Licap

Formação Cientista de Dados

PUC Minas

Licap

Preparação da base de dados

Luis Enrique Zárate

Agenda

Licap

- Limpeza dos dados
  - Dados irrelevantes
  - Granularidade nos dados
  - Inconsistência nos dados
  - Relação entre dados
  - Domínios condicionais
  - Defaults de valores
  - Integridade nos dados
  - Duplicações e redundâncias
  - Poluição

Agenda

Licap

- Preparação dos dados
  - Fusão de dados
  - Junção de dados
  - Blocagem de dados
  - Discretização de dados
    - Discretização supervisionada
    - Discretização não-supervisionada

Carga horária: 2 horas e 30 min.

Limpeza dos dados

Licap

- DADOS IRRELEVANTES

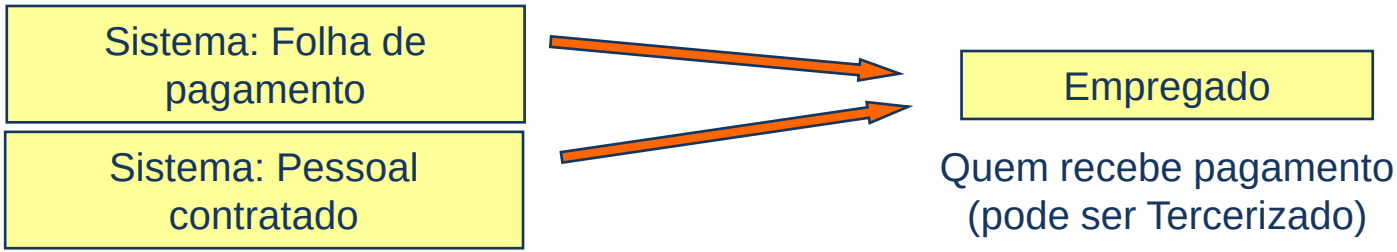
É necessário retirar dados irrelevantes que podem trazer conhecimento falso ou aumentar o tempo de processamento dos algoritmos de Data Mining.

• INCONSISTÊNCIAS NOS DADOS

Licap

Diferentes *“objetos”* representados pelo mesmo nome em diferentes sistemas.

- Empresas de RH buscam colaboradores que atendam um perfil específico



- Empresas internacionais de aluguel de automóveis

RELAÇÕES ENTRE DADOS

Licap

É importante observar e analisar a consistência das instâncias dos objetos da estrutura problema.

Maria da Silva; 15 anos; comprou Toyota Prius .....

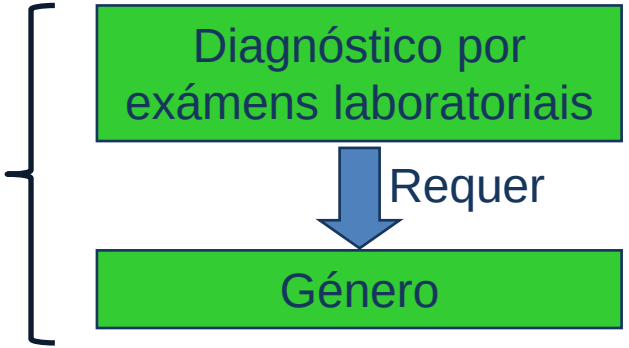
Provavelmente uma Inconsistência ou outlier.
Recomenda-se eliminar esse registro

• DOMÍNIOS CONDICIONAIS

Licap

Cada variável possui um domínio particular ou faixa de valores dentro de um contexto.

Domínios das variáveis podem ser condicionais:



Domínios baseados em regras:

DEFAULTS DE VALORES

Licap

- Alguns campos de um Sistema de Cadastros podem ter valores Default, que são inseridos na base de dados quando os valores reais não são fornecidos. Esses Campos devem ser identificados e cuidadosamente analisados.
  - Ex. Nível de escolaridade, *DEFAULT = "Ensino médio",* Valor real = "Ensino superior"
  - *Ex. Número de dependentes, DEFAULT = 0,* Valor real = 3
- Os valores Default tendem a contaminar a base de dados e o conhecimento a ser extraído.
- A falta do valor real, leva a um problem de tratamento de dados ausentes. Daí, seria interessante descobrir o mecânismo de ausência de forma a imputar valor coerente.

