos740/Bayesianas.pdf>>. Acesso em 3 de agosto de 2008.

MESTER, L. J. **What's the point of credit scoring?**. Business Review, p3, 14p, Set/Out 1997.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. Upper Saddle River: Pearson, 2004.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems**. Morgan Kaufmann, San Mateo, CA, 1988.

REZENDE, S. O. (Org.). **Sistemas inteligentes: Fundamentos e Aplicações**. Barueri-SP: Malone, 2005.

RICH, E.. **Inteligência Artificial**. São Paulo: McGraw-Hill, 1988.

RUSSEL, S. J.; NORVIG, P.. **Inteligência Artificial**. Editora Campus, 2004.

SAHAMI, M.. **Learning Limited Dependence Bayesian Classifiers**. In KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 335-338, Menlo Park, CA: AAAI Press, 1996.

SHANNON, C. E.. **A mathematical theory of communication**. Bell System Tech. J. 27, 379-423, 623-656. 1948.

SPIRITES, P.; GLYMOUR, C.; SCHEINES, R. **An algorithm for fast recovery of sparse causal graphs.** *Social Science Computer Review*, v. 9, p. 62-72, 1991.

TAFNER, M.A.; XERES M.; RODRIGUES-FILHO I.W.. **Redes Neurais Artificiais: Introdução e Princípios de Neuro-computação**, 1a ed.. Blumenau, EKO, Ed. da Furb, 1995.

VIEIRA FILHO, V.; ALBUQUERQUE, M. T. C. F. . **Abordagem Bayesiana para Simulação de Jogos Complexos.** In: SBGames, 2007, São Paulo. Proceedings of SBGames 2007, 2007.

ZWEIG, M. H.; CAMPBELL, G. **Receiver-operating characteristic (ROC) plots.** *Clin. Chem.*, 1993, N. 29, p. 561-577, 1993.