CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Conceitos

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Agenda

- Tarefas de mineração
- Tipos de aprendizado
- Instâncias e atributos
- Preparação e observação dos dados
- Desafios da mineração de dados

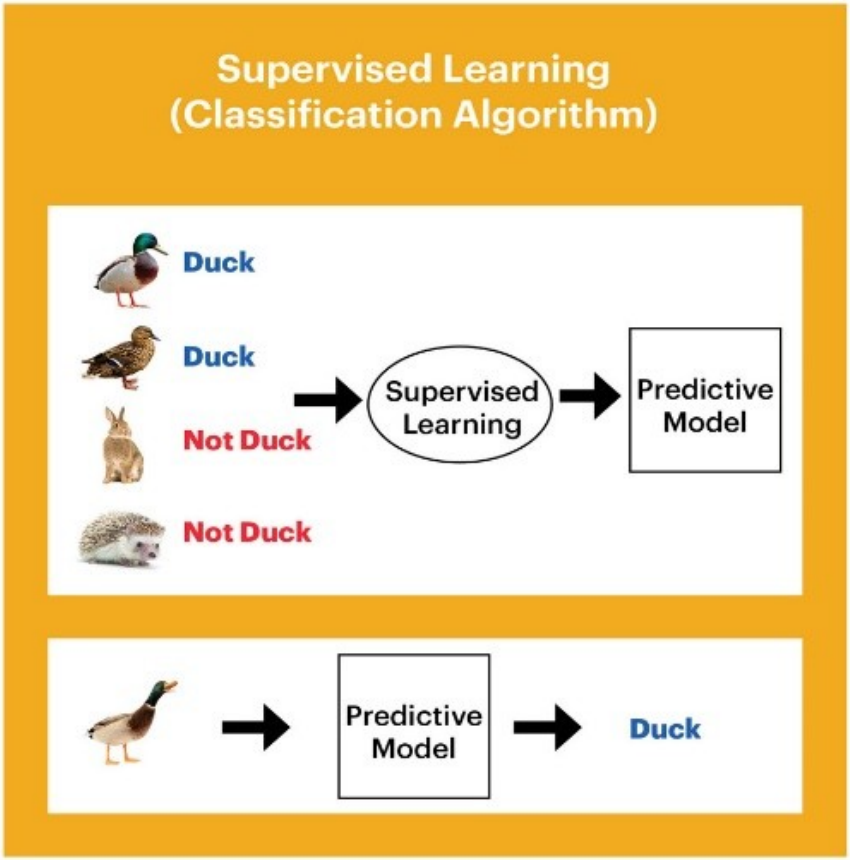
Conceitos/Tarefas da Mineração de Dados

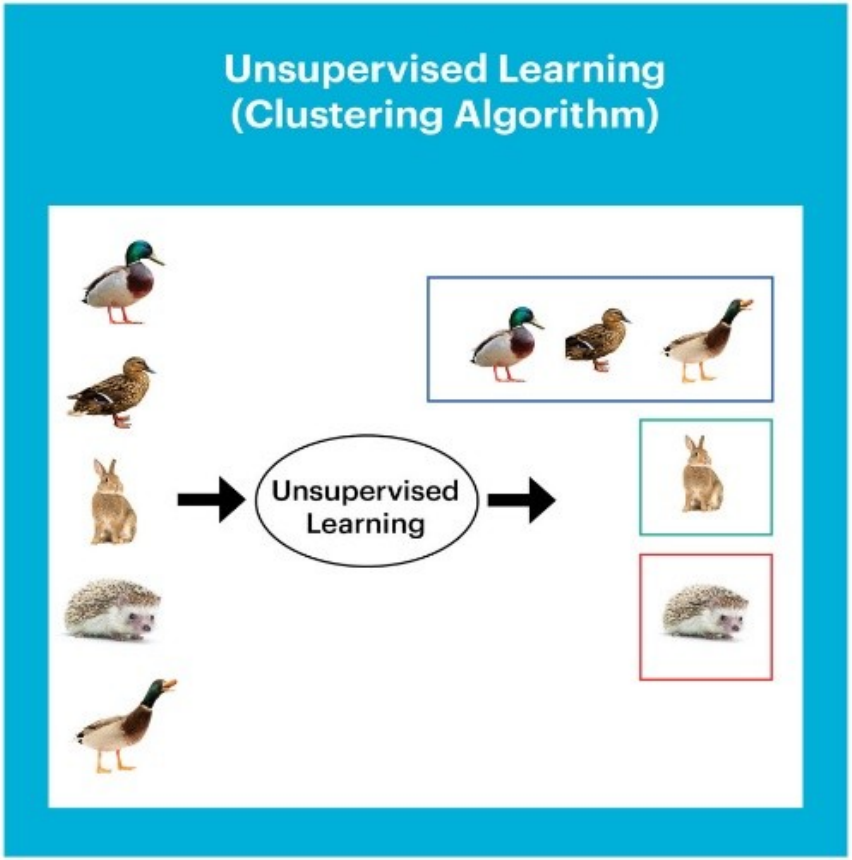
- **Conceito**
  - O que deve ser aprendido
- **Descrição do conceito**
  - Representação do aprendizado
- **Estilos de aprendizagem**
  - **Tarefas Preditivas**: têm a finalidade de prever o valor de um atributo alvo (variável dependente) baseando-se nos valores de outros atributos (variáveis independentes)
    - Classificação
    - Regressão
  - **Tarefas Descritivas**: têm o propósito de derivar padrões (e.g., regras, correlações, tendências, grupos) que expliquem relacionamentos nos dados
    - Clusterização
    - Regras de Associação