INTEGRIDADE NOS DADOS

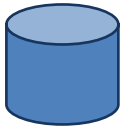
Licap

Deve ser observada a integridade das relações em registros complementares.

1 Empregado pode ter vários carros

1 Empregado não pode ter múltiplos registros funcionais

1 Empregado não pode ter múltiplas esposas



ID	EC	Esposa
09	C	Maria

ID	EC	Esposa
09	C	Silvia

DUPLICAÇÕES E REDUNDÂNCIAS

Licap

- Ocorre principalmente quando as instâncias dependem de diferentes fluxos de dados

Data de Nascimento => Idade Atual

*Preço Unitário **x** Quantidade => Valor da venda*

Peço, altura => IMC

- As variáveis Duplicadas ou Redundantes exigem maior esforço computacional e dependendo do caso podem ser reduzidas.

• POLUIÇÂO NOS DADOS

Licap

Os campos podem conter espaços em branco, estar incompletos, inexatos, inconsistentes ou não identificáveis.

Pessoa Física	EC	Data de Nasc.	Idade	Dependente	Escola/ salário	Telefone
Maria da Silva	C	28/02/13	15		30%	(xxx)4567890

Inconsistente

Não identificável

Ausente

Inexato

Incompleto

Preparação dos dados

Licap

- Os dados precisam ser transformados, discretizados e codificados de forma que possam servir de entrada para os algoritmos de Mineração de Dados.
- Normalmente se faz necessária a transformação ou mudanças de escala dos dados garantindo que as características dos valores originais sejam preservadas.
- A melhor forma de transformar os dados é verificar quais requisitos a solução precisa atender e quais são os requisitos que a técnica de mineração de dados impõe.

FUSÃO DE DADOS

Licap

- É o processo de integrar multiplos dados para produzir mais consistentes com informação útil.
- A Integração de dados pode trazer uma redução significativa na dimensionalidade da base, sem perdas de informação.

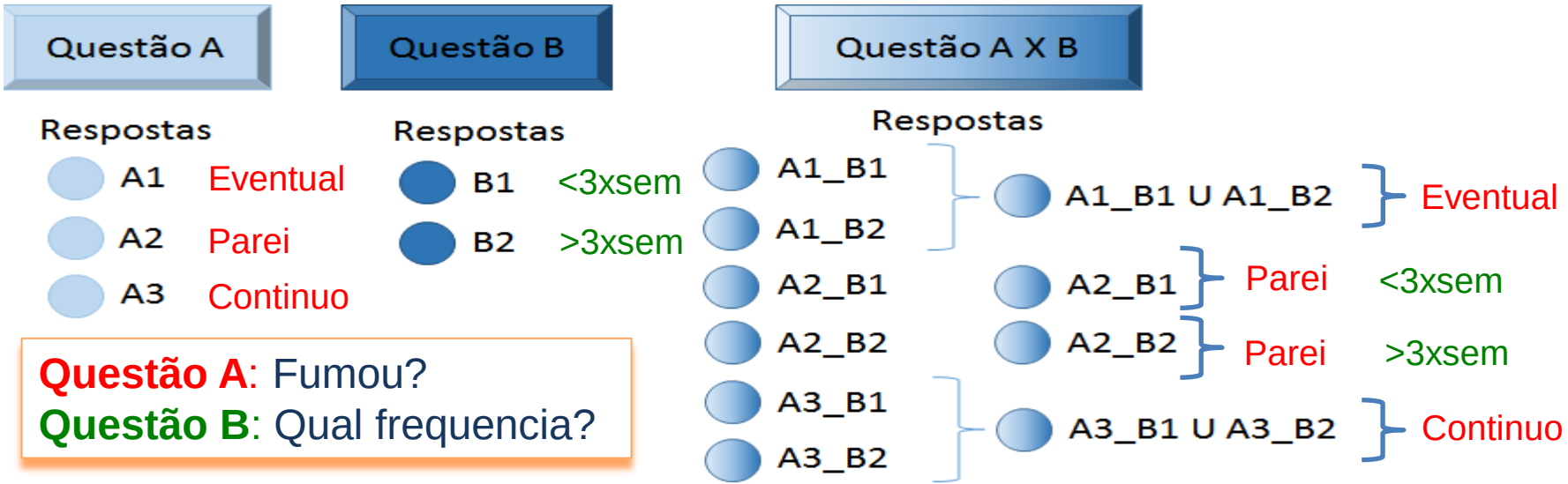
Empréstimo Janeiro	Empréstimo Fevereiro	Empréstimo Março	Empréstimo Abril	**Valor Médio do Empréstimo nos últimos meses**	**Desvio Padrão**	**Maior Emprêstimo em (65%) das vezes**
2000	3000	1000	2000	2000	700	2700

