

Artificial Intelligence/ Inteligência Artificial

Lecture 9: Supervised Learning/Aprendizagem Supervisionada (adaptado de Faria, 2018 e Castillo 2011)

Luís Paulo Reis

lpreas@fe.up.pt

Director of LIACC – Artificial Intelligence and Computer Science Lab.
Associate Professor at DEI/FEUP – Informatics Engineering Department,
Faculty of Engineering of the University of Porto, Portugal
President of APPIA – Portuguese Association for Artificial Intelligence

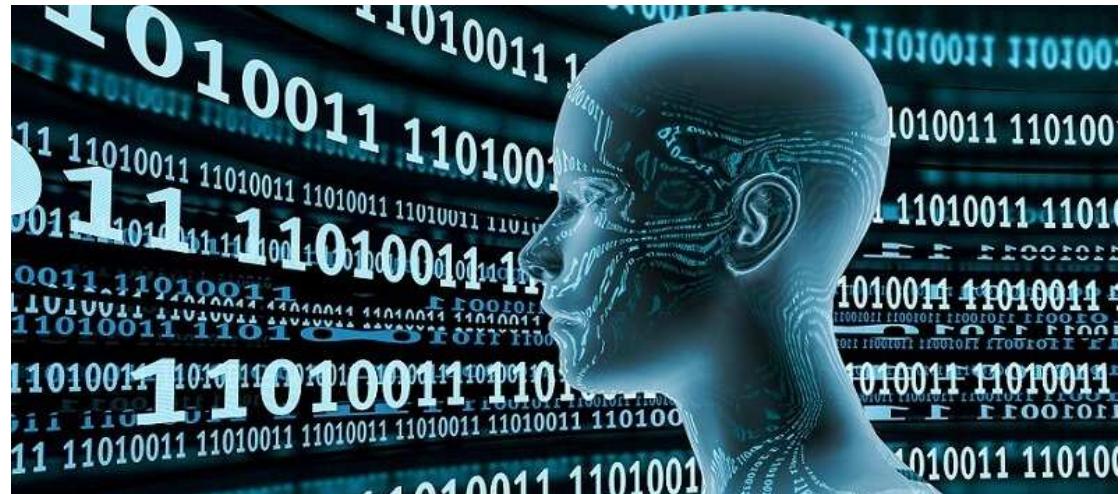
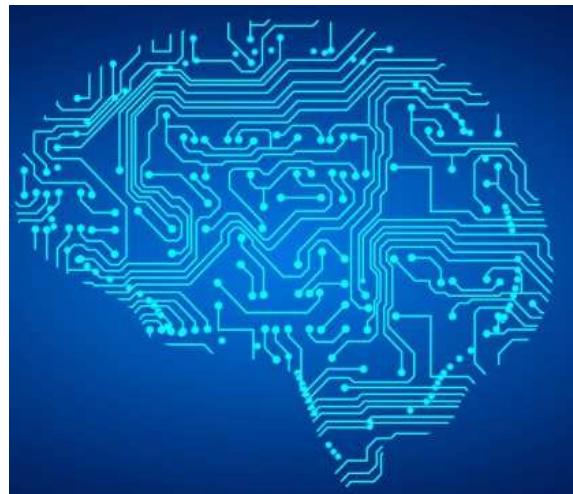


Artificial Intelligence: Machine Learning



Machine Learning

- Machine learning is a field of artificial intelligence that gives computer systems the **ability to "learn"** (e.g., progressively **improve performance** on a specific task) from data/results of their actions, without being explicitly programmed



Machine Learning

Herbert Simon (pioneiro da IA):

“Aprender implica mudanças adaptativas de um sistema no sentido de lhe permitir realizar, da próxima vez, a(s) mesma(s) tarefa(s), em circunstâncias análogas, de um modo mais eficiente e efetivo

Aprendizagem humana:

Aquisição de conhecimento (Knowledge acquisition)

Consiste na aquisição estruturada de nova informação simbólica assim como da capacidade de aplicar a situações novas.

Exemplo: aprender teoremas da Matemática, Ling. programação...

Refinamento de habilidades (skill refinement)

Processo, eminentemente Não (ou sub-) simbólico caracterizado por refinar um procedimento através de tentativas sucessivas até atingir uma proximidade desejada com o objetivo.

Exemplo: aprender a andar de bicicleta, nadar,...

Taxonomia dos Métodos de Aprendizagem Simbólica

- Taxonomia dos Métodos de Aprendizagem Simbólica
- Classificação quanto: à estratégia; à representação; ao domínio.
- Classificação quanto à estratégia usada:
 - “rote learning” (by being programmed)
 - aprendizagem por conselho ou por instrução (by being told or by instruction)
 - aprendizagem por analogia
 - aprendizagem por exemplos
 - aprendizagem por observação ou descoberta
 - aprendizagem por Reforço

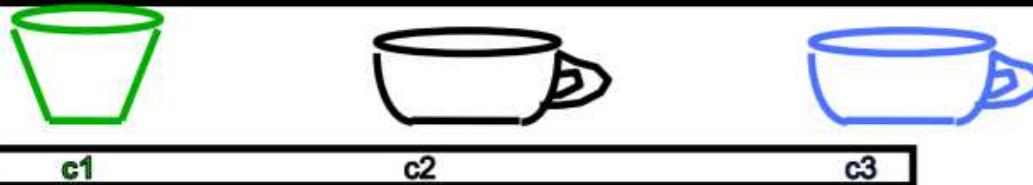
Taxonomia dos Métodos de Aprendizagem Simbólica

- Classificação quanto à representação do conhecimento adquirido
 - Parâmetros em expressões algébricas
 - Árvores de decisão (Ex ID3, C4.5)
 - Gramáticas formais
 - Regras de produção (ou Produções)
 - Expressões baseadas em lógica formal e formalismos semelhantes
 - Grafos e redes
 - Instâncias
 - Hiperplanos

Métodos Baseados em Explicações EBL

- Estratégia que inclui "aprendizagem por exemplos":
- Problema: Como identificar uma definição correta de um conceito a partir de exemplos positivos, tais que o conceito seja uma especialização de um conceito-genérico (target) definido pela "Teoria do Domínio" dada.
- 3 Versões do método (para a Especialização de Conceitos baseados na Teoria):
 - Generalização Baseada nas Explicações (EB)
 - EBG de exemplos múltiplos (mEBG)
 - uma versão diversa: Indução Sobre as Explicações (IOE)

Exemplo



Descrição dos objetos *C_i* (**exemplos treino-positivos**):

superficie(c1, s1).
feito-de(s1, plastico).
fundo(c1, f1).
feito-de(f1, plástico)
plano(f1).
tem(c1, p1).
concavidade(p1).
para-cima(p1).
pequeno(c1).
cilindrico(s1).

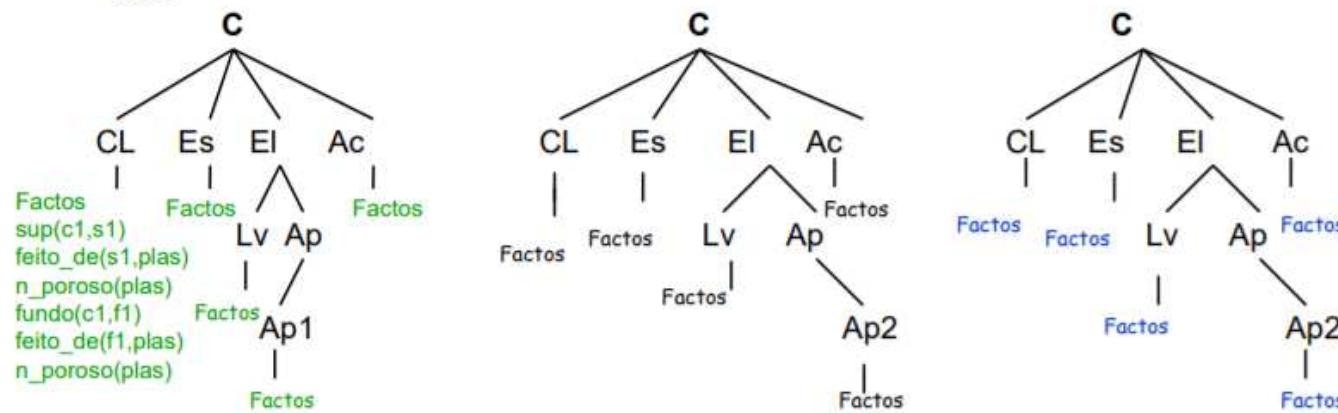
superficie (c2,s2)
feito-de(s2, plástico)
fundo(c2, f2).
feito-de(f2, alumínio)
plano(f2).
tem(c2, p2).
concavidade(p2).
para-cima(p2).
pequeno(c2).
tem(c2, p21).
asa(p21).

superficie(c3, s3).
feito-de(s3, porcelana).
fundo(c3, f3).
feito-de(f3, porcelana).
plano(f3).
tem(c3, p3).
concavidade(p3).
para-cima(p3).
pequeno(c3).
tem(c23, p31).
asa(p31).

Exemplo



As árvores de **explicação** construídas para provar cada exemplo a partir da Teoria do Domínio são:



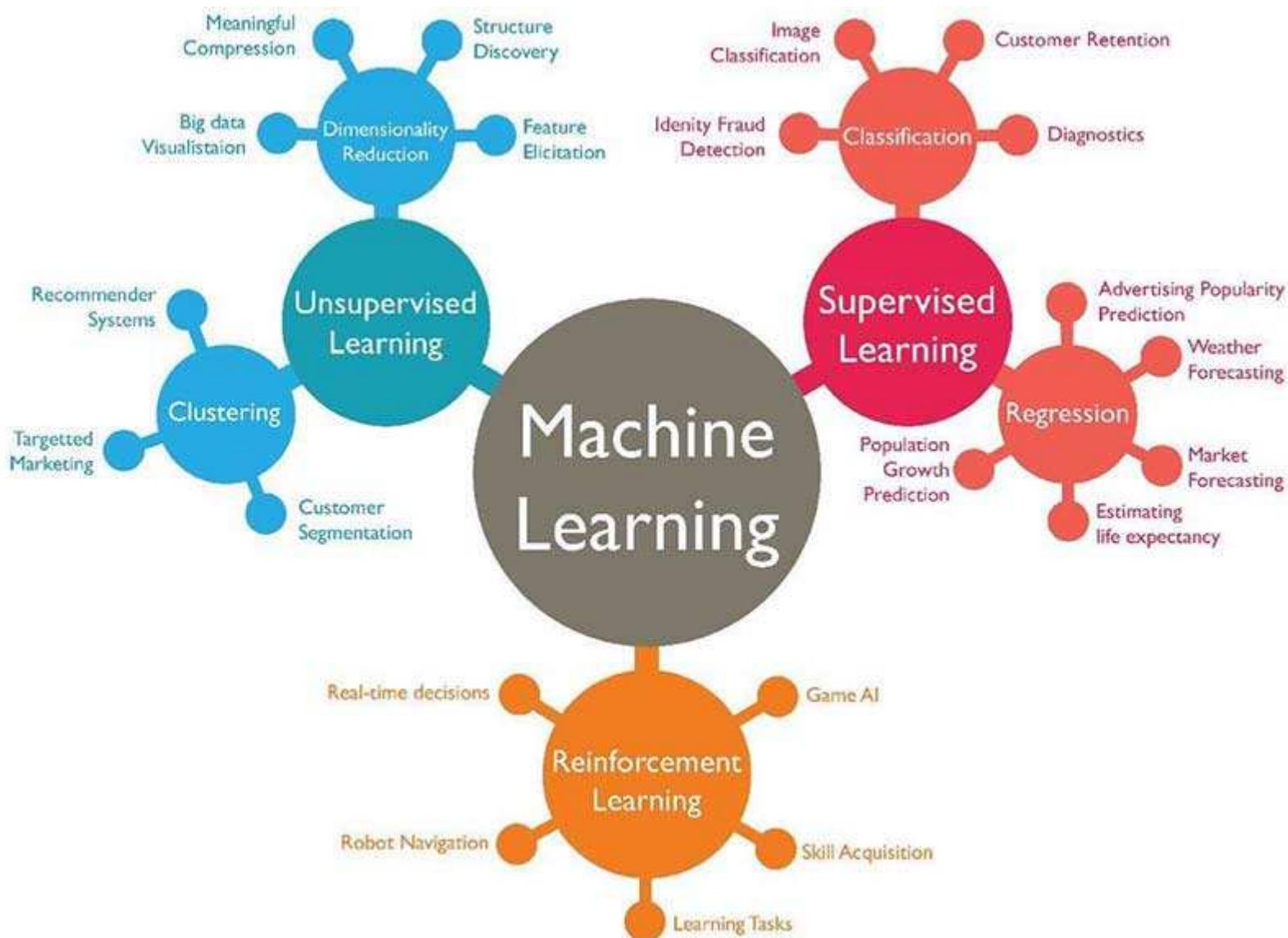
As folhas das árvores (não representadas) são os factos que descrevem os Exemplos

C1 "apanhável" porque pequeno, cilíndrico, isolante térmico; **C2 e C3** pequenos e tem_asa

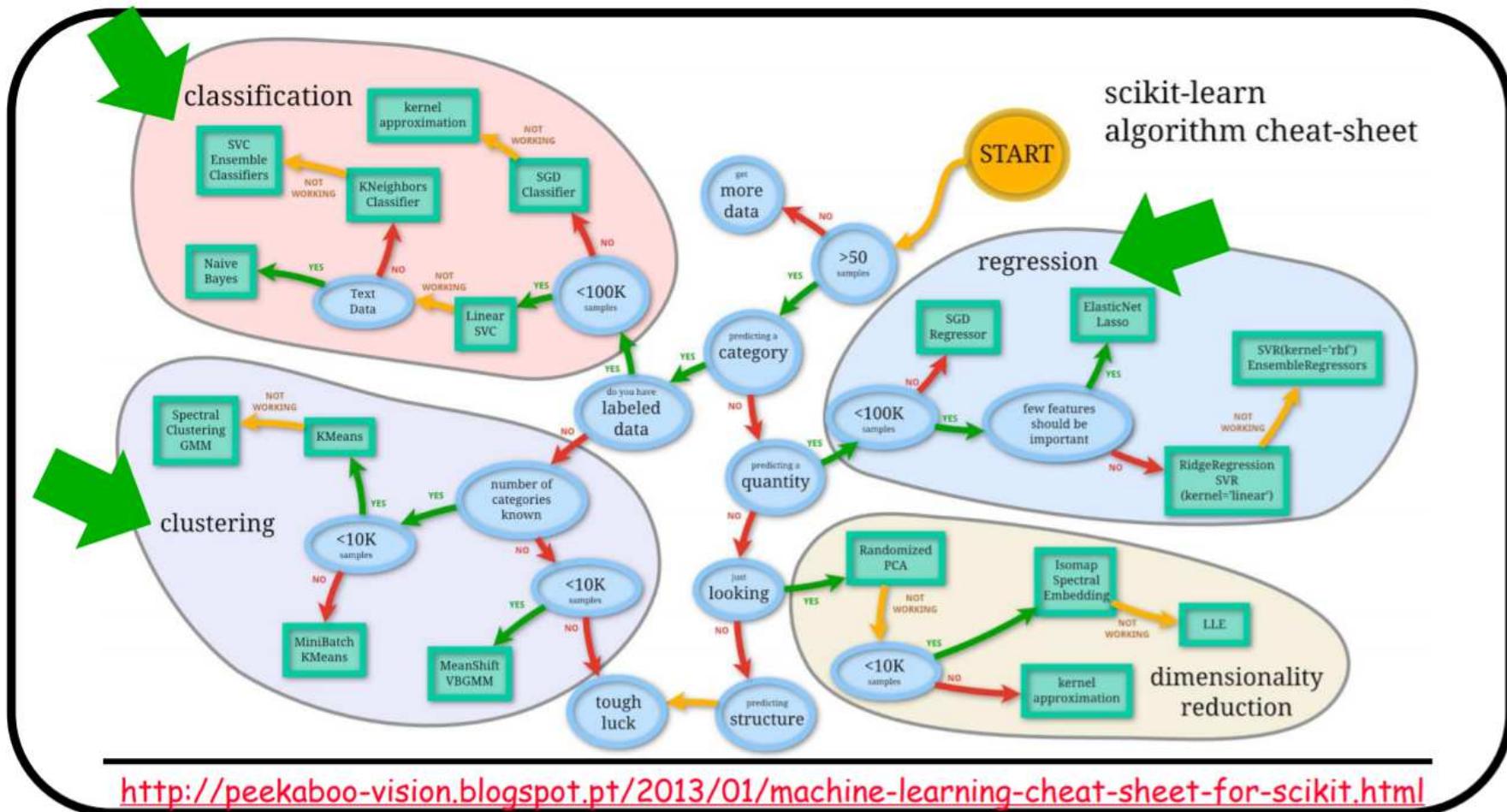
Machine Learning

- Machine Learning (ML) Tasks:
 - **Supervised learning:** Example inputs and desired outputs are available/given by a "teacher", and the goal is to learn how to map inputs to outputs (possibly semi-supervised)
 - Active learning: Computer has to obtain training labels/outputs for a limited set of instances and decide the objects to acquire
 - **Reinforcement learning:** Data (in form of rewards and punishments) are given only as feedback to the computer/agent actions in a dynamic environment
 - **Unsupervised learning:** No labels/outputs are given to the learning algorithm, leaving it on its own to find structure in its input

Machine Learning



Aprendizagem Máquina



Aprendizagem Supervisionada

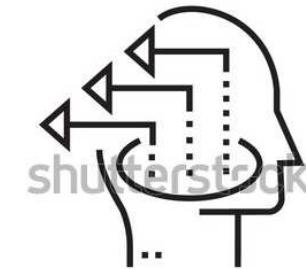
Extração de Conhecimento: Fases; Dados; Pré-processamento; Mineração de dados (Data Mining); Interpretação e Avaliação; Softwares

Aprendizagem Supervisionada: Avaliação de modelos; Fases de treino e teste; Métodos de estimação da taxa de erro (Holdout, Cross-validation, Leave-one-out, Bootstrap); Métodos estatísticos para comparar classificadores. Curva ROC.

Algoritmos de Aprendizagem: Árvores de Decisão; Conjunto de Regras; Baseados em Instâncias; Bayesiana; Redes Neuronais; Máquinas de Suporte Vetorial

Extração de Conhecimento

- Motivação
 - Acesso a dados de forma mais fácil e rápida
 - Extração de padrões e correlações
 - Objetivos
 - Conhecer algoritmos de aprendizagem
 - Construção de modelos de predição
 - Aplicação aos dados extraídos em diferentes
 - Big Data
 - Gerada por tudo em qualquer altura
 - Cada processo digital
 - troca de media
 - De várias fontes a grande velocidade, volume
 - Sistemas, sensores e dispositivos



Knowledge Extraction



Extração de Conhecimento

- Oportunidades e desafios
 - Captura, armazenamento, análise, procura, partilha, transferência, visualização, extração de informação, segurança e privacidade
 - Diversas áreas de aplicação
 - Economia e Finanças
 - Ciência e Tecnologia
 - Setor da Energia
 - Governos (transparência)
 - Segurança
 - Medicina e Cuidados de Saúde

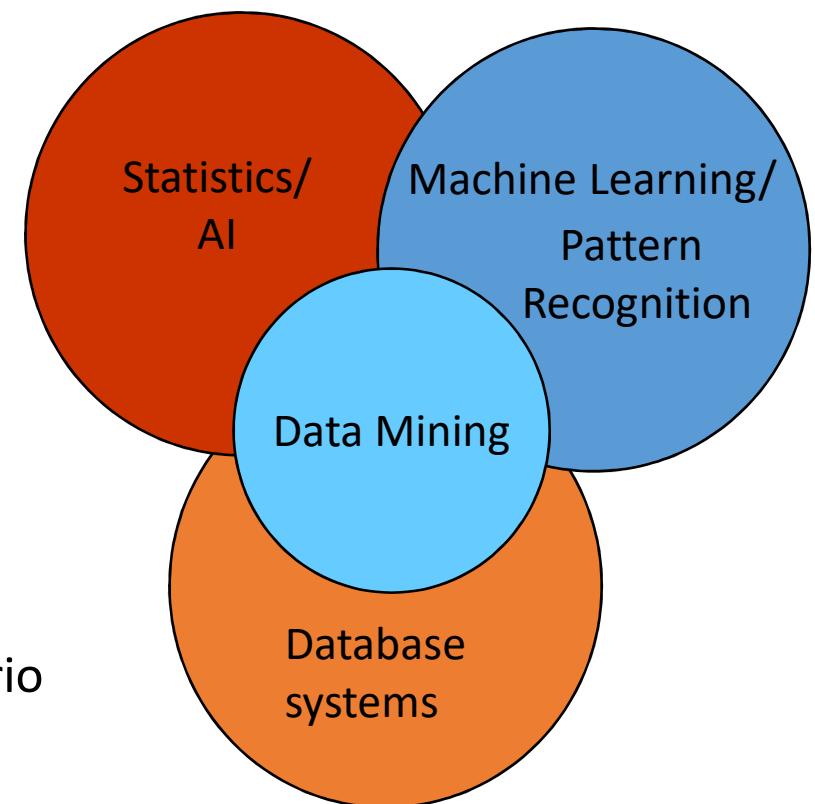


Extração de Conhecimento

- Técnicas tradicionais podem não ser solução para dados em bruto
- Novas abordagens podem ajudar os investigadores
 - na classificação e segmentação de dados
 - na formulação de hipóteses
- Há informação “escondida” nos dados que não é claramente evidente
- Analistas/estatísticos poderão demorar imenso a descobrir informação útil
- Muitos destes dados nunca são analizados

Extração de Conhecimento

- Extração de Conhecimento (Knowledge Discovery) e Mineração de Dados (Data Mining)
 - Áreas de crescente interesse
 - ciência, engenharia, saúde,...
 - Enorme volumes de dados que ao serem explorados podem dar mais informação
 - Exemplos
 - Classificação de Email/Spam
 - Criação de perfis na Amazon (sugestões...)
 - Atribuição de crédito bancário



Extração de Conhecimento

- Exemplos de Aplicações

Predição em Medical Data Mining - Aprender que tipo de pacientes apresentam alto risco de ser submetidas a cesariana a partir dos dados dos registos médicos

Patient 99 time=1
Age: 23
FirstPregnancy: no
Anemia: no; Diabetis: no
PreviousPrematureBirth: no
Ultrasound?:
Effective C-Section:?
...
Patient 99 time=2
Age: 23
FirstPregnancy: no
Anemia: no; Diabetis: yes
PreviousPrematureBirth: no
Ultrasound?:
Effective C-Section:?
...



9714 registos de mulheres grávidas, cada registo clínico contém 215 atributos

Modelo Preditivo

Patient 99 time=
Age: 23
FirstPregnancy: no
Anemia: no; Diabetis: no
PreviousPrematureBirth: no
Ultrasound?:
Effective C-Section: yes
...

Uma das 18 regras aprendidas

IF no previous vaginal delivery
AND Abnormal 2nd Trimester Ultrasound
AND Abnormal positioning at admission
THEN Probability of Emergency C-Section is 0.6

Extração de Conhecimento

• Exemplos de Aplicações

Tecnologias da Informação em Saúde – Aceder à informação sobre a qualidade de vida de pacientes recorrendo a dados clínicos, demográficos



- Avaliação
 - Qualidade de vida
 - Ganhos em Saúde
- Comparação de pacientes
 - Entre categorias
 - Dentro de cada categoria
 - Analisar valores desviantes
- Estratificação de pacientes recorrendo a técnicas de Data Mining
- Determinação de variação relevante no estado de saúde, fatores de sobrevivência e qualidade de vida
- Atualmente criar um instrumento de medida da Qualidade de Vida menos intrusivo

Extração de Conhecimento

- Exemplos de Aplicações

Tecnologias de Apoio – introdução de diferentes formas de comunicação indivíduo/computador(máquina)

Problema de classificação com dados de EEG

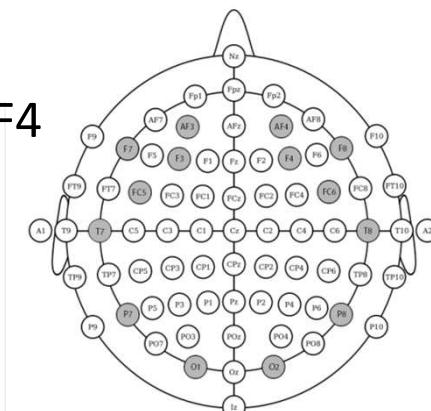
- Identificação da Expressão Facial
- Pacientes com Paralisia Cerebral

Dados adquiridos com uma Brain Computer Interface

- Permite a comunicação de indivíduos entre um computador/máquina através de sinais de EEG
- BCI Emotiv System
- Dados brutos de cada sinal do EEG
- AF3; F7; F3; FC5; T7; P7; O1; O2; P8; T8; FC6; F4; F8; AF4
- Valores do giroscópio
- Tempo



Fundação para a Ciéncia e a Tecnologia
MINISTÉRIO DA CIÉNCIA, TECNOLOGIA E ENSINO SUPERIOR



Extração de Conhecimento - Metodologias

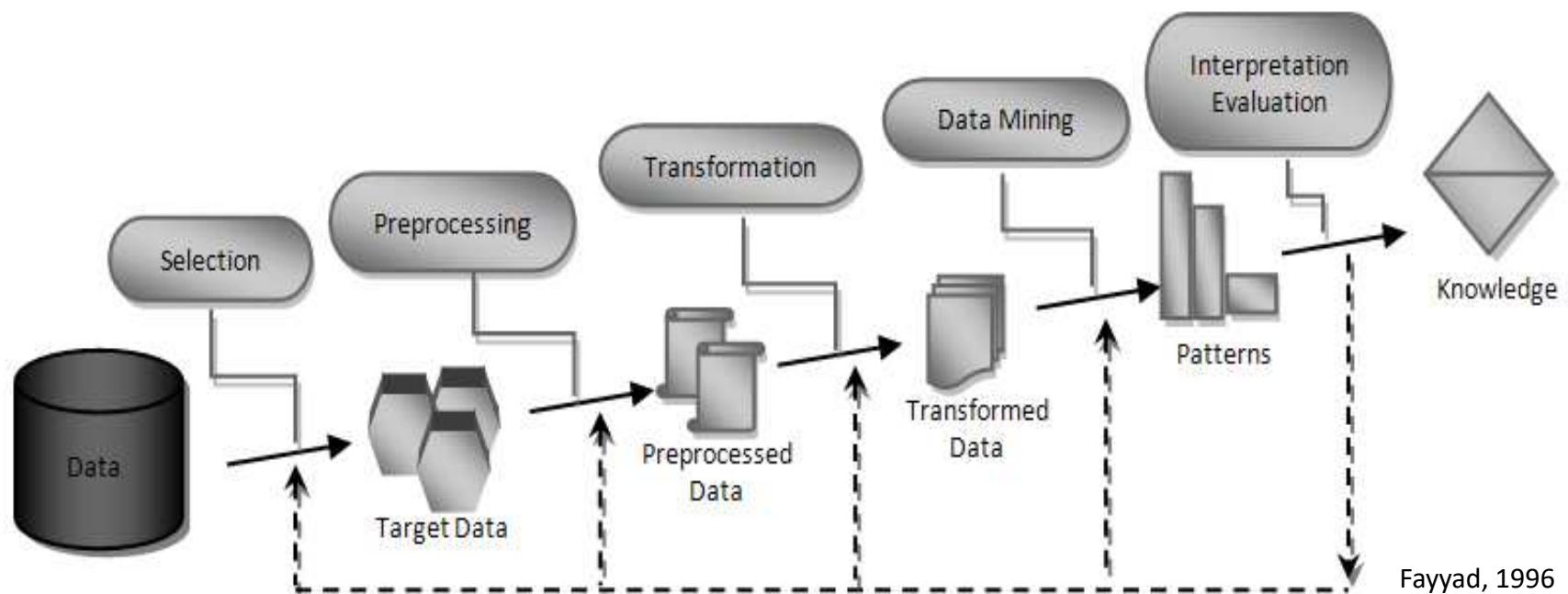
- Extração de Conhecimento (Knowledge Discovery) e Mineração de Dados (Data Mining)
 - Diferentes abordagens para extrair conhecimento dos dados
 - **KDD** (Knowledge Discovery in Databases)
 - **SEMMA** (Sample, Explore, Modify, Model and Access)
 - **CRISP-DM** (Cross-Industry Standard for Data Mining)

Extração de Conhecimento - Metodologias

Data Mining Process Models	KDD	CRISP-DM	SEMMA
No. of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application Creating a Target Data Set Data Cleaning and Pre-processing Data Transformation Choosing the suitable Data Mining Task Choosing the suitable Data Mining Algorithm Employing Data Mining Algorithm Interpreting Mined Patterns Using Discovered Knowledge	Business Understanding Data Understanding Data Preparation Modeling Evaluation Deployment	----- Sample Explore Modify Model Assessment -----

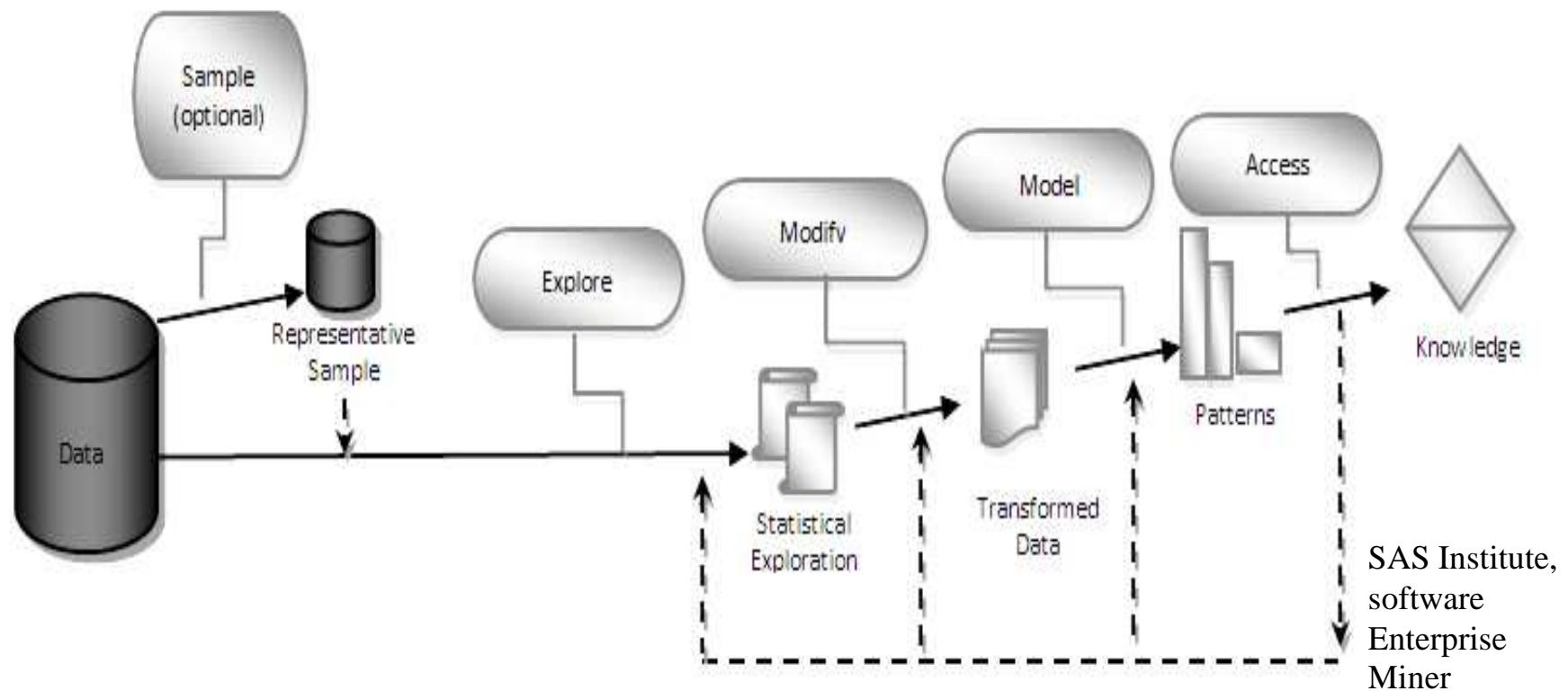
Fases do Processo KDD

- KDD (Knowledge Discovery in Databases)



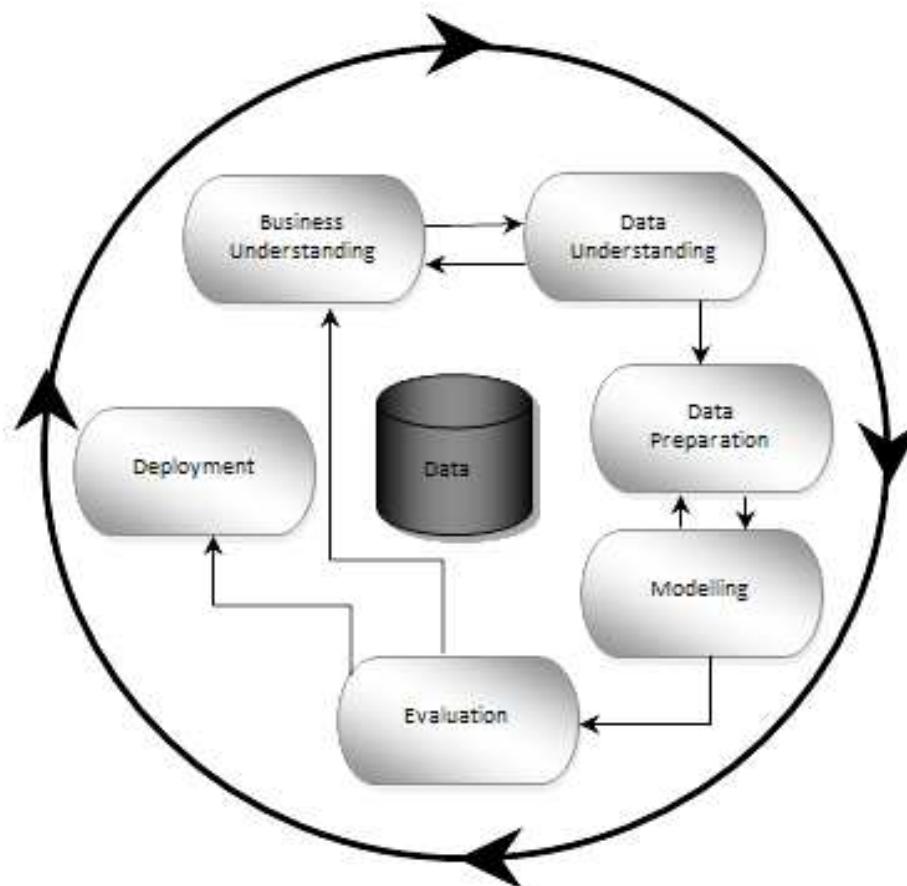
Fases do SEMMA

- SEMMA (Sample, Explore, Modify, Model and Access)



Fases do CRISP-DM

- CRISP-DM (Cross-Industry Standard for Data Mining)



União Europeia e 4 empresas:
Daimler-Benz (agora
DaimlerChrysler), Integral Solutions
Ltd. (ISL) (agora parte da IBM-SPSS),
NCR e OHRA, 2000

Dados Qualitativos vs Quantitativos

- Os indivíduos (sujeitos ou objetos) são caracterizados por variáveis (ou atributos)
- Variáveis Qualitativas
 - Nominais
 - Ordinais
- Variáveis Quantitativas
 - Racionais
 - Intervalares

Variáveis / atributos

indivíduos

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)						
Row No.	Play	Outlook	Temperature	Humidity	Wind	
1	no	sunny	85	85	false	
2	no	sunny	80	90	true	
3	yes	overcast	83	78	false	
4	yes	rain	70	96	false	
5	yes	rain	68	80	false	
6	no	rain	65	70	true	
7	yes	overcast	64	65	true	
8	no	sunny	72	95	false	
9	yes	sunny	69	70	false	
10	yes	rain	75	80	false	
11	yes	sunny	75	70	true	
12	yes	overcast	72	90	true	
13	yes	overcast	81	75	false	
14	no	rain	71	80	true	

Dados Nominais, Ordinais, Intervalares e Racionais

Tipo de variável	Descrição	Exemplos	Operações
Nominal	Variáveis nominais dão só informação para distinguir um indivíduo de outro. ($=$, \neq)	cor dos olhos, sexo	moda, entropia, contingência, correlação, testes χ^2
Ordinal	Os valores de variáveis ordinais dão informação para ordenar indivíduos. ($<$, $>$)	nível de escolaridade; classificação (<i>aprovado/reprovado</i>)	mediana, percentis, correlação de Spearman, testes do sinal
Intervalar	As diferenças entre valores são significativas, i.e., existe uma unidade de medida. (+, -)	datas, temperatura em Celsius ou Fahrenheit	média, desvio padrão, Correlação de Pearson, testes t e F
Racional	As diferenças e rácios entre valores são significativas. (*, /)	idade, peso, contagens, altura	média geométrica, media harmonica, variação

Dados – Variáveis Contínuas/Discretas

- Variáveis Quantitativas
 - Contínuas
 - Podem tomar qualquer valor real
 - Na prática os valores são representados com um número finito de dígitos
 - Exemplos: Altura, peso
 - Discretas
 - Tem só um número finito ou infinito numerável
 - Na prática são apresentados valores inteiros
 - Variáveis binárias são casos particulares de variáveis discretas
 - Exemplos: Número do cartão de cidadão, número de filhos

Dados – Tipos de Data Sets

- Tipos de data sets
 - Registo
 - Dados em matriz
 - Dados em documentos
 - Dados de transação
 - Grafos
 - World Wide Web
 - Estruturas moleculares
 - Ordenados
 - Dados espaciais
 - Dados temporais ou sequenciais
 - Dados de sequências genéticas

Dados - Registros

- Registro
 - Dados que consistem na coleção de dados com uma lista fixa de variáveis

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind	Compra	Itens
1	no	sunny	85	85	false		
2	no	sunny	80	90	true		
3	yes	overcast	83	78	false		
4	yes	rain	70	96	false		
5	yes	rain	68	80	false		
6	no	rain	65	70	true		
7	yes	overcast	64	65	true	1	Pão, Sumo, Leite
8	no	sunny	72	95		2	Cerveja, Pão
9	yes	sunny	69	70		3	Cerveja, Sumo, Fraldas, Leite
10	yes	rain	75	80		4	Cerveja, Pão, Fraldas, Leite
11	yes	sunny	75	70			
12	yes	overcast	72	90			
13	yes	overcast	81	75			
14	no	rain	71	80			

Dados – Em Matriz

- Dados em matriz
 - Cada data set pode ser representada por uma matriz $m \times n$, onde m pode representar o número de indivíduos e n o número de variáveis

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

Dados – Em Documentos

- Dados em documentos
 - Cada documento será um vetor de termos
 - Cada termo é uma componente (variável) do vetor
 - O valor de cada componente é o número de vezes que o termo correspondente aparece no documento

	“Cancro”	“Saúde”	“Cuidados”	“Terapia”	“Evidência”
Documento 1	1	0	1	2	2
Documento 2	3	1	0	1	1
Documento 3	0	2	2	1	1

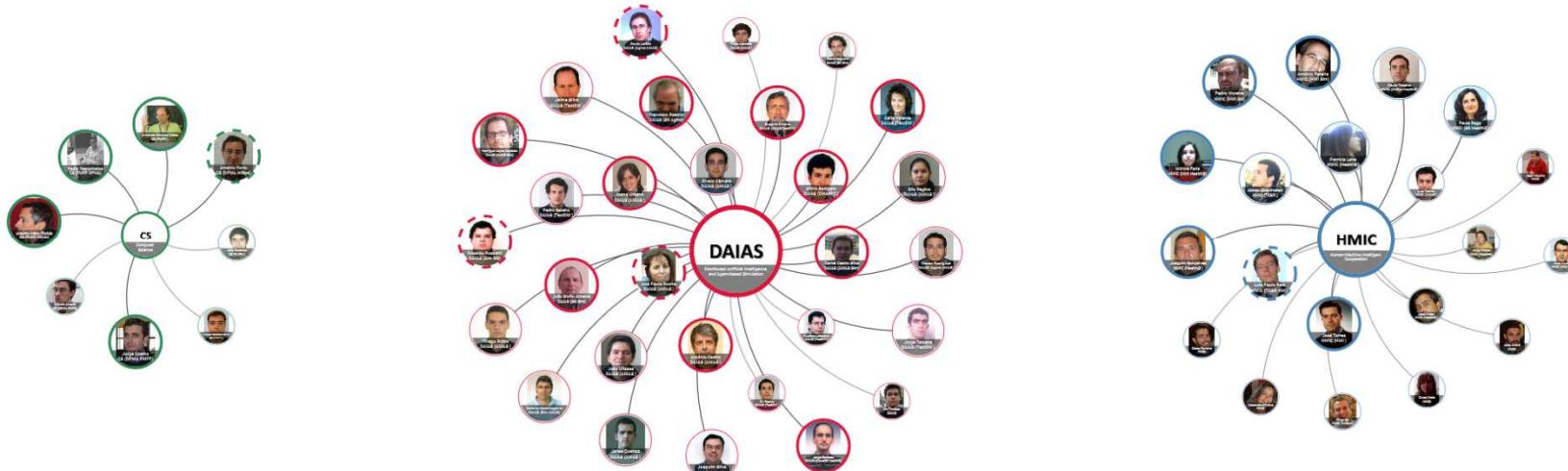
Dados – Em Transação

- Dados de transação
 - Cada transação envolve um conjunto de itens
 - Exemplo: conjunto de produtos (itens) adquiridos em diferentes compras (transação)