Conceitos/Tipos de Aprendizado

- **Aprendizado Supervisionado**
  - O apredizado acontece sob um esquema de supervisão, ou seja, um **conceito** é previsto a partir do treinamento de conceitos fornecidos por instâncias conhecidas, e posteriormente, testado com novas instâncias
  - A taxa de sucesso nos testes fornece uma medida objetiva que quão bem foi o aprendizado
- **Aprendizado Não Supervisionado**
  - O problema é descobrir um conceito interessante, ou seja, não existe a busca por um **conceito** específico
  - Pode ser uma associação entre conceitos ou um agrupamento em torno de alguns conceitos

Exemplos


Supervised Learning
(Classification Algorithm)
Duck
Duck
Not Duck
Not Duck
Supervised Learning
Predictive Model
Predictive Model
Duck


Unsupervised Learning
(Clustering Algorithm)
Unsupervised Learning

Supervisionado

- **Classificação**
  - Lentes de contato, íris, empréstimo
  - Conceito (atributo alvo) é chamado de classe
  - Avalia o sucesso em novos dados pelos quais os rótulos de classe são conhecidos (instâncias de teste)
  - O sucesso é frequentemente medido por uma taxa de acerto
- **Regressão**
  - Velocidade da CPU, altura de pessoas, preço de produtos
  - Variante da tarefa de classificação em que o valor do conceito (atributo alvo) é **numérico**
  - Também avalia o sucesso nas instâncias de teste
  - Na prática, o sucesso é medido frequentemente por uma taxa de erro