Histórico de empréstimos

Novas variáveis integradras

JUNÇÃO DE DADOS

- Em dados nominais, a Junção de vaiáveis altamente relacionadas pode trazer uma redução significativa na dimensionalidade sem perdas de informação.
- A Junção pode ser um simples produto cartesiano das opções de resposta a questões, ou uma re-codificação destas, criando novas opções.



JUNÇÃO DE DADOS POR ESCALA

Licap

Empréstimo consignado	Empréstimo não consignado	Financiamento bens	Financiamento automóveis	Financiamento imóveis	Cheque especial
0	1	1	1	0	0
...	...	...	...	...	...
1	0	0	0	1	0

Atributos	Pairwise Prejuízo financeiro
Empréstimo consignado	1º
Empréstimo não consignado	4º
Financiamento bens	3º
Financiamento automóveis	5º
Financiamento imóveis	6º
Cheque especial	2º

Atributos	Grupos de Prejuízo Financeiro
Empréstimo consignado	1º
Cheque especial	2º
Financiamento bens	3º
Empréstimo não consignado	4º
Financiamento automóveis	5º
Financiamento imóveis	6º

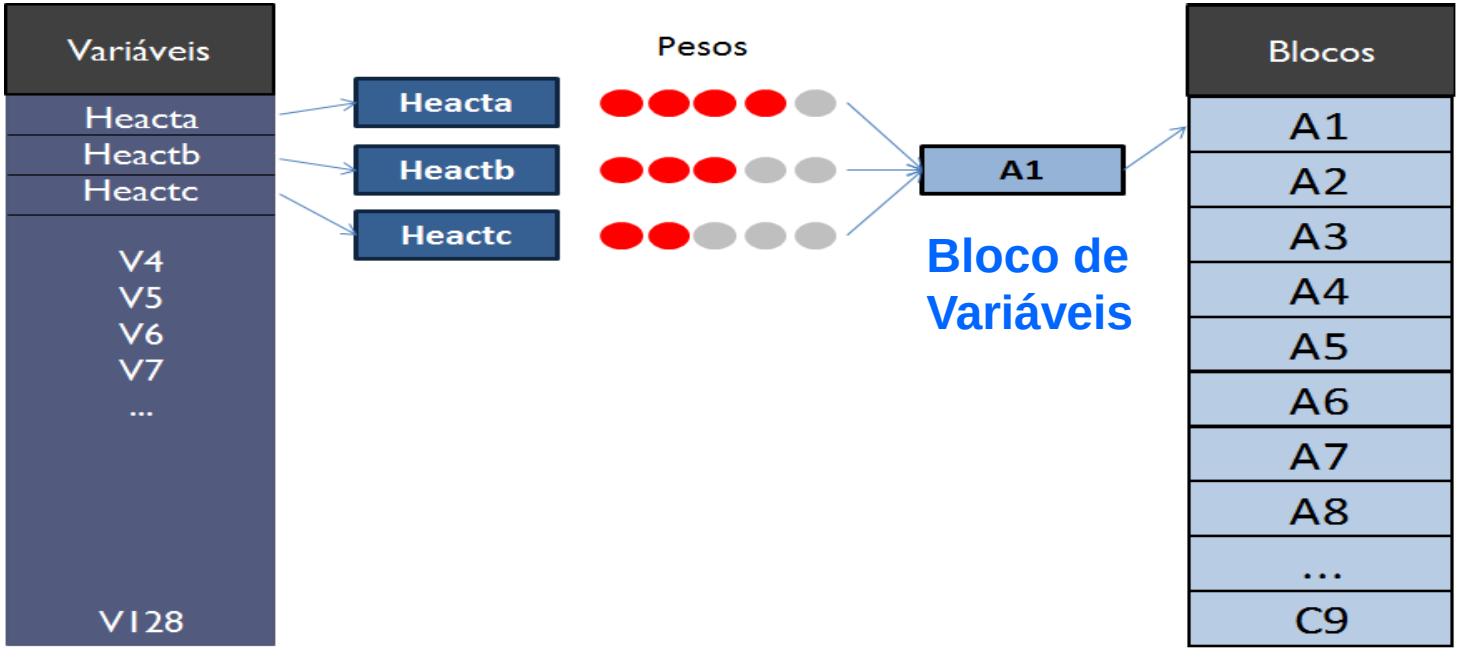
Novos Atributos	Prejuízo financeiro
Dívidas de Baixo Prejuízo Financeiro	1º 2º 3º
Dívidas de Alto Prejuízo Financeiro	4º 5º 6º

