**Universidade de São Paulo**
**Instituto de Matemática e Estatística**
**MAC 5789 - Laboratório de Inteligência Artificial**

# Exercício Programa 4: Redes bayesianas

Autor:

Walter Perez Urcia

São Paulo

Junio 2015

**Resumo**

Neste trabalho o objetivo foi analizar um conjunto de dados e construir manualmente tres redes bayesianas. Além disso, se dividirá o conjunto de dados em dados de treinamento e dados de test. Também se mostrará a construção de cada uma das redes bayesianas e suas assunções para estabelecer as diferentes relações entre as características dos dados. Por último, as redes serão comparados usando o Data-LogLikelihood.

# Sumário

# Lista de Figuras

# 1 Dados de entrada

Os dados de entrada foram extraídos da página Machine Learning Repository [2] e o conjunto de dados que será usado neste trabalho é Adulto [1]. Este conjunto tem 48842 instancias, mas algumas não tem sua data completa. Todos os dados no conjunto são de um censo no ano 1994 feito por Barry Becker para determinar se uma pessoa tem mais de 50 mil en dinheiro por ano.

## 1.1 Características

As características do conjunto de dados na página são:

1. Age (AG): a idade da pessoa
2. Workclass (W): tipo de empleado que a pessoa é
3. Fnlwgt (F): Não se da descrição sobre esta característica
4. Education (ED): Máximo nível de educação a pessoa tem
5. Education-num (EN): Máximo nível de educação a pessoa tem em forma numérica
6. Marital-status (M): Estado civil da pessoa
7. Occupation (O): A ocupação da pessoa
8. Relationship (RE): Relações de familia como esposa ou filho
9. Race (RA): Descrição da raza da pessoa
10. Sex (S): sexo biológico
11. Capital-gain (CG): Ganhos de capital
12. Capital-loss (CL): Perdas de capital
13. Hours-per-week (H): Horas trabalhadas por semana
14. Native-country (N): Pais de origem da pessoa
15. Annual-income (AI): Diz se a pessoa tem mais de 50 mil por ano

Das características anteriores, não será usada a terceira (Fnlwgt) por não ter suficiente informação sobre ela. Por outro lado, algumas características tem valores discretos, mas algumas como $Age$ e $Capital - gain$ são valores continuos e portanto tem que ser discretizados, o que será feito na subseção 1.2.

## 1.2 Discretização de características

As características numéricas são:

- Age (AG)
- Education-num (EN)

- Capital-gain (CF)
- Capital-loss (CL)
- Hours-per-week (H)

Para cada um de eles, se calculo sua mediana e foi mudado para uma variável booleana. Então para uma característica $X_i$, temos que $Median(X_i)$ é sua mediana e todos seus valores mudaram da forma $x_{ij} = (1$ se $x_{ij} > Median(X_i)$ e 0 no caso contrario ).

## 1.3 Dados de treinamento e dados de test

Para poder testar e comparar as redes bayesianas que serão construídas nas seguintes seções temos que dividir os dados em dois conjuntos. O primeiro será usado para treinar e calcular todas probabilidades da rede bayesiana. Enquanto o segundo conjunto será usado para testar cada uma das redes bayesianas construídas. Para este trabalho os dados serão divididos em um 65% para treinar e um 35% para testar as redes.

# 2 Construção manual de redes bayesianas

## 2.1 Primeira rede

A continuação se mostram as relações de dependencia para construir o grafo da rede bayesiana: Cada linha é da forma $X_i : Childs(X_i)$ que representa todas as relações de causalidade que tem cada característica com algumas outras.
De forma mais ilustrada, a Figura 1 mostra a rede bayesiana construida com as linhas anteriores.

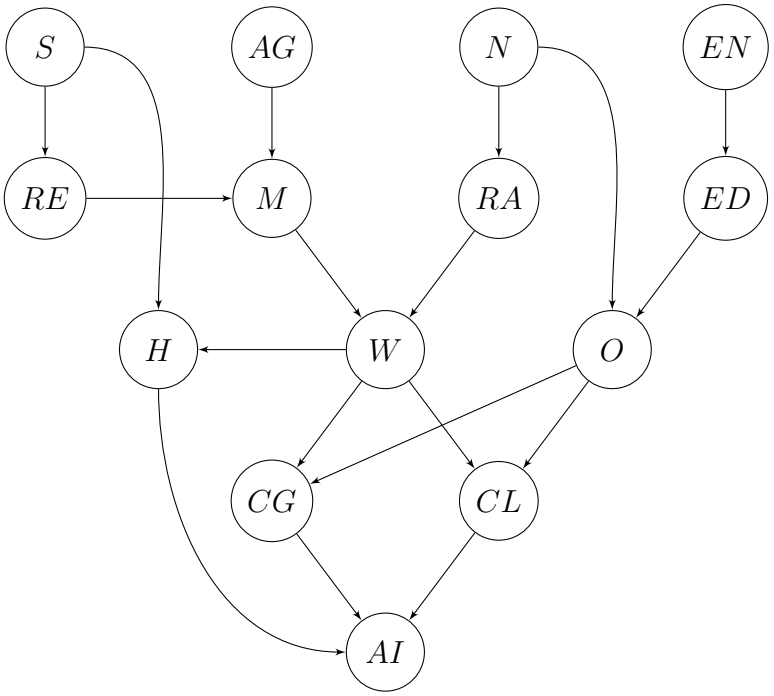


Figura 1: Primeira rede bayesiana

Para construir esta rede se levou em consideração as seguintes suposições:

- Sexo, idade e pais de origem não tem como causa alguma das outras características porque na vida real não dependem de nada
- O pais de origem da pessoa influi na raza, por exemplo é mais seguro que uma pessoa com rasgos asiáticos tenha um pais de origem asiático
- O pais de origem influi na ocupação da pessoa porque não existem todas as ocupações em todos os países do mundo (e.g. astronauta)
- O estado civil e a raza da pessoa influem no tipo de trabalho que tem (e.g. pessoas casadas podem ter trabalhos privados)
- O tipo de trabalho e o sexo influem nas horas de trabalho por semana porque existem trabalhos só para mulheres que não precisam muitas horas

- O tipo de trabalho e a ocupação influem nos ganhos e perdas de capital diretamente (e.g. algumas ocupações ganham mais dinheiro que outras em alguns casos, dependendo do tipo de trabalho)
- Para os ingresos anuales da pessoa, temos em consideração os ganhos, perdas de capital e as horas de trabalho semanais (e.g. uma pessoa que perde muito, tem poucos ingresos anuais, da mesma forma para alguém que ganha muito)

## 2.2 Segunda rede

Da mesma forma que para a anterior rede temos: A Figura 2 mostra todas as relações de dependencia que tem cada característica.

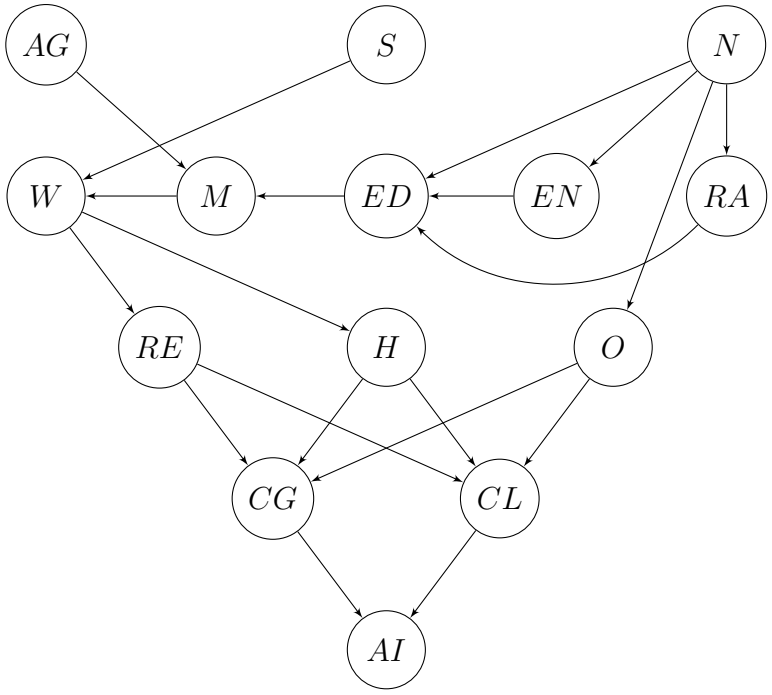


Figura 2: Segunda rede bayesiana

Tendo en consideração as seguintes suposições podemos construir esta rede:

- A educação, raza e a ocupação dependem do pais de origem da pessoa por as oportunidades que podem existir li
- O estado civil depende da idade e a educação que tem a pessoa (e.g. não muitos jovens se casam, gente que consegue nível universitario se casa)
- A educação depende da raza da pessoa em alguns lugares do mundo (com muito racismo)
- O sexo e o estado civil influi no tipo de trabalho pela mesma razão da rede anterior

- Dependendo o tipo de trabalho pode necessitar mais horas de trabalho e também influi no tipo de relação que a pessoa tem (e.g. solteiro, casado)
- Os ganhos e perdas de capital dependem das horas trabalhadas na semana porque mais horas trabalhadas podem significar mais ganhos
- Da mesma forma, uma melhor ocupação pode ganhar o perder mais que outras
- A relação também influi porque se gasta mais quando a pessoa está uma relação
- Os ingresos anuais dependem de quanto a pessoa ganha e perde capital

## 2.3 Terceira rede

Por último, a terceira rede construída manualmente é:

Na Figura 3 a continuação mostra a última rede bayesiana construída manualmente.