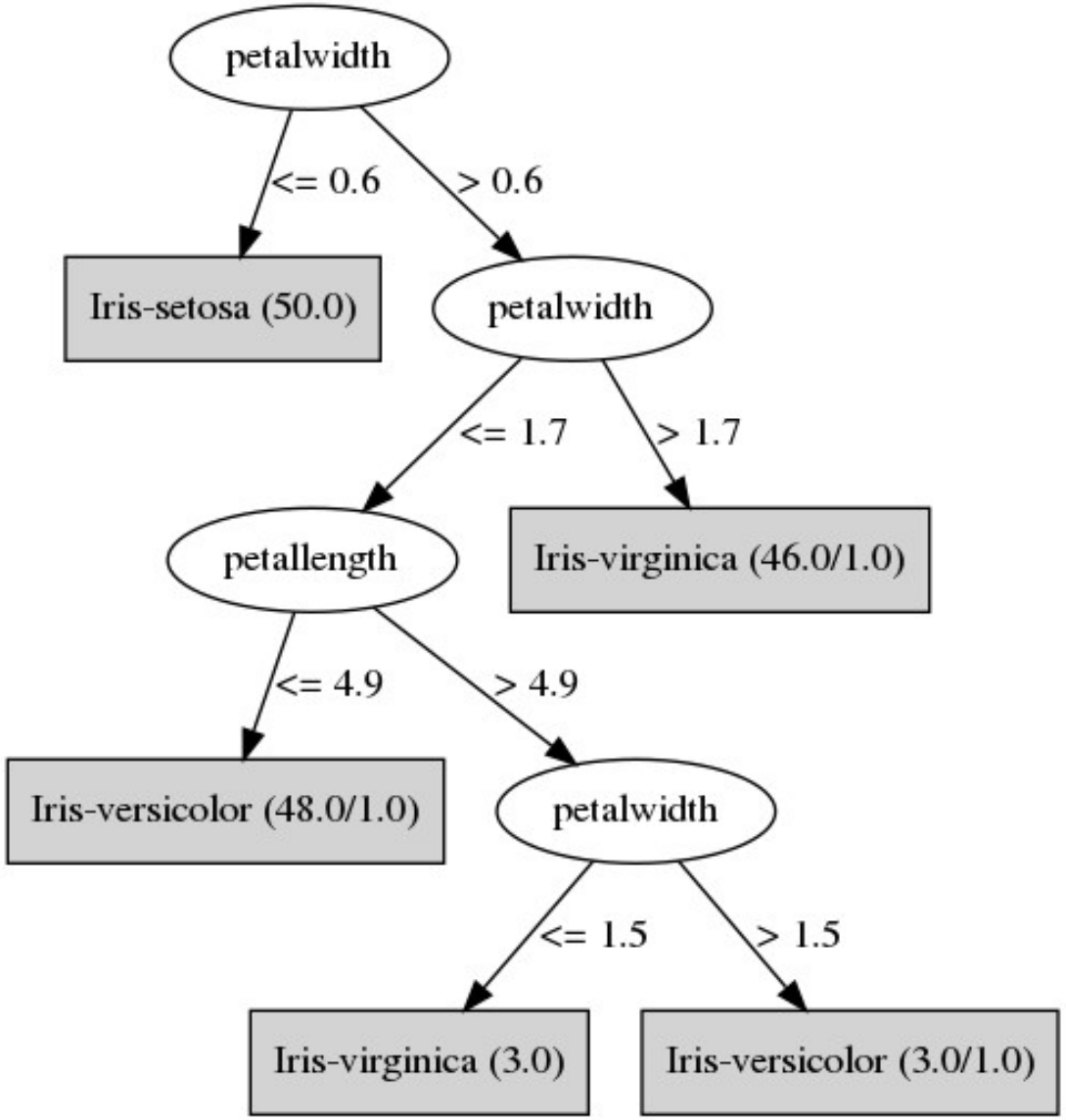
Não Supervisionado

- **Regras de Associação**
  - Pode ser aplicado se nenhuma classe for especificada
  - Pode descrever o valor de **qualquer atributo**, não apenas a classe, e o valor de mais de um atributo por vez
  - Regras de associação são mais numerosas que regras de classificação
  - São necessárias restrições (medidas de interesse)

- **Clusterização**
  - Localiza **grupos** de instâncias semelhantes
  - A classe de uma instância não é conhecida
  - O sucesso é frequentemente medido de forma subjetiva

