Mineração de Dados

GRIMALDO OLIVEIRA

Sobre Grimaldo

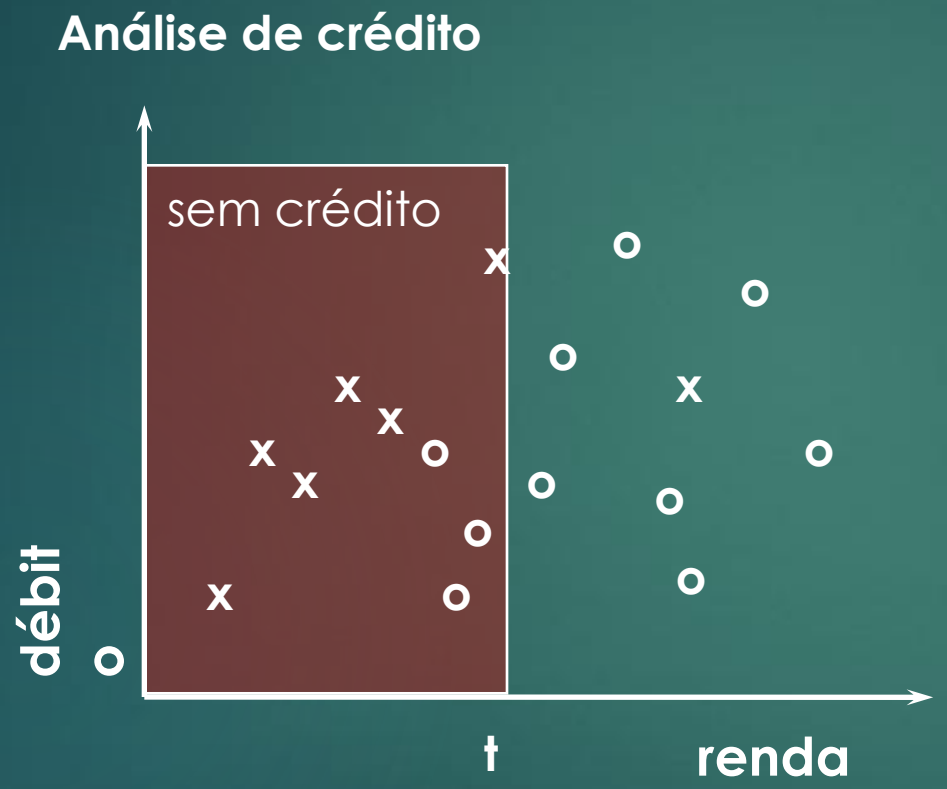
- Grimaldo Oliveira
  - grimaldo_lopes@hotmail.com
- Formação
  - Mestre em Tecnologias Aplicadas a Educação Universidade do Estado da Bahia.
  - Especialização em Análise de Sistemas pela Faculdade Visconde de Cairu.
  - Estatístico pela Universidade Federal da Bahia.
- Atividades
  - Mais de 10 anos atuando como Consultor de Business Intelligence.
  - Projetos Governos Maranhão, Mato Grosso e Bahia.
  - Idealizador do Blog : BI com Vatapá – bicomvatapa.blogspot.com.
  - Livro: BI Como Deve Ser – bicomodeveser.com.br

Agenda

- Tarefas de Mineração de Dados
  - Análise de Regras de Associação
  - Análise de Padrões Sequenciais

Quais Tarefas de Mineração são utilizadas?