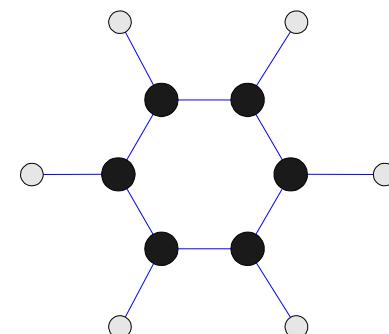
Compra	Itens
1	Pão, Sumo, Leite
2	Cerveja, Pão
3	Cerveja, Sumo, Fraldas, Leite
4	Cerveja, Pão, Fraldas, Leite

Dados - Grafos

- Grafos
 - Membros do LIACC (Lab. Inteligência Artificial e Ciência de Computadores) e contribuições para o grupo



- Dados em Química:
 - Molécula de Benzina (C_6H_6)



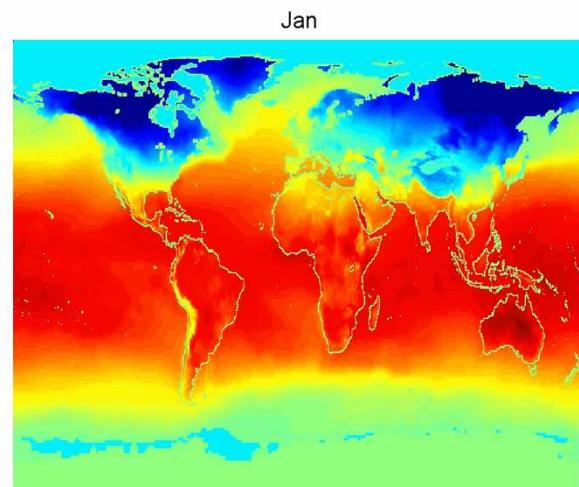
Dados – Ordenados, Espaciais/Temporais

- Dados Ordenados

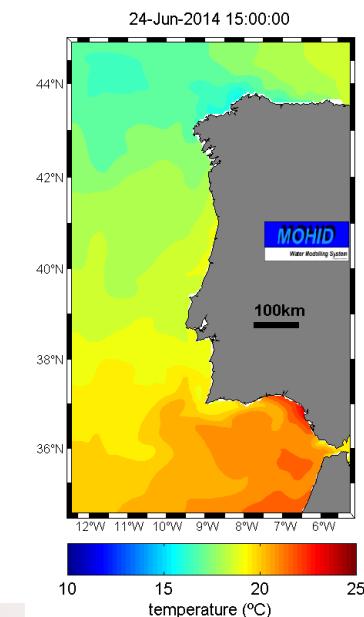
Dados com variáveis que têm relações entre elas e envolvem ordem no tempo e espaço

- Dados espaciais – objetos com atributos espaciais tais como posição ou área. Ex: dados meteorológicos (precipitação, temperatura, pressão) que são recolhidos em diferentes localizações

Temperatura média (mês) da terra e oceanos



Temperatura média do mar na costa Portuguesa



Dados Ordenados

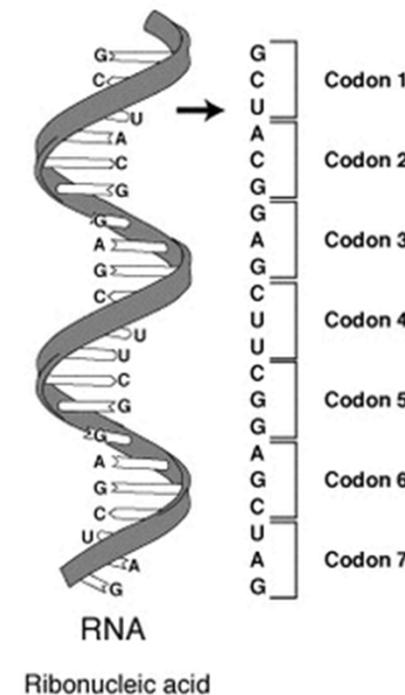
- Dados Ordenados
 - Dados temporais ou sequenciais – são dados de registo com o tempo associado.
 - Ex: lista de compras com o historial do cliente

Tempo	Cliente	Itens
t1	C1	Pão, Sumo, Leite
t2	C3	Cerveja, Pão
t3	C1	Cerveja, Sumo, Fraldas, Leite
t4	C2	Cerveja, Pão, Fraldas, Leite

Dados Ordenados

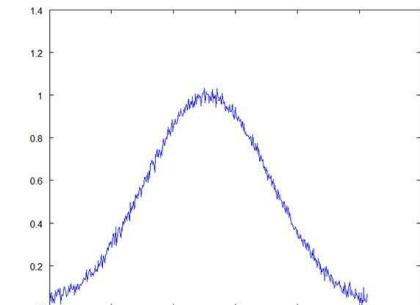
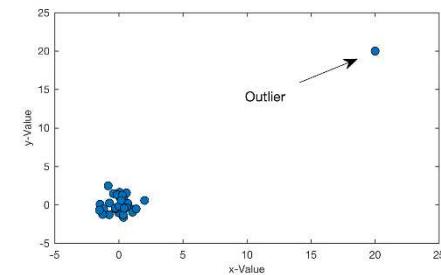
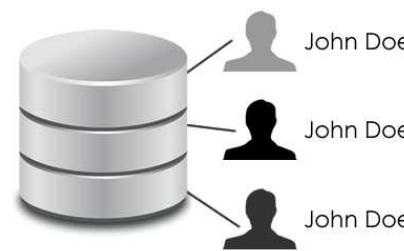
- Dados Ordenados
 - Dados de sequências genéticas

GGTTCCGCCTTCAGCCCCGCGGCC
CGCAGGGCCCCGCCCGCGCCGTC
GAGAAGGGCCCAGCCTGGCGGGCG
GGGGGAGGCAGGGGCCGCCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCAGCAGCGGACAG
GCCAAGTAGAACACCGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG



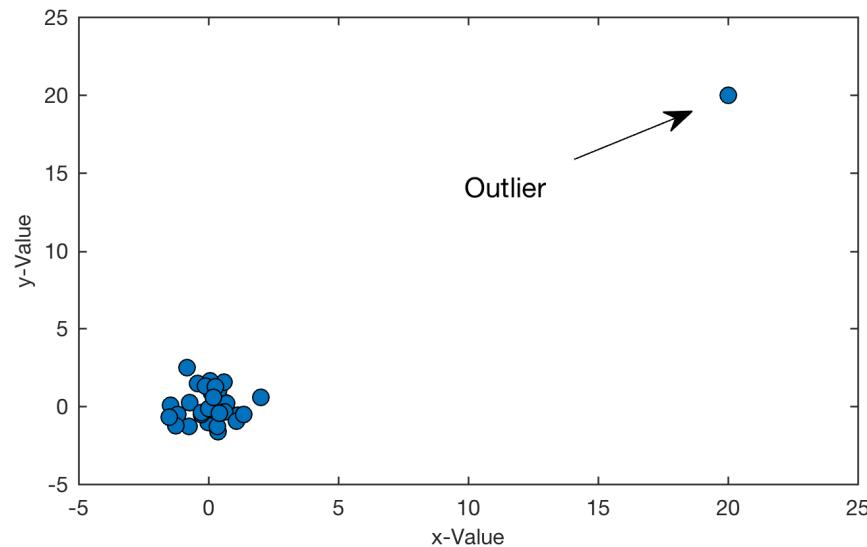
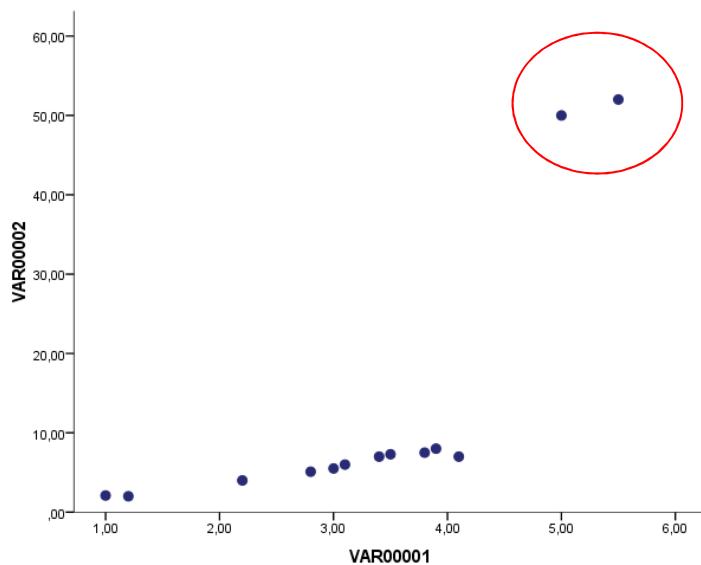
Qualidade dos Dados

- Qualidade dos Dados
 - Problemas que podem surgir
 - Outliers (valores extremos)
 - Ruído nos dados
 - Missing values (valores omissos)
 - Dados duplicados



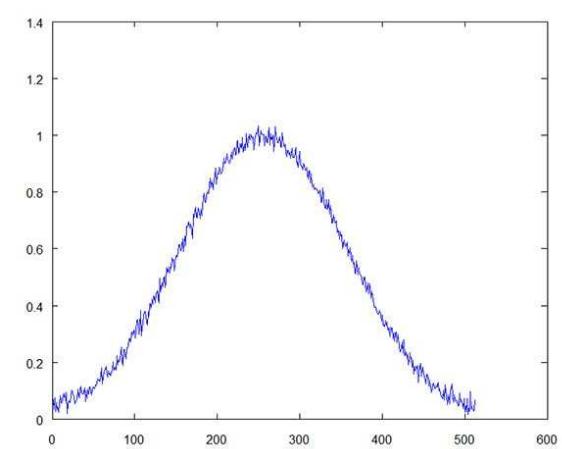
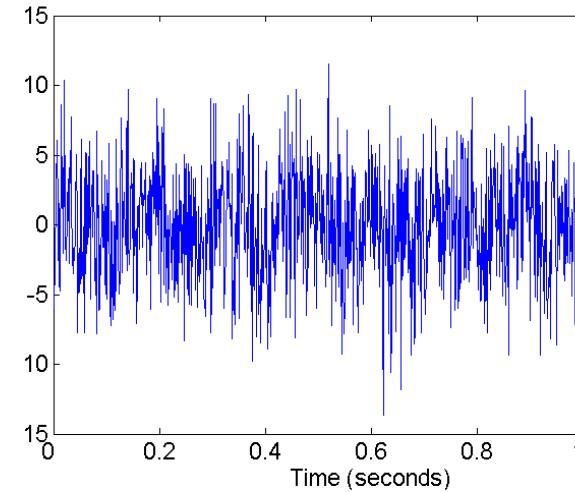
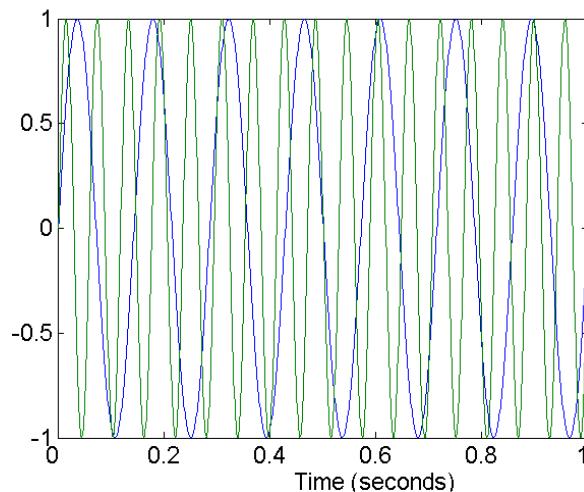
Dados - Outliers

- Qualidade dos Dados
 - Outliers
 - dados extremos
 - diferem da distribuição habitual dos restantes dados



Dados - Ruído

- Qualidade dos Dados
 - Ruído
 - modificações aos dados originais (erro aleatório)
 - Valores incorretos devido
 - Problemas com instrumentos de entrada de dados
 - Problemas de transmissão de dados
 - Limitações tecnológicas



Dados – Valores Omissos/Duplicados

- Qualidade dos Dados
 - Valores omissos
 - Informação que não foi recolhida (questões mais sensíveis que ficaram por preencher)
 - Variáveis que não são aplicáveis a todos os casos
 - Como gerir valores omissos:
 - Eliminar dados
 - Estimar valores omissos
 - Ignorar indivíduos/variáveis durante a análise
 - Substituir por todos os valores possíveis (pesar pelas suas probabilidades)
- Dados duplicados
 - Por exemplo quando se agrupam dados de diversas fontes
 - Exemplo: Mesmo indivíduo com diferentes emails

Dados - Exploração

- Exploração dos Dados

- Utilizada para melhor entender as suas características
- Permite selecionar de forma adequada o método para pré-processamento
- Permite ao avaliador reconhecer alguns padrões nos dados
 - Humanos conseguem reconhecer padrões que não são capturados automaticamente por ferramentas de análise de dados

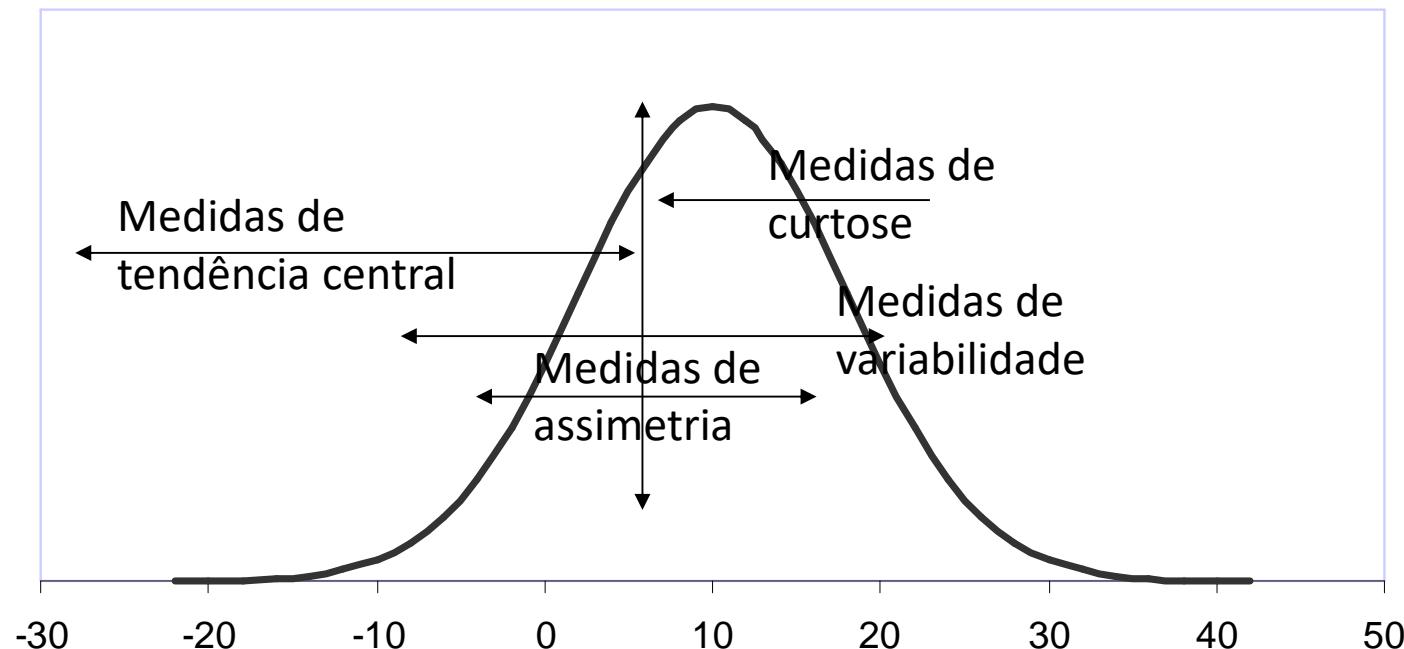
Dados - Exploração

- Exploração dos Dados
 - Medidas estatísticas sumárias
 - Medidas de Tendência Central: média; moda; mediana; percentis
 - Medidas de Dispersão: amplitude total; intervalo interquartis; desvio interquartis; variância; desvio padrão; coeficiente de variação
 - Medidas de Assimetria e de Curtose
 - Visualização
 - Online Analytical Processing (OLAP)

Dados - Exploração

- Exploração dos Dados
 - Medidas estatísticas sumárias

Curva de Gauss



Dados - Exploração

- Exploração dos Dados
 - Visualização
 - Conversão de dados num formato visual ou em tabela
 - Permite analisar e reportar características dos dados e suas relações
 - Modo fácil de exploração de dados
 - Os humanos têm habilidade de analisar grande quantidade de informação apresentada de forma visual
 - Pode ser possível detetar padrões e tendências
 - Podem detetar outliers e padrões incomuns

Dados - Exploração

- Exploração dos Dados
 - Visualização
 - Tabelas de Frequências - lista de todos os valores observados e quantas vezes esses valores aparecem (frequência absoluta F_i) e/ou a frequência relativa desses valores (f_i)

Grau de escolaridade	Alunos F_i	f_i (%)
4º ano	12	8,2
9º ano	13	8,8
12º ano	56	38,1
Licenciatura	43	29,3
Mestrado	23	15,6
Total	147	100

Dados – Tabelas de Contingência

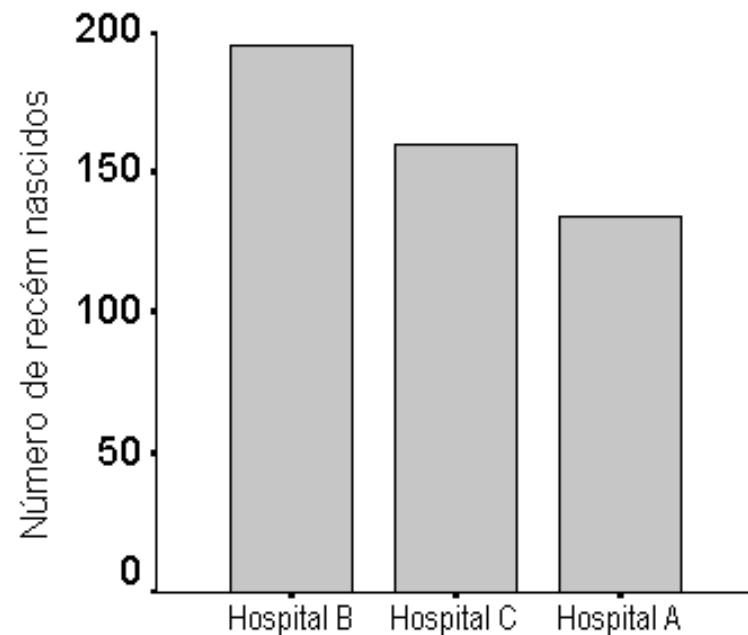
Tabelas de Contigência (Dupla entrada) - Listas nos cabeçalhos das colunas e linhas todos os valores observados e nas células o número de ocorrências de cada par

Tomou antibiótico 2 semanas antes	<i>Enterobacter multiresistente</i>		
	Sim	Não	Total
Sim	36	67	103
Não	1	25	26
Total	37	92	129

Dados – Gráficos

- **Gráfico de barras:**
 - eixo horizontal: categorias
 - eixo vertical: contagem e/ou percentagem
- **Gráfico circular:**

Se o número de categorias for baixo)



- **Gráfico circular:**

Se o número de categorias for baixo)



Dados - Diagramas

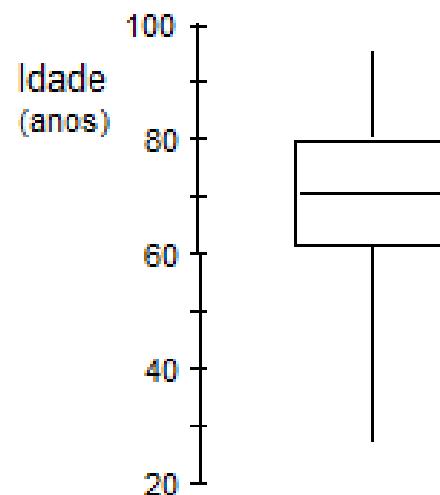
- **Diagrama de tronco e folha:** trata-se de uma maneira muito atrativa para apresentar diretamente uma contagem para dados numa escala racional ou intervalar

– Exemplo (idade de 17 pacientes):
22, 33, 35, 41, 43, 43, 44, 45, 47,
51, 56, 64, 72, 77, 81, 82, 96.

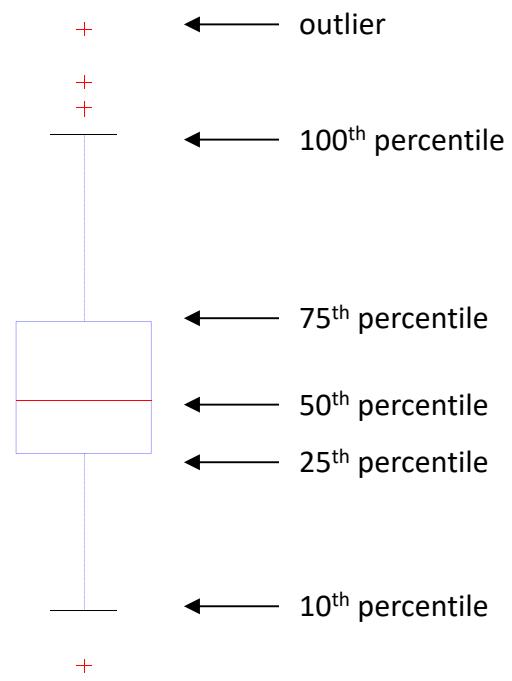
2	2
3	35
4	133457
5	16
6	4
7	27
8	12
9	6

Dados – Gráfico de Caixa e Fio

- **Gráfico de caixa e fio:** é uma representação gráfica cujo objectivo é mostrar algumas medidas de localização na distribuição: Mediana, 1º Quartil, 3º Quartil; apresenta ainda as observações extremas que podem ser consideradas outliers.



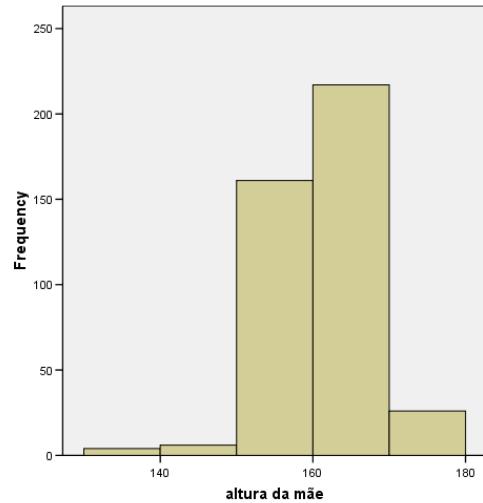
- Gráfico de caixa e fio para a idade dos doentes com cancro colorectal.



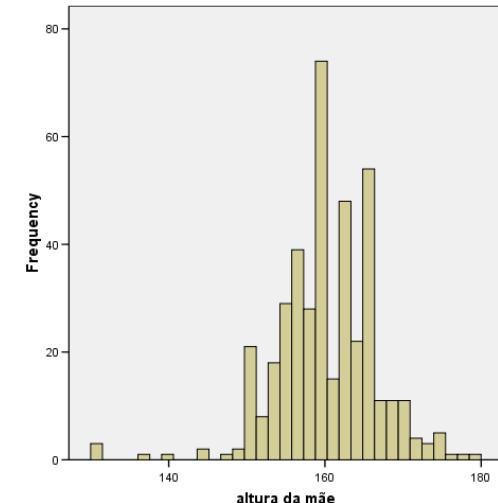
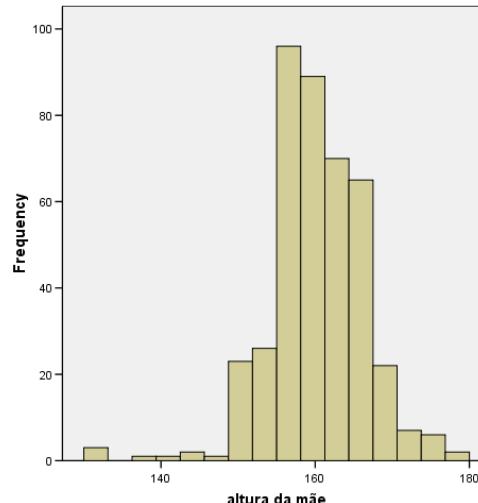
Dados - Histogramas

- **Histograma:** é um gráfico que reflecte a forma da distribuição de frequências da amostra. Também procura reflectir a estrutura (forma) da população de onde foi retirada a amostra.
 - Os histogramas são particularmente úteis para variáveis contínuas ou variáveis com poucos valores repetidos
 - Para construir um histograma é necessário primeiro repartir os dados por classes e depois calcular as respetivas frequências
 - Depende do número de classes considerado
 - Um número muito grande de classes produz um histograma com demasiada irregularidade, enquanto um histograma com um número demasiado reduzido de classes oculta a forma da distribuição (perde-se demasiada informação).

Dados - Histogramas



Poucas classes



Muitas classes

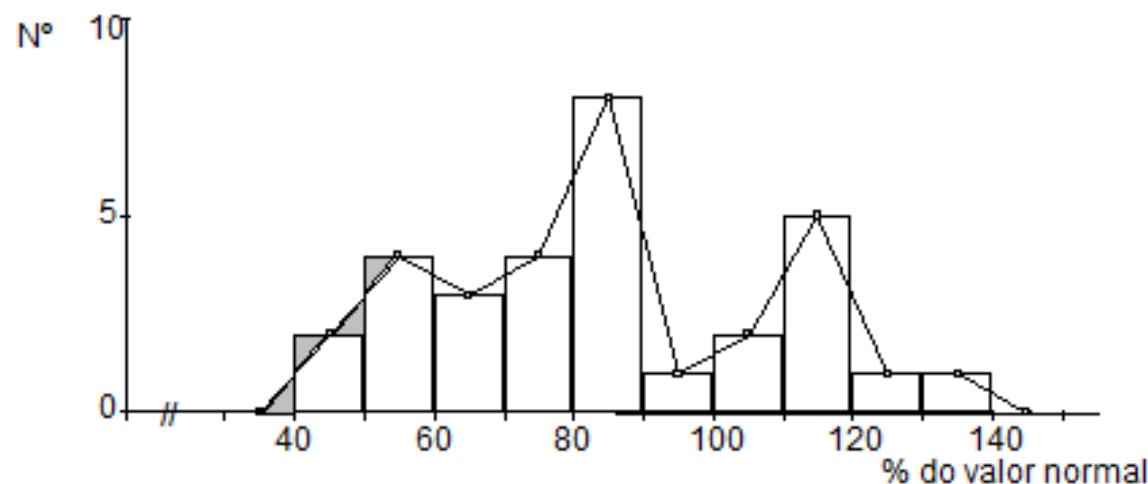
A expressão

$$\text{nº de classes} = 1 + 3.3 \times \log_{10} n$$

dá-nos uma aproximação para o nº de classes a usar, baseado no nº total de observações (n):

Dados

- **Polígono de frequências:** São gráficos de linha que se obtêm através do histograma, unindo os pontos médios de cada classe.



- Polígono de frequências para o valor da análise laboratorial (% do valor normal) nos indivíduos do sexo masculino.

Dados

- Duas variáveis medidas numa escala nominal ou ordinal
 - Gráfico de barras

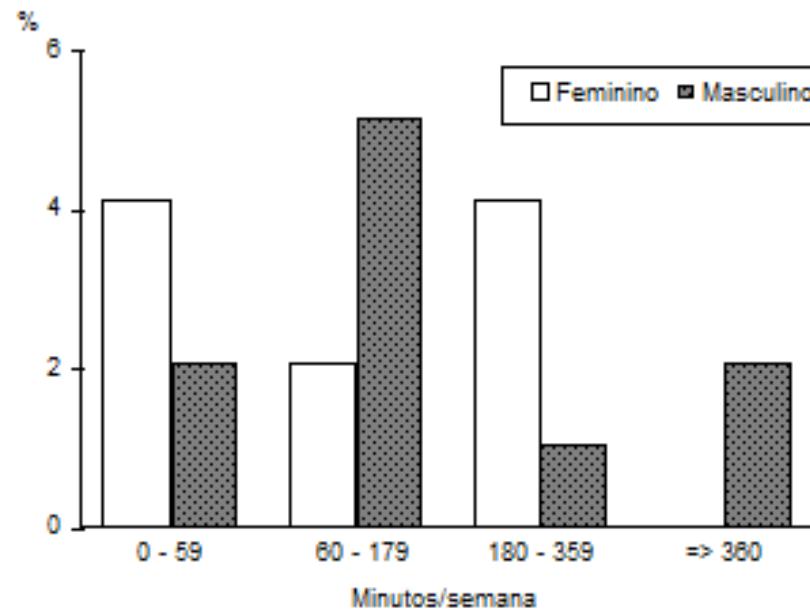
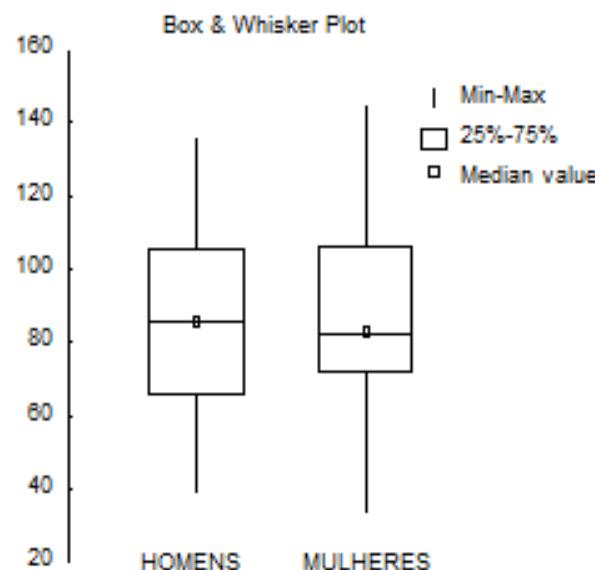


Diagrama de barras para a actividade física durante o último mês (minutos/semana) segundo o sexo.

Dados

- Uma variável na escala nominal ou ordinal e a outra na escala racional/intervalar

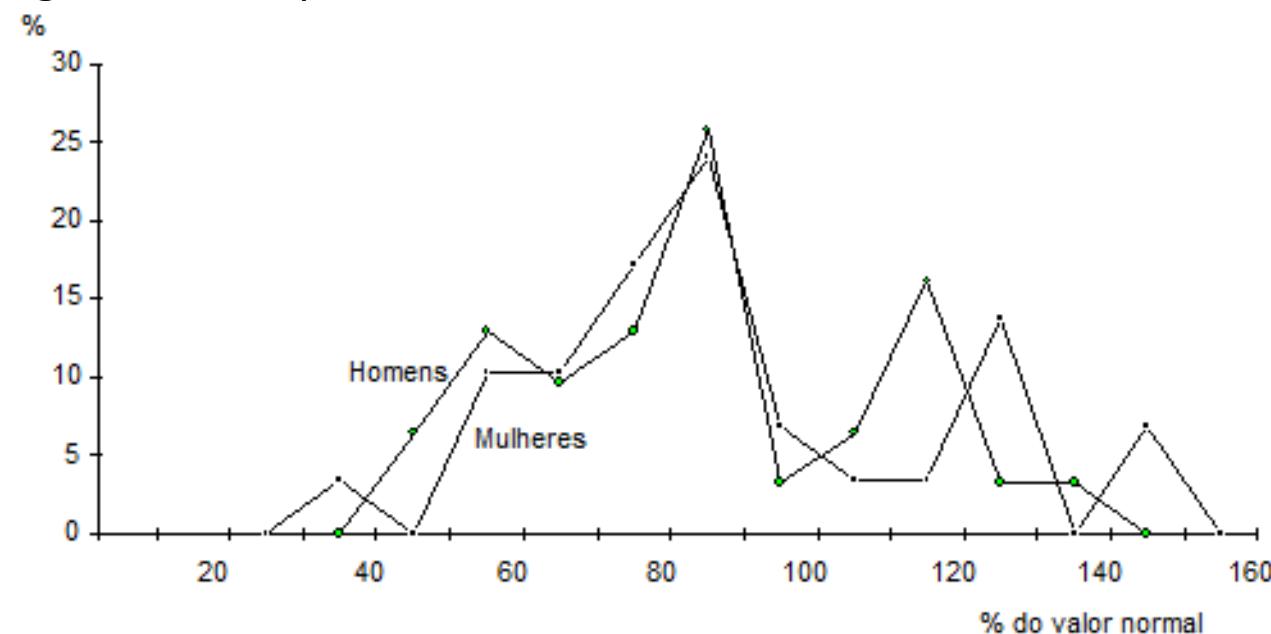
- Gráfico de caixa e fio



- Gráfico de caixa e fio para o valor da análise laboratorial (% do valor normal) segundo o sexo do doente.

Dados

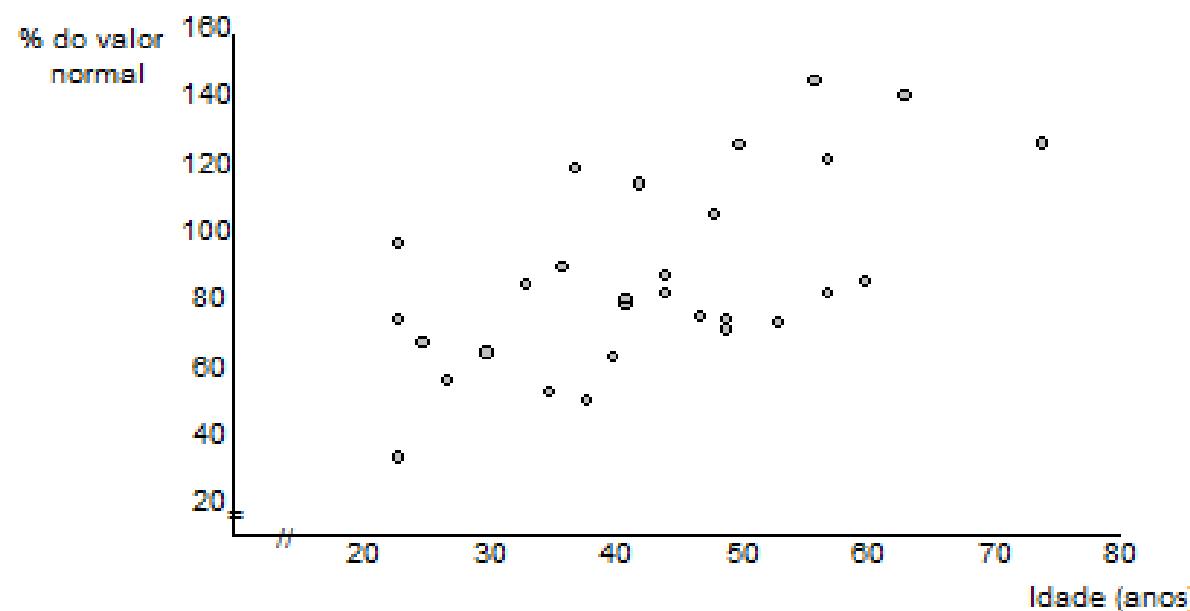
- Uma variável na escala nominal ou ordinal e a outra na escala racional/intervalar
 - Polígono de frequências



- Polígonos de frequências do valor de uma análise laboratorial (% do valor normal) para homens e mulheres.

Dados

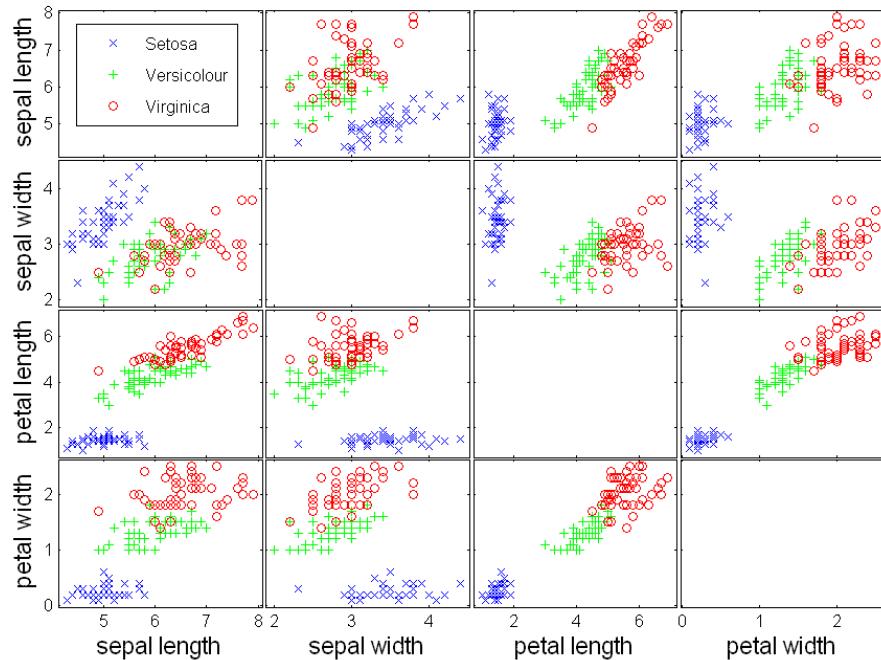
- Duas variáveis na escala racional/intervalar
 - Diagrama de Dispersão



- Diagrama de dispersão para o valor da análise laboratorial e idade nos indivíduos do sexo feminino.

