Data Mining

Felipe E. Barletta Mendes

21 de maio de 2008

Data Mining

O foco principal deste material não é apresentar em minúcia todo o contexto de Data Mining, muito menos o sobre o processo KDD em que a mineração de dados usualmente está inserida.
O objetivo é apresentar uma metodologia estatística para identificação de padrões nos dados e prever o valor de uma ação no futuro.

Sumário

Introdução

- As duas últimas décadas acompanharam um aumento na quantidade de informações ou dados que são armazenados;

Introdução

- As duas últimas décadas acompanharam um aumento na quantidade de informações ou dados que são armazenados;
- Valor dos dados está ligado à capacidade de extrair informações;

Introdução

- As duas últimas décadas acompanharam um aumento na quantidade de informações ou dados que são armazenados;
- Valor dos dados está ligado à capacidade de extrair informações;
- Informação útil que sirva para dar suporte a decisões;

Introdução

- As duas últimas décadas acompanharam um aumento na quantidade de informações ou dados que são armazenados;
- Valor dos dados está ligado à capacidade de extrair informações;
- Informação útil que sirva para dar suporte a decisões;
- Exploração e melhor entendimento do fenômeno gerador dos dados.

Introdução

- Identificar padrões ou tendências úteis.

Introdução

- Identificar padrões ou tendências úteis.
- Otimizar um processo de negócio em uma empresa;

Introdução

- Identificar padrões ou tendências úteis.
- Otimizar um processo de negócio em uma empresa;
- Ajudar no entendimento dos resultados de um experimento científico;

Introdução

- Identificar padrões ou tendências úteis.
- Otimizar um processo de negócio em uma empresa;
- Ajudar no entendimento dos resultados de um experimento científico;
- Ajudarem médicos a entender efeitos de um tratamento;

Data Mining

- Coleção de Técnicas e Métodos para extração de informações em grandes bases de dados.

Data Mining

- Coleção de Técnicas e Métodos para extração de informações em grandes bases de dados.
- Data mining foi popularmente tratado como sinônimo de descoberta de conhecimento em bases de dados.

Processo KDD

O termo KDD, (Knowledge Discovery in Databases) refere-se ao processo global de descobrimento de conhecimento em bases de dados.

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

1. Dados Alvo - acesso, seleção e extração dos dados;

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

1. Dados Alvo - acesso, seleção e extração dos dados;
2. Dados Pré-Processados - limpeza, formatação e alinhamento dos dados;

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

1. Dados Alvo - acesso, seleção e extração dos dados;
2. Dados Pré-Processados - limpeza, formatação e alinhamento dos dados;
3. Dados Transformados - codificação e importação dos dados ;

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

1. Dados Alvo - acesso, seleção e extração dos dados;
2. Dados Pré-Processados - limpeza, formatação e alinhamento dos dados;
3. Dados Transformados - codificação e importação dos dados ;
4. Data Mining - mineração dos dados, extração de conhecimento;

Diagrama do KDD

Dados Alvo → Dados Pré-Processados → Dados Transformados

Data Mining → Conhecimento

1. Dados Alvo - acesso, seleção e extração dos dados;
2. Dados Pré-Processados - limpeza, formatação e alinhamento dos dados;
3. Dados Transformados - codificação e importação dos dados ;
4. Data Mining - mineração dos dados, extração de conhecimento;
5. Conhecimento - resultados e resposta ao problema.

Etapas para DM

- Definição do problema;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;
- Preparação dos Dados;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;
- Preparação dos Dados;
- Partição dos Dados;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;
- Preparação dos Dados;
- Partição dos Dados;
- Desenvolvimento dos Modelos;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;
- Preparação dos Dados;
- Partição dos Dados;
- Desenvolvimento dos Modelos;
- Avaliação e comparação dos Modelos;

Etapas para DM

- Definição do problema;
- Seleção e aquisição dos dados;
- Exploração dos Dados;
- Preparação dos Dados;
- Partição dos Dados;
- Desenvolvimento dos Modelos;
- Avaliação e comparação dos Modelos;
- Escolha do modelo final.

Árvore de Classificação

Este modelo tem como objetivo estabelecer uma relação entre variáveis preditoras uma variável resposta e tem com principal característica o seu tipo de representação. Se a variável resposta é contínua se ajusta uma Àrvore de Regressão, caso contrário é uma Àrvore de Classificação.

A técnica consiste em particionar de forma recursiva o espaço das variáveis explicativas e assim ajustar modelos locais dento das partições. Deste modo, em cada nível de uma árvore, um problema mais complexo de predição ou classificação é decomposto em subproblemas mais simples. Isto traduz-se na geração de um modelo no qual a heterogeneidade da variável a ser explicada é mais atenuada através dos nós da árvore.