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Exemplo – Classificação



Exemplo – Regressão

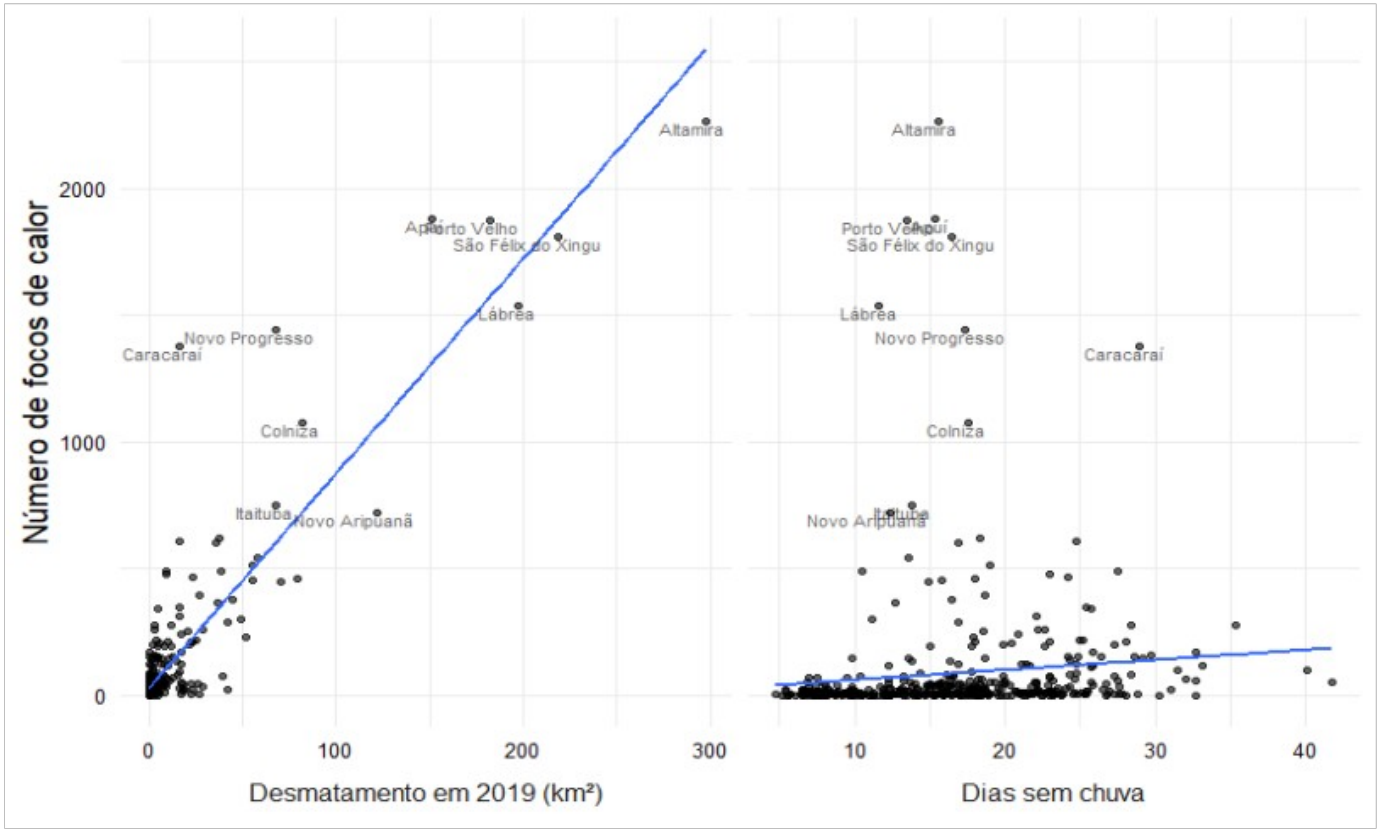


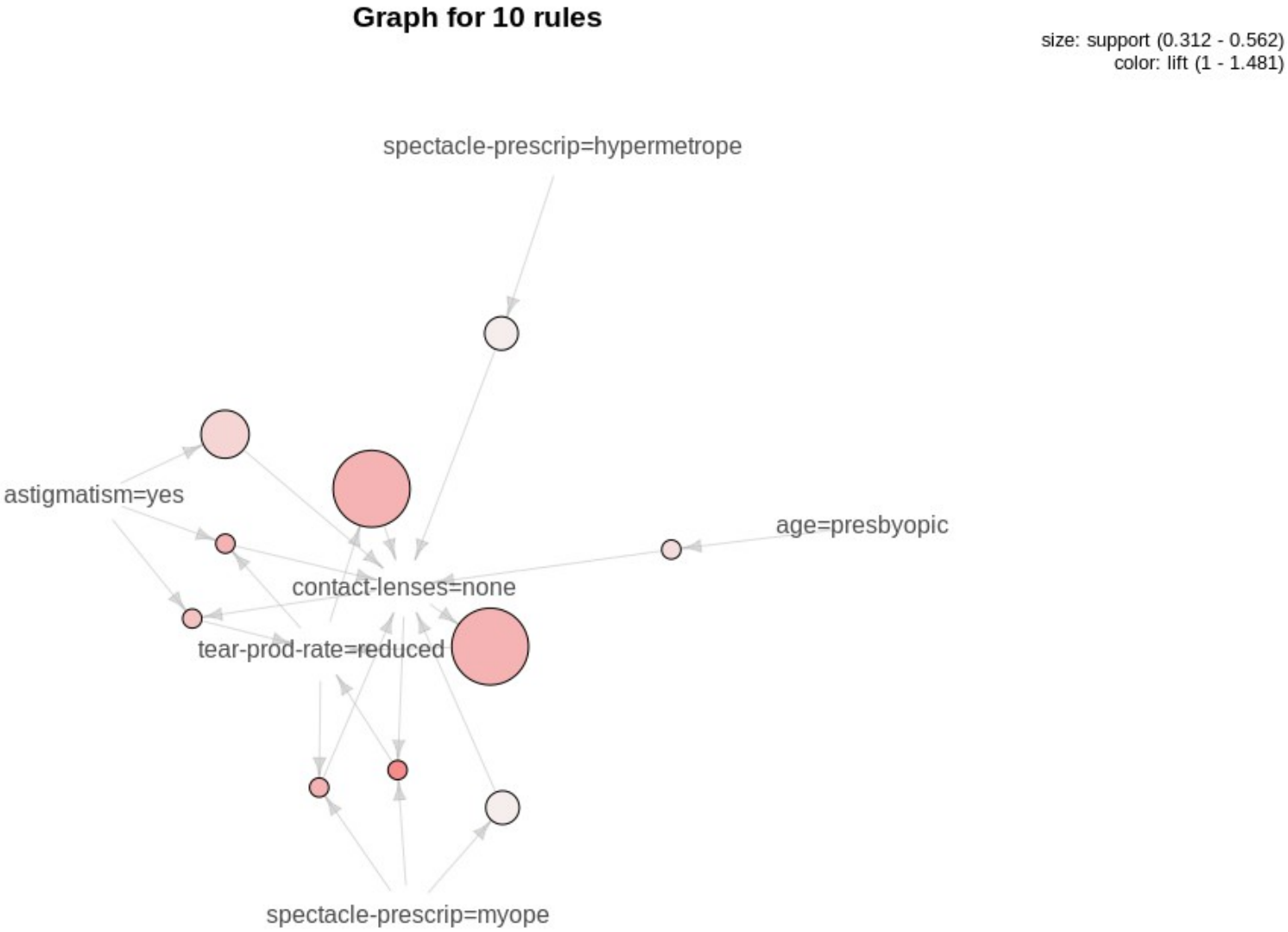
Figura 1 - Relação entre o número de focos de incêndios acumulados até 19 de agosto e área desmatada (esquerda) e número cumulativo de dias sem chuva (direita) para municípios do bioma Amazônia em 2019. Os municípios identificados no gráfico são aqueles onde se registrou um número particularmente elevado de focos de incêndios. Fonte: IPAM