Dados

- Múltiplas variáveis



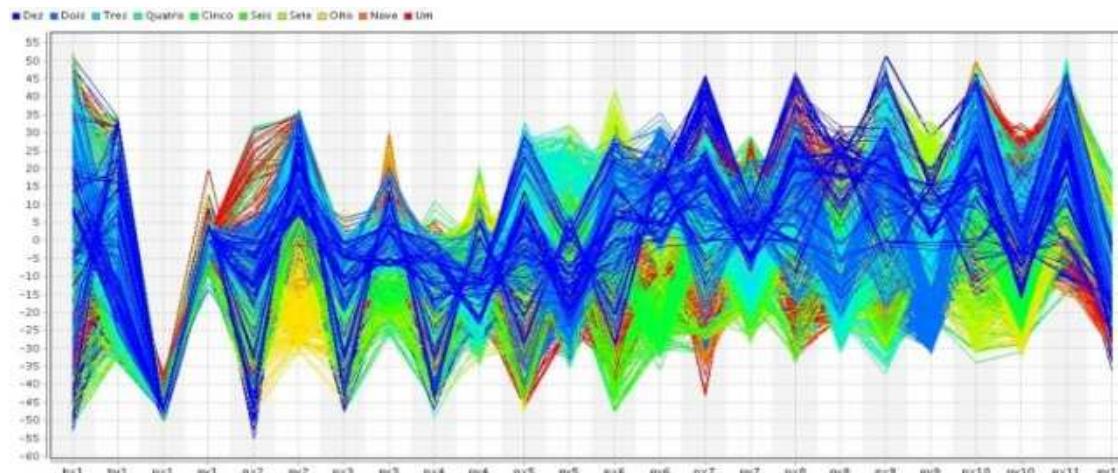
Scatter Plot (Diagramas de
Dispersão Múltiplos)



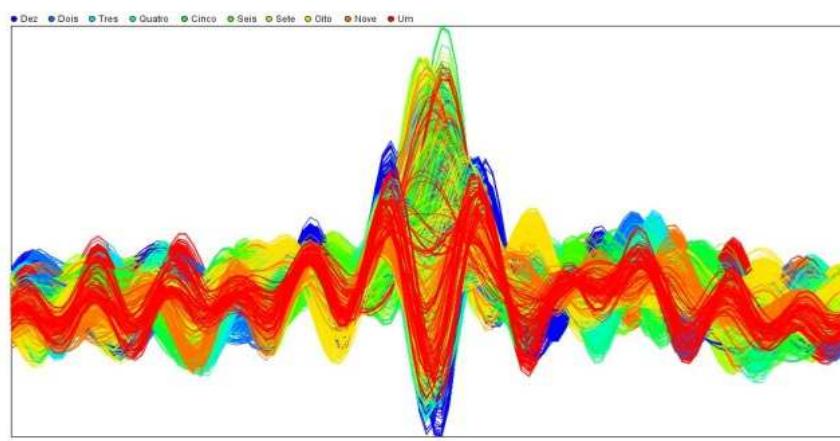
Survey Plot

Dados

- Múltiplas variáveis



Parallel coordinates plot



Andrews' Curves

Para cada observação $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ é transformada numa curva usando

$$f_i(t) = \begin{cases} \frac{x_{i1}}{\sqrt{2}} + x_{i2} \sin(t) + x_{i3} \cos(t) + \dots + x_{ip-1} \sin\left(\frac{p-1}{2}t\right) + x_{ip} \cos\left(\frac{p-1}{2}t\right) & \text{for } p \text{ odd} \\ \frac{x_{i1}}{\sqrt{2}} + x_{i2} \sin(t) + x_{i3} \cos(t) + \dots + x_{ip} \sin\left(\frac{p}{2}t\right) & \text{for } p \text{ even} \end{cases}$$

onde cada observação representa os coeficientes da série de Fourier e ($t \in [0, \pi]$)

Dados

- Exploração dos Dados
 - Online Analytical Processing (OLAP)
 - Proposto por E. F. Codd (pai das bases de dados relacionais)
 - Utiliza uma representação em vetores multidimensionais
 - Há um grande número de operações de análise de dados e exploração de dados que utiliza esta representação

Dados

- Exploração dos Dados - Online Analytical Processing (OLAP)
 - 2 Passos para converter dados de tabelas em vetores multidimensionais
 - **Primeiro**, identificar quais atributos devem ser as dimensões e qual deve ser o atributo alvo cujos valores aparecem como entradas no vetor multidimensional
 - Os atributos utilizados como dimensões devem ter valores discretos
 - O valor alvo é geralmente uma contagem ou valor contínuo, por exemplo, o custo de um item
 - Não pode ter nenhuma variável de destino, exceto a contagem de objetos que possuem o mesmo conjunto de valores de atributos
 - **Segundo**, encontrar o valor de cada entrada no vetor multidimensional, somando os valores (do atributo de destino) ou a contagem de todos os objetos que possuem os valores de atributo correspondentes a essa entrada

Dados

- Exemplo: Iris Data

- Como os atributos (comprimento da pétala, largura da pétala, tipo de espécie) podem ser convertidos num vetor multidimensional
 - Primeiro, discretiza-se a largura e o comprimento da pétala para ter valores categóricos: baixos, médios e altos.
 - Obtemos a tabela - observe o atributo de contagem

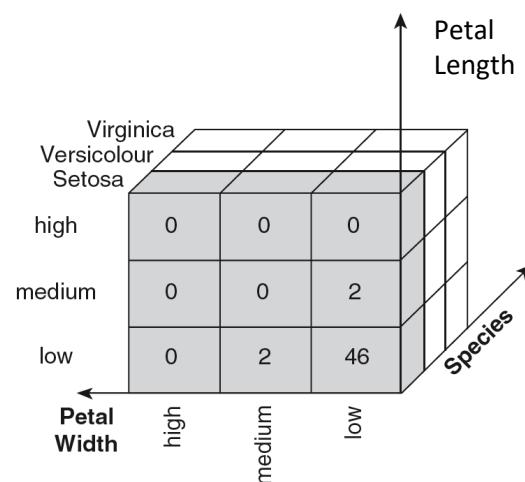
Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44



Dados

- Exemplo: Iris Data

- Cada tuplo (largura da pétala, comprimento da pétala, tipo de espécie) identifica um elemento da matriz
- Este elemento recebe o valor de contagem correspondente
- A figura ilustra o resultado
 - Todos os tuplos não especificados são 0
 - Fatias do vetor multidimensional são mostrados nas tabelas



		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

Dados

- Exemplo: Iris Data
 - A operação chave de um OLAP é a formação de um cubo de dados
 - Um cubo de dados é uma representação multidimensional dos dados, juntamente com todas as agregações possíveis
 - Todas as agregações possíveis que resultam ao selecionar um subconjunto adequado das dimensões e somar todas as dimensões restantes
 - Por exemplo, se escolhermos a dimensão do tipo de espécie dos dados Iris e somarmos todas as outras dimensões, o resultado será uma entrada unidimensional com três entradas, cada uma das quais dá o número de flores de cada tipo

Dados

- Repositório de data sets para Extração de Conhecimento e Data Mining
 - UCI Machine Learning Repository
<http://archive.ics.uci.edu/ml/index.php>



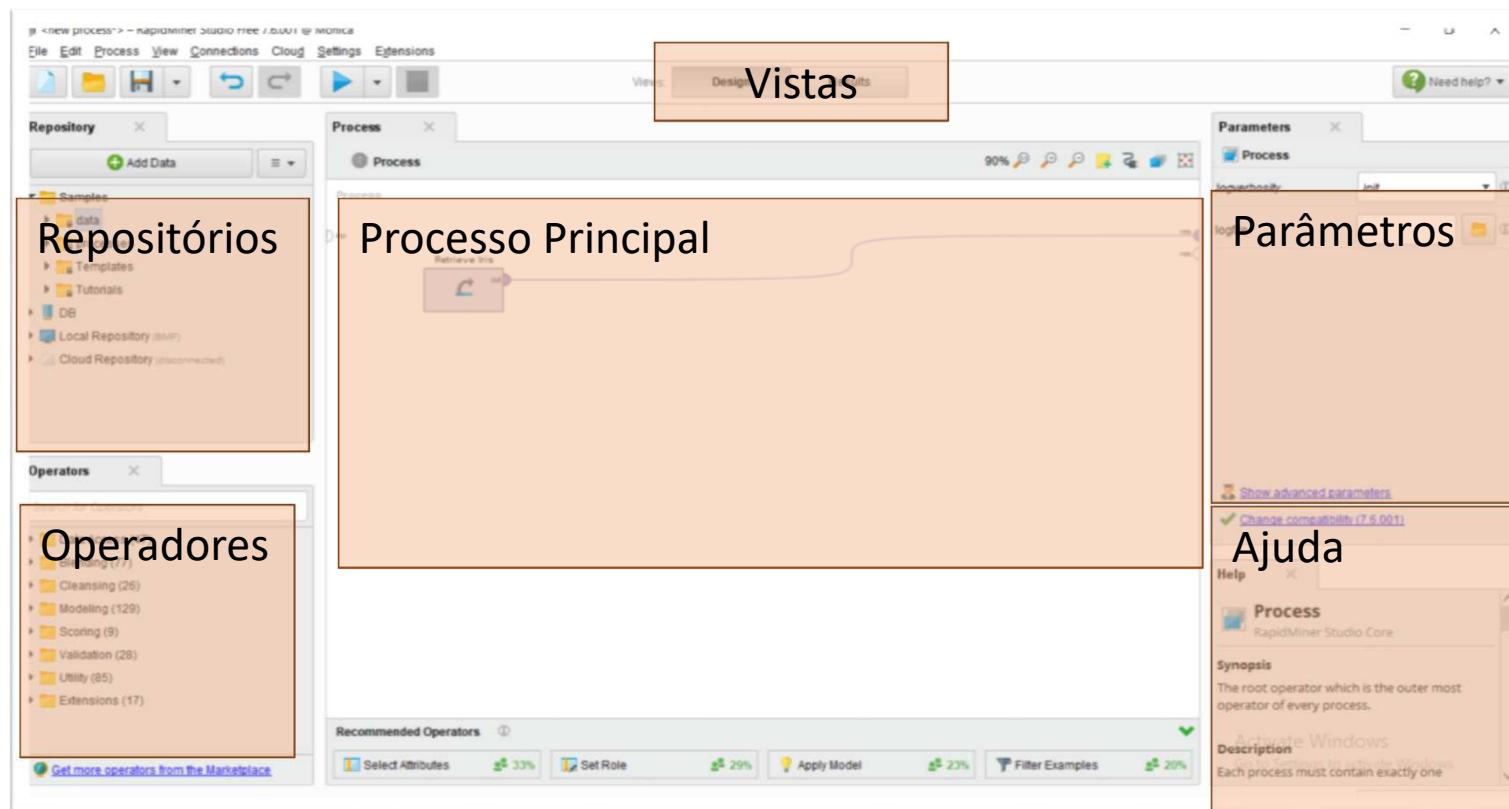
The screenshot shows the homepage of the UCI Machine Learning Repository. The header features the UCI logo and the text "Machine Learning Repository" and "Center for Machine Learning and Intelligent Systems". A search bar and navigation links for "About", "Citation Policy", "Donate a Data Set", and "Contact" are visible. Below the header, a sidebar on the left provides filters for "Default Task" (Classification, Regression, Clustering, Other), "Attribute Type" (Categorical, Numerical, Mixed), "Data Type" (Multivariate, Univariate, Sequential, Time-Series, Text, Domain Theory, Other), "Area" (Life Sciences, Physical Sciences, CS / Engineering, Social Sciences, Business, Game, Other), and "# Attributes" (Less than 10, 10 to 100). The main content area displays a table titled "Browse Through: 394 Data Sets" with columns for Name, Data Types, Default Task, Attribute Types, # Instances, # Attributes, and Year. The table lists several datasets, including Abalone, Adult, Annealing, Anonymous Microsoft Web Data, Arrhythmia, Artificial Characters, Audiology (Original), and Audioloav (Standardized).

Browse Through: 394 Data Sets						
	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes
	Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	6
	Adult	Multivariate	Classification	Categorical, Integer	48842	14
	Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38
	Anonymous Microsoft Web Data		Recommender Systems	Categorical	37711	294
	Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279
	Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7
	Audiology (Original)	Multivariate	Classification	Categorical	226	Windows
	Audioloav (Standardized)	Multivariate	Classification	Categorical	226	80

Dados

- RapidMiner

<https://rapidminer.com/>

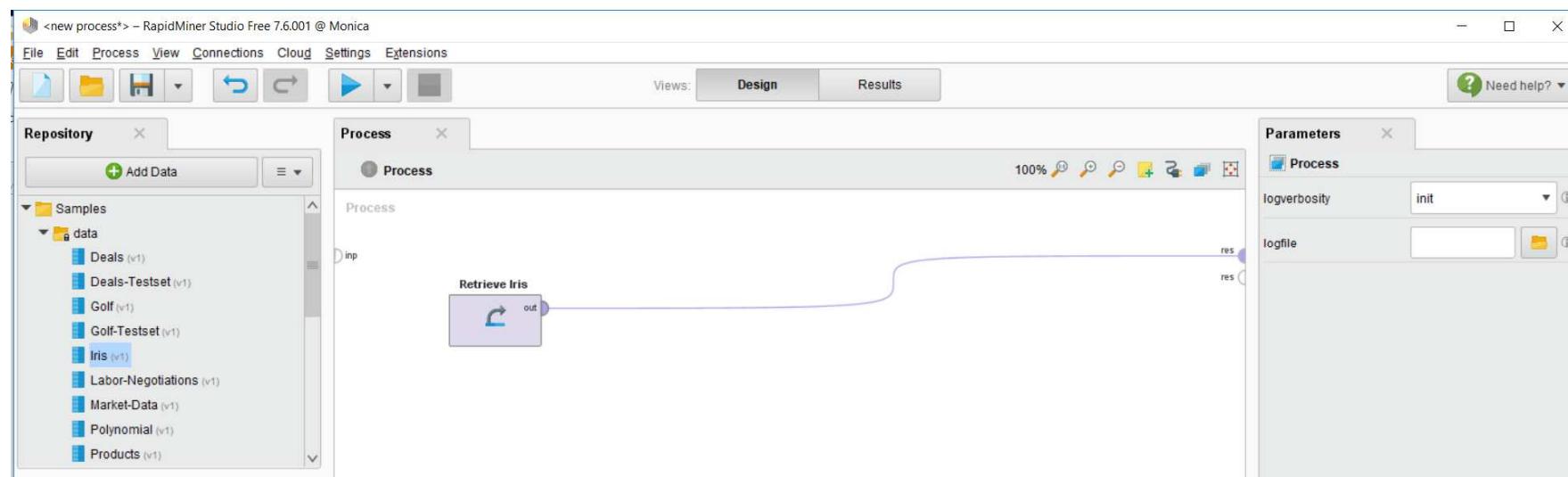


Dados

- Tarefas no RapidMiner
 - Importar dados de ficheiros de dados
 - Disponíveis no RapidMiner
 - Ficheiros pessoais
 - Visualizar estatísticas sumárias dos dados
 - Visualizar graficamente os dados
 - Agregar dois ficheiros de dados
 - Gerir valores omissos

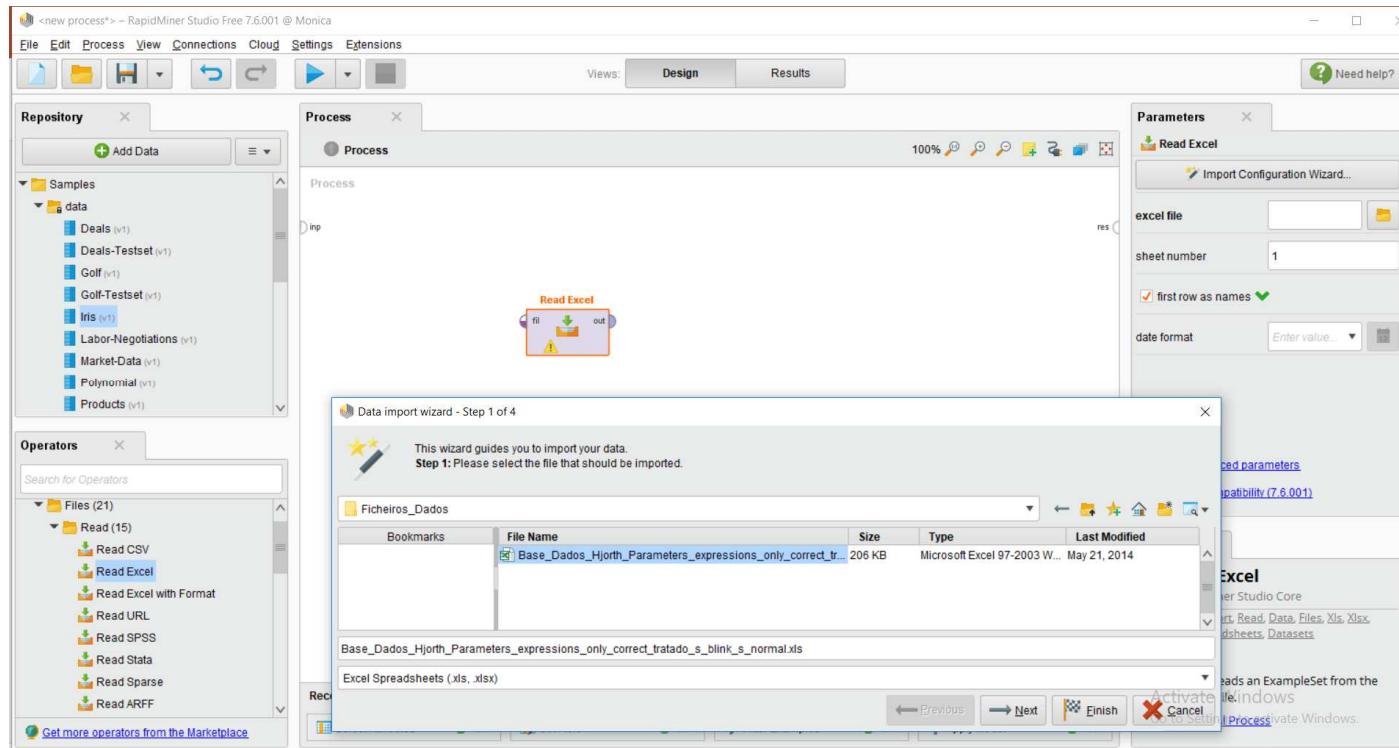
Dados

- Tarefas no RapidMiner
 - Importar dados de ficheiros de dados - Disponíveis no RapidMiner
 - Visualizar medidas estatísticas sumárias
 - Visualizar graficamente os dados



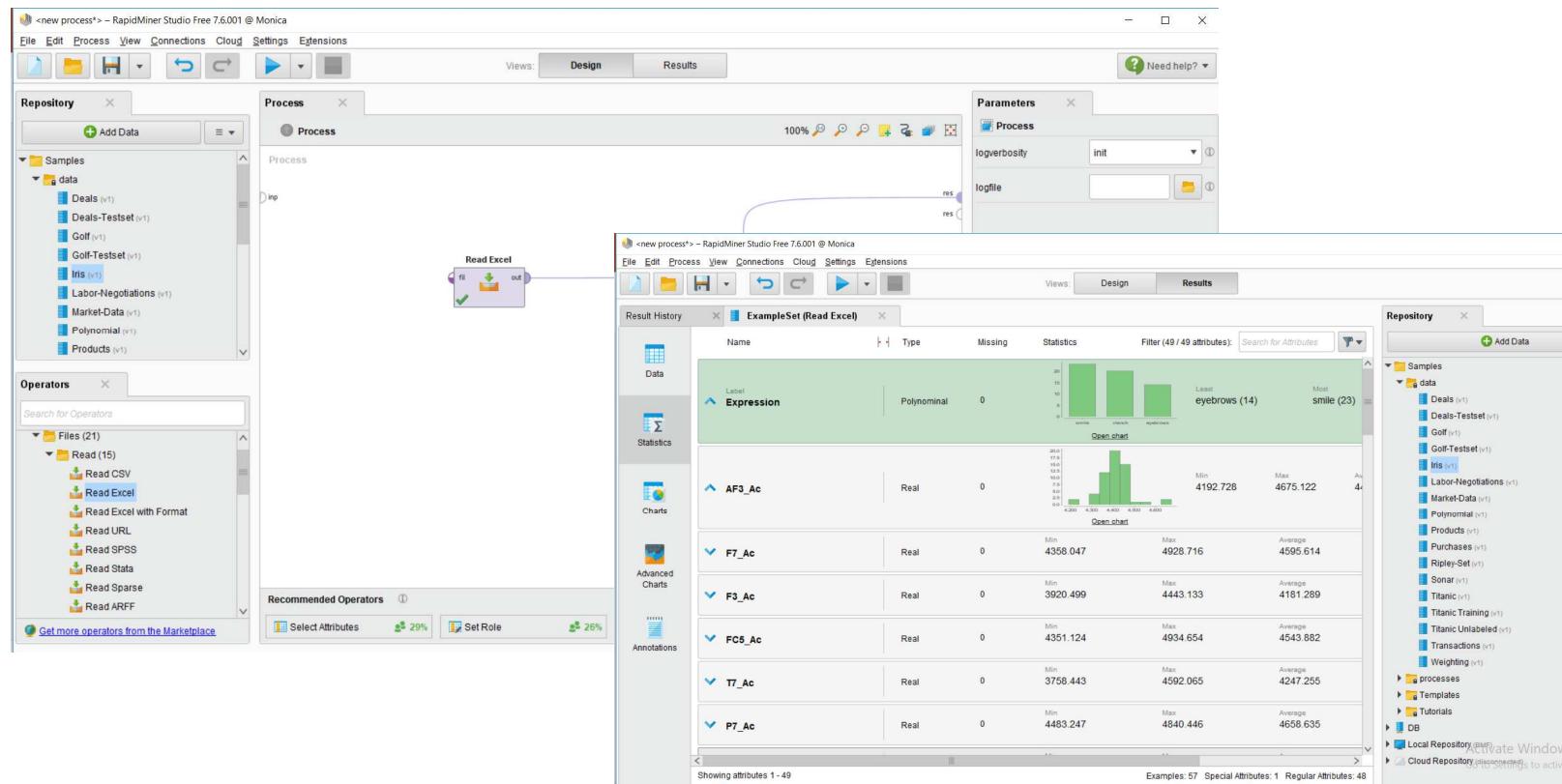
Dados

- Tarefas no RapidMiner
 - Importar dados de ficheiros de dados - Ficheiros pessoais



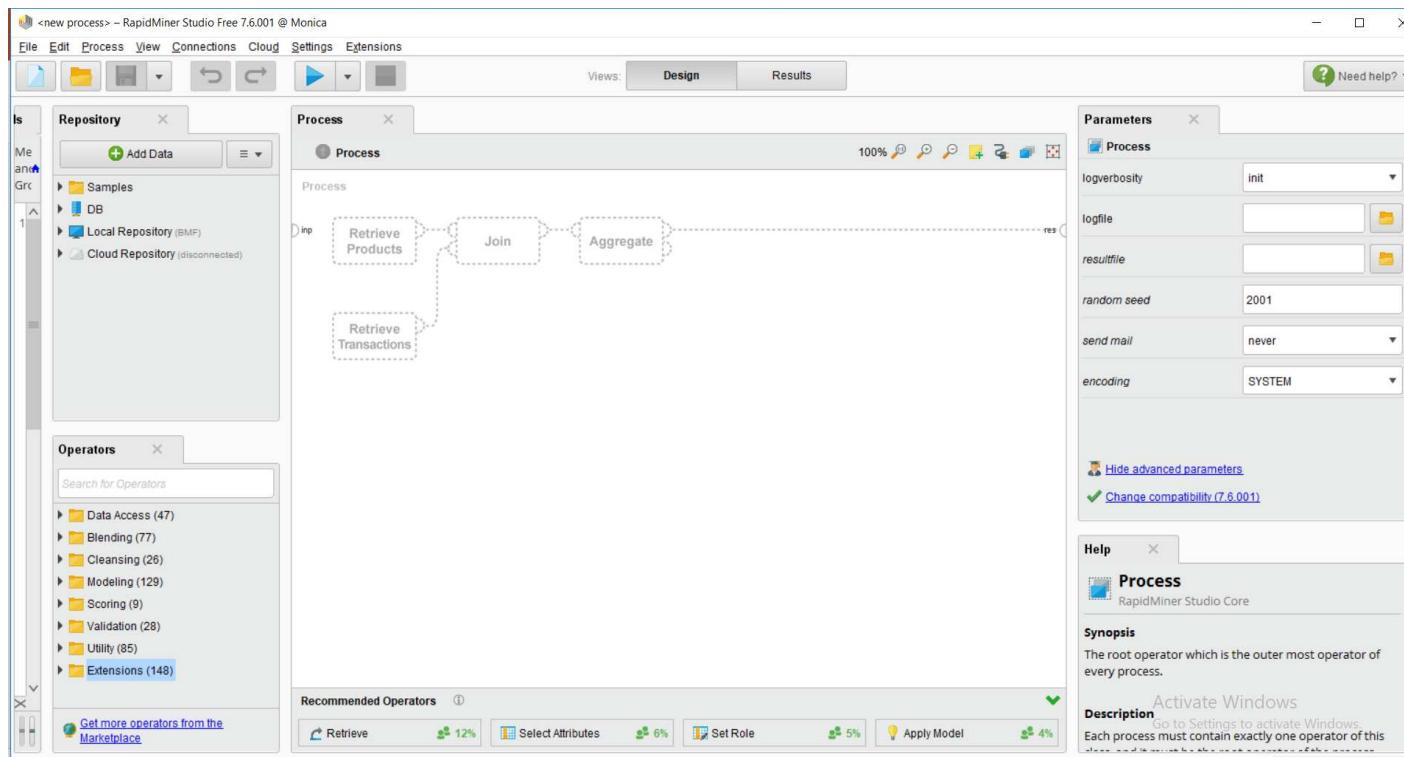
Dados

- Tarefas no RapidMiner
 - Visualizar medidas estatísticas sumárias
 - Visualizar graficamente os dados



Dados

- Tarefas no RapidMiner
 - Agregar dois ficheiros de dados

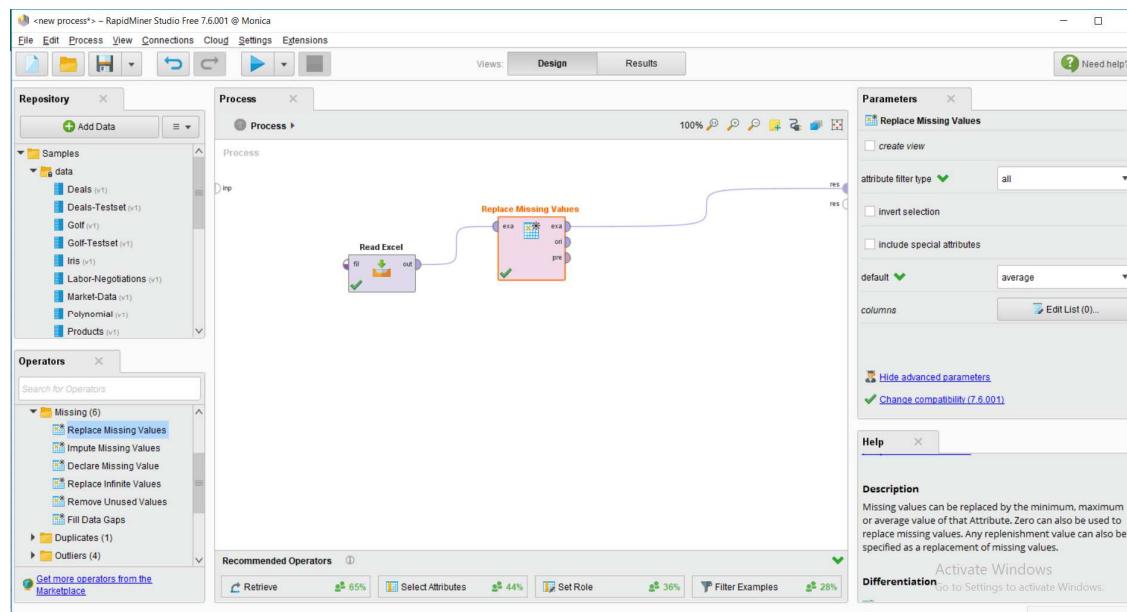


Dados

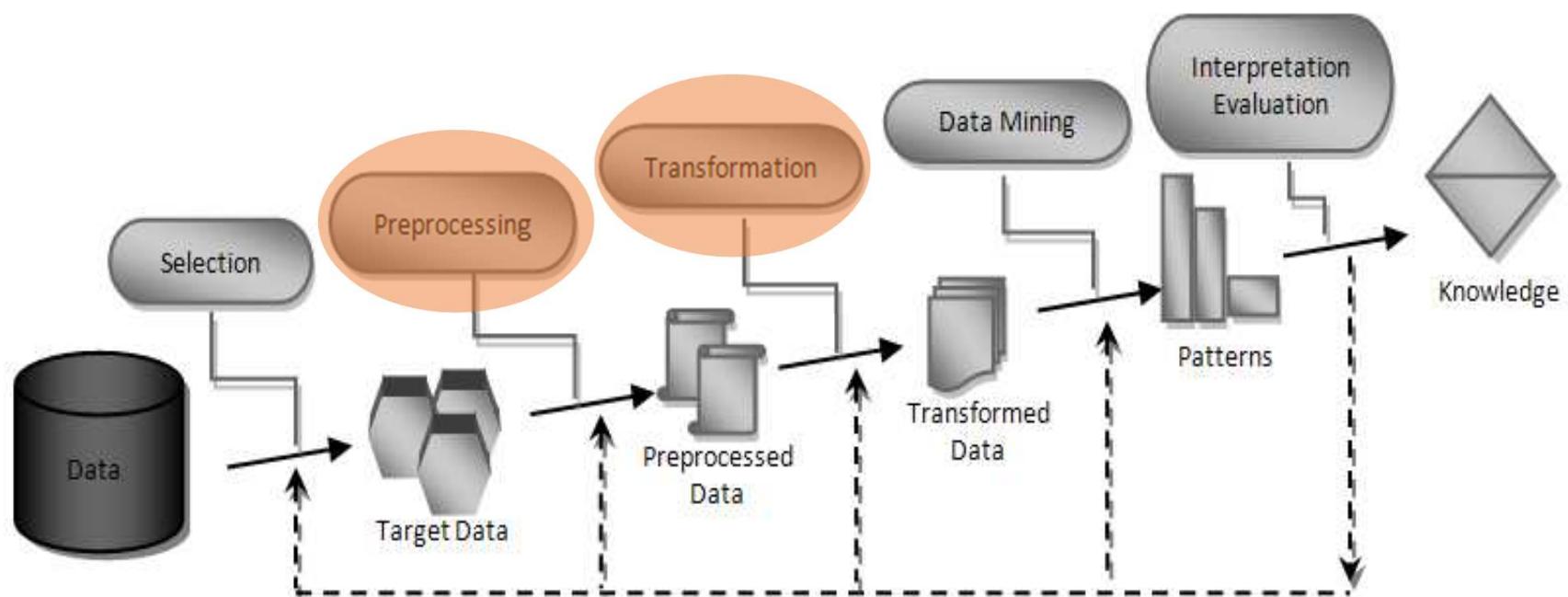
- Tarefas no RapidMiner

- Gerir valores omissos

- Eliminar os exemplos com valores omissos ou ignorá-los durante as análises
 - Substituir com possíveis valores (min, max, average, ...) – *Replace Missing Values*
 - Estimar valores omissos – *Impute Missing Values*



Pré-processamento dos dados



Fayyad, 1996

Pré-processamento dos dados

- Alterar os dados em bruto num formato mais fácil de os analisar
- Várias técnicas são utilizadas:
 - Agregação
 - Amostragem
 - Redução de dimensionalidade
 - Criação de variáveis
 - Discretização
 - Transformação

Pré-processamento dos dados

- Agregação – condensar informação em entradas mais gerais
 - Objetivos
 - Redução de dados
 - Ex: combinar dois atributos num só
 - Alteração da escala
 - Ex: em vez de se organizar em dias pode-se trabalhar com meses
 - Dados mais estáveis
 - Dados agregados tendem a ter menor variabilidade

Pré-processamento dos dados

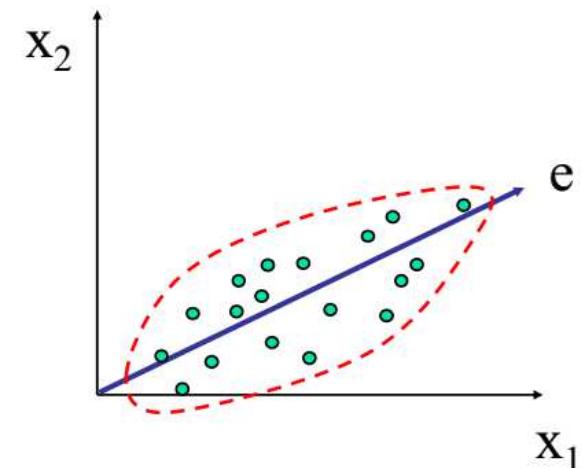
- Amostragem – uma amostra representativa da população permite obter bons resultados com custos computacionais mais baixos!
 - É utilizada na análise preliminar dos dados ou para a análise final
 - Trabalhar com todos os dados pode ter custos monetários e de tempo elevados
 - Princípios gerais para amostragem
 - Se a amostra for representativa utilizar a amostra poderá produzir os mesmos resultados ao utilizar todo o data set
 - Uma amostra é representativa se tiver as mesmas propriedades de interesse do data set original

Pré-processamento dos dados

- Tipos de Amostragem:
 - Amostragem aleatória
 - A probabilidade é a mesma de se escolher um determinado item
 - Amostragem sem reposição
 - À medida que determinado item é selecionado ele é retirado da população
 - Amostragem com reposição
 - Itens não são removidos da população se forem selecionados
 - O mesmo item pode ser escolhido mais do que uma vez
 - Amostragem estratificada
 - Dividir os dados em várias partições, depois escolher amostras aleatórias de cada partição

Pré-processamento dos dados

- Redução de dimensionalidade – bases de dados poderão conter muitas variáveis, mas é importante reduzir esse número por questões de interpretação do modelo, por questões de visualização e entendimento.
 - Análise de componentes principais
 - Objetivo: Encontrar a projeção que captura a maior parte de variação nos dados
 - Encontrar os vetores próprios da matriz de covariâncias para definir o novo espaço
 - Decidir quais são significativos
 - Formar um novo sistema de vetores (dimensão menor)
 - Fazer o mapeamento dos dados para o novo espaço



Pré-processamento dos dados

- **Análise de Componentes Principais** – em data mining
 - Aplicar ACP ao conjunto de treino para obter um novo sistema de coordenadas definido apenas pelos vetores próprios que são estatisticamente significativos (segundo o valor dos valores próprios)
 - Obter a representação de cada exemplo (objeto) do conjunto de treino no novo sistema de coordenadas de dimensão menor que o espaço original de atributos
 - Aprender um classificador usando a nova representação dos objetos do conjunto de treino
 - Na etapa de classificação cada novo objeto deve ser inicialmente representado no novo sistema de coordenadas ACP. Logo o classificador induzido é usado para a sua classificação

Pré-processamento dos dados

- Análise de componentes principais – aplicação num problema de reconhecimento de faces: Dada uma imagem de uma face humana, compará-la com faces em data sets e concluir quem é e se existe um exemplar igual



Algumas imagens utilizadas para construir o novo espaço de vetores próprios (Matthew Turk e Alex Pentland)

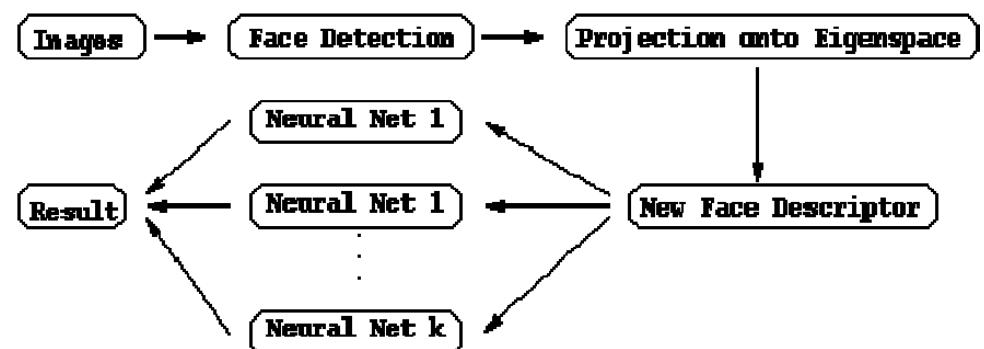
1. Encontrar a face na imagem (outro problema)
2. Cada face é uma matriz de 2 dimensões com os valores de intensidade (46×46 pixels) que pode ser tratado como um vetor ou ponto no espaço de dimensão 2116 -> cada face -> objeto descrito através de **2116 atributos**
3. ACP é aplicado para encontrar o aspetto de faces que são importantes para identificação. Um novo Sistema de coordenadas para as faces é criado utilizando os vetores próprios de um conjunto de imagens com faces.

Pré-processamento dos dados

- Análise de componentes principais – aplicação num problema de reconhecimento de faces: Dada uma imagem de uma face humana, compará-la com faces em data sets e concluir quem é e se existe um exemplar igual



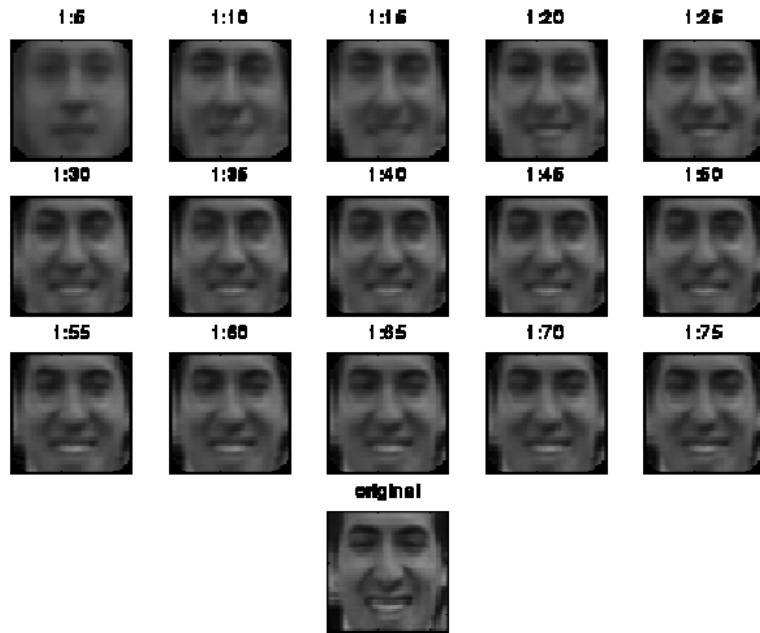
20 Imagens com valores
próprios mais significativos



Redes neurais separadas são construídas para cada pessoa para reconhecimento de faces através da aprendizagem da classificação correta do novo Sistema de coordenadas de ACP. A face de entrada é projetada no espaço de vetores próprios e são obtidos os pesos que são utilizados como variáveis de entrada para cada rede neuronal por indivíduo. Aquele com o output maior é selecionado.

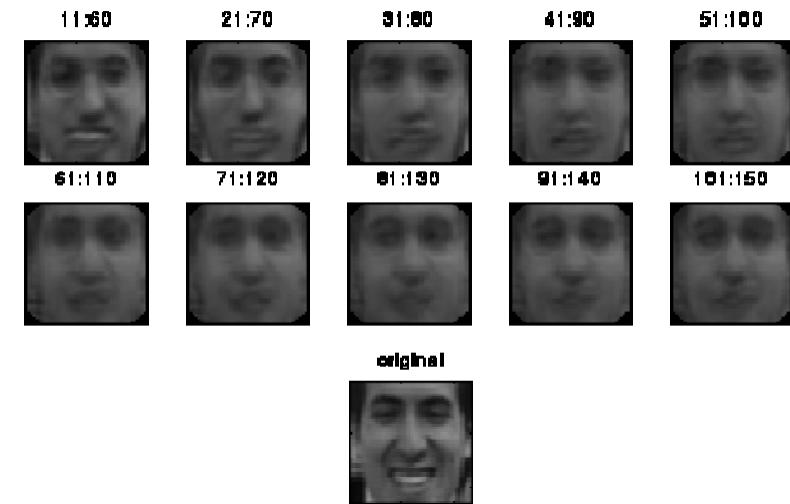
Pré-processamento dos dados

- Análise de componentes principais – aplicação



Faces reconstruídas utilizando os vetores próprios com os maiores valores próprios.

A legenda acima de cada face é o intervalo de vetores próprios utilizados.



Faces reconstruídas utilizando os vetores próprios com os menores valores próprios.