Árvore de Classificação

Além dessas características fundamentais, este método, devido a sua representação em uma estrutura hierárquica que é visualizada através de um gráfico de árvore invertida que se desenvolve da raíz para as folhas, permite ao analista ter uma comprensão de como os dados estão apresentados, identificar padrões e interaçãoes entre as variávies independentes, ou seja, este método pode realizar uma mineração dos dados [2] .

Árvore de Classificação

Há vários algoritmos para construção de Árvores de Classificação. No R é utilizado o algoritmo Cart que significa Classification and Regression Trees que consiste em uma árvore com partição binária sucessivas no conjunto de dados, para tornar os sub-conjuntos de dados cada vez mais homogeneos.

Árvore de Classificação

Para selecionar a melhor partição dos dados, procura-se minimizar a impureza dos nós folhas resultantes, para isso são utilizadas medidas de impureza. Por exemplo o Ganho de Informação, este critério baseia-se na medida da entropia que mede o quanto o espaço de probabilidade é homogêneo, por outro lado, quanto maior a entropia maior a desordem, ou ainda, mede a eficiência de transmissão de informação do nó.

Árvore de Classificação

A entropia pode ser traduzida na seguinte fórmula:

$$Entropia = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (1)$$

onde p_i é a probabilidade da classe C_i pertencer a um conjunto S.

Árvore de Classificação

Mas o que pretende-se é saber qual o ganho de informação da variável X, este ganho é dado pela seguinte fórnula em função da entropia:

$$Ganho(S, X) = Entropia(S) - \sum_{v \in Valores(X)} \frac{\mid S_v \mid}{\mid S \mid} Entropia(S_v) \quad (2)$$

em, que valores (X) é o conjunto de todos valores possíveis para a variável X, e $\mid S_v \mid$ é o sub-conjunto de S para qual a variável X tem valor v. Desta forma, o Ganho de Informação, mede a eficácia em classificar os dados de treino.

Regressão Logística

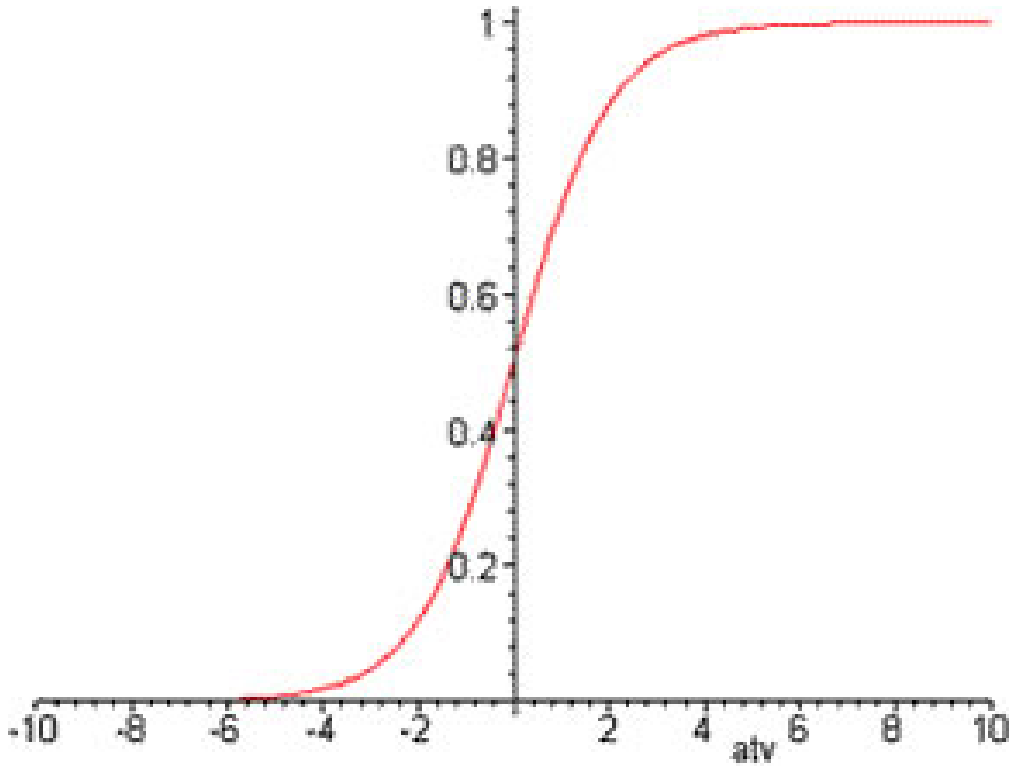
A Regressão Logística é um caso particular dos modelos lineares generalizados [1].
Este tipo de modelo é apropriado quando a variável resposta é categórica[2].
A variável resposta é usualmente dicotômica (0 e 1 ou sim e não), mas pode ser policotômica.
Assim como a Regressão Linear, a logística descreve a relação entre a variável resposta (variável dependente), e um conjunto de covariáveis(variáveeis independentes).