Dívidas de Baixo Prejuízo Financeiro	Dívidas de Alto Prejuízo Financeiro
1	1
...	...
1	0

BLOCAGEM DE DADOS

Licap

Os blocos são formados pela informação das questões que os compõem, e recebe valores diretamente delas, com uma ponderação de acordo com a relevância da questão.



DISCRETIZAÇÃO DE VARIÁVEIS

Licap

- Muitos algoritmos como C4.5, Apriori, Naive Bayes utilizam dados categóricos ou nominais ao invés de dados contínuos. A discretização é uma importante e comúm tarefa em Data Mining.

- Muitos problemas do mundo real, contendo dados numéricos contínuos precissam de métodos de discretização que os convertam para dados categóricos.

- O processo de discretização transforma dados quantitativos em dados qualitativos. Por exemplo, transforma atributos numéricos em atributos discretos ou nominais com um número finito de intervalos.

DISCRETIZAÇÃO DE VARIÁVEIS

Licap

Vantagens:

Algoritmo capazes de lidar com dados contínuos, podem ser menos eficientes para a compreensibilidade dos resultados.
A discretização também pode ser entendida como uma técnica para a redução de dimensionalidade.

Desvantagens:

Perda de informação
Reduzir a perda de informação é o principal objetivo dos métodos de discretização.

DISCRETIZAÇÃO DE VARIÁVEIS

Licap

Não-supervisionada:

A discretização é realizada sem levar em conta a informação dos grupos a que pertencem as instâncias de treinamento.

Supervisionada:

A discretização é realizada levando em conta os grupos a que pertencem as instâncias no conjunto de treinamento.

Não-Supervisionada: Mapeamento em intervalos

Licap

a) Intervalos com tamanho pré-definidos (univariada)

0 a 1 → 0 2 a 5 → 1 6 a 99 → 2

b) Intervalos de igual tamanho (não necessariamente com a mesma quantidade de valores)

2 intervalos / 5 valores: 0 a 4 → 0 5 a 9 → 1

c) Intervalos com o mesmo número de elementos

{1,2,3,6,7,8,9,13,15,16}

{1,2,3,6,7} →1 {8,9,13,15,16} →2

Não-Supervisionada: Mapeamento em intervalos

Licap

d) Número pré-determinado de intervalos uniformes (*equal-interval binning*)

No exemplo (idade):

64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

Considerando Bins (caixas) com largura **6**:

60 < x ≤ 66: 64, 65
66 < x ≤ 72: 68, 69, 70, 71, 72, 72
72 < x ≤ 78: 75, 75
78 < x ≤ 84: 80, 81, 83
84 < x ≤ 90: 85

- Como qualquer método não supervisionado, arrisca destruir diferenças úteis, devido a divisões muito grandes ou fronteiras inadequadas
- Distribuição de amostras muito irregular.

Não-Supervisionada: Mapeamento em intervalos

Licap

e) Número uniforme de amostras por intervalo (*equal-frequency binning*).