Pré-processamento dos dados

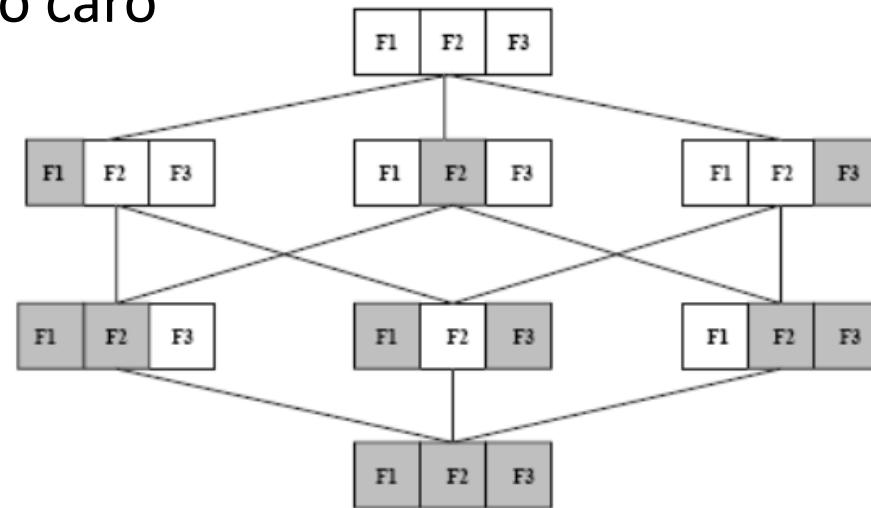
- Seleção de subconjunto de variáveis (Feature Subset Selection)
 - a dimensão é reduzida usando apenas um subconjunto de atributos
 - Tipo de atributos que podem ser removidos
 - **Atributos redundantes:** duplicam informação contida em um ou mais atributos
 - Ex: preço de um produto e valor total da venda
 - **Atributos irrelevantes:** não contêm informação útil para a tarefa de aprendizagem
 - Ex: Número do estudante é irrelevante na tarefa de prever o seu desempenho

A existência de atributos redundantes e irrelevantes pode afetar o desempenho de um classificador ou a qualidade dos clusters que podem ser construídos dos dados

Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) – **Abordagem de Força Bruta**
- Explorar todos os possíveis subconjuntos de atributos e escolher aquele que produz os melhores resultados para a tarefa de aprendizagem (para n atributos será necessário explorar $2^n - 1$ possíveis conjuntos)
- Computacionalmente muito caro

Ex: No caso representado na figura os subconjuntos possíveis para 3 atributos

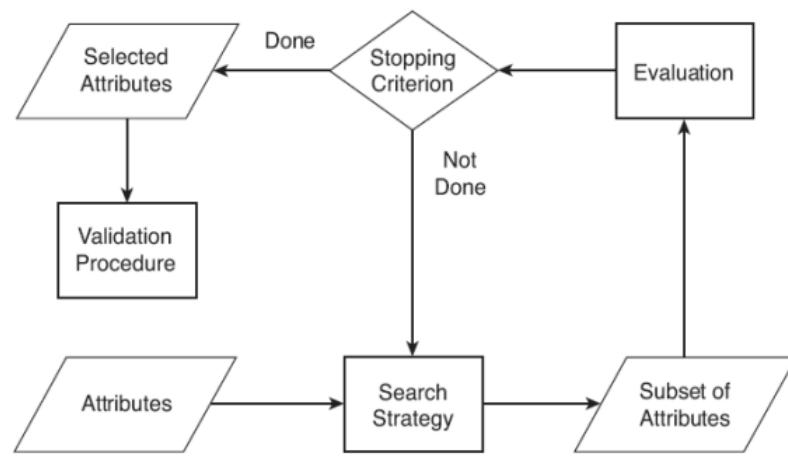


Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) –
Tipos de abordagens
 - Embedded approaches (Integradas)
 - Filter approaches (Filtro)
 - Wrapper approaches (Envolvidas)
 - Feature weighting

Pré-processamento dos dados

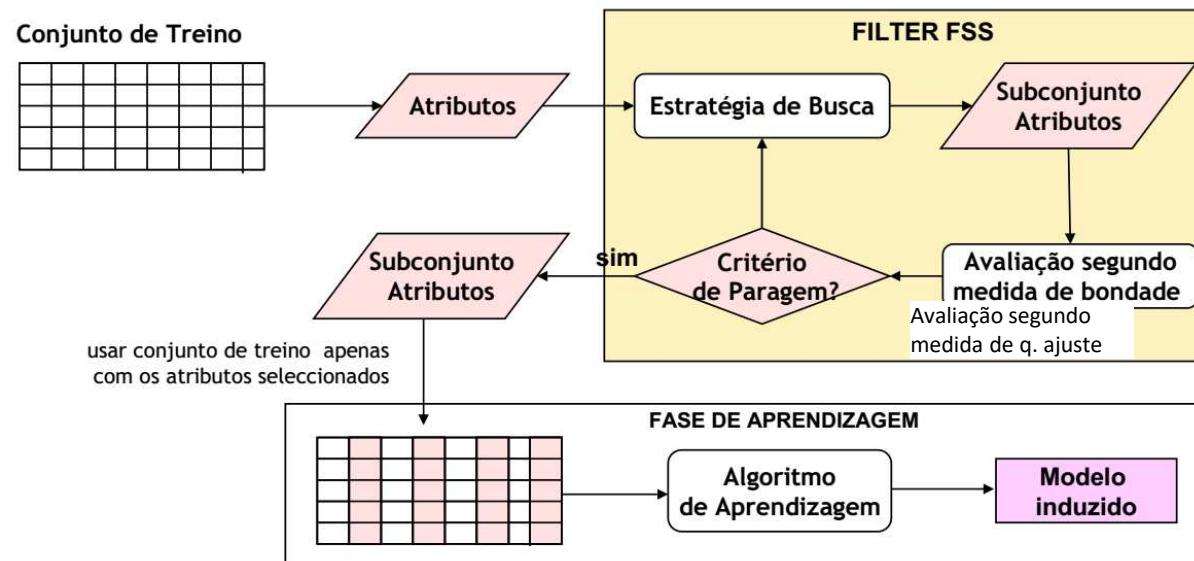
- Seleção de subconjunto de variáveis (Feature Subset Selection) – **Problema de Busca Integrada**
- Objetivo: encontrar o subconjunto de atributos no espaço de possíveis subconjuntos com o melhor valor de uma medida de qualidade de ajuste (*goodness of fit*)



1. **Medida para avaliar a qualidade de ajuste** - em relação à tarefa de aprendizagem (classificação, clustering,...) -> determinar a função objetivos que guia a busca
2. **Estratégia de busca** – força bruta vs aborgagens heurísticas (hill-climbing, best-first search,...)
3. **Critério de paragem** – nº máximo de iterações, max de elementos no conjunto, valor da função objetivo que ultrapassa um dado limite,...)
4. **Procedimento de validação** – ex: avaliar o desempenho do modelo induzido usando todos os atributos vs. usando apenas o subconjunto obtido (melhor? pior? fica na mesma?)

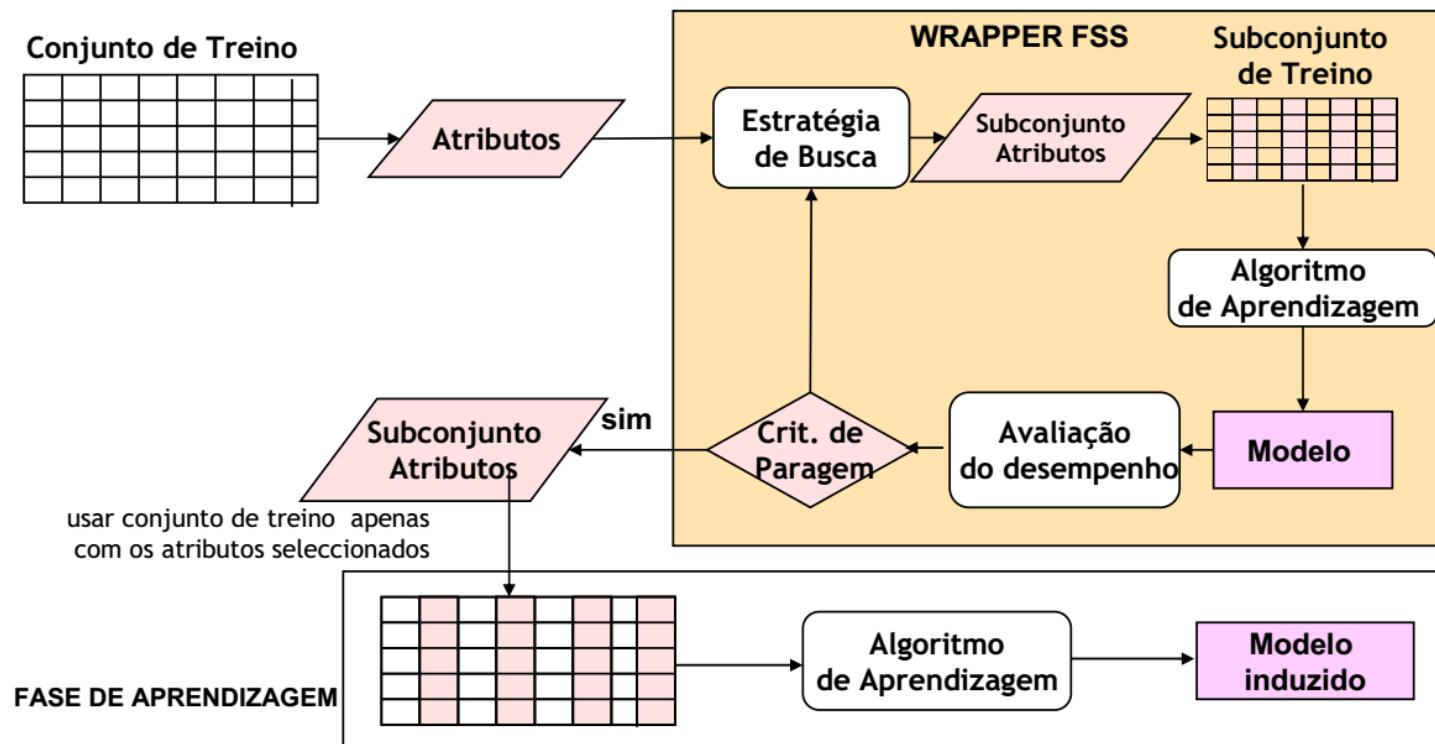
Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) – **Filtros**
- Os atributos são selecionados antes da fase de aprendizagem. Normalmente usam-se testes estatísticos ou medida de informação mútua para medir o grau de correlação entre cada atributo e a classe (se classificação) ou entre pares de atributos (se clustering), podendo-se escolher os pares cuja correlação é a menor possível.



Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) – **Wrapper**
- Usa o próprio algoritmo de aprendizagem como black-box (caixa negra) na tarefa de seleção de atributos.



Pré-processamento dos dados

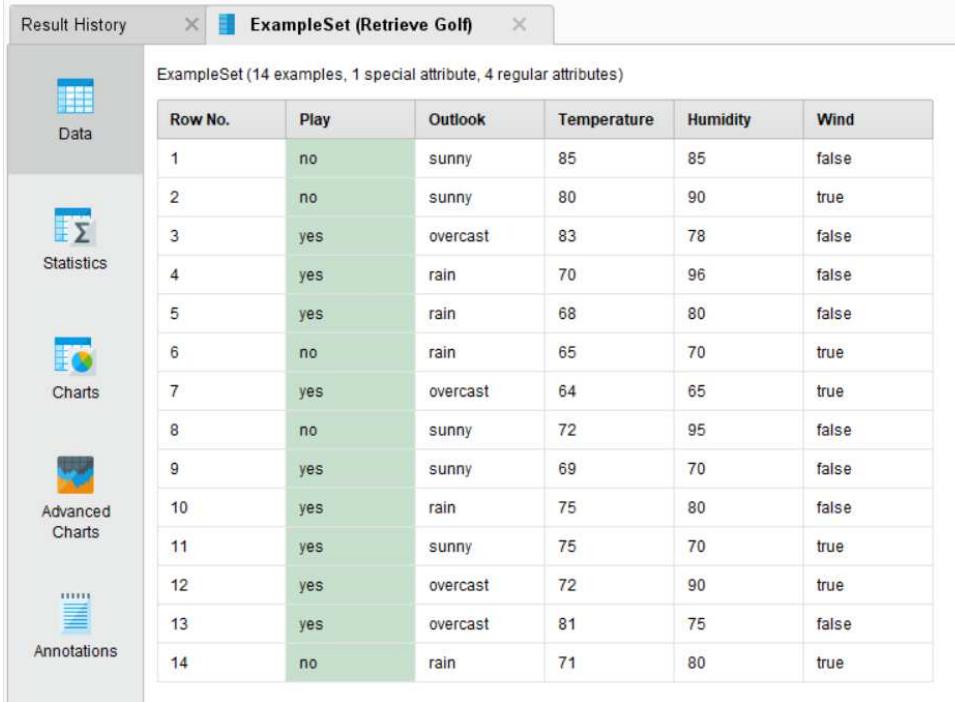
- Seleção de subconjunto de variáveis (Feature Subset Selection) –
Feature weighting
 - A cada atributo X_i é atribuído um peso w_i que reflete a sua importância na tarefa de aprendizagem
 - um atributo com maior peso é mais relevante do que um atributo com menor peso
 - Os atributos são ordenados segundo os pesos (normalmente os pesos são normalizados para que fiquem no intervalo [0, 1])
 - Podem ser selecionados os k atributos com maiores pesos, ou alternativamente, aqueles com peso superior a um dado limite (*threshold*)
 - O peso pode ser calculado por diferentes métodos

Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) –
Feature weighting
 - A estatística do chi-quadrado, χ^2 , é usada para medir o grau de dependência da cada atributo em relação à classe
 - para cada atributo X_i , $i = 1, \dots, n$ é calculado o valor da estatística de χ^2
 - os valores de χ^2 são normalizados e tomados como pesos
 - os atributos são ordenados decrescentemente pelo peso
 - são selecionados os k atributos com maior peso ou aqueles que satisfazem um dado critério de seleção (por exemplo, cujos pesos são maiores que um dado *threshold*)

Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) –
Feature weighting
 - Exemplo – Aplicação Teste Qui-Quadrado



Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

1º Criar tabela de contingência para Play x Wind com as frequências observadas

		Wind		Total
		false	true	
Play	no	2	3	5
	yes	6	3	9
Total		8	6	14

$$\text{FreqEsp}_{ij} = \frac{\text{TotalLinha}_i \times \text{TotalColuna}_j}{\text{Total}}$$

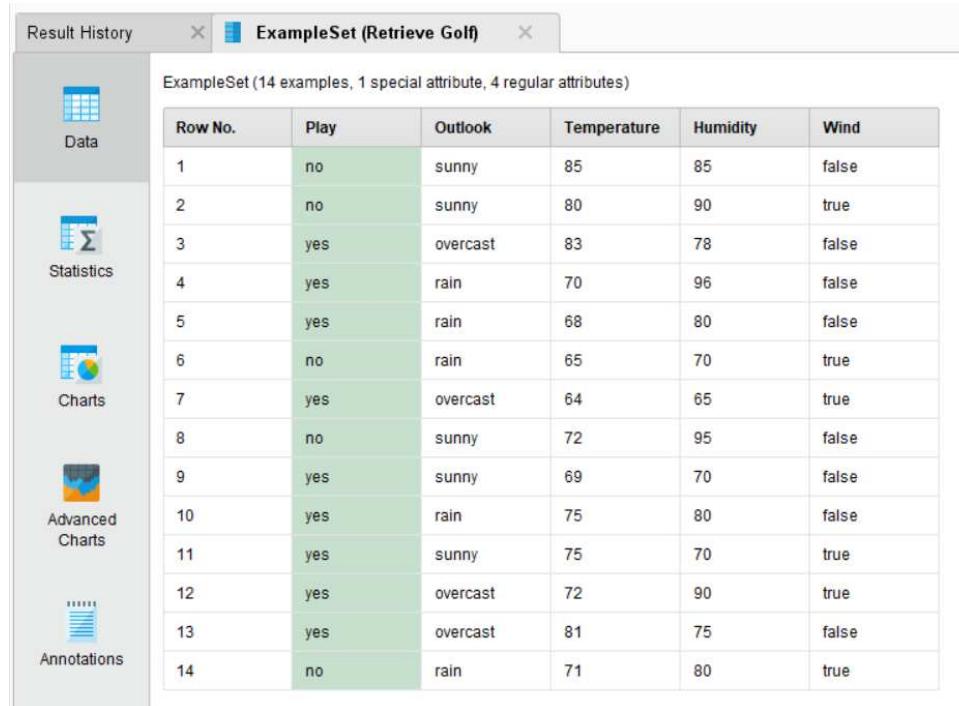
2º Calcular as frequências esperadas

		Wind			
		false		true	
Play	Obs	Esp	(O-E)^2/E	Obs	Esp
	2	2.8571	0.2571	3	2.1429
yes	6	5.1429	0.1428	3	3.8571
					0.1905

$$\text{Esp}_{11} = \frac{8}{14} \times 5 = 2.8571$$

Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) –
Feature weighting
 - Exemplo – Aplicação Teste Qui-Quadrado



Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

3º Calcular a estatística Chi-quadrado

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$



$$\chi^2 = 0.9332$$

Pré-processamento dos dados

- Seleção de subconjunto de variáveis (Feature Subset Selection) – **Feature weighting**

- Exemplo – Aplicação Teste Qui-Quadrado

- 1. Para cada atributo X_i , $i = 1, \dots, n$ é calculado o valor da estatística de χ^2

attribute	weight
Wind	0.933
Humidity	3.474
Outlook	3.547
Tempera...	6.741

Se o atributo não é discreto é preciso aplicar um método de discretização

- 2. Normalizar estes valores por forma a que fiquem no intervalo $[0, 1]$

A **temperatura** é o atributo mais relevante com relação à decisão se ir ou não a jogar golf

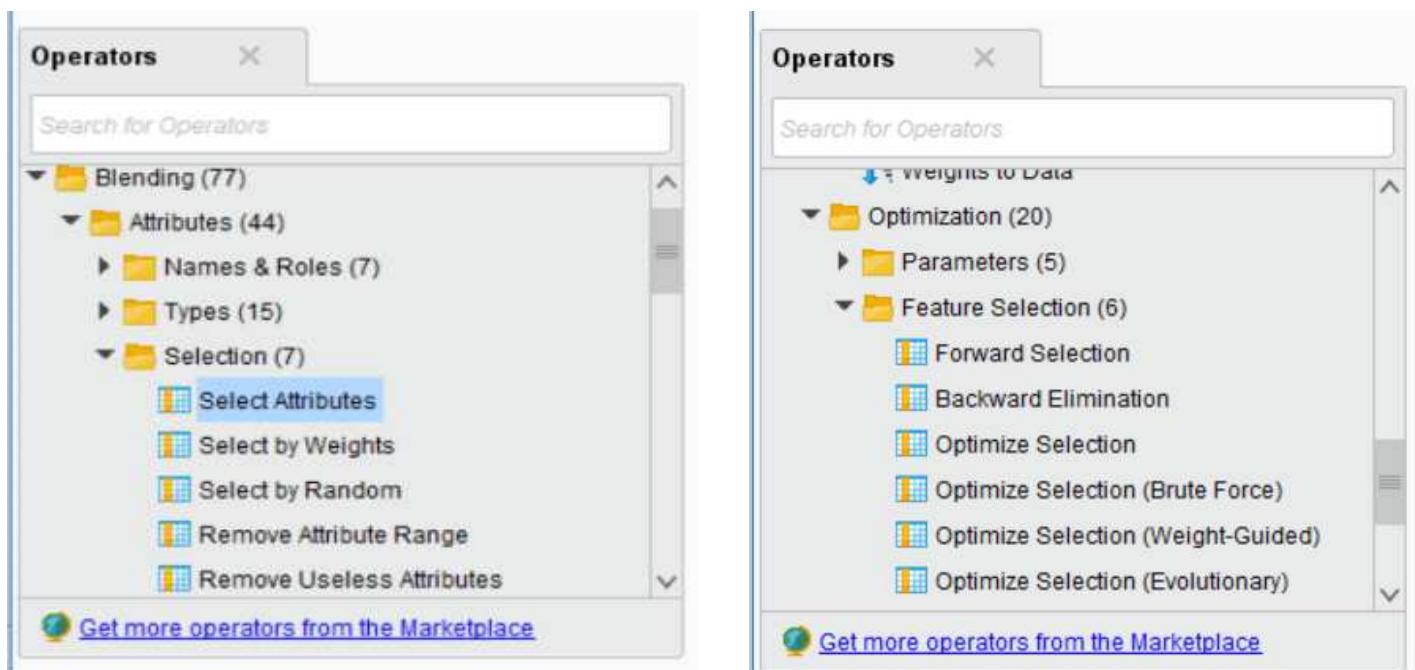
attribute	weight
Wind	0
Humidity	0.438
Outlook	0.450
Temperature	1

$$normValor = \frac{valor - min}{max - min} \times (b - a) + a$$

Podemos normalizar para obter valores num intervalo $[a, b]$ aplicando o método de min-max

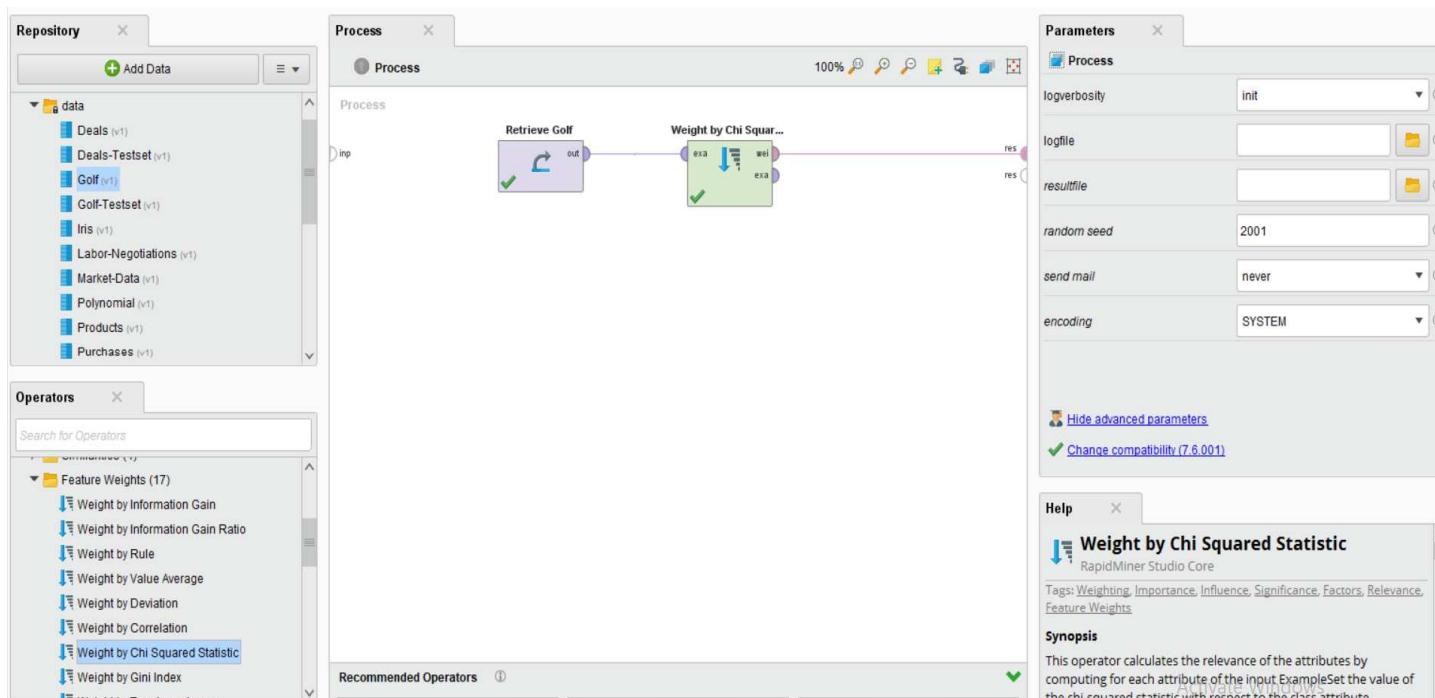
Pré-processamento dos dados

- Seleção de subconjunto de variáveis – **No RapidMiner**
 - Path: Blending – Selection – Select Attributes
 - Path: Optimization – Feature Selection



Pré-processamento dos dados

- Seleção de subconjunto de variáveis – **No RapidMiner**
 - Path: Modeling – Feature weights – Weight by Chi Square Statistics



Pré-processamento dos dados

- Criação de variáveis – através das variáveis iniciais podem ser criadas novas variáveis
 - Ex: Através das variáveis peso e altura podemos criar a variável do Índice de Massa Corporal (IMC)

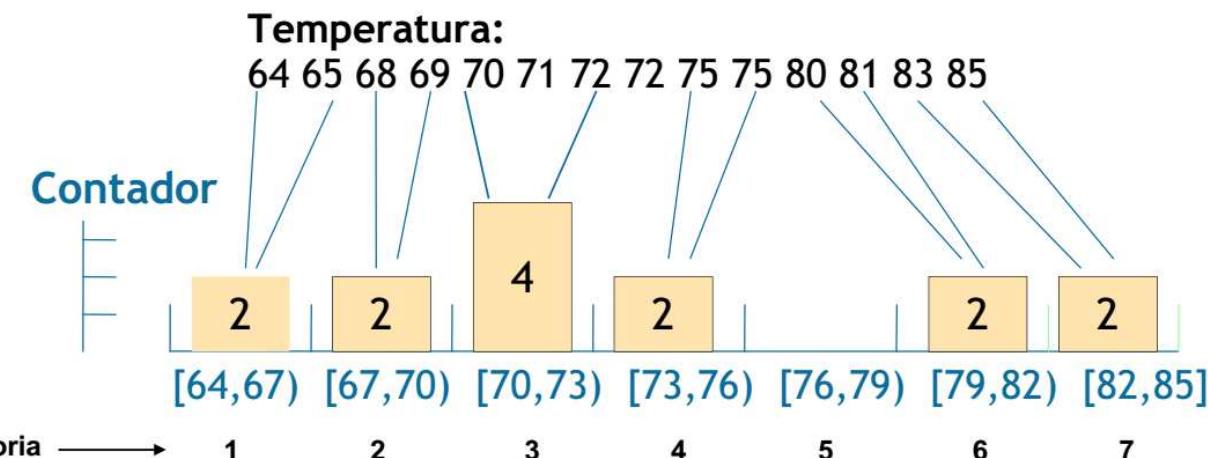
$$\text{IMC} = \frac{\text{peso(kg)}}{\text{altura}^2(\text{m})}$$

Pré-processamento dos dados

- Discretização – muitos algoritmos têm como requisito que as variáveis sejam qualitativas e processos de transformação de variáveis contínuas sejam discretizadas
 - Duas abordagens consoante a classe é ou não usada no processo de discretização
 - **Não supervisionada** – não considera a classe
 - Intervalos de largura fixa (Equal width method)
 - Intervalos com igual frequência (Equal frequency method)
 - K-means (algoritmo para criar grupos (clusters))
 - **Supervisionada** – considera a classe
 - Baseada na entropia (Minimal Entropy Partitioning)

Pré-processamento dos dados

- Discretização – Intervalos de Largura Fixa (divide o intervalo de valores de um atributo num número dado de intervalos de largura fixa – não é apropriado se existem outliers)



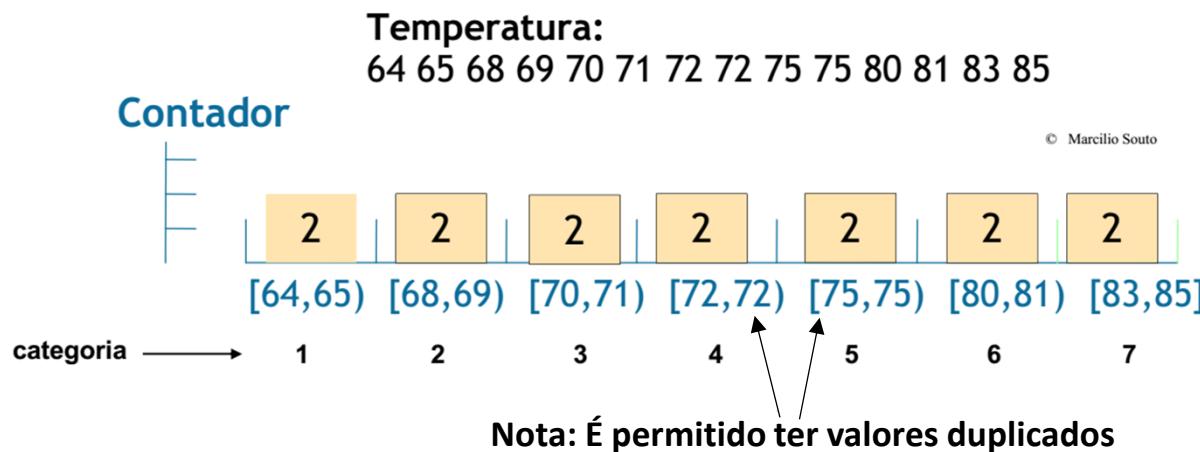
Dividir em 7 intervalos (bins)

- K=7
- Largura = $(\text{max-min}+1)/k = (85-64+1)/7 = 3$

Não supervisionada

Pré-processamento dos dados

- Discretização – Intervalos com Frequência Fixa (atribuição do mesmo número de valores a cada intervalo)



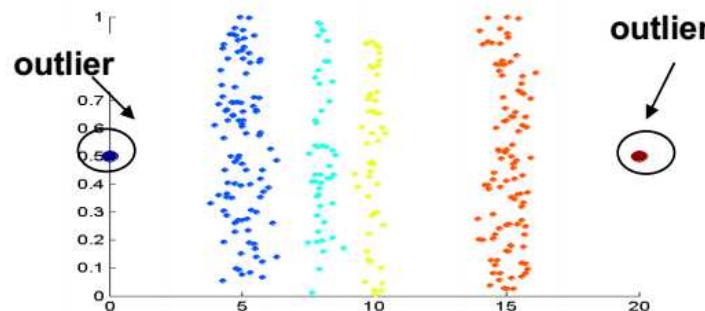
Dividir em 7 intervalos (bins)

- $K=7$
- Frequência = $N/k = 14/7 = 2$

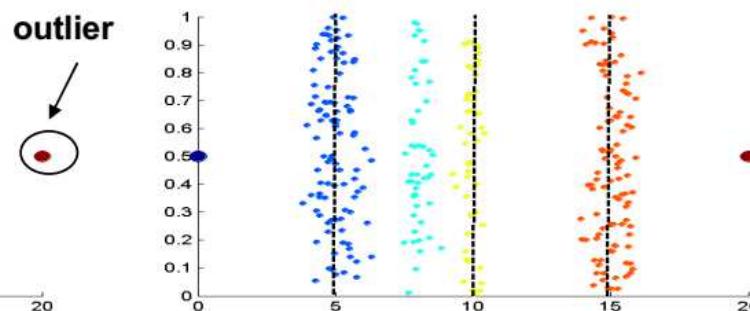
Não supervisionada

Pré-processamento dos dados

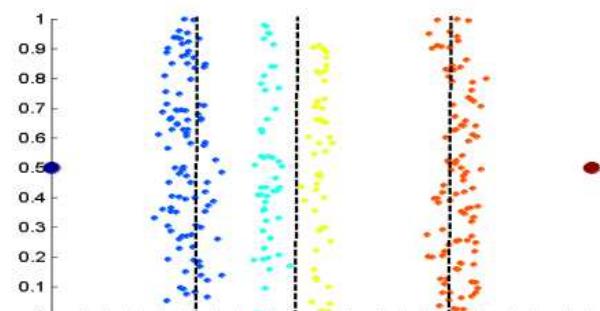
- Discretização – k-means (forma grupos nos dados)



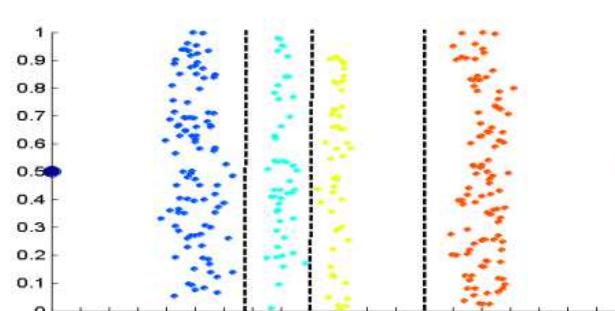
Data



Equal interval width



Equal frequency



K-means

Neste caso foram aplicadas as 3 abordagens considerando $k=4$ intervalos.

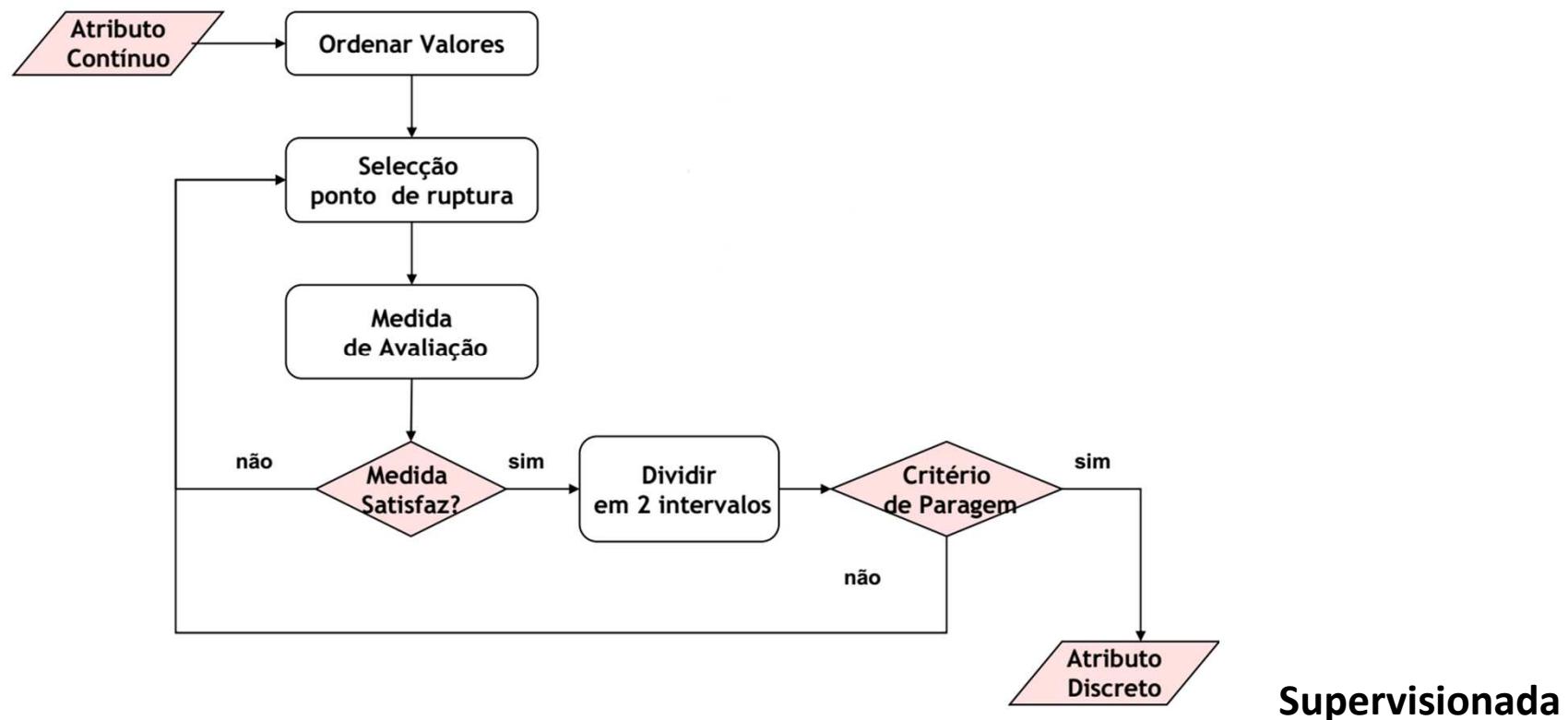
Se for medido o sucesso pela forma como os objetos de uma mesma classe ficam atribuídos a um intervalo, então o melhor método é o k-means seguido do Equal Frequency.