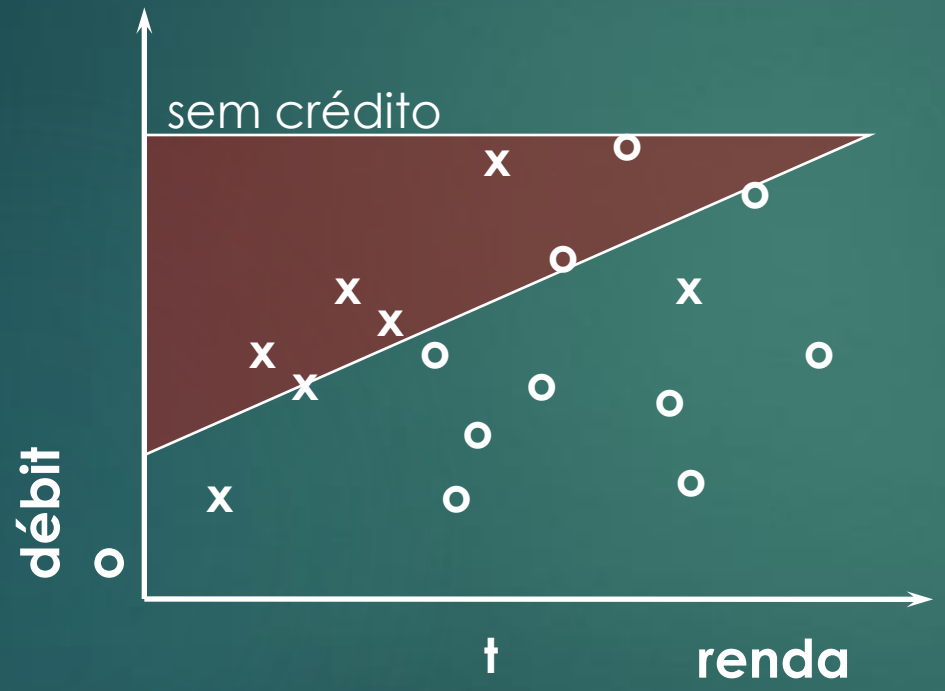
Exemplo de previsão (I)



- Um hiperplano paralelo de separação: pode ser interpretado diretamente como uma regra:
  - se a renda é menor que *t*, então o crédito não deve ser liberado
- Exemplo:
  - árvores de decisão;
  - indução de regras

Exemplo de previsão (II)

Análise de crédito



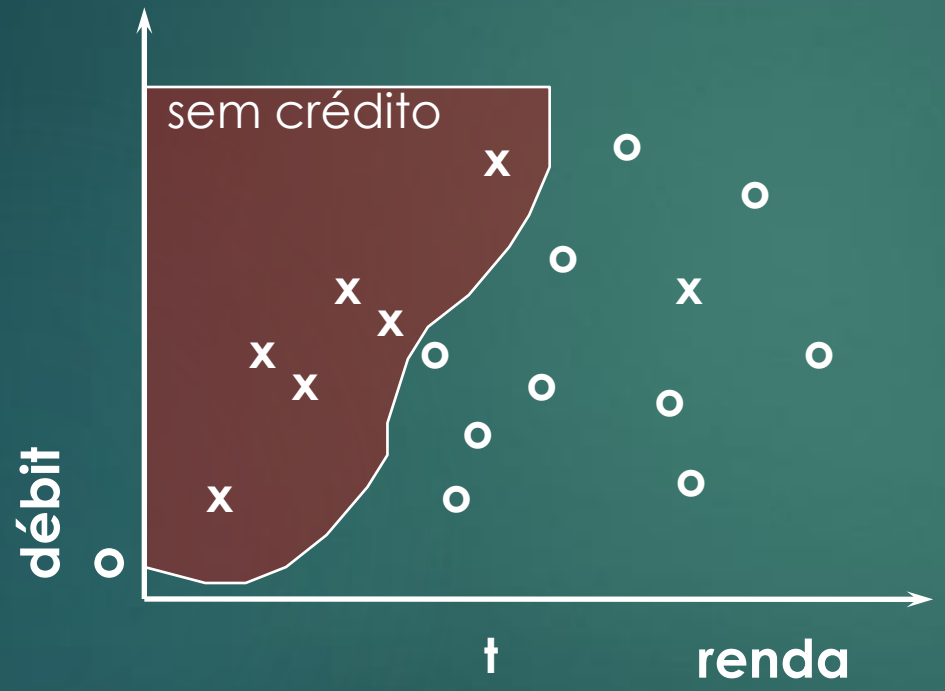
x: exemplo recusado
o: exemplo aceito

- Hiperplano oblíquo: melhor separação:
- Exemplos:
  - regressão linear;
  - perceptron;

Métodos

Exemplo de previsão (III)