No exemplo (idade):

64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

64 65 68 69|70 71 72 72 |75 75 80|81 83 85

14 amostras: **4** *Bins*

x ≤ 69,5:	**64, 65, 68, 69**
69,5 < x ≤ 73,5:	**70, 71, 72, 72**
73,5 < x ≤ 80,5:	**75, 75, 80**
x > 80,5:	**81, 83, 85**

- Também chamado de equalização do histograma.
- Cada *bin* tem o mesmo número aproximado de amostras

Não-Supervisionada: Mapeamento em intervalos

Licap

a) Intervalos com tamanho pré-definidos (univariada)

0 a 1 → 0 2 a 5 → 1 6 a 99 → 2

b) Intervalos de igual tamanho (não necessariamente com a mesma quantidade de valores)

2 intervalos / 5 valores: 0 a 4 → 0 5 a 9 → 1

c) Intervalos com o mesmo número de elementos

{1,2,3,6,7,8,9,13,15,16}

{1,2,3,6,7} →1 {8,9,13,15,16} →2

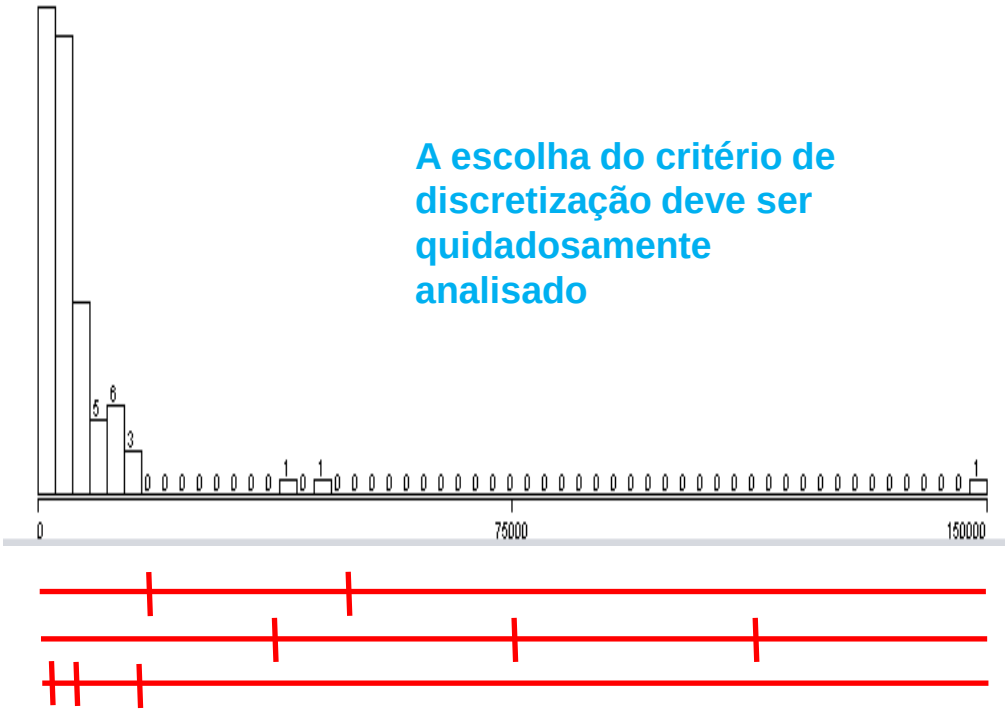
d) Número pré-determinado de intervalos uniformes (*equal-interval binning*)

Considerando Bins (caixas) com largura **6**:

60 < x ≤ 66: 64, 65
66 < x ≤ 72: 68, 69, 70, 71, 72, 72
72 < x ≤ 78: 75, 75
78 < x ≤ 84: 80, 81, 83
84 < x ≤ 90: 85

e) Número uniforme de amostras por intervalo (*equal-frequency binning*).

x ≤ 69,5: 64, 65, 68, 69
69,5 < x ≤ 73,5: 70, 71, 72, 72
73,5 < x ≤ 80,5: 75, 75, 80
x > 80,5: 81, 83, 85