© Tan, Steinbach, Kumar

Não supervisionada

Pré-processamento dos dados

- Discretização – Por partição [Esquema Geral]



Pré-processamento dos dados

- Discretização – Por partição [Baseada na Entropia]
- Medida de Avaliação: **Entropia** (mede o grau de impureza (desordem) de uma partição dos valores contínuos de um atributo)
- Entropia para um intervalo

$$e_i = -\sum_{j=1}^m p_{ij} \log p_{ij}, \quad p_{ij} = \frac{N_{ij}}{N_i}$$

N_{ij} - número de valores da classe c_j no intervalo I_i
 N_i - número de valores total no intervalo I_i

e_i – entropia associada à distribuição da variável classe C no intervalo I_i

- Entropia total para k intervalos

$$e = \sum_{i=1}^k w_i e_i, \quad w_i = \frac{N_i}{N}$$

N_i - número de valores total no intervalo I_i
 N - número total de valores = número de exemplos

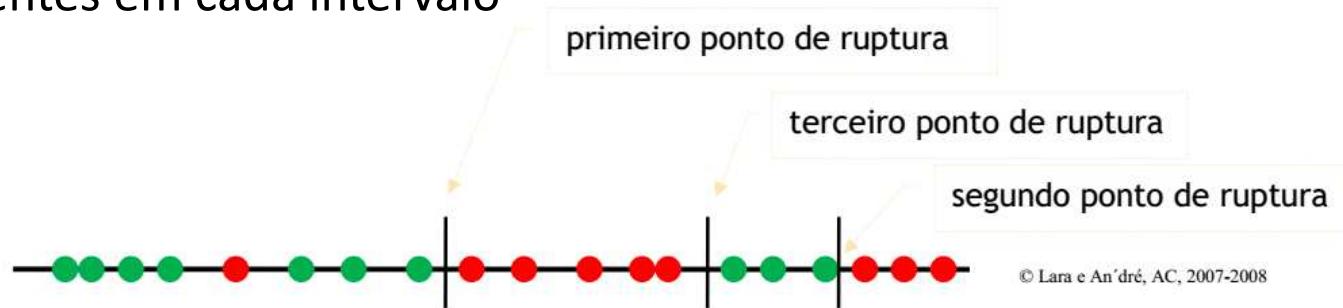
- Processo de partição

- Em cada passo selecionar o ponto de ruptura de forma a que os futuros dois intervalos produzam o **menor valor da entropia**

Supervisionada

Pré-processamento dos dados

- Discretização – Minimal Entropy Partition
- Objetivo: minimizar a entropia em cada intervalo, de forma a minimizar a entropia total \Rightarrow minimizar o número de objetos de classes diferentes em cada intervalo



Para o 1º ponto de rutura:

$$e_1 = -\left(\frac{1}{8} \log \frac{1}{8} + \frac{7}{8} \log \frac{7}{8}\right) = 0.3768 \quad e_2 = -\left(\frac{8}{11} \log \frac{8}{11} + \frac{3}{11} \log \frac{3}{11}\right) = 0.5860 \quad e = \frac{8}{19} e_1 + \frac{11}{19} e_2 = 0.4979$$

Para a partição encontrada: $e_1 = 0.3768, e_2 = e_3 = e_4 = 0$

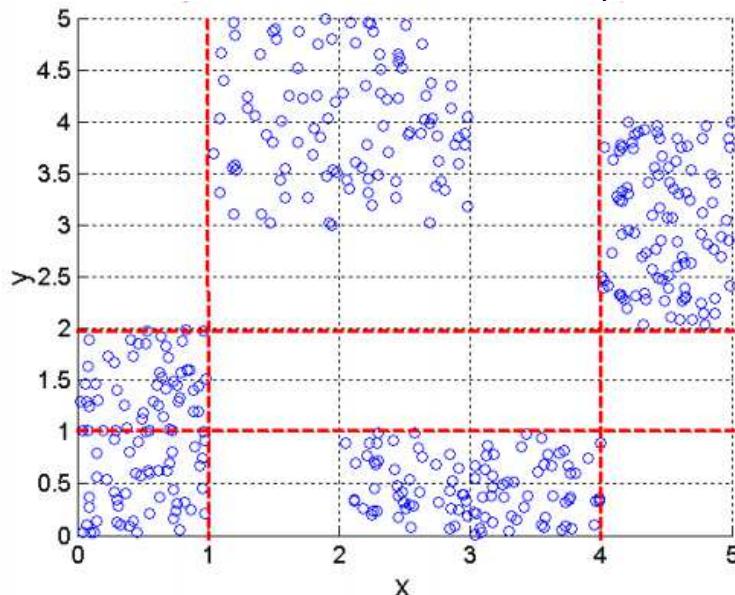
$$e = \frac{8}{19} e_1 + \frac{5}{19} e_2 + \frac{3}{19} e_3 + \frac{3}{19} e_4 = \frac{8}{19} \times 0.3768 = 0.1586$$

Supervisionada

Pré-processamento dos dados

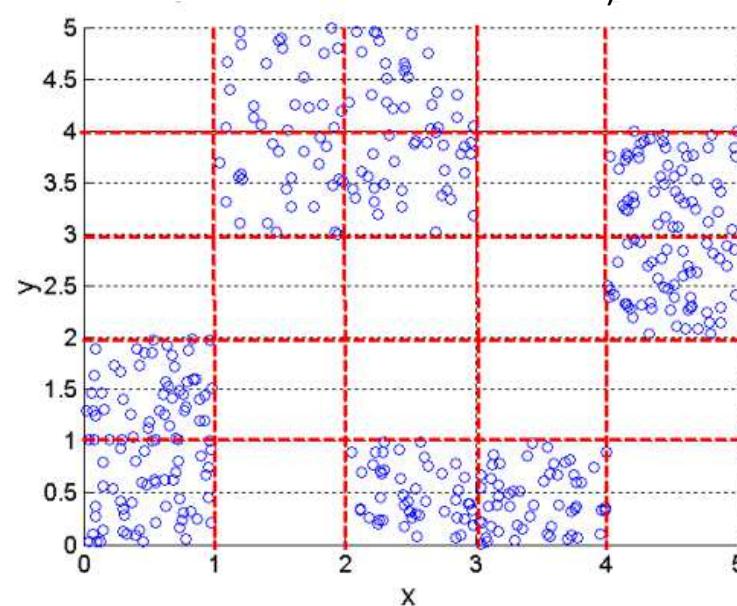
- Discretização – Minimal Entropy Partition

Para $k = 3$ (os valores dos atributos x e y são divididos em 3 intervalos)



Em 2 dimensões as classes ficam bem separadas, mas não numa só dimensão. Ou seja, se for discretizado 1 só atributo só garantimos resultados sub-ótimos

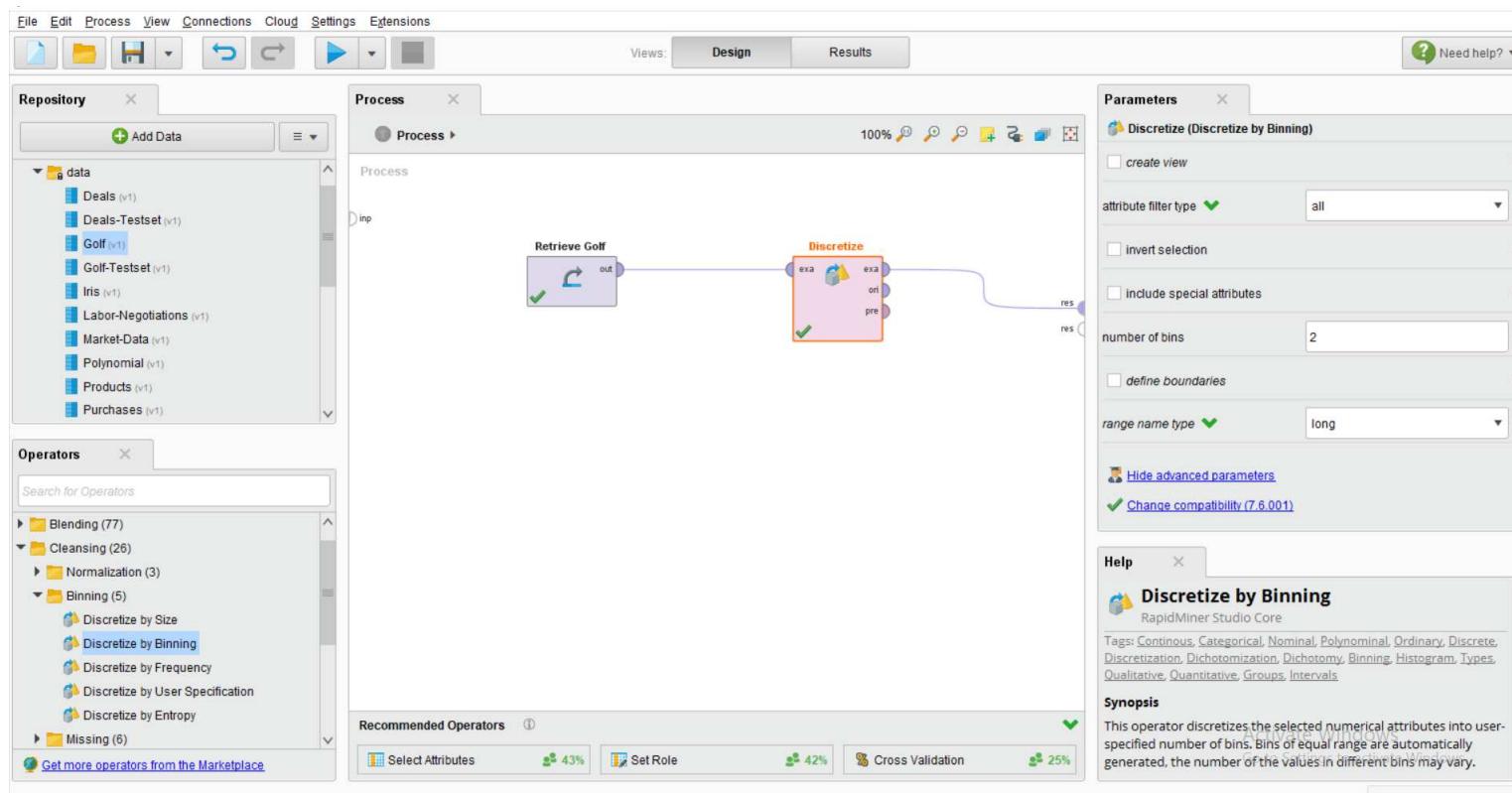
Para $k = 5$ (os valores dos atributos x e y são divididos em 5 intervalos)



5 intervalos produzem melhores resultados que 3, mas quando se usam 6 o resultado permanece igual em termos de entropia. É necessário o uso de um critério de paragem para determinar o nº apropriado de partições

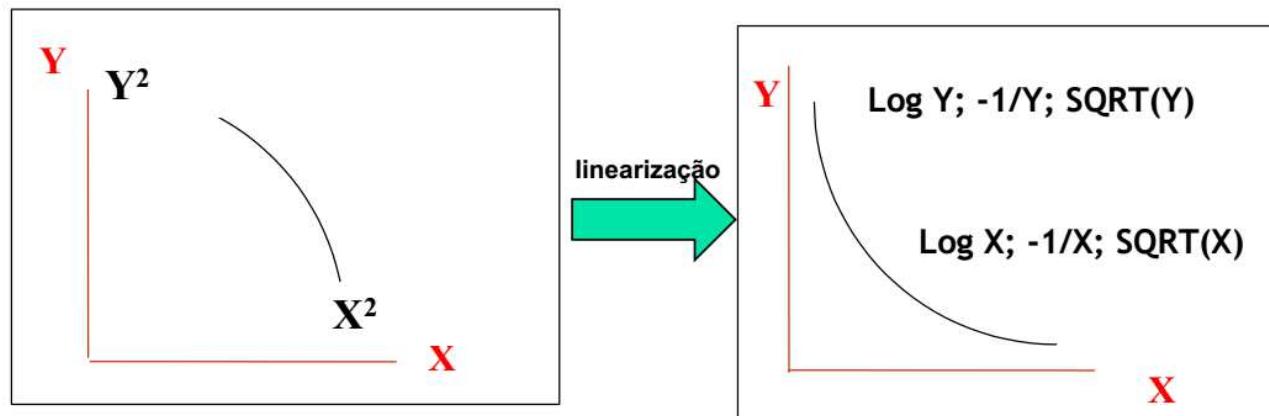
Pré-processamento dos dados

- Discretização – No RapidMiner
- Path: Cleansing – Binning



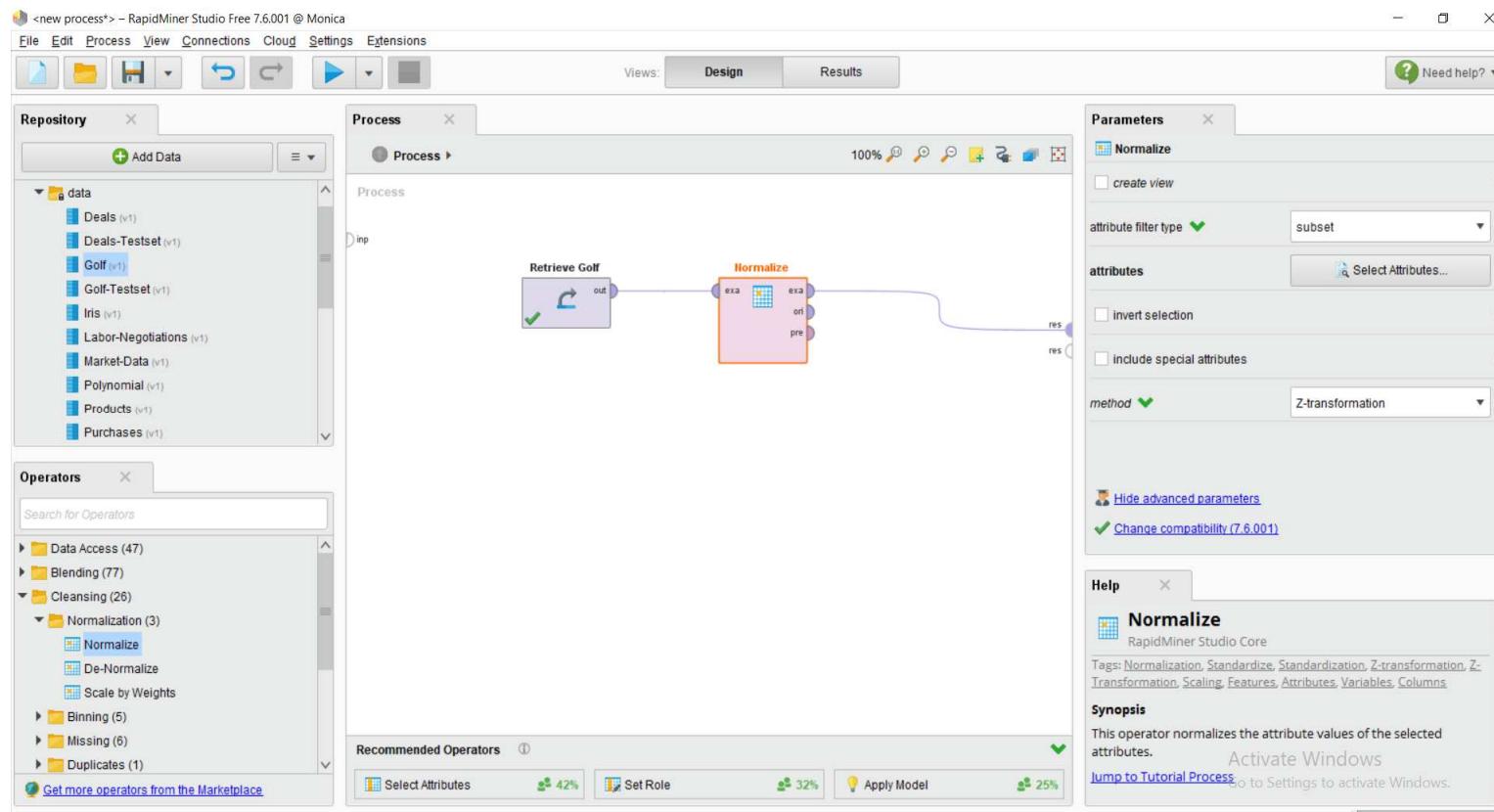
Pré-processamento dos dados

- Transformação - refere-se à alteração que é aplicada a todos os valores de uma dada variável
 - Normalização (score Z; média = zero; desvio padrão = 1)
 - Aplicação de funções tais como transformações logarítmicas e de raiz quadrada (x^k ; $\log(x)$; $|x|$)

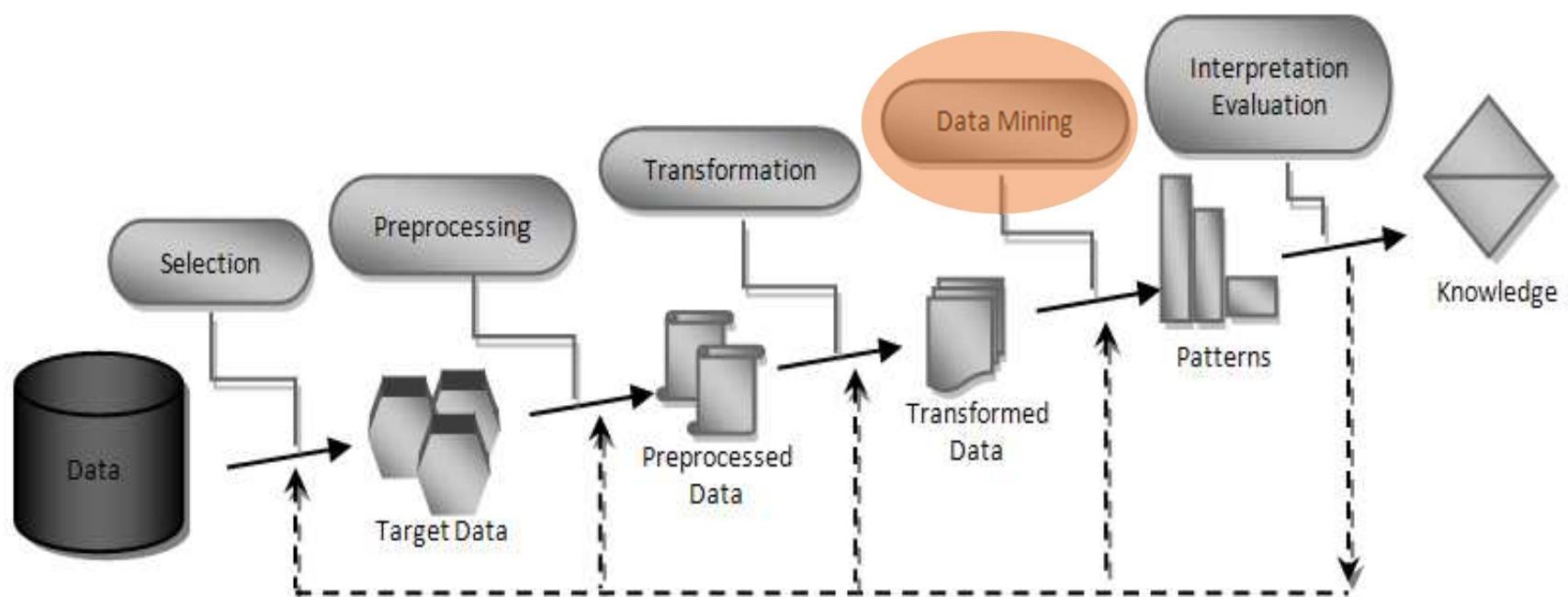


Pré-processamento dos dados

- Transformação – No RapidMiner
- Path: Cleansing – Normalize



Data Mining



Fayyad, 1996

Data Mining

- Tarefas e problemas
 - Preditivos – prever um determinado valor de uma variável baseado nos valores das outras variáveis
 - Descritivos – apresentar os padrões e relações existentes nos dados
- Classificação e Regressão

V1	V2	V3	...	Classe

Data Mining

- Exemplos de tarefas de Data Mining:
 - Classificação [Preditivos]
 - Clustering [Descritivos]
 - Association Rule Discovery [Descritivos]
 - Sequential Pattern Discovery [Descritivos]
 - Regressão [Preditivos]
 - Deviation Detection [Preditivos]

Data Mining

- Definição

- Um classificador (modelo) é uma função que para um novo indivíduo retorna uma classe

$$f : X \rightarrow C$$

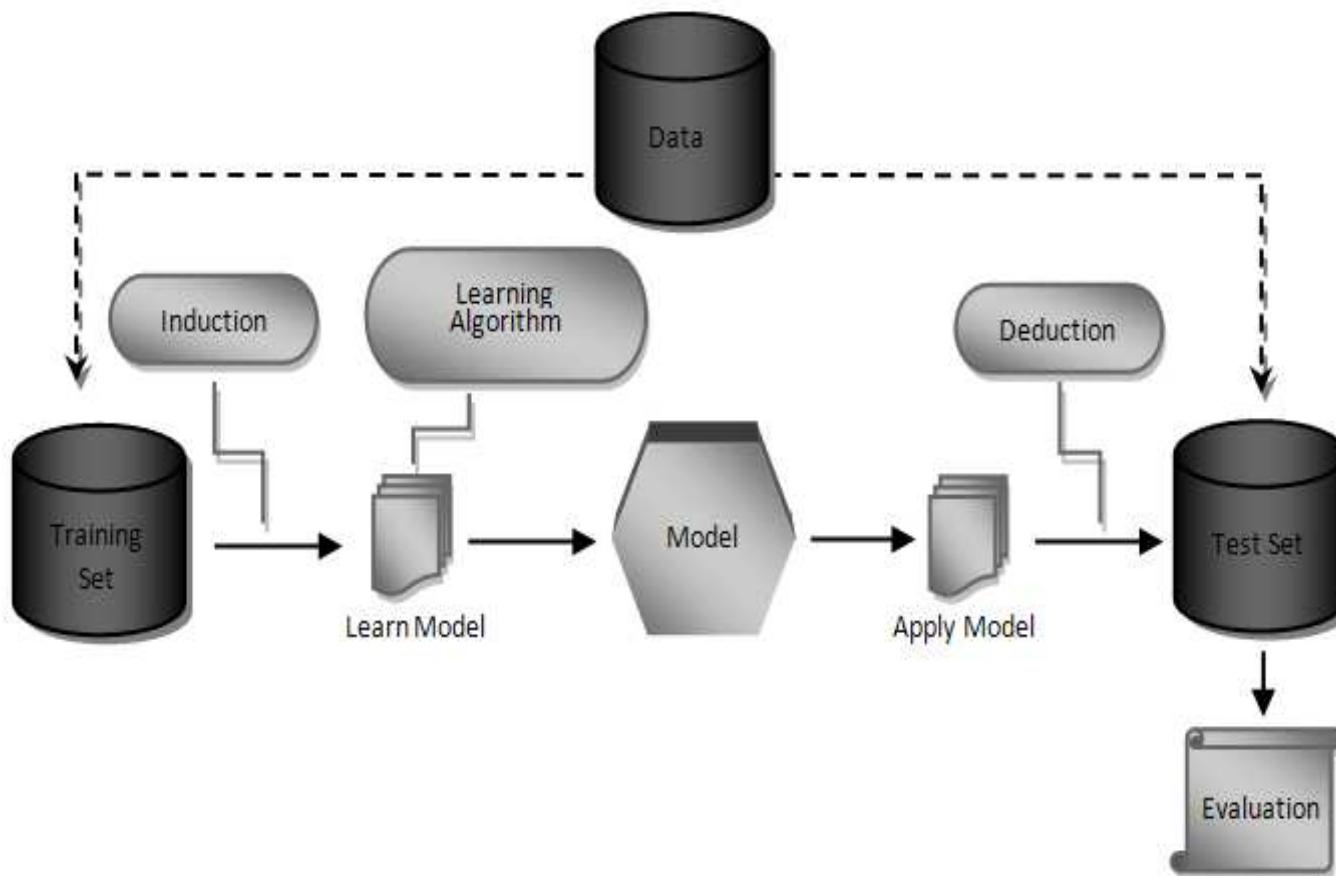
- Técnicas de aprendizagem supervisionadas
 - Induzir um classificador que poderá prever as classes de novos exemplos através do conjunto de treino com indivíduos previamente classificados
 - Técnicas de aprendizagem não supervisionadas
 - Dados padrões descobrir semelhanças agrupando-os
 - Clustering é exemplo deste tipo de aprendizagem

Data Mining

- Abordagem geral para construir um modelo
 - Dado um conjunto de dados (**training set**)
 - Cada objeto contém um conjunto de atributos, um dos quais é uma classe
 - Encontrar um modelo de forma a que um novo objeto sem classe tenha uma alocação o mais precisa possível
 - Um conjunto de teste (**test set**) é utilizado para determinar a taxa de acerto do modelo
 - Dado um conjunto de dados, este é habitualmente dividido em conjunto de treino para construir o modelo e conjunto de teste para o validar

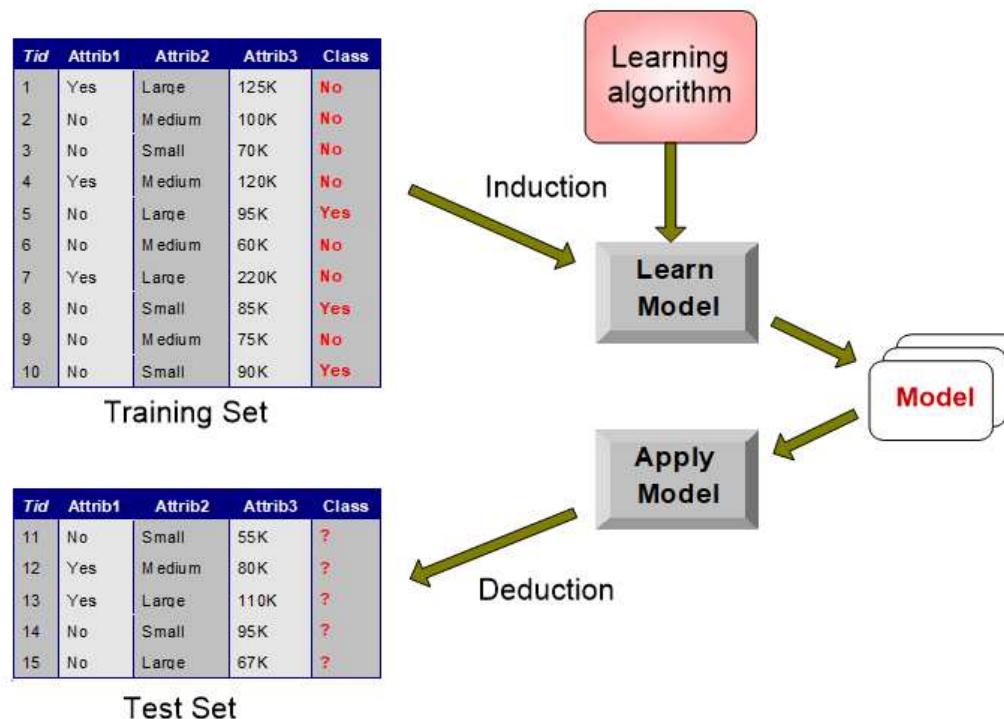
Data Mining

- Abordagem geral para construir um modelo



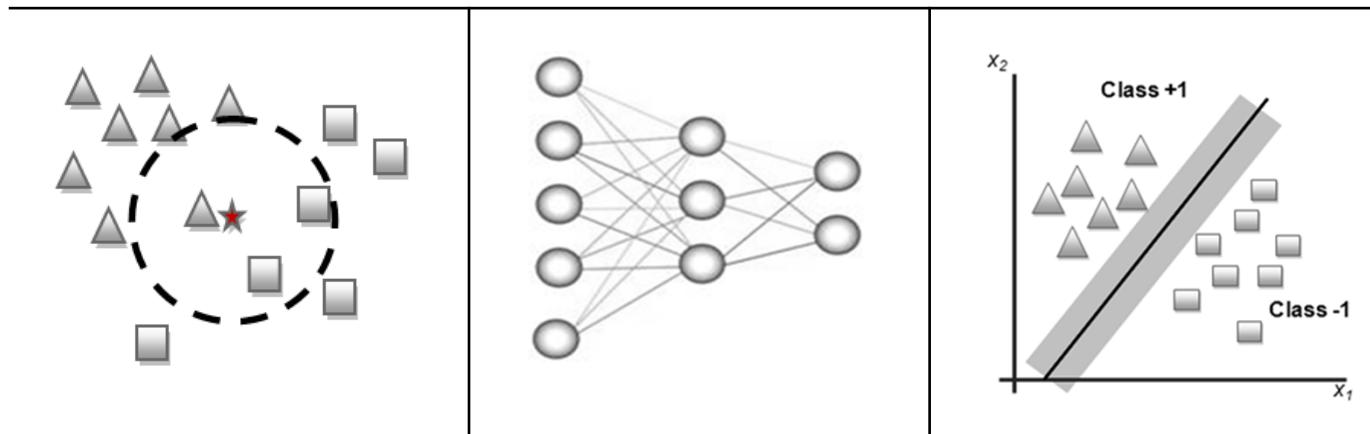
Data Mining

- Abordagem geral para construir um modelo
 - Exemplos:
 - Prever se células de um tumor são ou não cancerígenas
 - Classificar atribuição de crédito

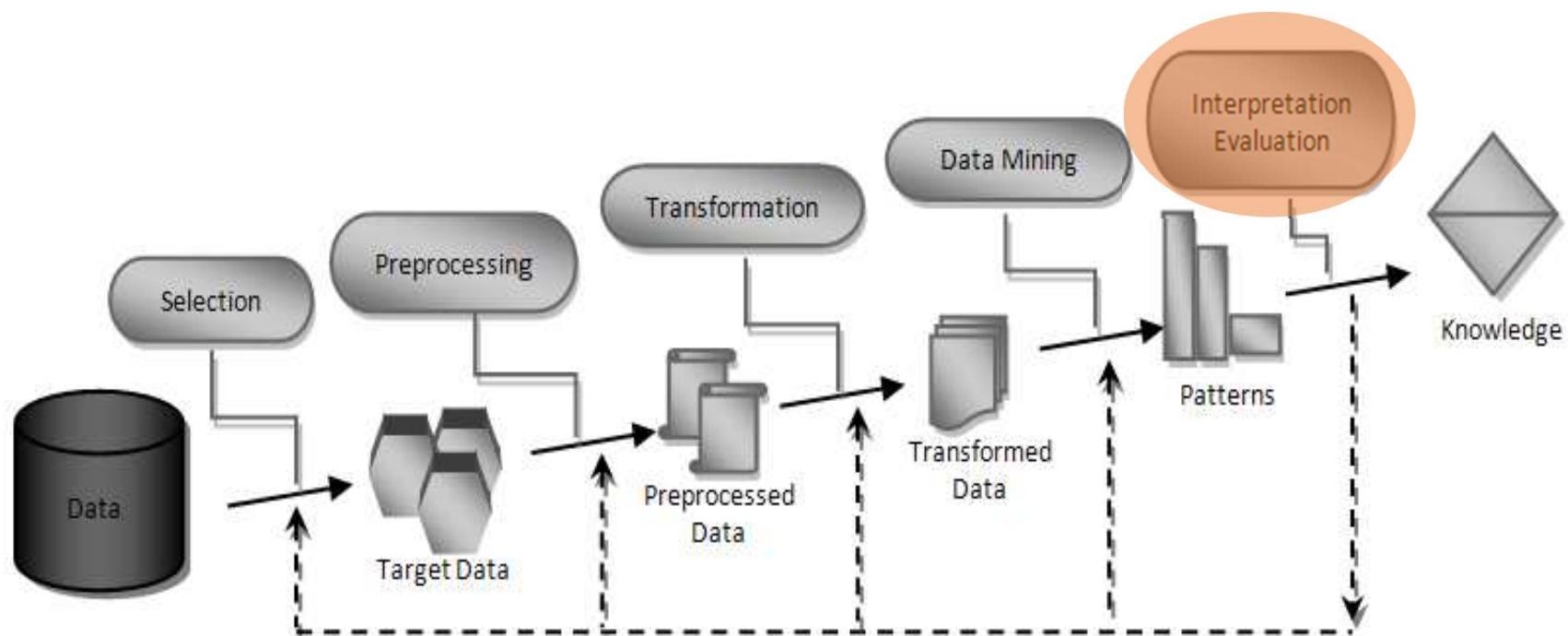


Data Mining

- Algoritmos de aprendizagem
 - Identificar o modelo que melhor ajusta os dados à sua respetiva classe
 - Decision Trees; Rule Based Classifier; Bayesian Classifiers; Nearest Neighbour; Neural Networks; SVM



Interpretação e Avaliação



Fayyad, 1996

Interpretação e Avaliação

- Última fase para entendimento e avaliação de resultados
- Após esta fase poderá finalizar-se o processo ou voltar a fases anteriores
- Após a obtenção dos resultados será importante interpretar e avaliar
 - Parâmetros
 - Modelos obtidos através de diferentes métodos

Interpretação e Avaliação

- **Métricas** para a Avaliação do Desempenho
- **Métodos** para a Avaliação do Desempenho
- Métodos de **Comparação** de Modelos

Interpretação e Avaliação

- Métricas para a Avaliação do Desempenho
 - Objetivo de prever a capacidade de um modelo
 - Mais do que saber a velocidade de construção do modelo
 - Matriz de Confusão

		CLASSE PREVISTA	
		Classe=Sim	Classe=Não
CLASSE VERDADEIRA	Classe=Sim	a	b
	Classe=Não	c	d

a: VP (verdadeiro positivo) TP (*true positive*)

b: FN (falso negativo)
 FN (*false negative*)

c: FP (falso positivo)
 FP (*false positive*)

d: VN (verdadeiro negativo) TN (*true negative*)

Interpretação e Avaliação

- Taxa de acerto

		CLASSE PREVISTA	
CLASSE VERDADEIRA		Classe=Sim	Classe=Não
	Classe=Sim	a (VP)	b (FN)
	Classe=Não	c (FP)	d (VN)

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{VP + VN}{VP + VN + FP + FN}$$

Interpretação e Avaliação

- **Taxa de acerto – Limitações**

- Considerar um problema com 2 classes
 - Número de exemplos da Classe 0 = 9990
 - Número de exemplos da Classe 1 = 10
- Se o modelo prevê todos os objetos na Classe 0, a taxa de acerto é $9990/10000 = 99.9\%$
 - A taxa de acerto é elevada contudo o modelo não deteta nenhum objeto na classe 1

Interpretação e Avaliação

- Precisão

		CLASSE PREVISTA	
CLASSE VERDADEIRA		Classe=Sim	Classe=Não
	Classe=Sim	a (VP)	b (FN)
	Classe=Não	c (FP)	d (VN)

$$\text{Precision (p)} = \frac{a}{a + c}$$

Interpretação e Avaliação

- **Sensibilidade**

		CLASSE PREVISTA	
CLASSE VERDADEIRA		Classe=Sim	Classe=Não
	Classe=Sim	a (VP)	b (FN)
	Classe=Não	c (FP)	d (VN)

$$\text{Recall (r)} = \frac{a}{a + b}$$

Interpretação e Avaliação

- **Medida F**

- relaciona a sensibilidade e a precisão através da média harmónica
- Valores altos da medida F são obtidos só quando a precisão e a sensibilidade têm valores altos

		CLASSE PREVISTA	
		Classe=Sim	Classe=Não
CLASSE VERDADEIRA	Classe=Sim	a (VP)	b (FN)
	Classe=Não	c (FP)	d (VN)

$$F\text{-measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

Interpretação e Avaliação

- **Métodos para Avaliação de Desempenho**

- O desempenho de um modelo pode depender de outros fatores para além do algoritmo de aprendizagem

- Distribuição da classe
- Custo da classificação errada
- Dimensão do conjunto de treino e de teste

- **Curva de Aprendizagem**

- Mostra a variação da taxa de acerto variando a dimensão do conjunto de treino

- **Métodos de estimação da taxa de erro:**

- Holdout; Cross-validation; Leave-one-out; Bootstrap

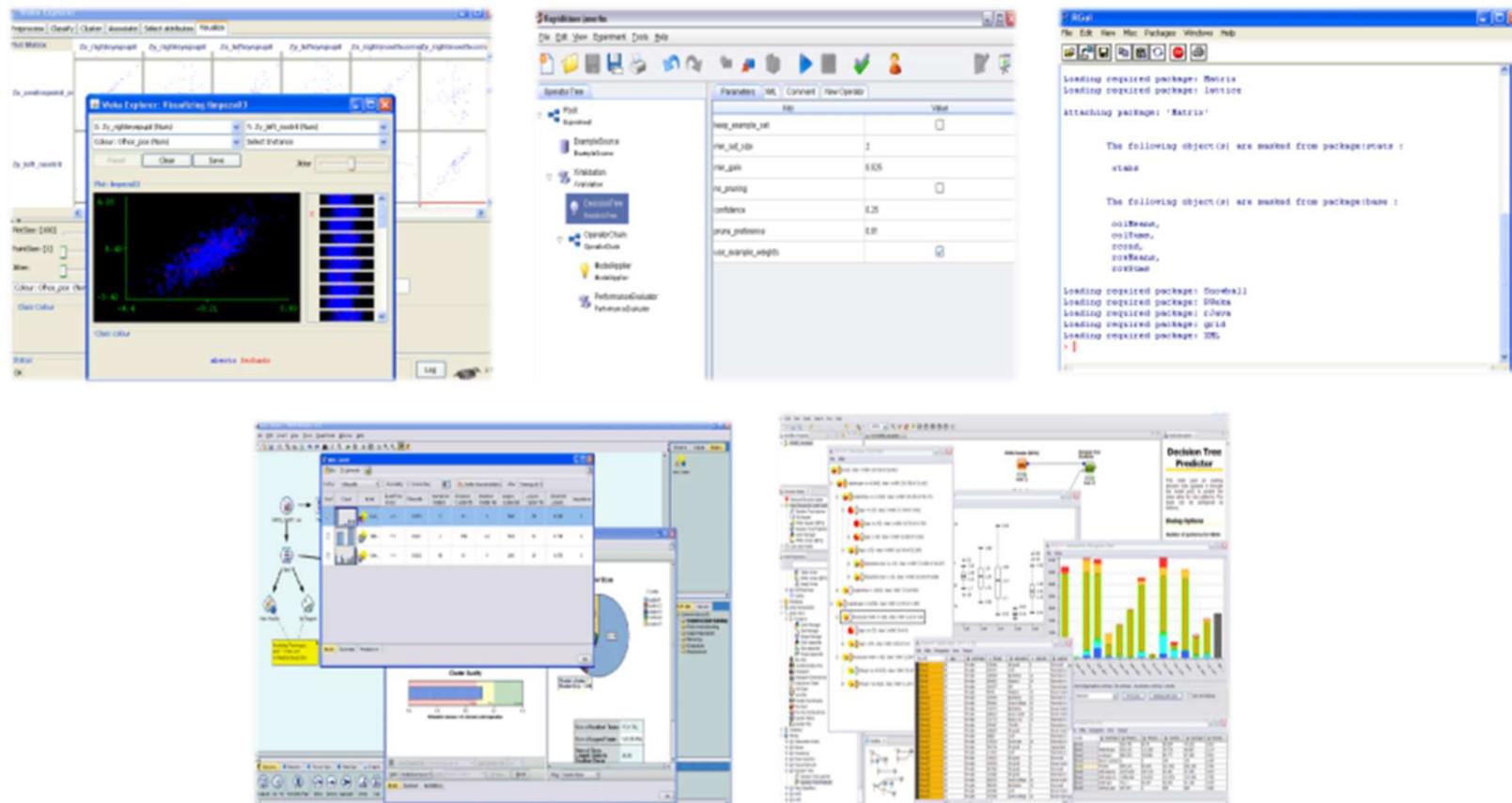
- **Métodos para Comparação de Modelos**

- Testes estatísticos
- Intervalos de Confiança
- Curva ROC (Receiver Operating Characteristic)

Pacotes de Software/Bibliotecas

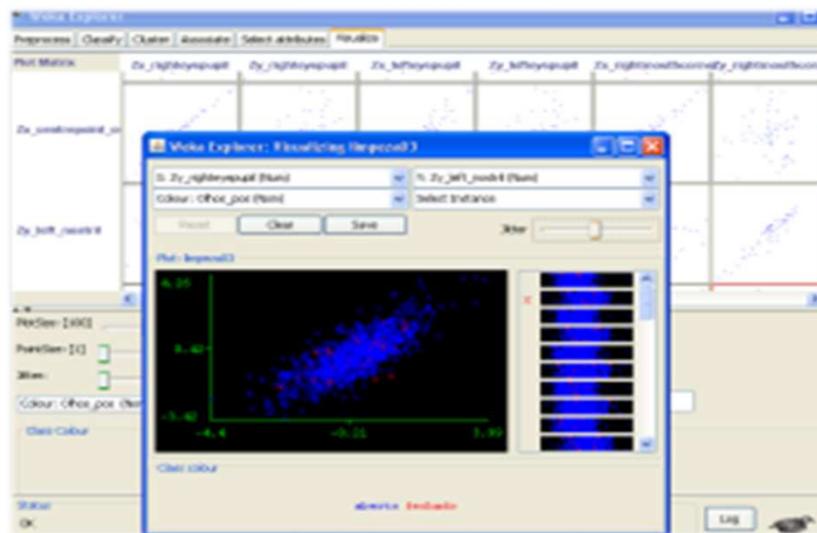
- Ferramentas/Linguagens/Bibliotecas

RapidMiner; WEKA; R; Python; SPSS; KNIME; SAS Enterprise Miner; Insightful Miner



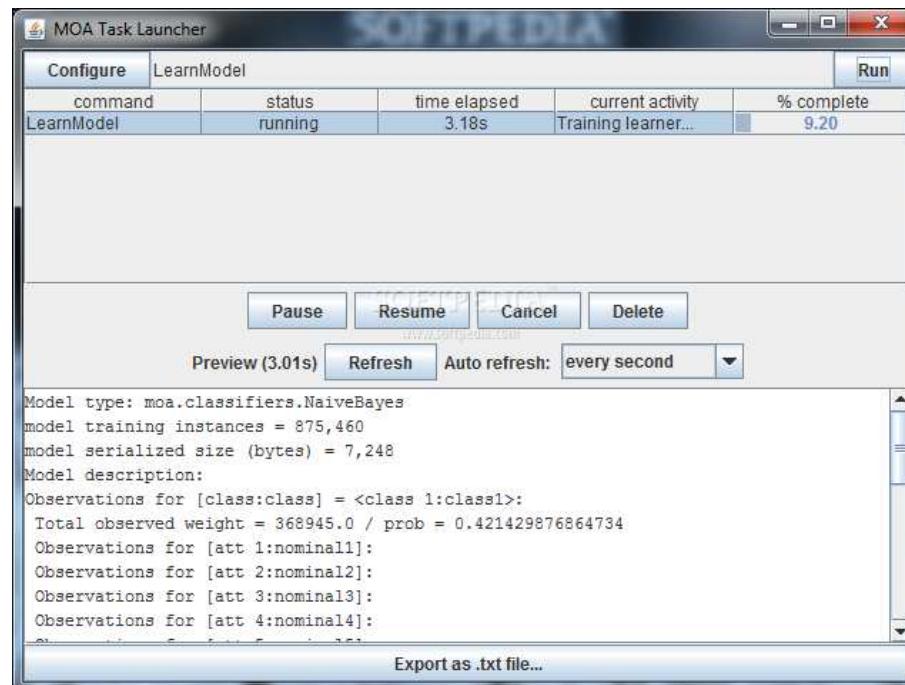
Pacotes de Software

- WEKA – Waikato Environment for Knowledge Analysis (<https://www.cs.waikato.ac.nz/ml/weka/>)
 - um dos mais antigos e um dos mais utilizados neste campo
 - é fácil de integrar outros produtos de software
 - para integrar diferentes processos no Weka é necessário reconstruir os dados



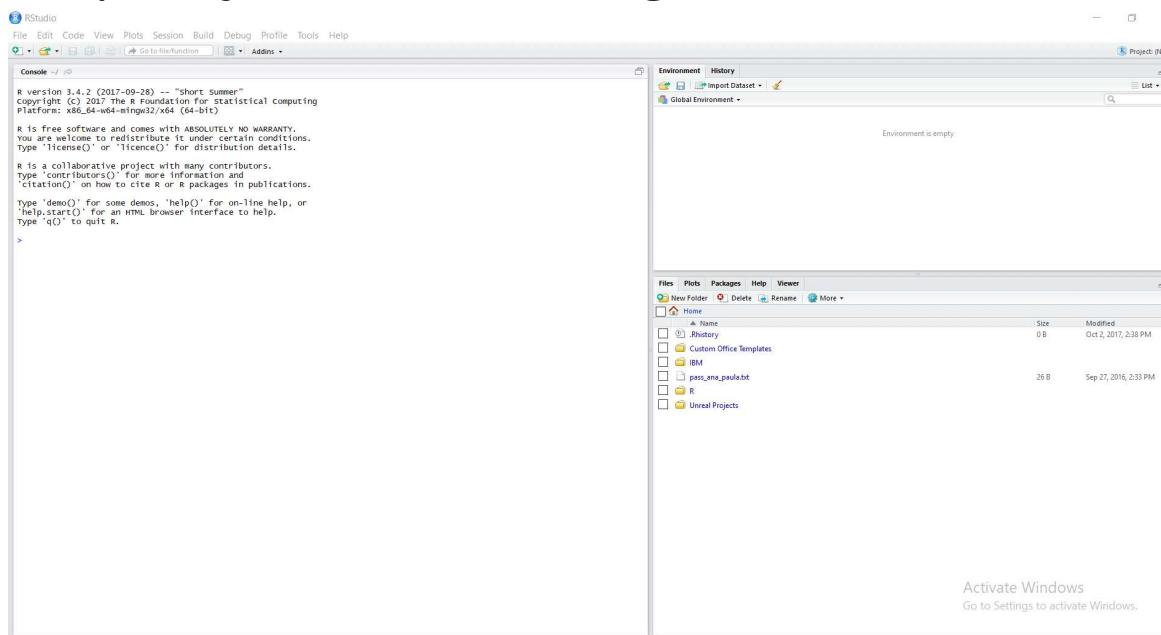
Pacotes de Software

- MOA – Massive Online Analysis
(<https://moa.cms.waikato.ac.nz/downloads/>)
 - Relacionado com o projeto WEKA
 - Open source para data stream mining