Análise de crédito

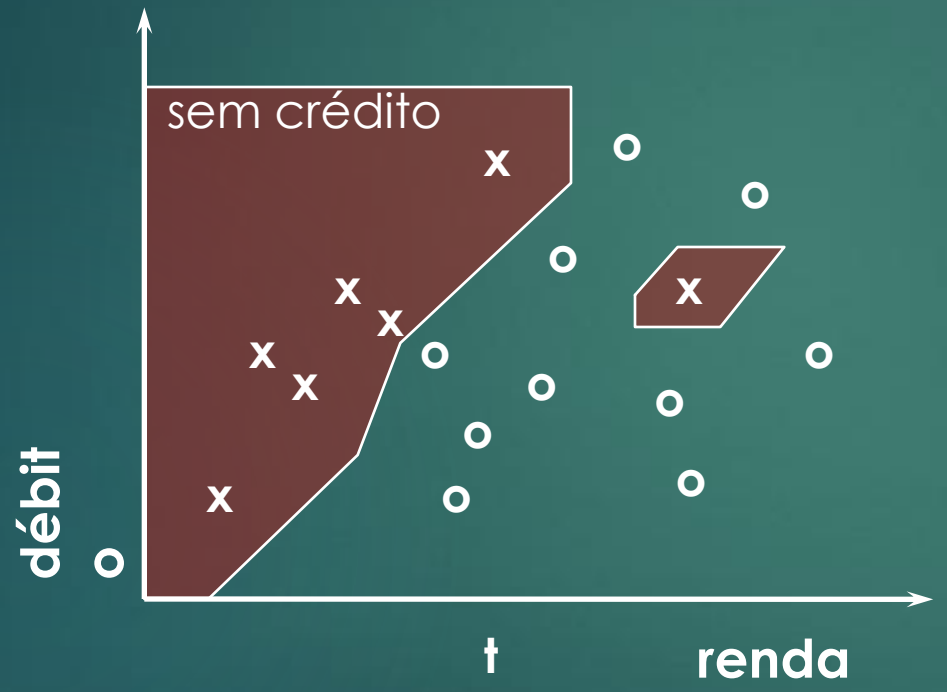


x: exemplo recusado
o: exemplo aceito

- Superfície não linear: melhor poder de classificação, pior interpretação;
- Exemplos:
  - perceptrons multicamadas;
  - regressão não-linear;

Exemplo de previsão (IV)

Análise de crédito



x: exemplo recusado
o: exemplo aceito

- Métodos baseado em exemplos;
- Exemplos:
  - k-vizinhos mais próximos;
  - raciocínio baseado em casos;

Análise de Clusters (agrupamentos) – Segmentação

- Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos

Análise de Outliers (exceções)

- Identificação de dados que não apresentam o comportamento geral

Estimativa (ou regressão)

- Usada para definir um valor para alguma variável contínua desconhecida

Sumarização

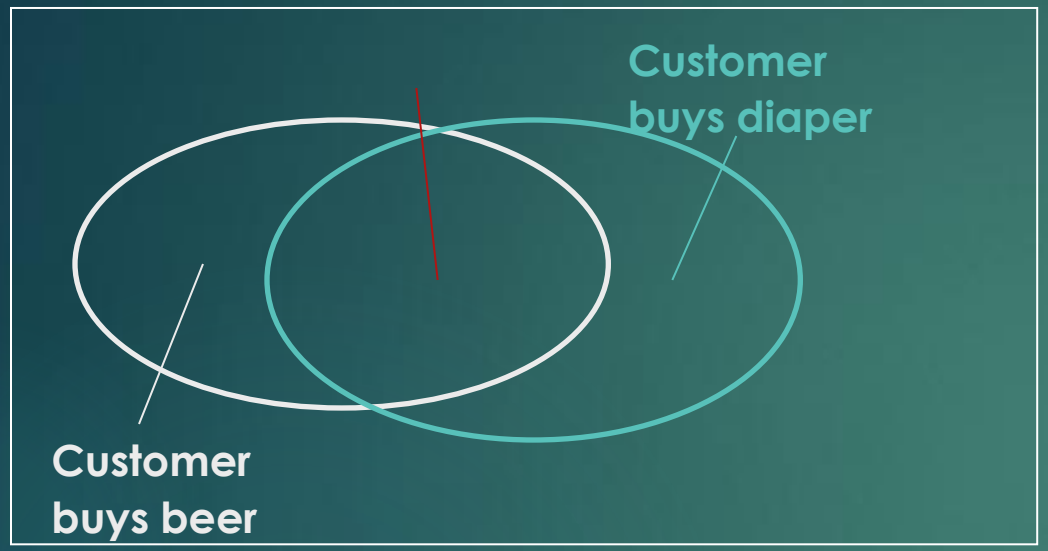
- Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados

Regras de Associação

Regras de Associação

- Mineração de associações ou de regras de associação:
  - Encontrara padrões frequentes, associações, correlações, ou estruturas causais a partir de conjuntos de itens ou objetos em DB de transações, relacionais, ou em outros repositórios de informações.
- Aplicações:
  - Análise de cestas de dados (basket data), marketing cruzado, projeto de catálogos, agrupamento, etc.

Regras de Associação



Encontrar regras $X \& Y \Rightarrow Z$ **com** suporte e confiança mínimos

- **Suporte, s**, é a proporção de transações que contém os itens {X ∩ Y ∩ Z}
- **Confiança, c**, é a proporção que os itens {X∩Y∩Z} aparecem nas transações que contém o item {X∩Y} ou pode ser calculado através do s**uporte {**X∩Y∩Z**}/ suporte {**X∩Y**}**.

Transação	Itens
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Para um suporte mínimo de 50%, e confiança mínima de 50%, tem-se:

- $A \Rightarrow C$ **(50%, 66.6%)**
- $C \Rightarrow A$ **(50%, 100%)**

Regras de Associação

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

Análise de Regras de Associação

ID	Compras
1	Pão, **Leite** **Manteiga**
2	**Leite** ,Açucar
3	**Leite**, **Manteiga**
4	Manteiga, Açúcar

Leite → Manteiga

Suporte = $\frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{Total de clientes}}$ = 50%

Confiança = $\frac{\text{número de clientes que compraram Leite, Manteiga}}{\text{número de clientes que compraram Leite}}$ = 66,6%

Análise de Padrões Sequenciais

Itens = { TV, Vídeo , DVD, FitaDVD, ... }

ITEMSET >> ITEMSET >> ITEMSET >> ... >>ITEMSET

Análise de Padrões Sequenciais

1	{TV , Rádio} >> {DVD}
2	{Computador}
3	{TV} >> {Rádio, DVD}
4	{Rádio} >>{Comp}
5	{Comp} >> {Impressora}

< {TV} , {DVD} >

Suporte = número de clientes que compraram TV, DVD em seqüência / Total de clientes = 40%

Próximos vídeos...

Finalidade: Coleta de dados com os gestores para a construção do BI.

Fatos / Dimensões		Diária
Hóspede		✓
Tipo Quarto		✓
Código Tipo Quarto		
Tipo Quarto	HISTÓRICO	
Classe Quarto		✓
Tempo (Data Registro Primeira Diária)		✓

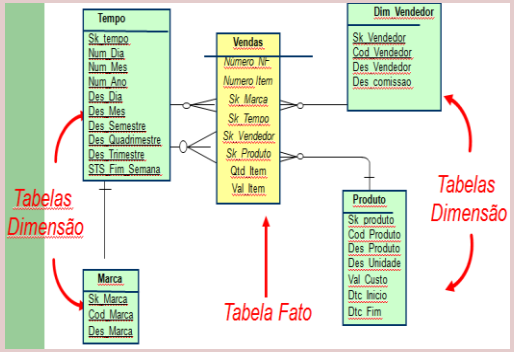
Tarefas de Mineração de Dados -Parte01

Finalidade: Levantamento dos relacionamentos e objetos que armazenam os dados da empresa.

DIMENSÕES	ORIGEM	
	TABELA/VISÃO	CAMPO
Hóspede		
Nome Hóspede	HOSPEDE	NOM_HOSPEDE
Cidade Hóspede	CIDADE_ORIGEM	NOM_CIDADE
País Hóspede	PAIS_ORIGEM	NOM_PAIS
Aeroporto Hóspede	AEROPORTO_SAIDA	DES_AEROPORTO
Local Aeroporto Saída	AEROPORTO_SAIDA	NOM_LOCALIDADE
Código Hóspede	HOSPEDE	COD_HOSPEDE

Tarefas de Mineração de Dados -Parte02

Finalidade: Modelo adequado para realizar as consultas nas bases que servirão ao BI



Mineração Visual

contato@bicomodeveser.com