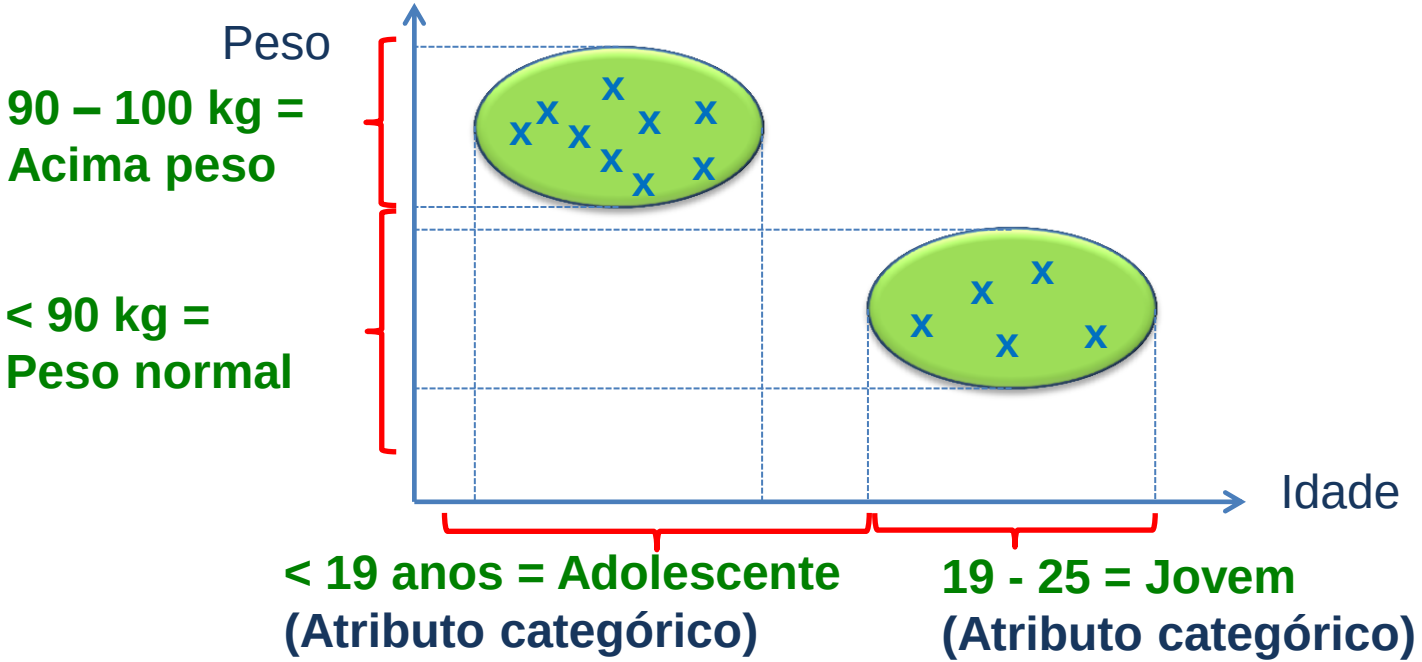


A escolha do critério de discretização deve ser quidadosamente analisado

Supervisionada: Mapeamento por grupos

Licap

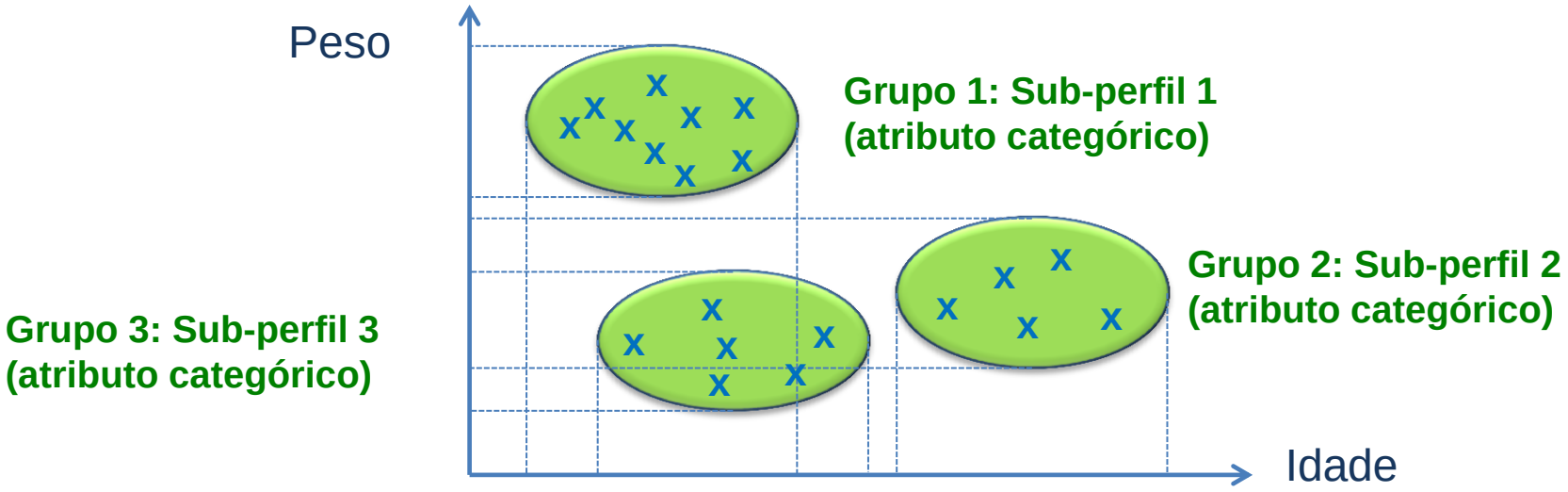
f) Intervalos por meio de clusterização (multivariada)