Pacotes de Software

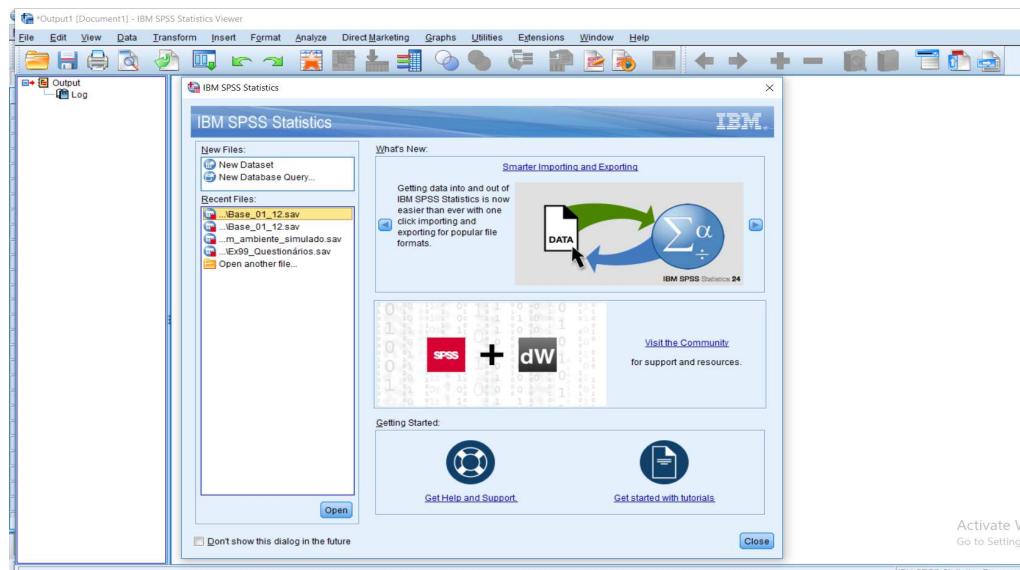
- R (<https://www.r-project.org/>)
- O R é um ambiente computacional e uma linguagem de programação que vem progressivamente se especializando na manipulação, análise e visualização gráfica de dados. um sistema de computação estatística e gráficos



Pacotes de Software

- Ferramentas

- IBM SPSS (<https://www.r-project.org/>)
 - software para análise estatística
 - módulos que permitem, entre outros, estatísticas avançadas, bootstrapping, tabelas personalizadas, preparação de dados, árvores de decisão, testes estatísticos, valores omissos, redes neurais e regressão



Pacotes de Software - Python



NumPy

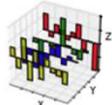


matplotlib

IP[y]: IPython
Interactive Computing

pandas

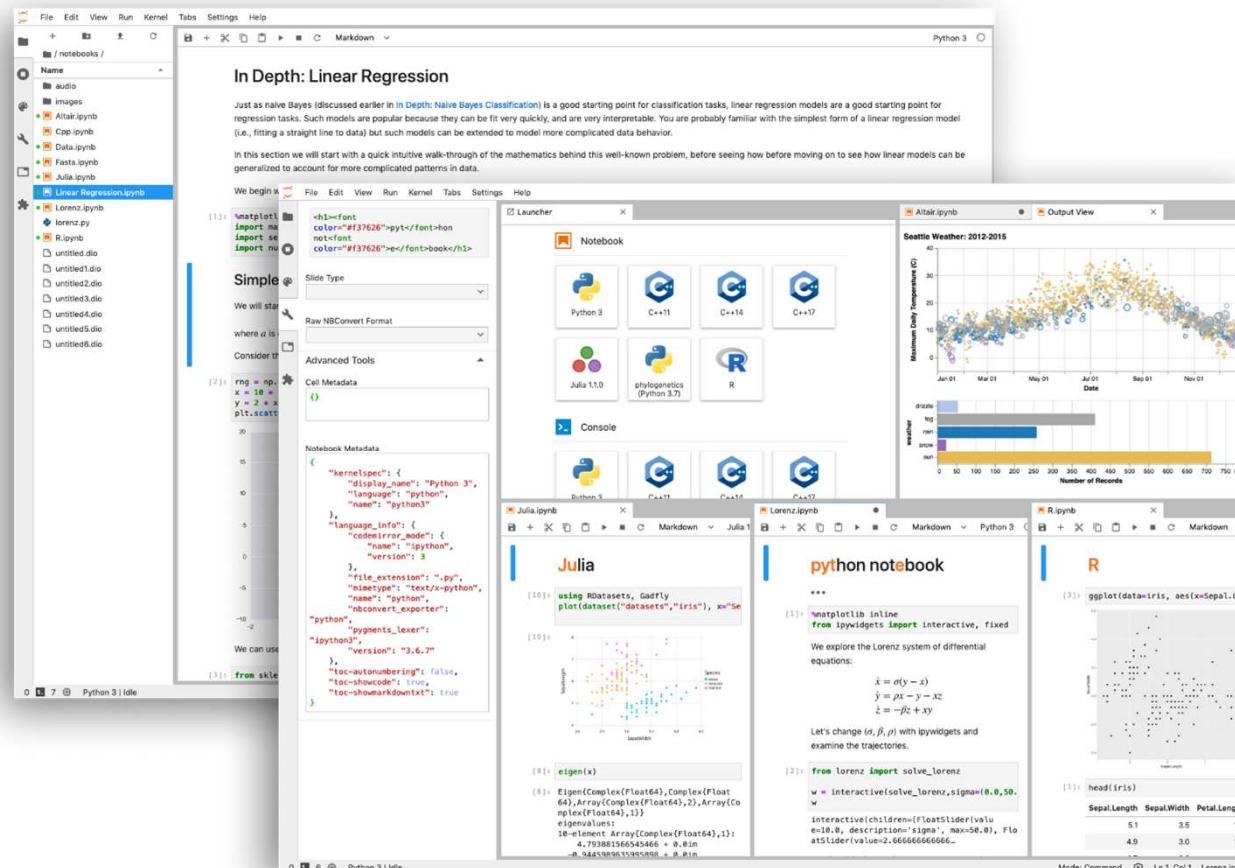
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Pacotes de Software - Jupyter

Jupyter Lab/Notebook - <https://jupyter.org/>

Very lightweight IDE that is a favorite among data analysts.



Python Libraries - NumPy

NumPy

NumPy is shortened from Numerical Python, it is the most universal and versatile library both for pros and beginners. Using this tool you are up to operate with multi-dimensional arrays and matrices with ease and comfort. Such functions like linear algebra operations and numerical conversions are also available.

Documentation: <https://docs.scipy.org/doc/numpy/user/>

Quick Start: <https://docs.scipy.org/doc/numpy/user/quickstart.html>



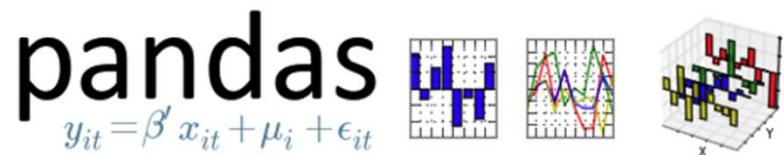
Python Libraries - Pandas

Pandas

Pandas is a well-known and high-performance tool for presenting data frames. Using it you can load data from almost any source, calculate various functions and create new parameters, build queries to data using aggregate functions akin to SQL. Various matrix transformation functions, a sliding window method and other methods for obtaining information from data. It is totally an indispensable thing in the arsenal of a good specialist.

Documentation: <https://pandas.pydata.org/pandas-docs/stable/>

Quick Start: https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html



Python Libraries – Matplotlib and Seaborn

Matplotlib

Matplotlib is a flexible library for creating graphs and visualization. It is powerful but somewhat heavy-weight.

Documentation: <https://matplotlib.org/contents.html>

Quick Start: <https://matplotlib.org/tutorials/introductory/pyplot.html>



Seaborn: statistical data visualization

<https://seaborn.pydata.org/>

Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics.

Python Libraries – Scikit-Learn

Scikit-Learn

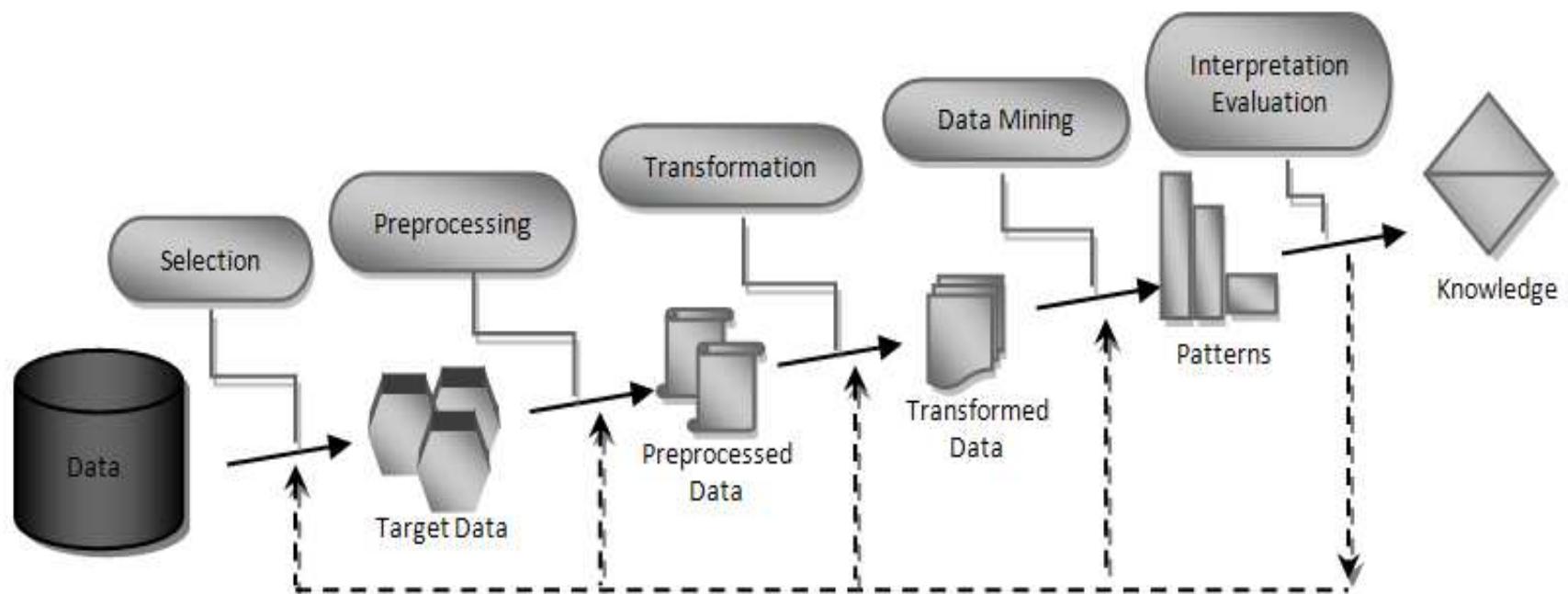
Very well-designed ML package that implements a wide-range of machine-learning algorithms and makes it comfortable to plug them into actual applications. Whole slew of functions like regression, clustering, model selection, preprocessing, classification and more. Very high speed of work. Used by leading platforms like Spotify, Booking.com, J.P.Morgan.

Documentation: <https://scikit-learn.org/stable/index.html>

Quick Start: <https://elitedatascience.com/python-machine-learning-tutorial-scikit-learn>



Métodos de Aprendizagem Supervisionada

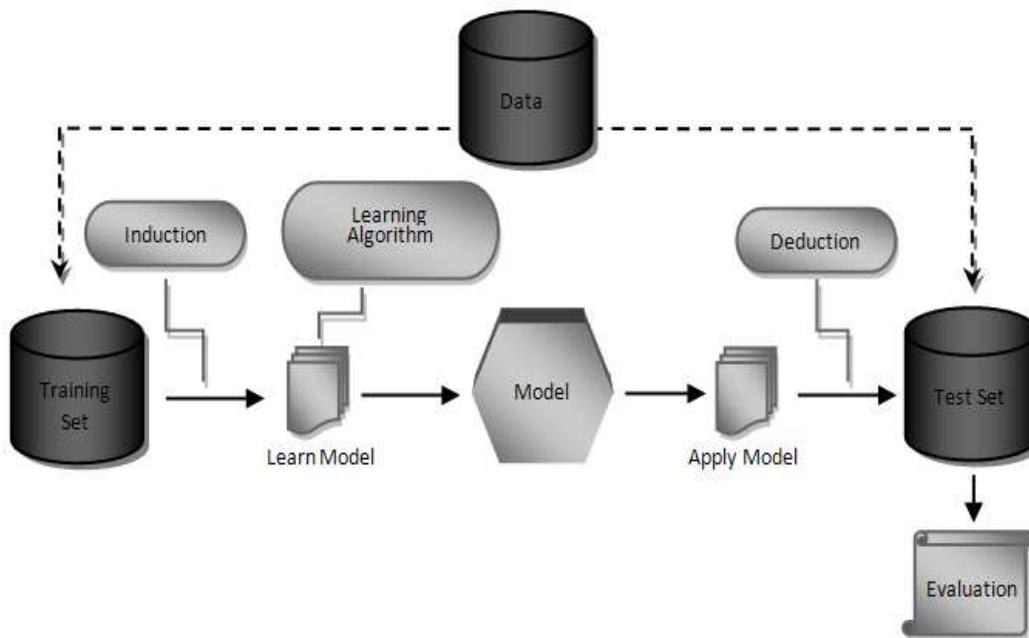


Fayyad, 1996

Métodos de Aprendizagem Supervisionada

- Induzir um classificador que poderá prever as classes de novos exemplos através do conjunto de treino com indivíduos previamente classificados

Abordagem geral para construir um modelo – **Fases de Treino e de Teste**



Métodos de Aprendizagem Supervisionada

Um **Conjunto de treino** com exemplos rotulados é usado para treinar o classificador. Cada objeto é caracterizado pelos seus atributos e a sua classe.

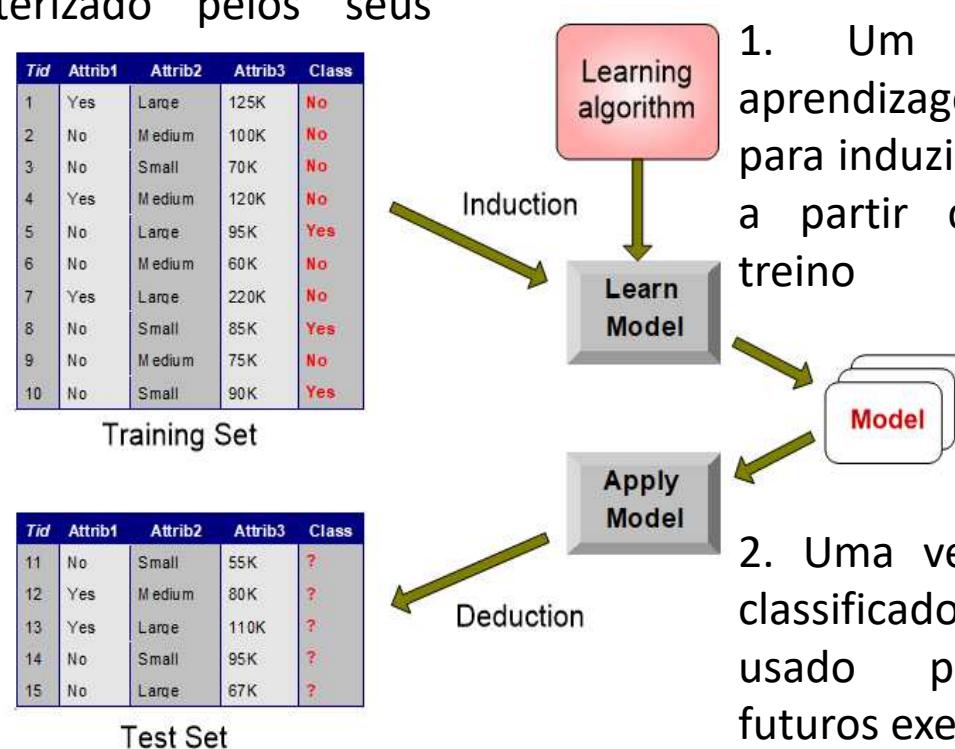
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Conjunto de teste com objetos sem classe



1. Um algoritmo de aprendizagem é executado para induzir um classificador a partir do conjunto de treino

2. Uma vez construído o classificador, ele poderá ser usado para classificar futuros exemplos

Avaliação de Modelos

- Métricas para a Avaliação do Desempenho

- Matriz de Confusão

- Taxa de acerto
- Precisão e Sensibilidade

- Medida F

		CLASSE PREVISTA	
CLASSE VERDADEIRA		Classe=Sim	Classe=Não
	Classe=Sim	a (VP)	b (FN)
	Classe=Não	c (FP)	d (VN)

Avaliação de Modelos

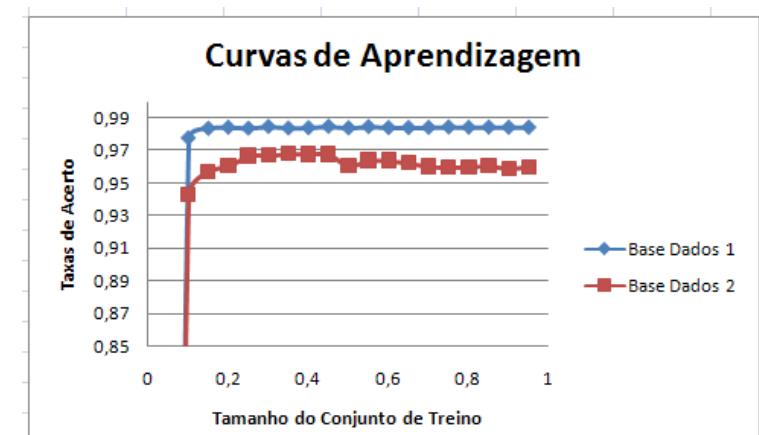
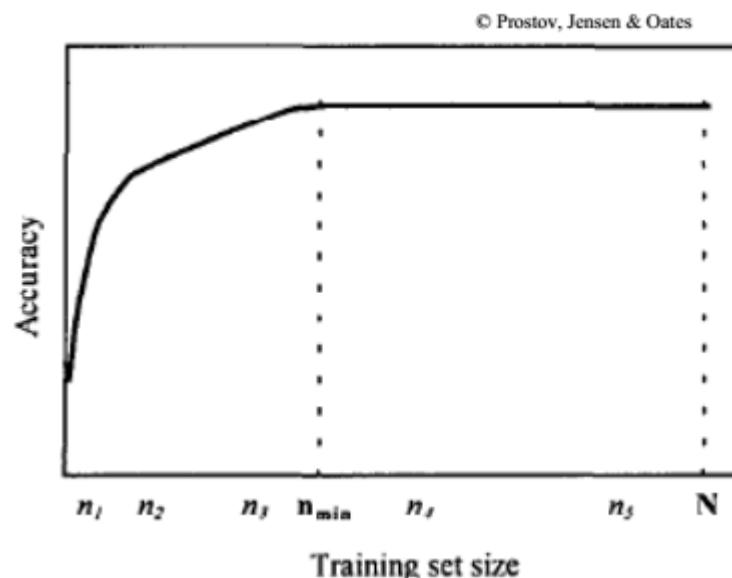
- Métodos para Avaliação de Desempenho
 - O desempenho de um modelo pode depender de outros fatores para além do algoritmo de aprendizagem
 - Distribuição da classe
 - Custo da classificação errada
 - Dimensão do conjunto de treino e de teste
 - Curva de Aprendizagem
 - Mostra a variação da taxa de acerto variando a dimensão do conjunto de treino
 - Métodos de estimação da taxa de erro:
 - Holdout; Cross-validation; Leave-one-out; Bootstrap

Avaliação de Modelos

- Métodos para Comparação de Modelos
 - Testes estatísticos
 - Intervalos de Confiança
 - Curva ROC (Receiver Operating Characteristic)

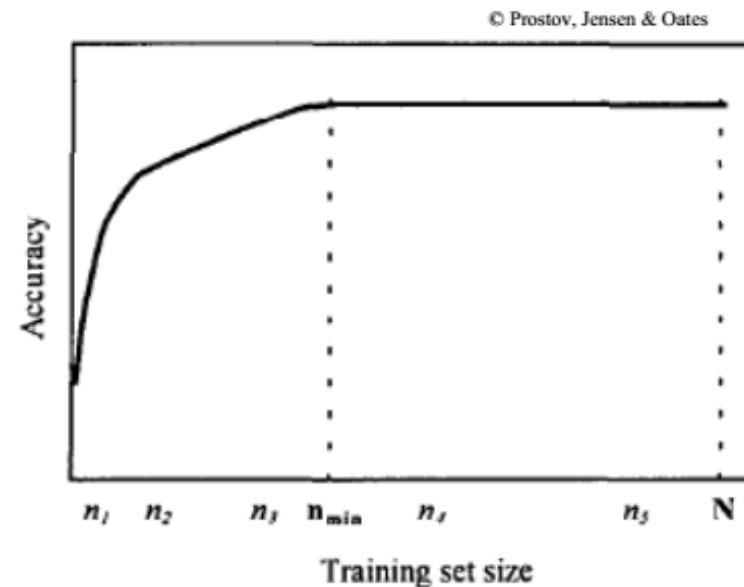
Avaliação de Modelos

- Curva de Aprendizagem - Mostra a relação entre tamanho do conjunto de treino e a taxa de acerto
 - **Eixo x:** número de exemplos do conjunto de treino (varia entre 0 e N – nº de exemplos disponível)
 - **Eixo y:** taxa de acertos do classificador induzido por um algoritmo de aprendizagem usando um conjunto de treino de tamanho n



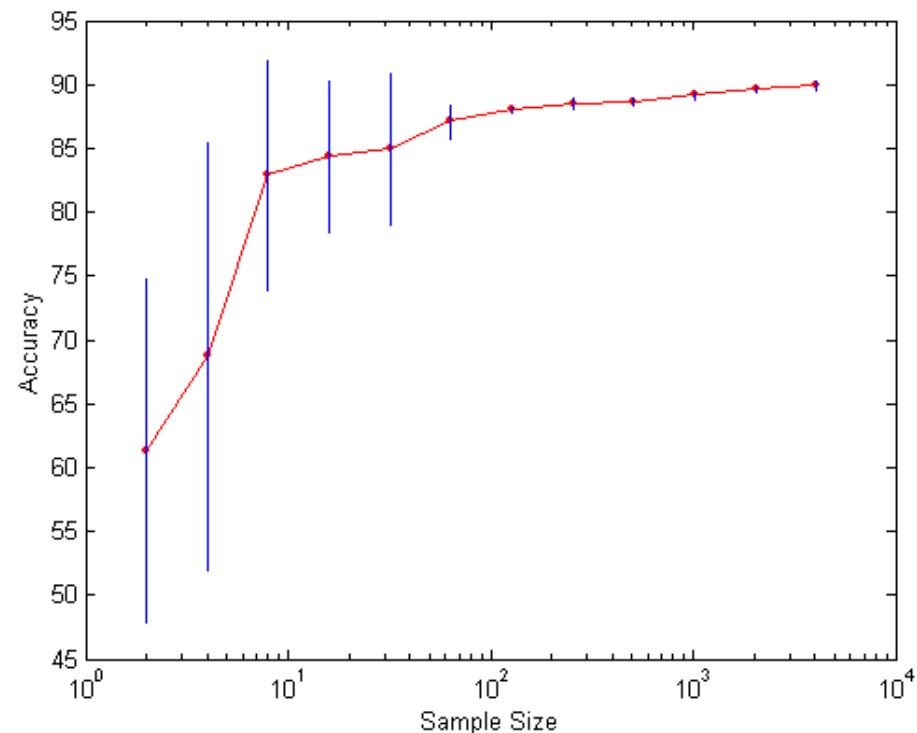
Avaliação de Modelos

- Curva de Aprendizagem
 - Uma curva de aprendizagem tipicamente está composta por 3 partes
 - Parte inicial (mais pequena) que mostra um crescimento muito rápido
 - Parte intermédia (maior) que mostra um crescimento desacelerado
 - Parte final que mostra um plateau \Rightarrow a adição de novos exemplos de treino não melhora mais o desempenho
 - Uma curva de aprendizagem converge quando atinge o seu plateau
 - n_{\min} representa o tamanho do conjunto de treino onde a curva converge



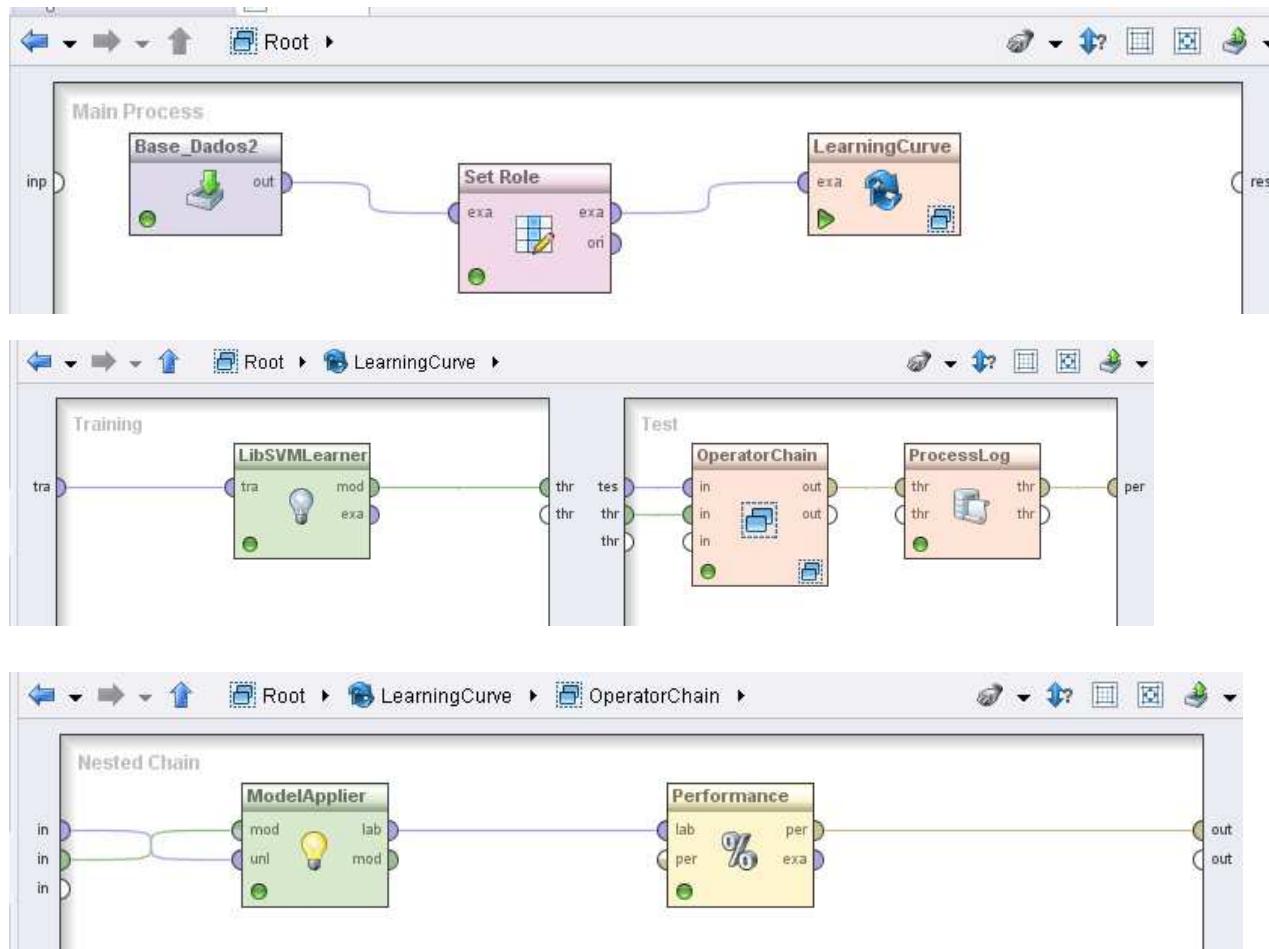
Avaliação de Modelos

- Curva de Aprendizagem
- Para construir uma curva de aprendizagem tem-se de definir uma estratégia de amostragem
 - Aritmética
 - Geométrica
- Se tamanho do conjunto de treino pequeno: maior variância
- À medida que aumenta tamanho do conjunto de treino \Rightarrow a variância diminui



Avaliação de Modelos

- Curva de Aprendizagem – No RapidMiner



Path: ver no repository Samples->processes->06_Meta

O operador [LearningCurve](#) é usado para construir uma curva de aprendizagem. Iterativamente divide o conjunto de dados em **treino e teste**, incrementando o tamanho do conjunto de treino segundo o valor indicado em [step_fraction](#) enquanto o conjunto de teste fica com tamanho fixo. O parâmetro "[training_ratio](#)" indica qual a percentagem máxima para o conjunto de treino ficando o resto como teste. Tomando como valor 0.7, o conjunto de teste vai sempre conter 30% dos exemplos, enquanto o conjunto de treino começa com o 5% (ver [step_fraction](#)) e vai aumentando o tamanho em 5% até atingir o tamanho máximo (70% dos exemplos).

Avaliação de Modelos

- Seja $\mathbf{X} = X_1 \times X_2 \times \dots \times X_n$ o espaço de atributos para um problema dado e $C = \{c_1, c_2, \dots, c_m\}$ um conjunto de classes definidas em X

Dado: um conjunto $\mathcal{D} = \{ \langle \mathbf{x}^{(1)}, c^{(1)} \rangle, \langle \mathbf{x}^{(2)}, c^{(2)} \rangle, \dots, \langle \mathbf{x}^{(N)}, c^{(N)} \rangle \}$ com N exemplos previamente classificados tal que $\mathbf{x}^{(k)} \in \mathbf{X}, c^{(k)} \in C, k=1\dots N$

Induzir: um classificador (hipótese) $h : \mathbf{X} \rightarrow C$ capaz de predizer o melhor possível a classe de futuros exemplos de \mathbf{X}

classificação aprendizagem

\mathcal{D}	X_1	...	X_n	C
$\langle \mathbf{x}^{(1)}, c^{(1)} \rangle$	$x_1^{(1)}$		$x_n^{(1)}$	$c^{(1)}$
$\langle \mathbf{x}^{(2)}, c^{(2)} \rangle$	$x_1^{(2)}$		$x_n^{(2)}$	$c^{(2)}$
...
$\langle \mathbf{x}^{(N)}, c^{(N)} \rangle$	$x_1^{(N)}$		$x_n^{(N)}$	$c^{(N)}$

Entrada: exemplo sem classificar

$x_1 \quad x_2 \quad \dots \quad x_n$



Supervised Learning Algorithm

classificador Induzido

h



h

Saída: um classificador que aproxima o melhor possível a função alvo $f: \mathbf{X} \rightarrow C$ "oculta" nos dados

classificador Induzido

Saída: a classe de x



c

para cada futuro exemplo x a classe atribuída é $c = h(x) \approx f(x)$

Avaliação de Modelos

- *True Error* – Erro verdadeiro

O desempenho de um classificador é medido em termos da sua capacidade preditiva nos futuros exemplos

O erro verdadeiro de um classificador h é a probabilidade de classificar erradamente um exemplo selecionado aleatoriamente, i.e.,

$$Err_{true}(h) = P_{\mathbf{x} \in \mathbf{X}} [h(\mathbf{x}) \neq c], \quad \langle \mathbf{x}, c \rangle : \mathbf{x} \in \mathbf{X}, c = f(x)$$

- Se tivéssemos um nº ilimitado de exemplos
 - Poderíamos com exatidão determinar o valor do erro verdadeiro de cada classificador h e escolher aquele com menor erro
- Na prática: há poucos dados
 - Como estimar o erro verdadeiro usando um limitado conjunto de exemplos?

Avaliação de Modelos

- *Função de Perda 0-1*

Assumindo que a cada exemplo $x \in \mathbf{X}$ é atribuído uma classe verdadeira $c = f(x)$

- Um classificador h induzido de dados pode **falhar** ou **acertar** a classe c de x

Função de perda 0-1 de um classificado $h(x)$ induzido a partir dos dados D com relação a um exemplo $x \in \mathbf{X}$

zero-one loss function

$$\delta(\mathbf{x}, f(\mathbf{x}), h(\mathbf{x})) = \begin{cases} 1, & \text{se } f(\mathbf{x}) \neq h(\mathbf{x}) \quad \text{erro} \\ 0, & \text{se } f(\mathbf{x}) = h(\mathbf{x}) \quad \text{sucesso} \end{cases}$$

- A função 0-1 mede a **perda** ou **custo** de atribuir a um exemplo x a classe $c' = h(x)$ (**classe predita**) quando a sua **classe verdadeira** é $c = f(x)$
 - Se a classificação do exemplo é correta (a classe predita é igual à classe verdadeira) então **não existe perda**; caso contrário, existe perda de 1
- Podemos usar a função perda 0-1 para redefinir a aprendizagem supervisionada

Dado: um conjunto de treino D com N exemplos previamente classificados

Induzir: um classificador $h: \mathbf{X} \rightarrow \mathbf{C}$ que **minimize** a **função de perda 0-1**

Avaliação de Modelos

- *Taxa de Erro* – medida de desempenho

A taxa de erro de um classificado $h(\mathbf{x})$ num conjunto de N exemplos

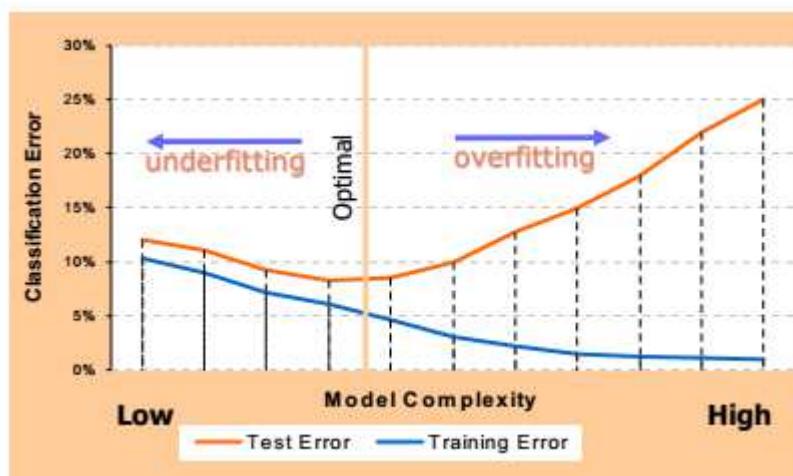
$$Err(h(\mathbf{x}), D) = \frac{1}{N} \sum_{\langle \mathbf{x}^{(k)}, c^{(k)} \rangle \in D} \delta(\mathbf{x}^{(k)}, c^{(k)}, h(\mathbf{x}^{(k)})) = \frac{\text{\#exemplos incorrectamente classificados}}{\text{total exemplos}}$$

Assim

- A **taxa de erro** de um classificador relativamente a um conjunto de exemplos é definida como a proporção dos exemplos incorretamente classificados
- A **taxa de acerto** (accuracy) é a proporção dos exemplos corretamente classificados

Avaliação de Modelos

- *Estimação da Taxa de Erro*
- Método trivial: Dado o conjunto de dados disponível D
 - Usar D para induzir o classificador (fase de treino)
 - Usar D para estimar a taxa de erro (fase de avaliação)
- Podem ocorrer dois problemas
 - O modelo induzido sobre ajusta-se (overfit) ao conjunto de treino (especialmente se o modelo é muito complexo com muitos parâmetros)
 - A taxa de erro obtida no conjunto de treino é um estimador otimista do erro verdadeiro (obtém-se por vezes um valor muito menor do que o erro verdadeiro)



- Se um classificador é muito simples pode fazer **underfit**
- Se é muito complexo pode fazer **overfit**
- Existe sempre um classificador de complexidade ótima que permite obter o melhor desempenho (menor taxa de erro num conjunto de exemplos que não foram usados para construir o classificador)

Avaliação de Modelos

- *Estimação da Taxa de Erro*
- Dada uma quantidade de exemplos limitada, qual a melhor estratégia que devemos seguir por forma a obter uma estimativa fiável do desempenho?
- Um algoritmo de aprendizagem deve ser avaliado tendo em conta o seu desempenho (capacidade de generalização) naqueles **exemplos que não foram usados para construir o classificador**

Ideia básica: Particionar o conjunto de dados disponível em dois conjuntos

- **conjunto de treino**
 - exemplos que são usados pelo algoritmo de aprendizagem para induzir o classificador
- **conjunto de teste**
 - exemplos que são usados para estimar a taxa de erro

Avaliação de Modelos

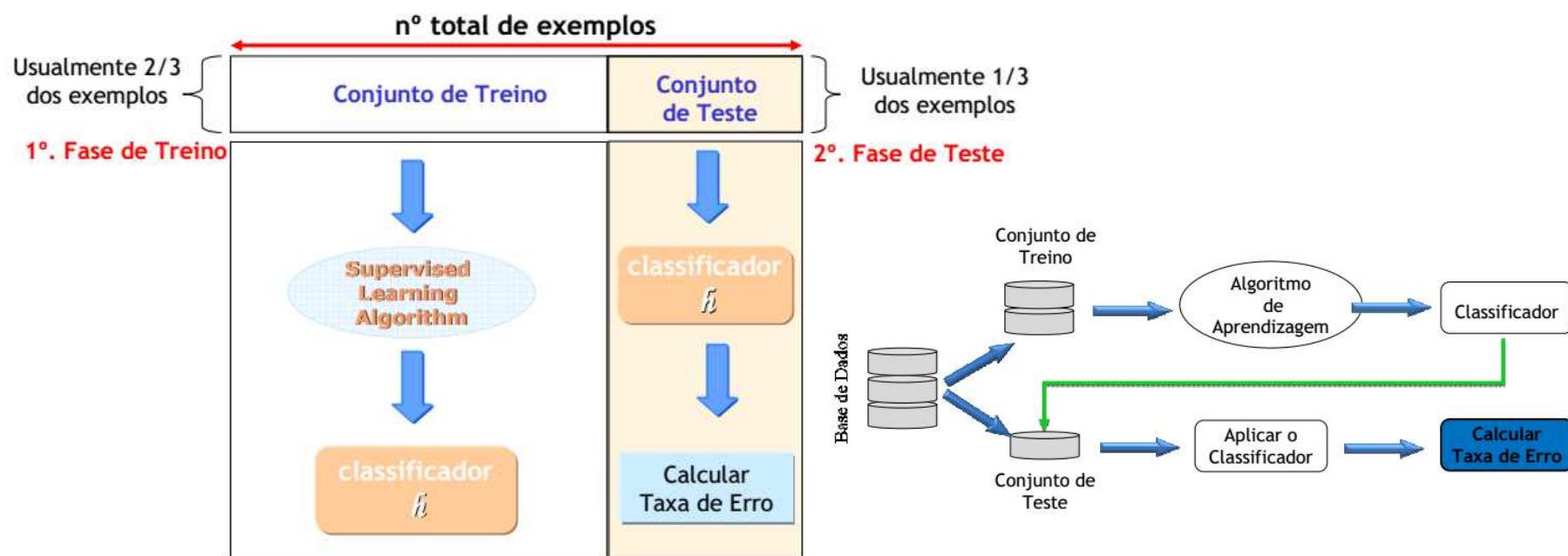
- *Estimação da Taxa de Erro* - Existem vários métodos para estimar a taxa de erro baseados em diferentes partições do conjunto de dados
- **Holdout** (se houver muitos exemplos, > 1000 exemplos)
 - Dividir os dados 2/3 para treino e 1/3 para teste

Métodos de Reamostragem:

- **Cross validation** (para conjuntos de tamanho intermédio, aprox. 1000 exemplos)
 - Dividir os dados em k partições disjuntas
 - k-fold: k-1 partições para treino, 1 para teste
 - Leave-one-out: k=N (nº de exemplos)
- **Bootstrap** (para conjunto pequenos, < 500 exemplos)
 - Amostragem com reposição

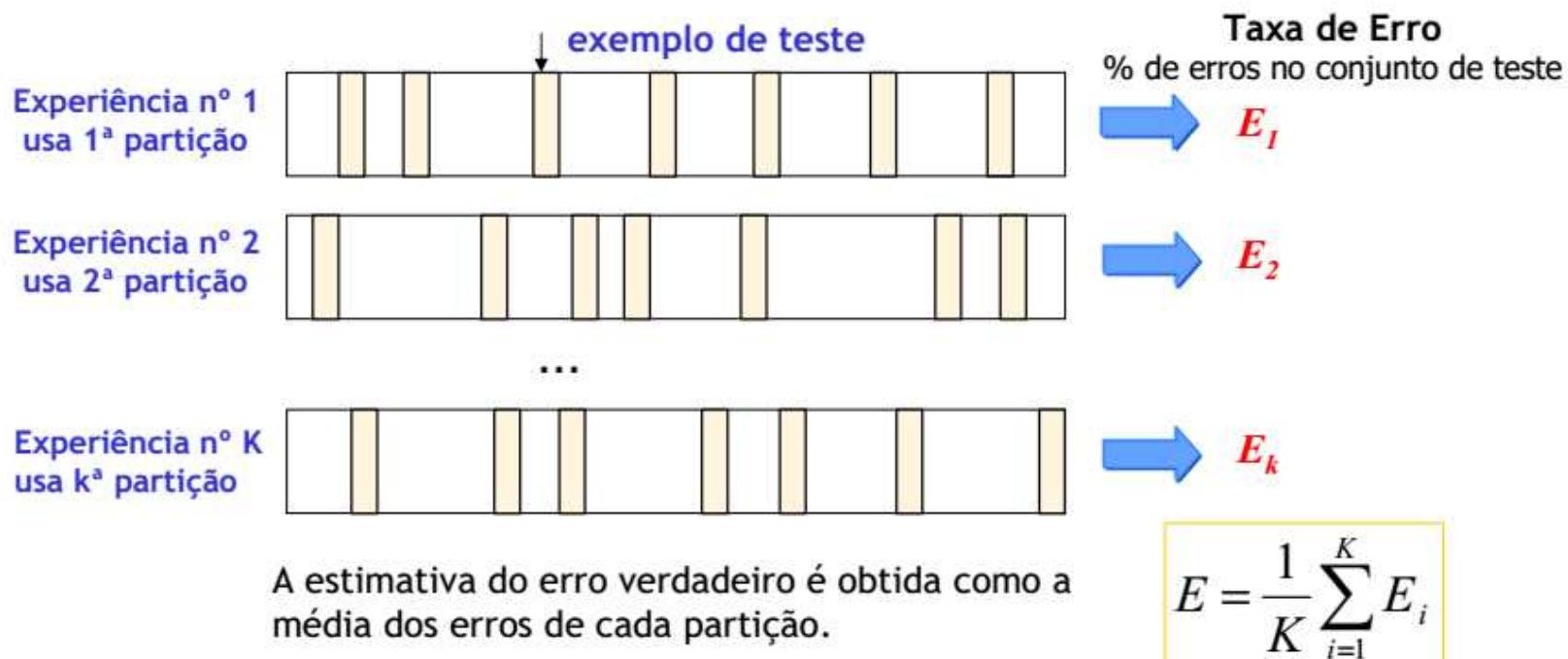
Avaliação de Modelos

- Método Holdout
- Dividir o conjunto de dados em dois
 - Um conjunto de treino para induzir o classificador
 - Um conjunto de teste para avaliar o desempenho (ex. calcular a taxa de erro, acerto,...)