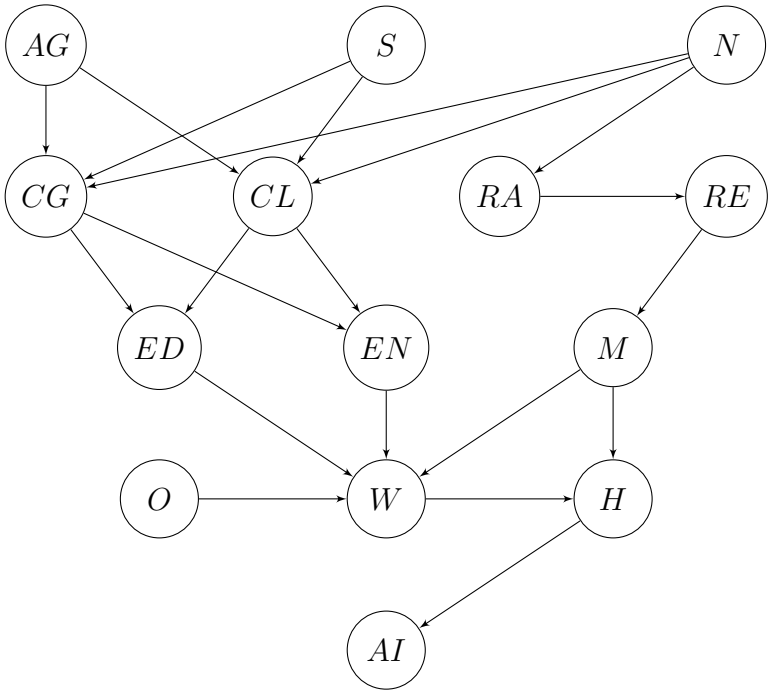


Figura 3: Terceira rede bayesiana

Para esta última rede se levou em consideração o seguinte:

- A idade, o sexo e o pais de origem influem nos ganhos e perdas de capital porque uma pessoa mais joven não tem muitas perdas, uma mulher pode ter muitas perdas por muitas compras e o pais de origem pode ter uma crise que gere muitas perdas as pessoas.
- Dependendo dos ganhos e perdas de capital da pessoa pode conseguir um nível de educação maior

- O tipo de trabalho é afeitado pela educação, o estado civil e a ocupação porque, por exemplo, algumas ocupações não são comumente trabalhadas de forma privada
- A ocupação da pessoa não depende de nada porque pode trabalhar em qualquer coisa que ele quizer
- As horas de trabalho semanais tem que ver com o estado civil (e.g. pessoas casadas trabalham mais que as solteiras) e o tipo de trabalho (e.g. trabalhos privados o independentes não tem um cronograma feito)
- Os ingresos anuais só dependem das horas trabalhadas semanalmente (e.g. se trabalha mais horas, pode ganhar mais)

# 3 Experimentos e resultados

Para cada uma das redes se calculou todas suas probabilidades usando a propriedade seguinte:

$$P(X_i, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i))$$

Onde $Pa(X_i)$ são todas características que são pais de $X_i$ na rede bayesiana. Além disso, a probabilidade condicional pode ser calculada da seguinte forma:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Mas para calcular todas as probabilidades, se tem que evaluar cada variável $X_i$ com os valores que aparecem os dados de entrada sendo a fórmula da seguinte forma:

$$P(A = a \mid B = b) = \frac{N(A = a \cap B = b) + \alpha_{ab}}{N(B = b) + \alpha_b}$$

Para este trabalho os valores de $\alpha$ foram uma estimação com base nos dados calculada com a formula:

$$\alpha_a = \frac{1}{|\Omega_a|}$$

Na formula anterior, se $a$ fosse um conjunto de variáveis, o fator debaixo da fração muda para uma produtora de $\Omega_i$. Tendo em consideração todas as fórmulas anteriores se calcularam todas as probabilidades de cada uma das redes.
Por último, com a rede já treinada pelos dados de treinamento, se tinha que testar a robustez de cada uma e fazer uma comparação entre elas usando os dados de test. Para comparar as redes foi calculado o parâmetro Data Log-Likelihood com os dados de test para cada uma. Os resultados de cada test estão na Tabela 1.

| Rede 1 | Rede 2 | Rede 3 |
|---|---|---|
| 163794.41 | 178879.85 | 180462.16 |

Tabela 1: Data Log-Likelihood para cada rede

# 4 Dificuldades no projeto

Durante a execução do projeto se tiveram algumas dificuldades mencionadas a continuação:

- Muitos conceitos novos
- Cálculo de probabilidades com base nos dados de treinamento
- Cálculo de Data-Loglikelihood

# 5 Conclusões

Pode-se concluir em geral que:

- A terceira rede é a melhor rede manual que foi construída comparando seu Data Log-Likelihood com as outras

# Referências

[1] Barry Becker. Adult Data Set, 1996.

[2] Center for Machine Learning and Intelligent Systems. UCI Machine Learning Repository, 2007.