Disponível em: https://ipam.org.br/bibliotecas/nota-tecnica-amazonia-em-chamas/

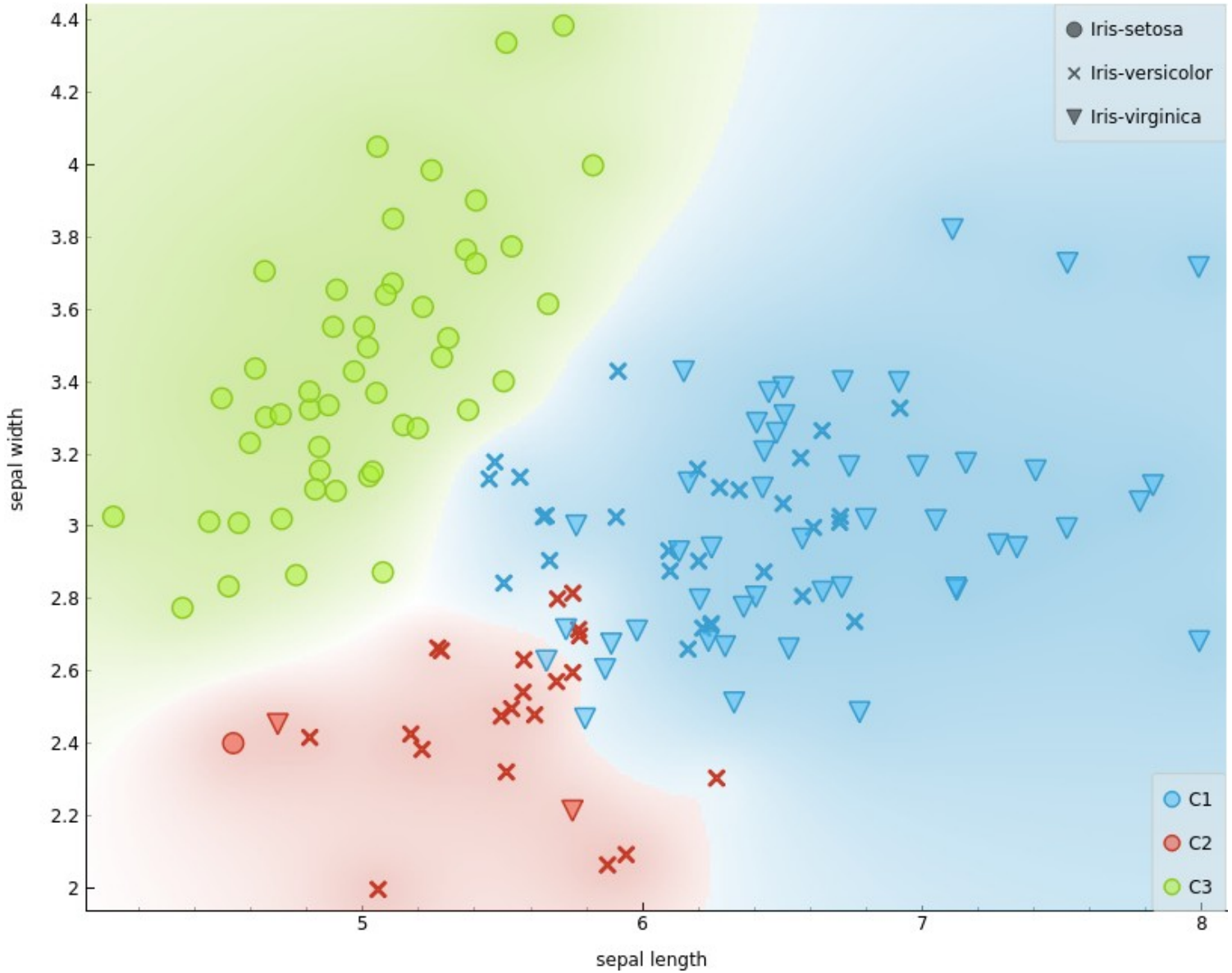
Exemplo – Regras de Associação

			Support	Confidence	Lift
[1]	{tear-prod-rate=reduced}	=> {contact-lenses=none}	0.5625	1.0000000	1.333333
[2]	{contact-lenses=none}	=> {tear-prod-rate=reduced}	0.5625	0.7500000	1.333333
[3]	{astigmatism=yes}	=> {contact-lenses=none}	0.4375	0.8750000	1.166667
[4]	{spectacle-prescrip=hypermetrope}	=> {contact-lenses=none}	0.3750	0.7500000	1.000000
[5]	{spectacle-prescrip=myope}	=> {contact-lenses=none}	0.3750	0.7500000	1.000000
[6]	{age=presbyopic}	=> {contact-lenses=none}	0.3125	0.8333333	1.111111
[7]	{astigmatism=yes,tear-prod-rate=reduced}	=> {contact-lenses=none}	0.3125	1.0000000	1.333333
[8]	{astigmatism=yes,contact-lenses=none}	=> {tear-prod-rate=reduced}	0.3125	0.7142857	1.269841
[9]	{spectacle-prescrip=myope,tear-prod-rate=reduced}	=> {contact-lenses=none}	0.3125	1.0000000	1.333333
[10]	{spectacle-prescrip=myope,contact-lenses=none}	=> {tear-prod-rate=reduced}	0.3125	0.8333333	1.481481

Exemplo – Regras de Associação



Exemplo - Clusterização



Instâncias e atributos

- **Instâncias**
  - Exemplos individuais e independentes de um conceito a ser aprendido
  - Caracterizado por um conjunto de atributos
  - Usadas como entrada para o esquema de aprendizado (um conjunto de instâncias)
- **Atributos**
  - Características que medem aspectos de uma instância
  - A quantidade pode variar muito
  - Problema relacionado: a existência de um atributo pode depender do valor de outro
  - Tipos
    - Nominais
    - Numéricos

Atributos

- **Numéricos**
  - Assumem valores reais ou inteiros e as vezes são chamados de contínuos. No entanto, valores inteiros certamente não são contínuos no sentido matemático
- **Nominais**
  - Assumem valores em um conjunto finito de possibilidades e as vezes são chamados de categóricos
  - Os próprios valores servem apenas como rótulos ou nomes, daí o termo nominal (origem da palavra latina nome)
  - Nenhuma relação está implícita entre os valores nominais (medida ou distância)
  - Somente testes de igualdade podem ser realizados