Avaliação de Modelos

- Métodos de Reamostragem (aleatória)
- Executa k experiências, uma por cada partição sobre o conjunto de dados
 - Em cada partição é selecionado aleatoriamente um número (fixo) de exemplos de teste
 - O classificador é induzido dos exemplos de treino e avaliado nos exemplos de



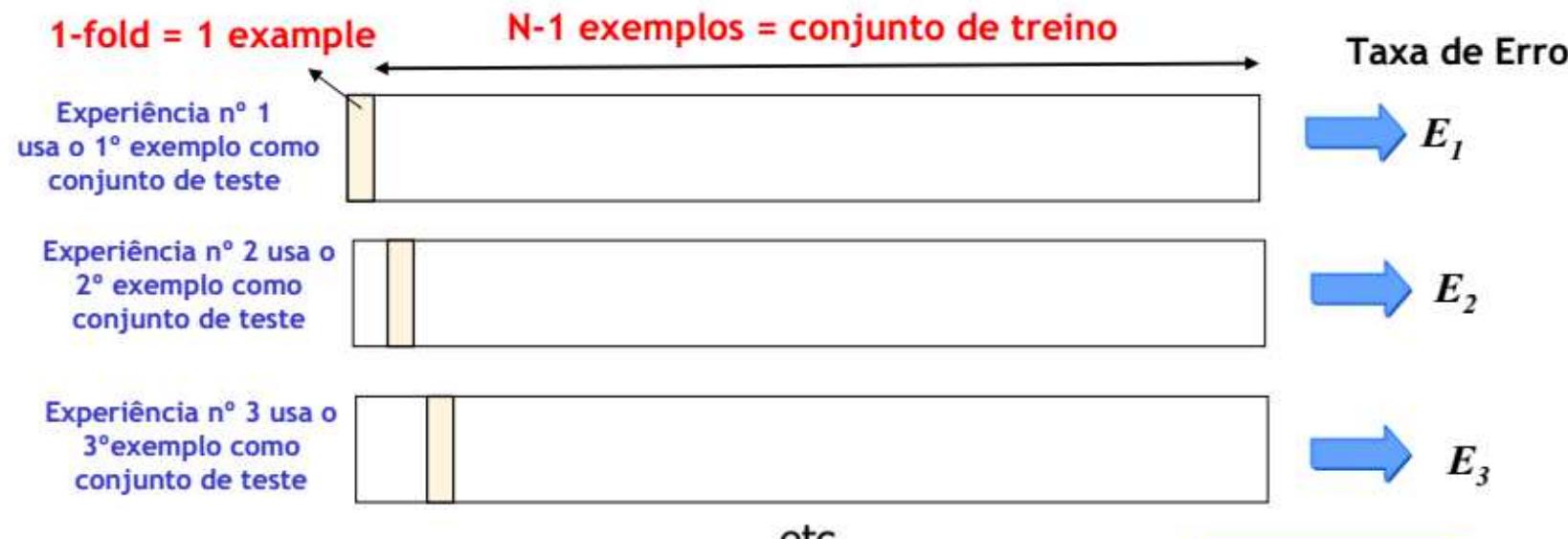
Avaliação de Modelos

- Método k-Fold Cross Validation
- Este é um dos métodos de validação mais populares em ML
 - O conjunto de dados é dividido em k-folds (k-partições)
 - Em cada experiência (run) usa: k-1 folds como conjunto de treino e 1-fold (o que sobra) como conjunto de teste



Avaliação de Modelos

- Método Leave-one-out
- Caso especial de k-fold cross validation quando k=N
 - Para um conjunto de dados com N exemplos executa N experiências (runs)
 - Para cada experiência (run): usa N-1 exemplos como conjunto de treino e apenas 1 exemplo (o que sobra) como conjunto de teste



A estimativa do erro verdadeiro é obtida como a média dos erros de cada experiência

$$E = \frac{1}{N} \sum_{i=1}^N E_i$$

Avaliação de Modelos

- Quantas partições (folds) escolher?
- Maior número de folds
 - Mais exata a estimativa do erro verdadeiro, menor desvio (bias)
 - Mais número de runs ⇒ maior tempo de computação
- Menor número de folds
 - Menos exata a estimativa do erro verdadeiro, maior desvio
 - Menos número de runs ⇒ menor tempo de computação
- Na prática: a escolha de k depende de N (nº de exemplos)
 - se tamanho muito grande ⇒ com 3-fold cross-validation podemos obter uma estimativa precisa
 - se dados muito esparsos ⇒ usar one-leave-out para poder obter o maior número possível de exemplos de treino
 - Como regra K=10

Avaliação de Modelos

- Bootstrapping

- Método de estimação baseado em re-amostragem com reposição e é usado quando dispomos de poucos exemplos
- Dado um conjunto de dados **D** com **N** exemplos é gerado um número **B de amostras** (bootstraps) de tamanho **N**
 - cada amostra é gerada usando amostragem com reposição
 - cada vez que um exemplo é adicionado aleatoriamente este é logo reposto
⇒ alguns exemplos podem aparecer mais do que uma vez, enquanto outros podem nunca aparecer
- Em cada experiência: (no total são efetuadas **B** experiências)
 - uma amostra bootstrap é gerada e usada como conjunto de treino
 - os exemplos do conjunto **D** que não pertencem à amostra são usados como conjunto de teste, e é obtida uma estimativa da taxa de erro
- A estimativa da taxa de erro é a média das taxas de erros obtidas por cada uma das amostras

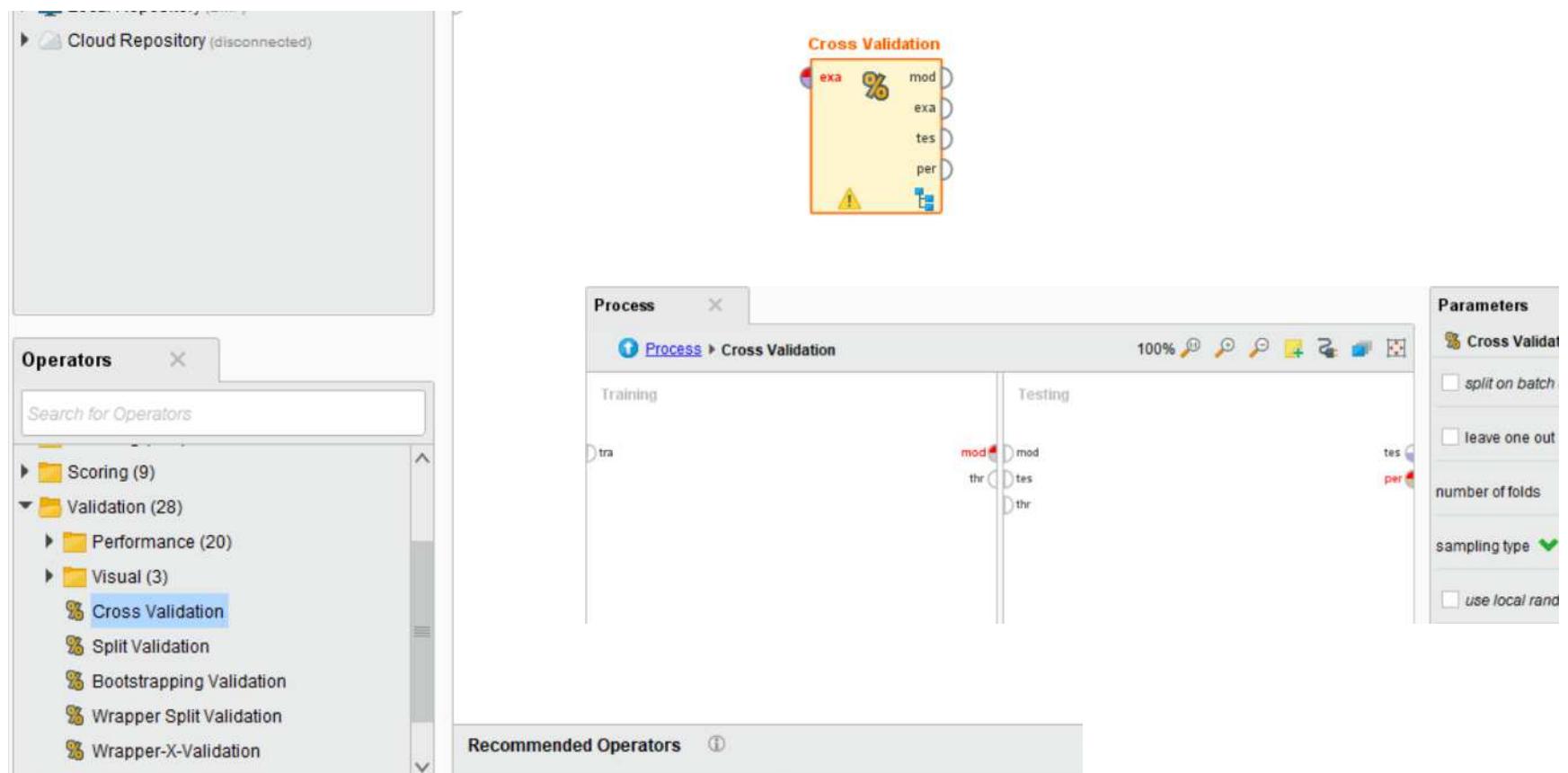
Avaliação de Modelos

- Bootstrap 0.632
- Um exemplo tem uma probabilidade de $1 - 1/N$ de não ser selecionado
 - a probabilidade que fique no conjunto de teste é de $(1-1/N)^N \approx e^{-1} = 0.368$
 - o conjunto de treino vai conter aproximadamente 63.2% dos exemplos de D
 - taxa de erro E_{teste} é um estimador muito pessimista (usa 36.8% dos exemplos)
 - ⇒ solução: usar também a taxa do erro E_{treino} obtida no conjunto de treino

	$\approx 63.2\% \text{ dos exemplos de } D$	$\approx 37\% \text{ de } D$	Estimativas da Taxa de Erro
Experiência n° 1 1º bootstrap sample	Conjunto de Treino: D_1	Conjunto de Teste: $D \setminus D_1$	$E_1 = 0.632 E_{\text{teste}} + 0.368 E_{\text{treino}}$
Experiência n° 2 2º bootstrap sample	Conjunto de Treino: D_2	Conjunto de Teste: $D \setminus D_2$	$E_2 = 0.632 E_{\text{teste}} + 0.368 E_{\text{treino}}$
	..., etc.		
Experiência n° B Bº bootstrap sample	Conjunto de Treino: D_B	Conjunto de Teste: $D \setminus D_B$	$E_B = 0.632 E_{\text{teste}} + 0.368 E_{\text{treino}}$
	A estimativa do erro verdadeiro é obtida como a média dos erros de cada experiência		$E = \frac{1}{B} \sum_{i=1}^B E_i$

Avaliação de Modelos

- Estimação da taxa de erro – **No RapidMiner**
 - Path: Validation – Visual



Avaliação de Modelos

- Comparação dos Métodos de Avaliação
 - Holdout: para N grande
 - Reamostragem Aleatória: melhora a estimativa de holdout, mas não existe controlo sobre os exemplos usados para treino e para teste
 - K-fold Cross Validation: para N intermédia
 - Estimação imparcial do erro verdadeiro, mas com elevada variância
 - 0.632 Bootstrapping: para N pequena
 - Estimação imparcial no limite e com pouca variância

Avaliação de Modelos

- Outros fatores que afetam desempenho
 - O desempenho de um classificador não depende apenas do algoritmo de aprendizagem; este depende também de outros fatores
 - A distribuição da classe
 - Disparidade do conjunto de dados
 - Custo associado ao facto de ter classificado erradamente um exemplo
 - Dimensão de conjunto de treino e de teste

Avaliação de Modelos

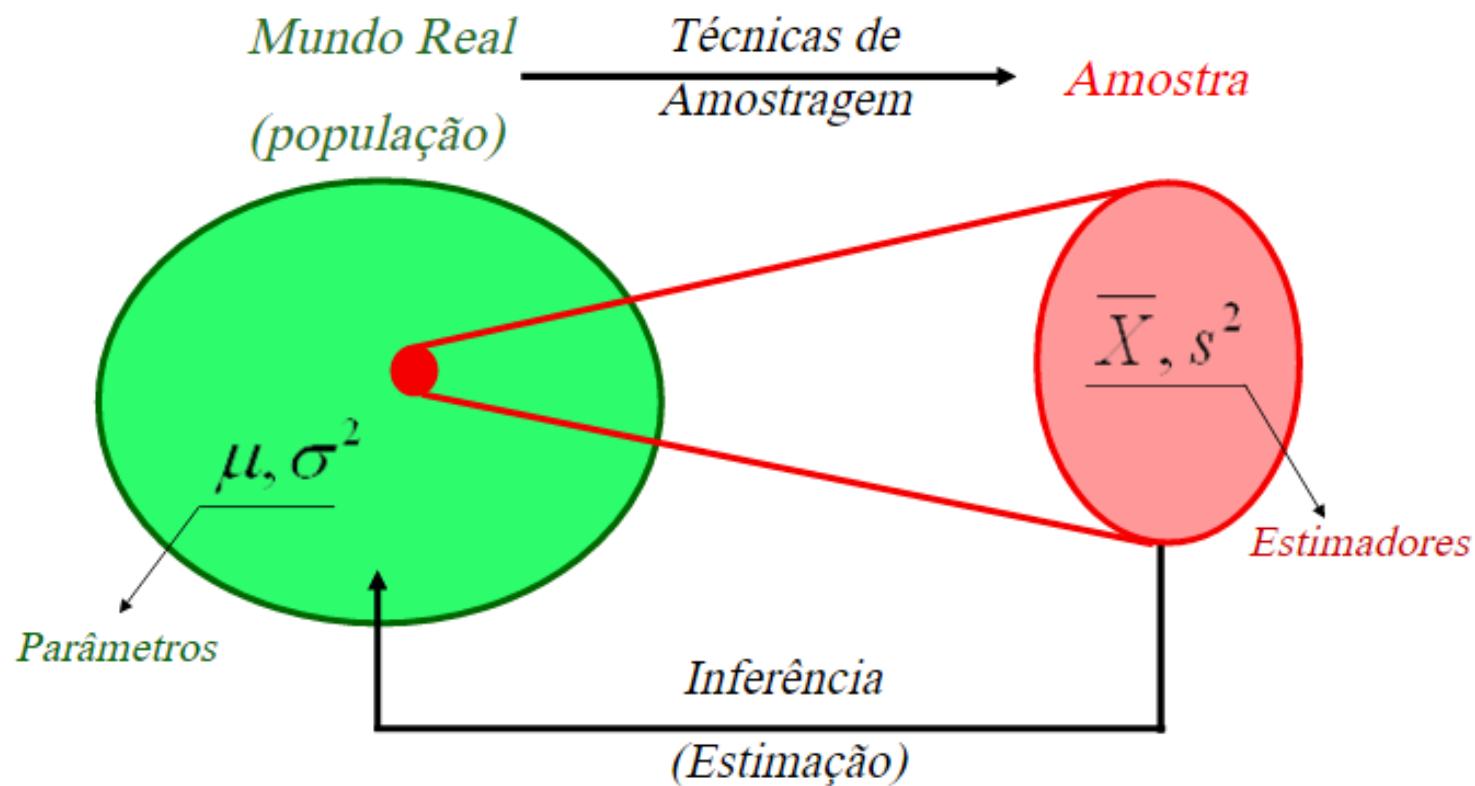
- Outras Medidas de Avaliação
- Velocidade
 - Tempo de construção do modelo (tempo de treino)
 - Tempo de utilização do modelo (tempo de predição/classificação)
- Ser robusto
 - Tratar do ruído e valores omissos
- Escalabilidade
 - Eficiência a lidar com um número crescente de dados
- Interpretação
 - Compreensão produzida pelo modelo
- Outras medidas
 - Qualidade de ajuste, tamanho das árvores de decisão, regras de classificação compactas

Avaliação de Modelos

- Avaliação de Algoritmos de Aprendizagem
- Dois tipos de abordagens
 - Dados um algoritmo e um conjunto de dados
 - Quanta confiança podemos ter na taxa de erro/taxa de acerto estimada?
 - Calcular Intervalos de Confiança
 - Dados dois algoritmos e um conjunto de dados
 - Qual o algoritmo que tem melhor desempenho (capacidade de generalização)
 - Realizar Testes de Hipóteses

Avaliação de Modelos

- Estimação de Parâmetros



Avaliação de Modelos

- Estimação de Parâmetros
 - Pontual (estatísticas)
 - Por Intervalo (intervalos de confiança)
- **estatística**: função de uma amostra aleatória que não depende de parâmetros desconhecidos
- **estimador**: estatística que estima (pontualmente) um parâmetro populacional
- **estimativa pontual**: valor obtido por um estimador para uma amostra específica

Avaliação de Modelos

- Parâmetro vs Estatística

- Parâmetro – medida usada para descrever a distribuição da população
 - A média μ e o desvio padrão σ são parâmetros de uma distribuição normal $[N(\mu, \sigma)]$
 - A probabilidade de sucesso p é um parâmetro da distribuição Binomial $[B(n,p)]$
- Estatística – função de uma amostra aleatória que não depende de parâmetros desconhecidos

$$\text{Média amostral: } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{Variância amostral: } V(X) = S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

- Exemplo: se para uma amostra se observar uma média de 6.2 então esse valor é uma estimativa pontual da média da população μ

Avaliação de Modelos

- Intervalo de confiança

Um intervalo de confiança para um parâmetro θ a um grau de confiança $1-\alpha$ é uma concretização de um intervalo aleatório $[L_{\text{inf}}, L_{\text{sup}}]$ para o qual se tem

$$P(L_{\text{inf}} < \theta < L_{\text{sup}}) = 1 - \alpha, \alpha \in]0,1[$$

Onde α deve ser um valor reduzido de forma a ter confianças elevadas

Valores usuais para o grau de confiança: 95%, 99% e 90%

Avaliação de Modelos

- Taxa de Acerto – Distribuição Binomial

- Uma prova de Bernoulli associa uma variável aleatória com 2 resultados possíveis: sucesso ou insucesso \Rightarrow resultados para classificação: correto ou incorreto
- Uma sequência (soma) de n provas de Bernoulli tem distribuição Binomial:
 - $X \sim \text{Binomial}(n, p)$; X - número de sucessos; p – probabilidade de sucesso
 - Ex: X - nº de caras saídas em 50 lançamentos de uma moeda
valor esperado de sair “cara”: $E(X) = n \times p = 50 \times 0.5 = 25$
- $X \sim \text{Binomial}(N, p)$
 - **X - # de classificações corretas, N - # de exemplos de teste**
 - **p – probabilidade de classificar corretamente um exemplo**
- **Estimar:** p – taxa de acerto verdadeira (para toda a população)

A taxa de acerto, $acc = X/N$, é um estimador pontual de p

Avaliação de Modelos

- Taxa de Acerto – Intervalo de Confiança (amostras grandes)
- Para conjunto de teste com $N > 30$, a taxa de acerto pode ser aproximada, pelo Teorema do Limite Central, a uma distribuição Normal de média p e variância $p(1-p)/N$
- Intervalo de confiança (IC) para a taxa de acerto verdadeira p (desconhecida)

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{acc} - p}{\sqrt{p(1-p)/N}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

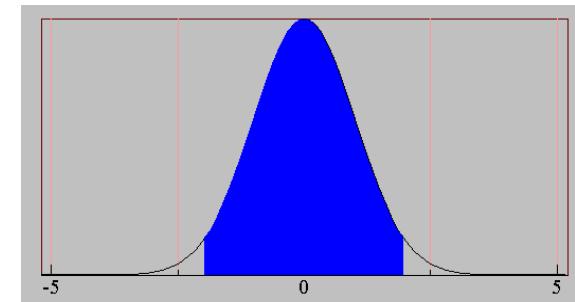
- Um IC aproximado para p , a um grau de confiança $1 - \alpha$

$$IC_{(1-\alpha)}(p) \approx [\hat{acc} - z_{\alpha/2} s_p, \hat{acc} + z_{\alpha/2} s_p]$$

onde

$$\hat{acc} = \frac{x}{N} \quad s_p = \sqrt{\frac{\hat{acc}(1-\hat{acc})}{N}}$$

usamos \hat{acc} como uma estimativa pontual de p para calcular o desvio padrão amostral e a letra minúscula x pois estamos a representar uma concretização da v.a. X



Avaliação de Modelos

- Taxa de Acerto – Intervalo de Confiança (amostras grandes)
- Exemplo: Dado um classificador com taxa de acerto 80% num conjunto de teste com 100 exemplos
 - $N=100$; $\hat{acc} = 0.8$
 - Para $1-\alpha = 0.95$ então $z_{\alpha/2}=1.96$
- Um IC aproximado para a taxa de acerto a 95%

$$\begin{aligned}IC_{(95\%)}(p) &\approx [\hat{acc} - z_{\alpha/2} s_p, \hat{acc} + z_{\alpha/2} s_p] \\&= [0.8 - 1.96 \times 0.04, 0.8 + 1.96 \times 0.04] \\&= [0.7216, 0.8784]\end{aligned}$$

$1-\alpha$	z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

A taxa de erro verdadeira situa-se entre 72.16% e 87.84%

sendo

$$s_p = \sqrt{\frac{\hat{acc}(1-\hat{acc})}{N}} = \sqrt{\frac{0.8 \times 0.2}{100}} = 0.04$$

Avaliação de Modelos

- Taxa de Acerto – Intervalo de Confiança (amostras grandes)
- Exemplo: Dado um classificador com taxa de acerto 80% num conjunto de teste com 100 exemplos
 - $N=100$; $\hat{acc} = 0.8$
 - Para $1 - \alpha = 0.95$ então $z_{\alpha/2} = 1.96$
- Usando a fórmula para determinar o intervalo de confiança obtemos os seguintes limites variando o nº de exemplos

N	50	100	500	1000	5000
p(lower)	0.689	0.722	0.765	0.775	0.789
p(upper)	0.911	0.878	0.835	0.825	0.811

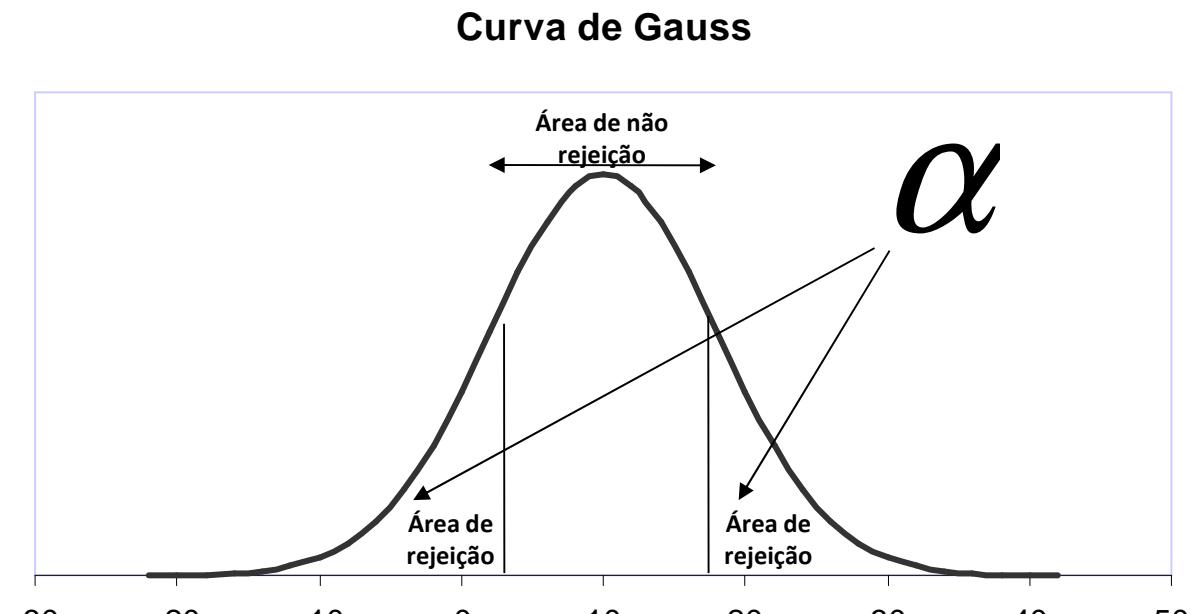
A amplitude do intervalo de confiança vai diminuindo à medida que aumentamos o nº de exemplos de teste

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- Procedimento estatístico que permite averiguar hipóteses
 - 2 hipóteses
 - Hipótese nula (H_0) - (usar sempre o sinal =)
 - Hipótese alternativa (H_1)
 - 2 tipos de testes
 - Unilateral
 - Bilateral
 - 2 tipos de decisão
 - Rejeitar a hipótese nula H_0
 - Não rejeitar a hipótese nula H_0

Avaliação de Modelos

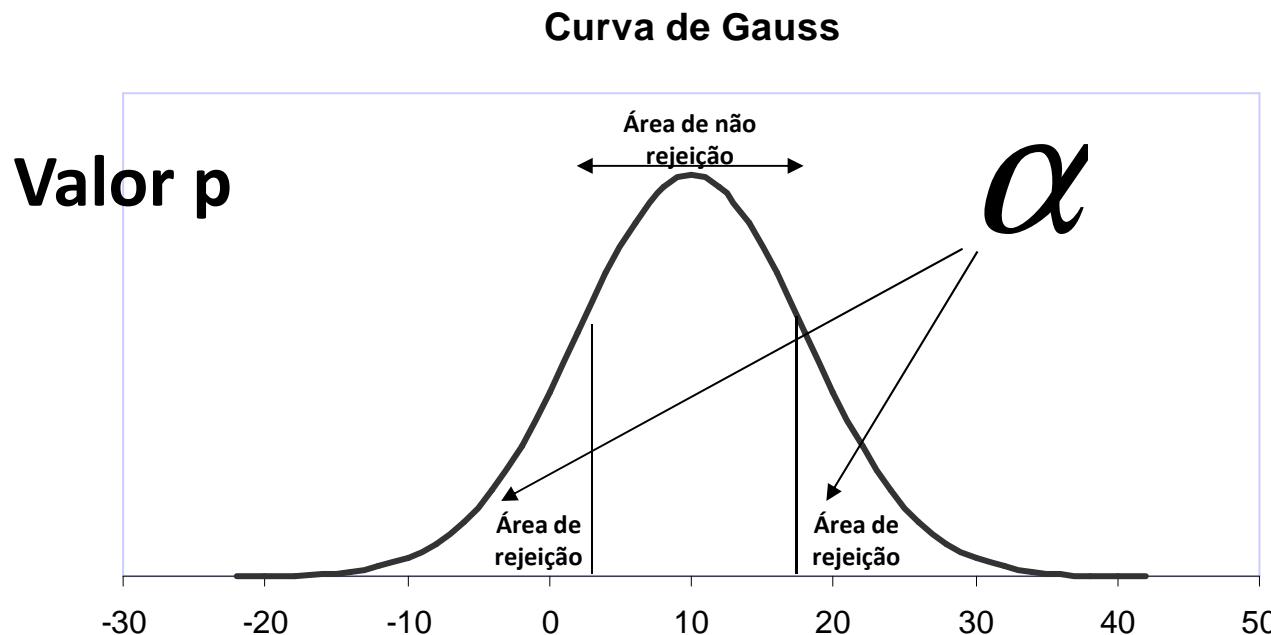
- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)



Probabilidade de ocorrência de um erro do tipo I
= $P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira})$

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)



Probabilidade de obter um resultado tão extremo ou mais do que o observado, assumindo a hipótese nula verdadeira = Valor p

Portanto se Valor p > α não se rejeita H_0

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)

Passos num teste de hipóteses

- 1º passo: Definição de hipóteses
- 2º passo: Indicação do nível de significância
- 3º passo: Escolha do teste a usar
- 4º passo: Valor observado
- 5º passo: Valor tabelado
- 6º passo: Conclusões

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)

Passos num teste de hipóteses (com valor p)

- 1º passo: Definição de hipóteses
- 2º passo: Indicação do nível de significância
- 3º passo: Escolha do teste a usar
- 4º passo: Comparação do valor prova com o nível de significância
- 5º passo: Conclusões

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- Ambos os algoritmos devem
 - Aprender nos mesmos conjuntos de treino
 - Avaliar os modelos induzidos nos mesmos conjuntos de teste
- Testar: existe uma diferença significativa no desempenho
 - Hipótese nula- H0: não há diferença significativa
 - Hipótese alternativa- H1: há diferença significativa
- Usar o teste t para amostras emparelhadas/Wilcoxon
 - permite inferir sobre a igualdade das médias/distribuições de duas amostras emparelhadas
 - se as amostras têm dimensão inferior a 30 \Rightarrow as amostras devem provir de populações normalmente distribuídas
 - se é violada a normalidade dos dados \Rightarrow usar testes não paramétricos
 - Teste de Wilcoxon

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- Teste t para amostras emparelhadas
 - **Amostras emparelhadas:** se pares de observações (x_i, y_i) são dependentes sendo todos os restantes pares (x_i, y_i) , $i \neq j$ independentes
 - Para obter duas amostras emparelhadas usar validação cruzada k-fold
 - Para cada fold j ($j=1, \dots, k$)
 1. estimar valor de medida de desempenho c_{ij} para cada algoritmo i ($i=1, 2$) (taxa de erro, taxa de acerto, precisão, sensibilidade, área AUC,...)
 2. Calcular as diferenças no desempenho: $d_j = c_{1j} - c_{2j}$

para $k=10$

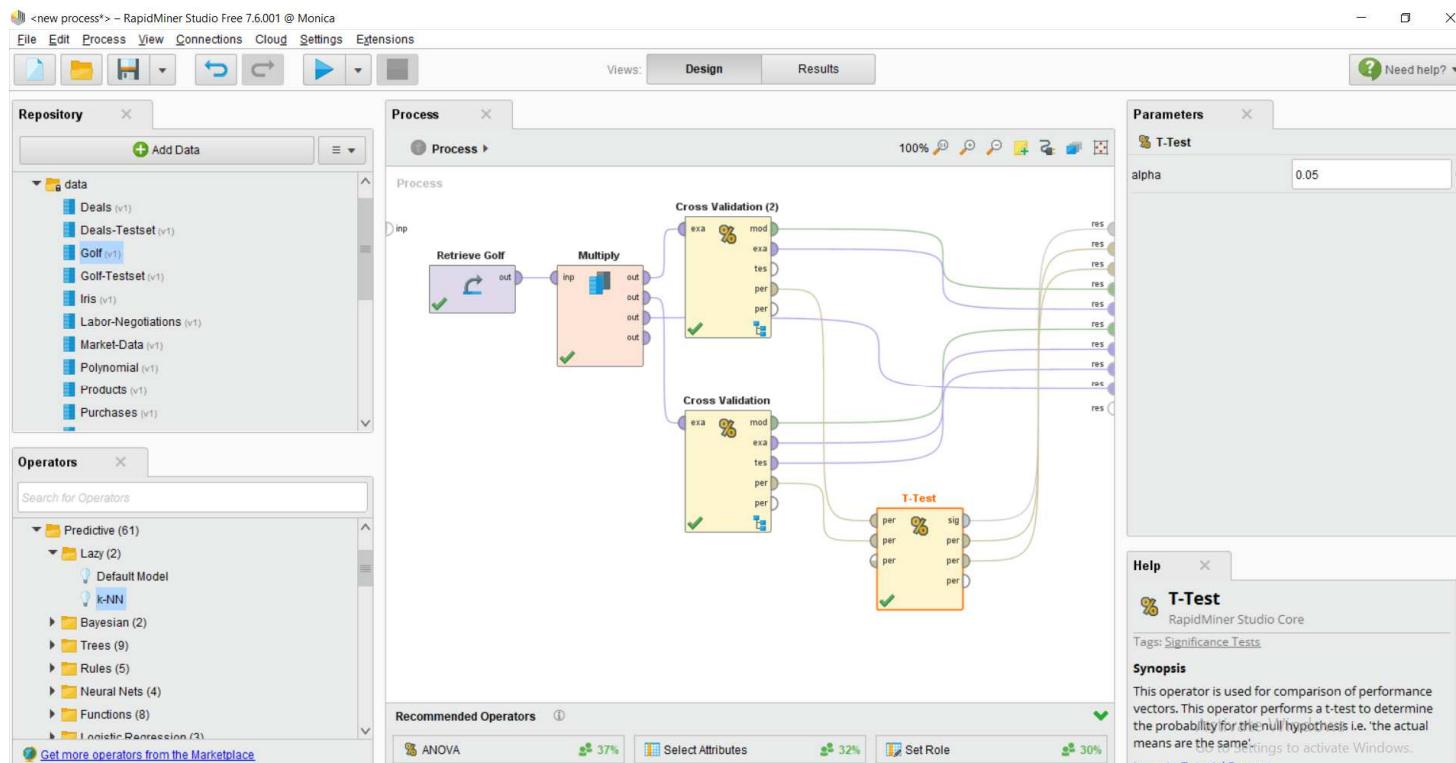
Fold	1	2	3	4	5	6	7	8	9	10
Algoritmo1	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{16}	c_{17}	c_{18}	c_{19}	c_{110}
Algoritmo2	c_{21}	c_{22}	c_{23}	c_{24}	c_{25}	c_{26}	c_{27}	c_{28}	c_{29}	c_{210}
Diferenças	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}

Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- Teste t para amostras emparelhadas – **Problemas de Implementação**
 - Pressupõe que as diferenças de desempenho $d_j = c_{1j} - c_{2j}$ provenham de uma distribuição Normal o que é difícil de provar pois há poucos dados (se $k=10$, a amostra apenas contém 10 elementos)
 - Os conjuntos de testes são independentes, mas os conjuntos de treino não
⇒ Elevada probabilidade de ocorrência do erro de Tipo I
 - Incorretamente deteta que existe diferença significativa no desempenho dos dois algoritmos quando realmente esta diferença não existe
- Alternativas
 - Testes não paramétricos: Wilcoxon
 - 10×10 cross validation = 10 iterações de 10 – fold CV ⇒ gera amostra de tamanho 100 (pelo TLC aproxima-se à Normal)
 - 5x2 Cross validation ⇒ Foi provado que reduz o erro tipo I

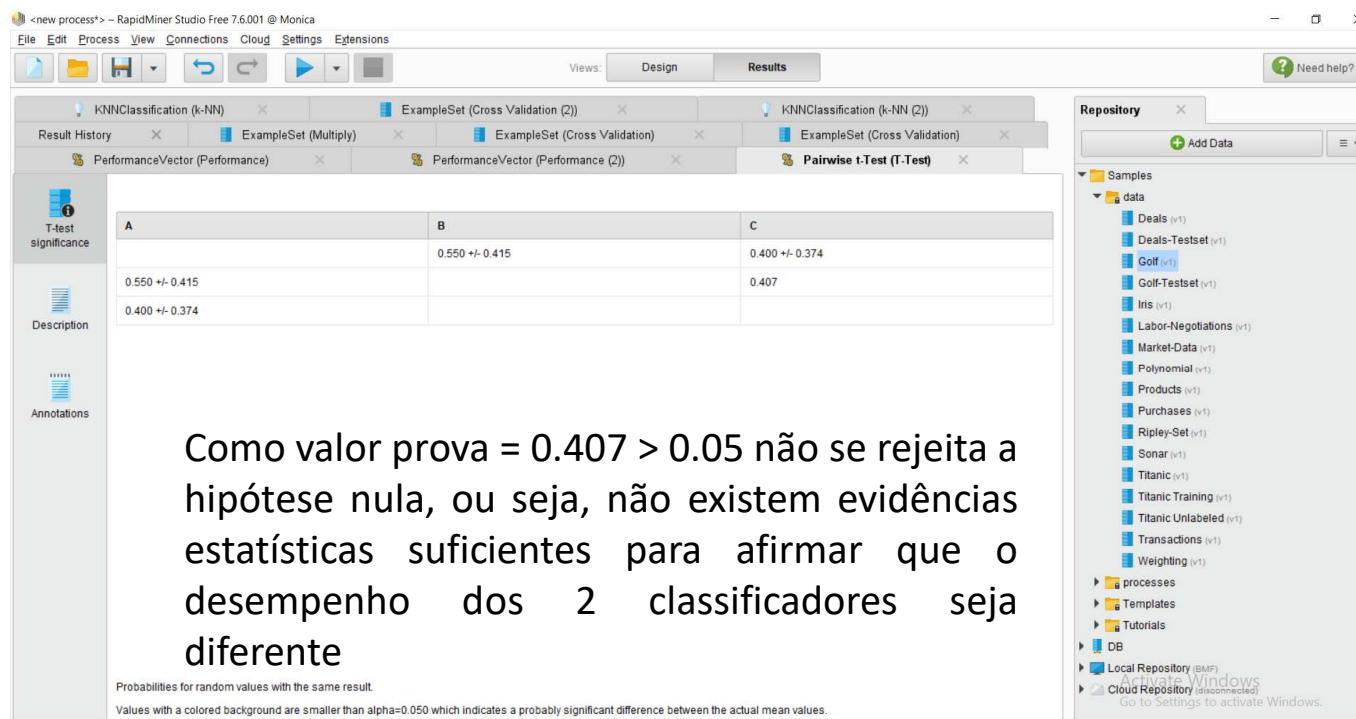
Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- No rapidMiner (pressupondo que os dados provém de distribuições normais)