Regressão Logística

A fumdamental diferença entre a linear e a logística está nas suas respostas. A regressão estima o valor médio da variável resposta dados os valores das covariáveis. Na linear este valor estimado varia de $-\infty$ e ∞, na logística este valor varia entre 0 e 1.

- Regressão Linear $-\infty < E(Y \mid x) < \infty$
- Regressão Logística $0 \leq E(Y \mid x) \leq 1$

Gráfico da Função Logística



Modelo Geral da Regressão Logística

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \qquad (3)$$

Redes Neurais para classificação

Análise de Cluster

Análise Discriminante

Redes Neurais para agrupamento

Análise de Componentes Principais

Regressão Linear

Regressão não-paramétrica

Redes Neurais para predição

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;
2. Dados: Amostra de 150 mil clientes dos quais se mediram as seguintes variáveis: idade, renda, variáveis demográficas, lucratividade, nível do depósito, frequência de investimentos, ocasião das aplicações, etc;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;
2. Dados: Amostra de 150 mil clientes dos quais se mediram as seguintes variáveis: idade, renda, variáveis demográficas, lucratividade, nível do depósito, frequência de investimentos, ocasião das aplicações, etc;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;
2. Dados: Amostra de 150 mil clientes dos quais se mediram as seguintes variáveis: idade, renda, variáveis demográficas, lucratividade, nível do depósito, frequência de investimentos, ocasião das aplicações, etc;
3. Exploração e Realce: Considerar apenas as variáveis relacionadas ao tempo decorrido desde a última aquisição, frequência e fator monetário;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;
2. Dados: Amostra de 150 mil clientes dos quais se mediram as seguintes variáveis: idade, renda, variáveis demográficas, lucratividade, nível do depósito, frequência de investimentos, ocasião das aplicações, etc;
3. Exploração e Realce: Considerar apenas as variáveis relacionadas ao tempo decorrido desde a última aquisição, frequência e fator monetário;

Exemplo

1. Problema: Identificar clientes que se interessariam em comprar CDB's;
2. Dados: Amostra de 150 mil clientes dos quais se mediram as seguintes variáveis: idade, renda, variáveis demográficas, lucratividade, nível do depósito, frequência de investimentos, ocasião das aplicações, etc;
3. Exploração e Realce: Considerar apenas as variáveis relacionadas ao tempo decorrido desde a última aquisição, frequência e fator monetário;
4. Modelo: Árvore de Classificação;

Exemplo

5 Avaliação do Modelo: A árvore explicou 80% do comportamento dos clientes;

Exemplo

5 Avaliação do Modelo: A árvore explicou 80% do comportamento dos clientes;

Exemplo

5 Avaliação do Modelo: A árvore explicou 80% do comportamento dos clientes;

6 Implementação: Baseado na àrvore foram enviados convites para parte da totalidade dos clientes no Banco propondo a aplicação em CDB's;

Exemplo

5 Avaliação do Modelo: A árvore explicou 80% do comportamento dos clientes;

6 Implementação: Baseado na àrvore foram enviados convites para parte da totalidade dos clientes no Banco propondo a aplicação em CDB's;

Exemplo

5. Avaliação do Modelo: A árvore explicou 80% do comportamento dos clientes;
6. Implementação: Baseado na àrvore foram enviados convites para parte da totalidade dos clientes no Banco propondo a aplicação em CDB's;
7. Retorno do Investimento: Gastou 30% a menos em divulgação pois o contato só foi feito com parte dos clientes. A resposta foi 50% melhor do que em promoções anteriores.

Referências


PAULA G. A. **Modelos de Regressão com Apoio Computacional**. São Paulo, Universidade de São Paulo - Intituto de Matemática e Esttística, 2007.


BRAGA, L. P. V. **Introdução à Mineração de Dados**. Rio de Janeiro, E-papers Serviços Editoriais, 2005


RODRIGUES, M. A. S. **Árvores de Classificação**. Açores, 2005. Monografia Departamento de Matemática, Universidade dos Açores


Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J.. **Classification and Regression Trees**. Wadsworth International: California, 1984

Referências

- GIOLO, S. R.. **Análise de Regressão Linear**. Curitiba, Universidade Federal do Paraná - Departamenmto de Estatística, 2007.
- GIOLO, S. R.. **Introdução à Análise de Dados Categóricos**. Curitiba, Universidade Federal do Paraná - Departamenmto de Estatística, 2006.