Supervisionada: Mapeamento por grupos

Licap

- A técnica de mapeamento por intervalos pode levar a sobre-posição de variáveis categóricas o que pode levar a um processo de discretização com incertezas. Frente a este problema pode ser dado um valor categórico para todo um grupo de variáveis.



Tipos de Discretizadores

Licap

•**Direto versus Incremental:**

- Os métodos de discretização **diretos** dividem os valores do atributo em um **número de intervalos** previamente definido pelo usuário.

- Nos métodos **incrementais**, o processo de discretização é realizado iterativamente até que **um critério de parada** seja alcançado.

Tipos de Discretizadores

Licap

- **Estático versus dinâmico:**

-Um discretizador **estático** é executado **antes** do processo de aprendizagem e independe do algoritmo de aprendizagem.

-Um discretizador **dinâmico** é executado **no momento** da construção do modelo, ou seja, o discretizador é embutido no algoritmo de aprendizagem.

Tipos de Discretizadores

Licap

•**Univariada versus Multivariada:**

- Métodos **univariados** consideram **um atributo** contínuo por vez, não levando em conta a relação entre os atributos.

- Multivariados podem considerar **simultaneamente todos os atributos** e a relação de dependência entre eles.

Tipos de Discretizadores

Licap

•**Supervisionado versus não-supervisionado:**

- Discretizadores **não-supervisionados** não consideram o rótulo da classe, enquanto os **supervisionados** o fazem.

- A maioria dos discretizadores propostos na literatura é supervisionada e teoricamente, usando informações de classe, deve determinar automaticamente o melhor número de intervalos para cada atributo.

Tipos de Discretizadores

Licap

- **Divisão versus Fusão:**

- Métodos de **Divisão** realizam a discretização por meio de um processo iterativo de subdivisão do intervalo de valores inicial que é executado até que uma condição de parada seja satisfeita.

- Os métodos de **Fusão** iniciam com os valores do atributo contínuo particionados e, iterativamente, realizam a junção dessas partições enquanto um critério de parada não é alcançado

- Além disso, alguns discretizadores podem ser considerados **híbridos** devido ao fato de que eles podem alternar as divisões.

Tipos de Discretizadores

Licap

- **Global versus Local:**

Para tomar uma decisão, um discretizador pode considerar todos os dados disponíveis do atributo ou usar apenas informações parciais.

- Um discretizador **global** utiliza todas as informações disponíveis. Um discretizador **local** faz uso apenas de parte das informações

- Alguns algoritmos seguem o esquema de divisão e conquista e quando é encontrada uma divisão, os dados são recursivamente divididos, restringindo o acesso a dados parciais.

Tipos de Discretizadores

Licap

- **Medidas de avaliação:**

- Essas medidas de avaliação são utilizadas pelos discretizadores para comparar intervalos de valores durante o processo de discretização.

São consideradas 5 medidas:

- **Informação**: utiliza a entropia
- **Estatística**: dependência/correlação entre os atributos
- **Conjuntos aproximados**: uso de medidas de conjuntos aproximados
- ***Wrapper****:* execução de algoritmos de classificação.
- ***Binning****:* quantidade pré-determinada de intervalos

Licap

Formação Cientista de Dados

Obrigado!

PUC Minas