Nominal *versus* Ordinal

- O tipo nominal não implica nenhuma ordem nos valores
- Exemplo
  - Atributo "*temperature*"
  - Valores: "*hot*" > "*mild*" > "*cool*"
- Não é possível realizar operação de adição ou subtração
- A distinção entre nominal e ordinal nem sempre é muito clara
- Exemplo
  - Atributo: "*outlook*"
  - Valores: "*sunny*", "*overcast*", "*rainy*"

Na prática

- Atributos nominais frequentemente são gerados por um processo de **discretização**
- Por isso, os atributos nominais também são chamados de **categóricos** ou **discretos**
- Discreto também tem conotação de ordenação porque define intervalos para valores numéricos
- Caso especial para os atributos que assumem apenas dois valores (*boolean*)
- Muitos algoritmos funcionam apenas com atributos do tipo nominal e numérico
- Atributos podem conter metadados, conhecimento implícito e difícil de ser capturado

Preparação dos dados

- Preparar a entrada de dados geralmente consome muito tempo e esforço
- Os dados reais são, muitas vezes, decepcionantemente baixos em qualidade
- Preocupações
  - Diferentes estilos e fontes de dados
  - Setores e empresas
  - Bases de dados e ambientes computacionais
  - Integração com outros sistemas
  - Transformação e limpeza dos dados

O arquivo CSV (Comma Separated Values)

```
outlook,temperature,humidity,windy,play
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Preparando os dados

- A interpretação dos tipos de atributos depende da estratégia de aprendizado dos algoritmos
- Alguns usam comparações menor que e maior que entre os valores (árvores de decisão)
- Outros tratam como escalas de proporção e usam cálculos de distância (vizinho mais próximo)
- Transformações
  - **Normalização**: distribuir os valores em um intervalo fixo [0,1], dividindo todos os valores pelo valor máximo encontrado ou subtraindo o valor mínimo e dividindo pelo intervalo entre os valores máximo e mínimo.
  - **Padronização**: calcular a média e o desvio padrão dos valores, subtrair a média de cada valor e dividir o resultado pelo desvio padrão

Preparando os dados – Valores ausentes

- Muitas bases de dados contêm valores ausentes
- O tratamento da ausência de valores deve ser executado de forma cuidadosa
- Razões
  - Mau funcionamento do equipamento, mudanças ou impossibilidade durante a coleta de dados
  - Muitas vezes não há significado particular no fato de que uma determinada instância tem um valor de atributo ausente: o valor simplesmente não é conhecido
  - No entanto, pode haver uma boa razão pela qual o valor do atributo é desconhecido
  - Em algumas circunstâncias, pacientes podem ser diagnosticados apenas a partir dos testes que um médico decide fazer, independentemente do resultado

Preparando os dados – Valores Incorretos

- Razão
  - Os dados não foram coletados para fazer mineração
- Consequência
  - Valores incorretos podem afetar diretamente o resultado obtido
- Erros de digitação em atributos nominais precisam ser corrigidos
- Erros de digitação ou medição em atributos numéricos devem ser identificados (*outliers*)
- Outros problemas
  - Duplicidade de instâncias
  - Dados obsoletos

Preparando os dados – Dados desbalanceados

- Na tarefa de classificação, é comum que uma classe seja muito mais frequente que as outras
  - Por exemplo, ao prever se o **Flamengo** ganhará do **Vasco**, é bastante seguro prever que a resposta é **verdadeira**
- Problema
  - A previsão da classe majoritária pode ser bastante alta, mas só isso pode não ser muito útil
- Prever que nenhum paciente tem a doença rara fornece alta precisão de classificação
- Classificar incorretamente um paciente afetado pode ser muito mais caro do que classificar incorretamente um paciente saudável

Estude os dados

- Ferramentas que mostram histogramas da distribuição de valores de atributos nominais e gráficos dos valores de atributos numéricos são muito úteis
- As visualizações gráficas dos dados facilitam a identificação de *outliers*, valores incorretos e ausentes
- Consultar *experts* no domínio de aplicação pode ser interessante
- O tempo que você gasta olhando para os seus dados é sempre um bom **investimento**

Desafios na Mineração de Dados

- Teoria unificada de Mineração de Dados
- Alta dimensionalidade dos dados
- Problemas relacionados ao processo de mineração de dados
- Privacidade e integridade de dados
- Dados não estáticos, desequilibrados e sensíveis

CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Conceitos

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com