Avaliação de Modelos

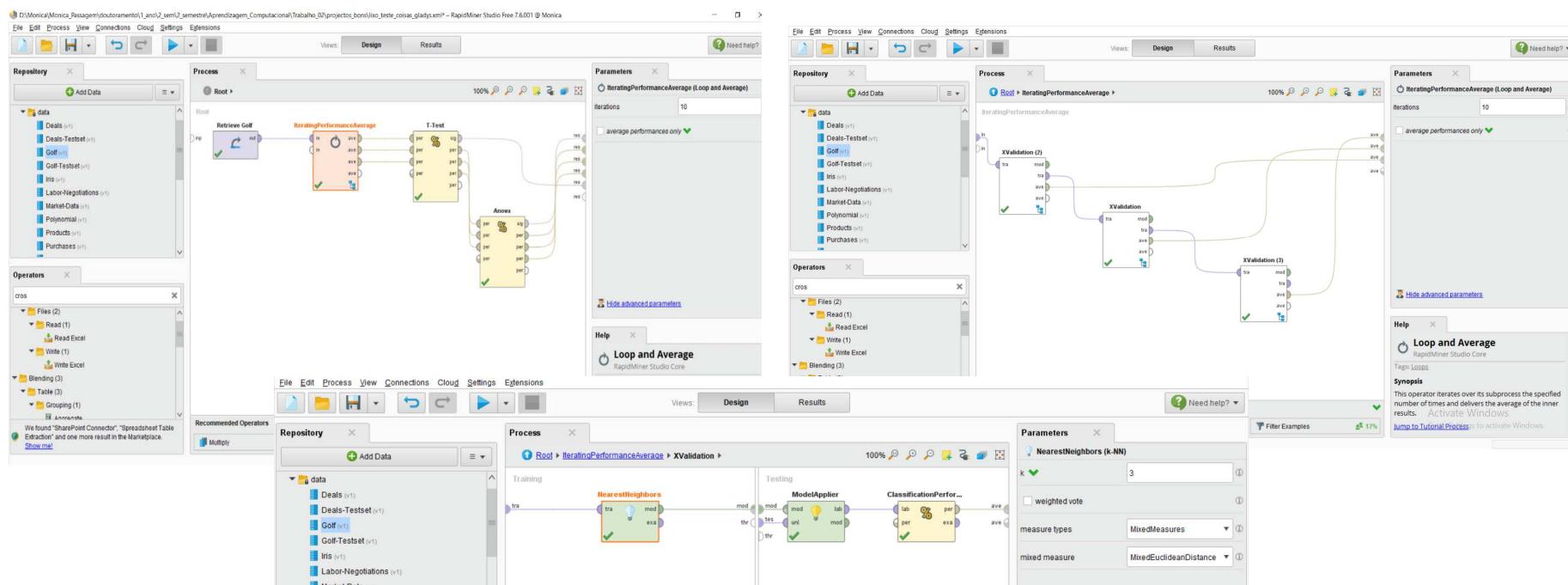
- Comparação do Desempenho de modelos – Teste de Hipóteses (2 algoritmos e um conjunto de dados)
- No rapidMiner (presupondo que os dados provém de distribuições normais)



Como valor prova = $0.407 > 0.05$ não se rejeita a hipótese nula, ou seja, não existem evidências estatísticas suficientes para afirmar que o desempenho dos 2 classificadores seja diferente

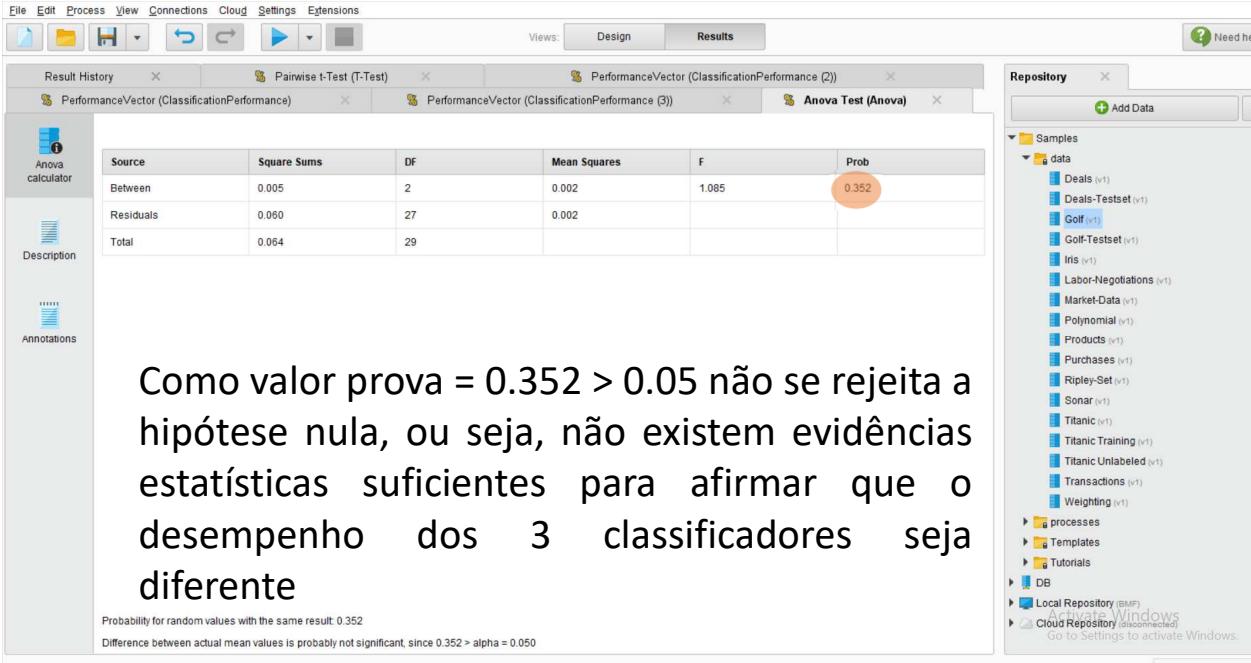
Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (3 algoritmos ou mais e um conjunto de dados) - ANOVA
- No rapidMiner – testar se é estatisticamente significativa a diferença de vários algoritmos - 10x10 cross validation (utilizar o operador Loop and Average)



Avaliação de Modelos

- Comparação do Desempenho de modelos – Teste de Hipóteses (3 algoritmos ou mais e um conjunto de dados) - ANOVA
- No rapidMiner - testar se é estatisticamente significativa a diferença de vários algoritmos - 10x10 cross validation



The screenshot shows the rapidMiner software interface. The top menu bar includes File, Edit, Process, View, Connections, Cloud, Settings, and Extensions. The toolbar contains icons for file operations like Open, Save, and Import. The main workspace has tabs for Result History, Pairwise t-Test (T-Test), PerformanceVector (ClassificationPerformance (2)), PerformanceVector (ClassificationPerformance (3)), and Anova Test (Anova). The central area displays a table from the 'Anova calculator' process:

Source	Square Sums	DF	Mean Squares	F	Prob
Between	0.005	2	0.002	1.085	0.352
Residuals	0.060	27	0.002		
Total	0.064	29			

The 'Prob' column for the 'Between' row is highlighted with an orange circle. Below the table, a note states: "Probability for random values with the same result: 0.352". Another note below it says: "Difference between actual mean values is probably not significant, since 0.352 > alpha = 0.050". To the right, a 'Repository' sidebar lists various datasets and processes. A message at the bottom right says: "Go to Settings to activate Windows".

Como valor prova = $0.352 > 0.05$ não se rejeita a hipótese nula, ou seja, não existem evidências estatísticas suficientes para afirmar que o desempenho dos 3 classificadores seja diferente

Avaliação de Modelos

- Comparação do Desempenho de modelos – **Avaliação sensível à distribuição das classes e ao custo**
- Tomada de decisão. Podemos tomar decisões erradas?

Tomada de decisão numa central nuclear: Um classificador h prediz se deve abrir ou fechar a válvula do módulo de refrigeração num dado momento

- Avaliamos desempenho num conjunto de teste = 100 000 dados acumulados no último mês; a classe é o resultado da decisão tomada por um operário em cada momento
- Número de exemplos da classe “fechar”: 99 500
- Número de exemplos da classe “abrir”: 500
- Suponhamos h prediz sempre “fechar” (classe maioritária). A taxa de erro é muito pequena:

$$\text{Err} = \frac{500}{100000} \times 100 = 0.5\%$$

Será que h é um bom classificador?

Avaliação de Modelos

- Comparação do Desempenho de modelos – **Avaliação sensível à distribuição das classes e ao custo**
- Tomada de decisão. Podemos tomar decisões erradas?

Exemplo com conjunto de teste de 100 000 instâncias

		Preditas	
		h_1	h_2
Real	abrir	abrir	abrir
	300	200	0
	fechar	500	99500
ERRO: 0,7%		Preditas	
		h_2	h_3
Real	abrir	abrir	abrir
	0	500	400
	fechar	99500	5400
ERRO: 0,5%		Preditas	
		h_3	h_1
Real	abrir	abrir	abrir
	100	100	500
	fechar	94100	5400
ERRO: 5,5%			

Avaliação de Modelos

- Comparação do Desempenho de modelos – **Avaliação sensível à distribuição das classes e ao custo**

Em muitas situações os erros produzidos por um modelo não têm as mesmas consequências

Tomada de decisão numa central nuclear: Deixar fechada uma válvula quando é necessário abri-la pode provocar uma **explosão**, enquanto abrir uma válvula quando pode se manter fechada pode provocar uma **paragem**

- **Matriz de custos**

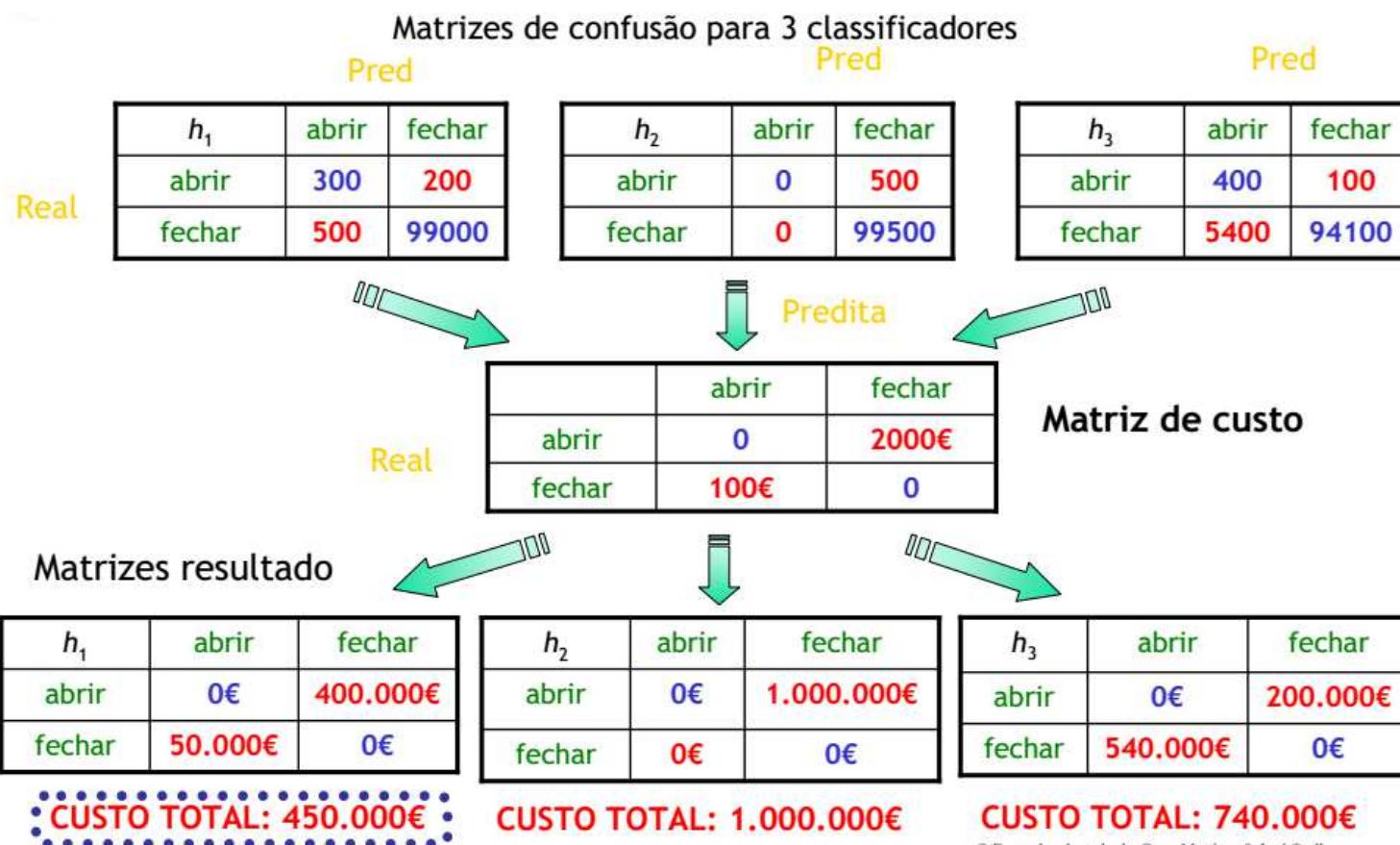
		Predita	
		abrir	fechar
Real	abrir	0	2000€
	fechar	100€	0

O importante não é obter um classificador que erre o menos possível mas aquele que tenha um **menor custo**

A partir da matriz de custo avalia-se cada classificador e selecionamos o classificador com menor custo

Avaliação de Modelos

- Comparação do Desempenho de modelos – **Avaliação sensível à distribuição das classes e ao custo**



Avaliação de Modelos

- Comparação do Desempenho de modelos – **Avaliação sensível à distribuição das classes e ao custo**

De que depende o custo final?

- Para problemas de duas classes depende de um contexto (o skew)
 - proporção do custo dos FP e FN
 - proporção de exemplos negativos e positivos
- Para o exemplo anterior calculamos o “slope”

Proporção dos custos dos erros

$$\frac{FPcost}{FNcost} = \frac{100}{2000} = \frac{1}{20}$$

Proporção das classes

$$\frac{Neg}{Pos} = \frac{99500}{500} = 199$$

$$slope = \frac{1}{20} \times 199 = 9,95$$

- o “slope” é suficiente para determinar qual classificador é o melhor

h_1 : FNR= 40%, FPR= 0,5%

Custo unitário =
 $1 \times 0,40 + 9,95 \times 0,005 = 0,45$

h_2 : FNR= 100%, FPR= 0%

Custo Unitário =
 $1 \times 1 + 9,95 \times 0 = 1$

h_3 : FNR= 20%, FPR= 5,4%

Custo Unitário =
 $1 \times 0,20 + 9,95 \times 0,054 = 0,74$

Menor custo unitário = melhor classificador

© Exemplo adaptado de Cesar Martines & José Orallo

Avaliação de Modelos

Exercício: Calcule as seguintes Medidas de Avaliação dos 3 cenários anteriores

True Positive Rate = recall (sensitivity):

proporção de positivos verdadeiros do total de positivos

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate: proporção positivos falsos (incorrectamente classificados como positivos) do total de negativos

$$FPR = \frac{FP}{FP + TN}$$

True Negative Rate:

proporção de negativos verdadeiros do total de negativos

$$TNR = \frac{TN}{FP + TN}$$

False Negative Rate: proporção de negativos falsos (incorrectamente classificados como negativos) do total de positivos

$$FNR = \frac{FN}{TP + FN}$$

Precision (precisão): proporção de positivos verdadeiros do total dos exemplos classificados como positivos

$$\text{precision} = \frac{TP}{TP + FP}$$

		predita	
		+	-
actual	+	TP	FN
	-	FP	TN
predita		+	-
actual	+	TP	FN
	-	FP	TN
predita		+	-
actual	+	TP	FN
	-	FP	TN
predita		+	-
actual	+	TP	FN
	-	FP	TN

Avaliação de Modelos

- Comparação do Desempenho de modelos – **Curva ROC**

O desempenho de um classificador também depende

- do contexto
 - distribuição das classes (nem sempre todas as classes têm a mesma proporção, podem não estar balanceadas)
 - custos de cada tipo de erro
 - tamanho dos conjuntos de treino e teste

PROBLEMA: Em muitas aplicações não se conhece à priori a distribuição das classes no conjunto de teste

- Torna-se difícil estimar a matriz de custos
- Portanto comparar classificadores usando análises ROC

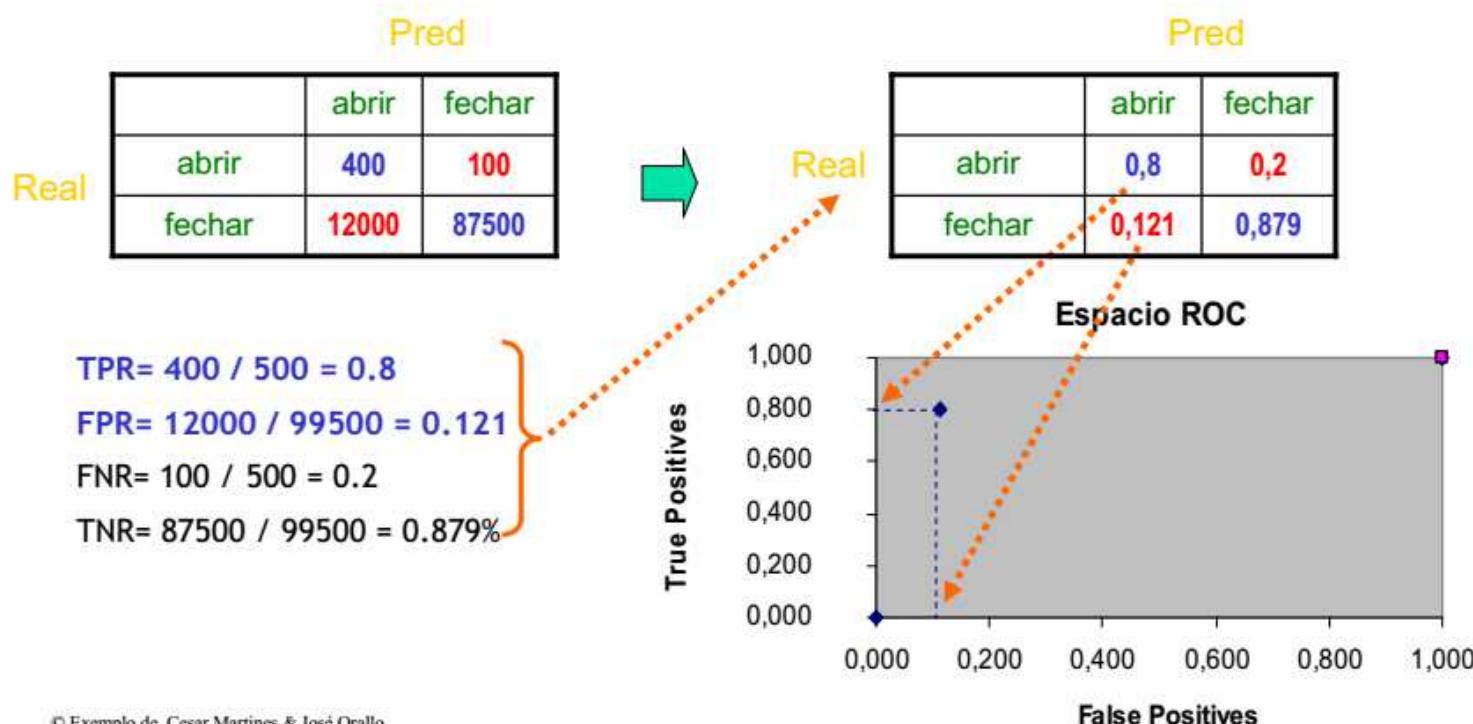
Avaliação de Modelos

- **Curva ROC – Receiver Operating Characteristics**
- Desenvolvido por primeira vez para avaliar radares na 2ª guerra mundial (ex: detecção de sinais ruidosos)
- Nos 70's: usado em aplicações de diagnóstico médico
- Finais dos 90 - começa a usar-se em data mining
- Caracteriza o trade-off: verdadeiros positivos vs. falsos positivos
 - ⇒ benefícios vs. custos
- Distingue dois tipos de classificadores:
 - **Discretos (crisp)** – predizem apenas uma classe entre as possíveis
 - **Contínuos (soft)** – predizem uma classe, mas também produzem um valor de confiança (e.x. uma probabilidade)

Avaliação de Modelos

- Curva ROC – Classificadores Discretos (“Crisp”)

O desempenho de cada classificador é representado como um único ponto (FPR, TPR) no gráfico ROC

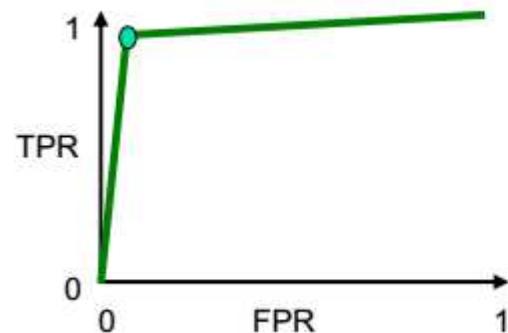


© Exemplo de Cesar Martines & José Orallo

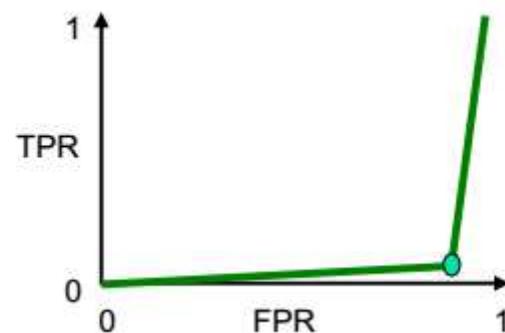
Avaliação de Modelos

- Curva ROC – Classificadores Discretos (“Crisp”)

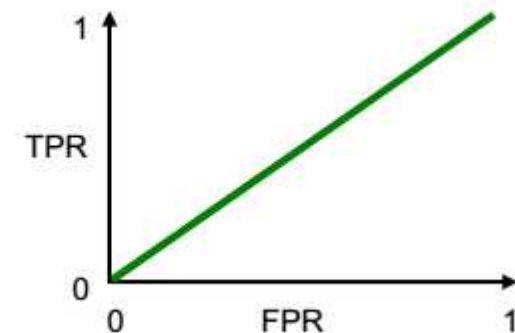
Bons e Maus classificadores



Bom classificador
TPR – alto
FPR - baixo



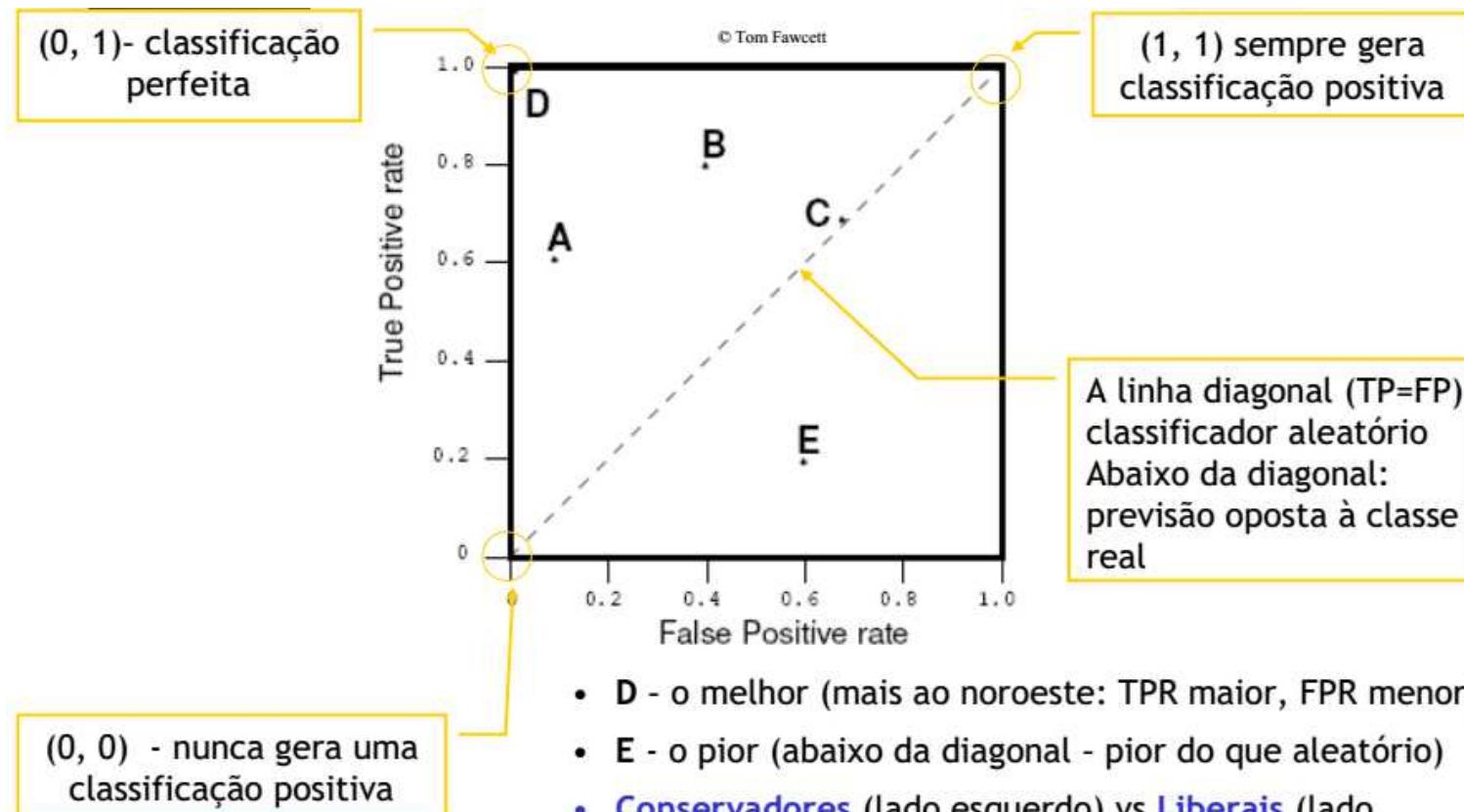
Mau classificador
TPR – baixo
FPR - alto



Mau classificador

Avaliação de Modelos

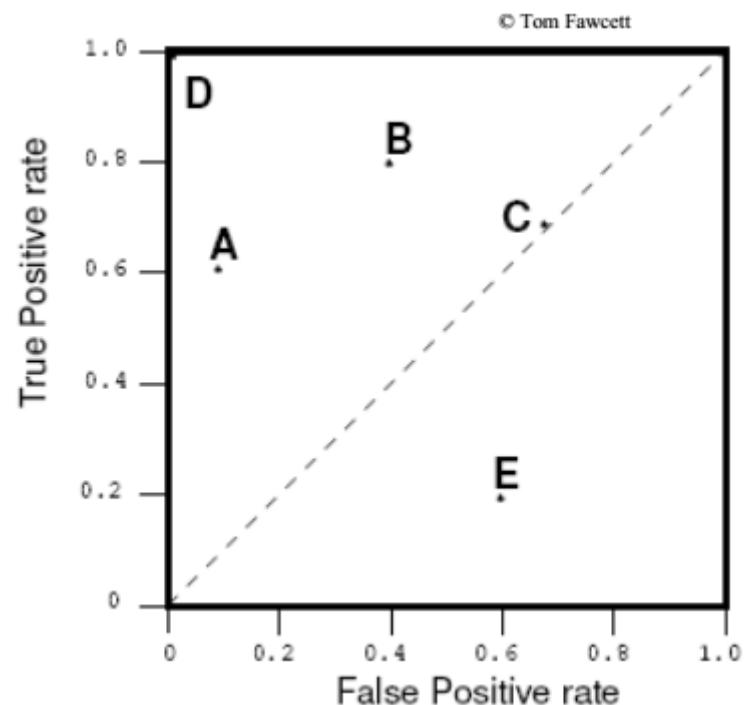
- Curva ROC – Classificadores Discretos (“Crisp”)



- D - o melhor (mais ao noroeste: TPR maior, FPR menor)
- E - o pior (abaixo da diagonal - pior do que aleatório)
- **Conservadores** (lado esquerdo) vs **Liberais** (lado direito): A é mais conservador que B (classifica como positivo com mais precaução)

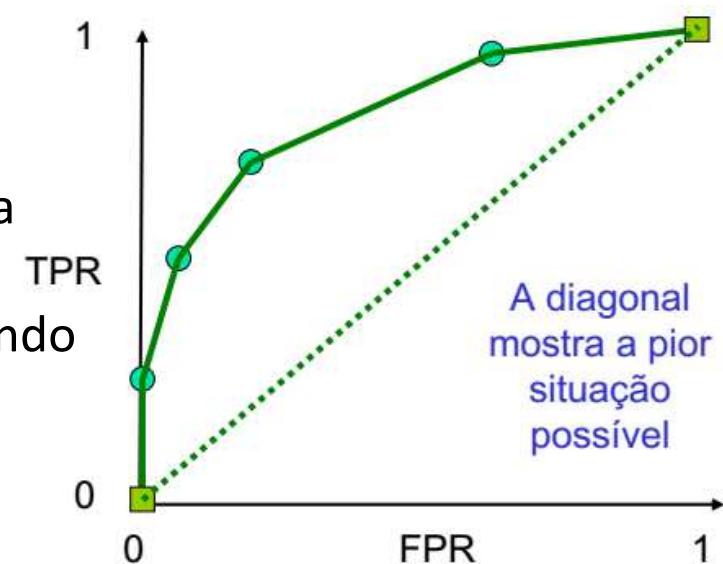
Avaliação de Modelos

- Curva ROC – Classificadores Discretos (“Crisp”)
- D – o melhor (mais ao noroeste: TPR maior, FPR menor)
- E - o pior (abaixo da diagonal – pior do que aleatório)
- Classificadores conservadores (lado esquerdo) vs. liberais (lado direito)
- A é mais conservador que B (menos TP, menos FP \Rightarrow classifica um exemplo como positivo com mais precaução do que B, só quando existir uma forte evidência)
- O triângulo inferior está geralmente vazio. Se um classificador produz pontos abaixo da diagonal, pode-se negá-lo para produzir pontos acima dela (ex: B é igual a E negado)



Avaliação de Modelos

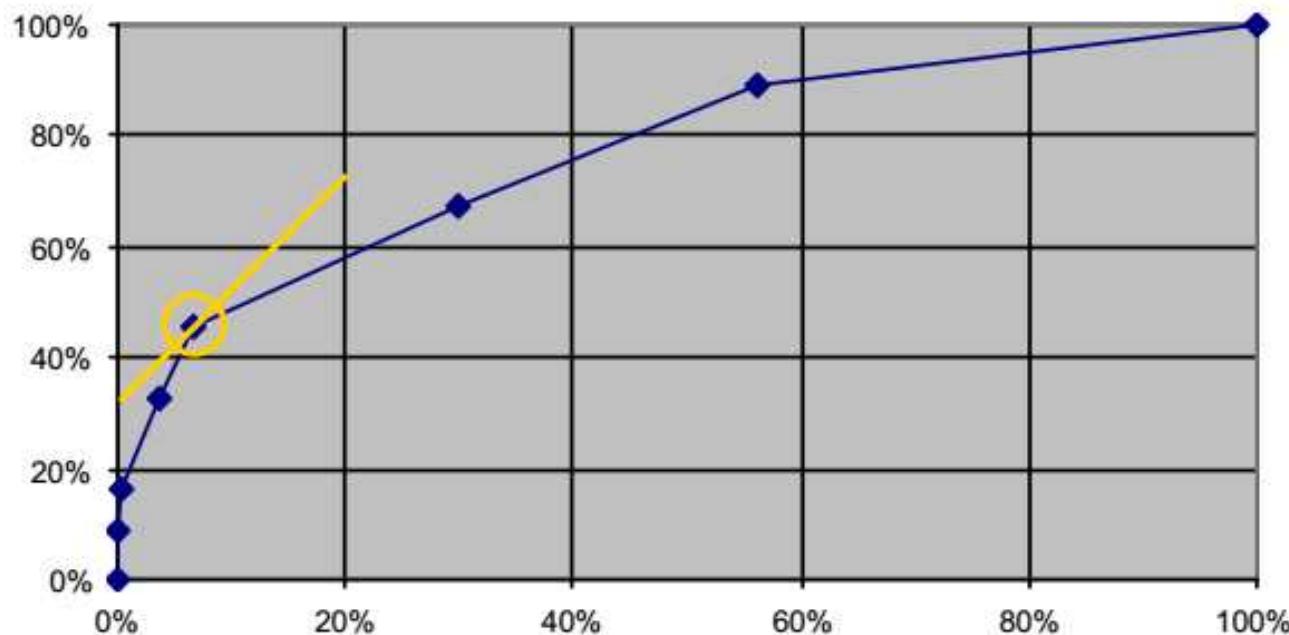
- **Curva ROC – Classificadores Discretos (“Crisp”) – Convex Hull**
- Construímos o invólucro convexo (convex hull) com os pontos de cada classificador e adicionamos os pontos dos classificadores triviais $(0,0)$ e $(1,1)$
- Os classificadores que caem debaixo da curva ROC descartam-se
- O melhor classificador vai ser selecionado tendo em conta o *slope* segundo o contexto da aplicação (distribuição de classes, matriz de custos)



Podemos descartar os classificadores que estão abaixo do invólucro convexo porque não há nenhuma combinação de distribuição de classes para a qual possam ser ótimos

Avaliação de Modelos

- Curva ROC – Classificadores Discretos (“Crisp”) – Convex Hull
- No contexto de aplicação selecionamos o classificador ótimo entre os que foram mantidos



Situação 1: Contexto

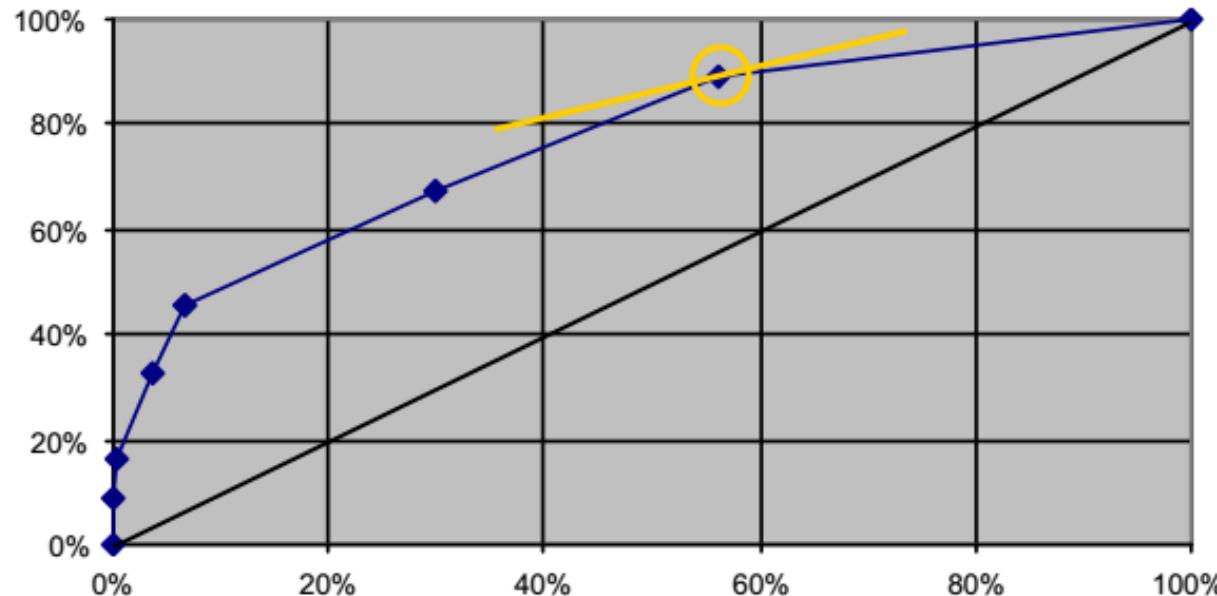
$$\frac{FPcost}{FNcost} = \frac{1}{2}$$

$$\frac{Neg}{Pos} = 4$$

$$slope = \frac{4}{2} = 2$$

Avaliação de Modelos

- Curva ROC – Classificadores Discretos (“Crisp”) – Convex Hull
- No contexto de aplicação selecionamos o classificador ótimo entre os que foram mantidos



Situação 2: Contexto

$$\frac{FPcost}{FNcost} = \frac{1}{8}$$

$$\frac{Neg}{Pos} = 4$$

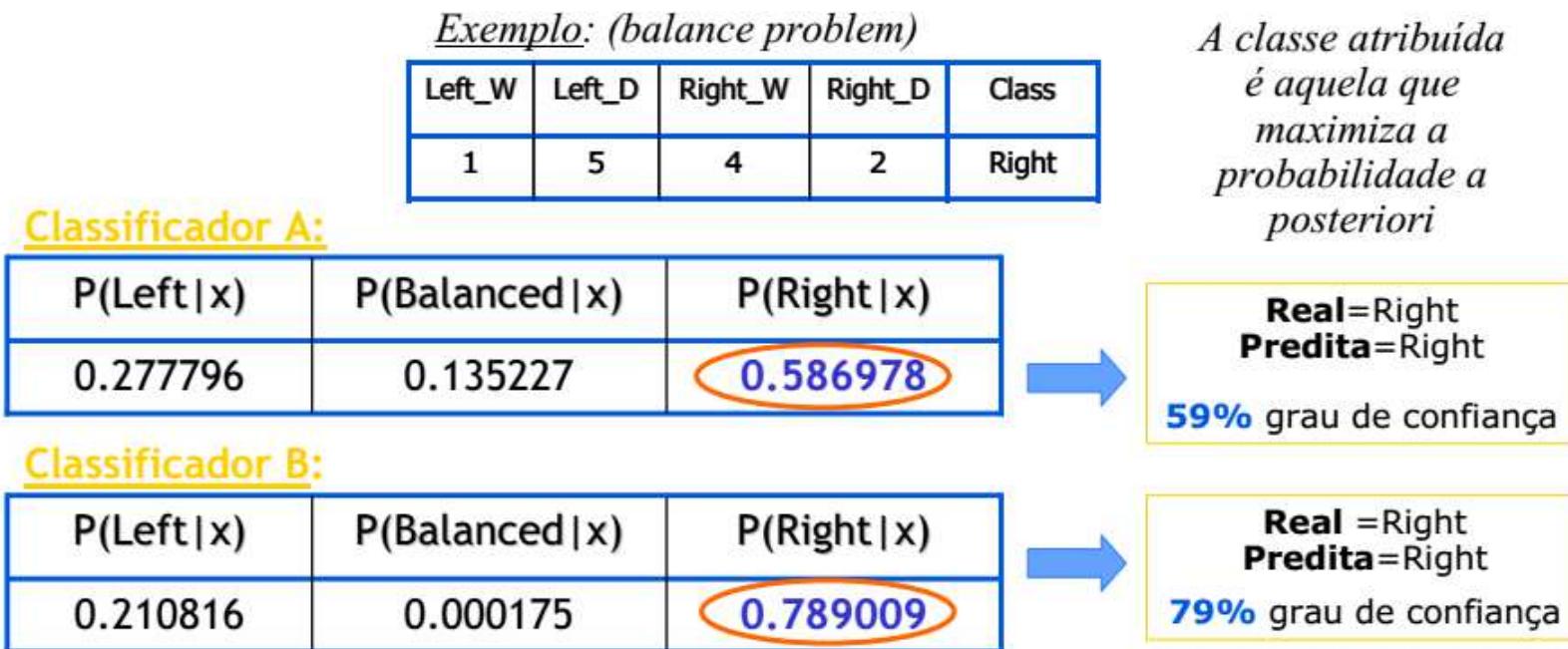
$$slope = \frac{4}{8} = .5$$

Avaliação de Modelos

- **Curva ROC – Classificadores Discretos (“Crisp”) – Conclusões**
- Um classificador ser ótimo depende da distribuição das classes e dos custos dos erros
- A partir deste contexto de aplicação podemos calcular o “slope” (“skew”)
- Se conhecemos o contexto
 - podemos selecionar o melhor classificador, multiplicando a matriz de confusão pela matriz de custos
- Se desconhecemos o contexto
 - usando a análises ROC podemos eleger um subconjunto de classificadores, entre os quais vai estar o classificador ótimo para qualquer contexto possível

Avaliação de Modelos

- Curva ROC – Classificadores Contínuos (“Soft”)
 - Classificadores probabilísticos
- se input: exemplo $x \Rightarrow$ output: $P(c_j|x)$ para cada classe



Os dois predizem “Right” mas o classificador B está mais seguro

Avaliação de Modelos

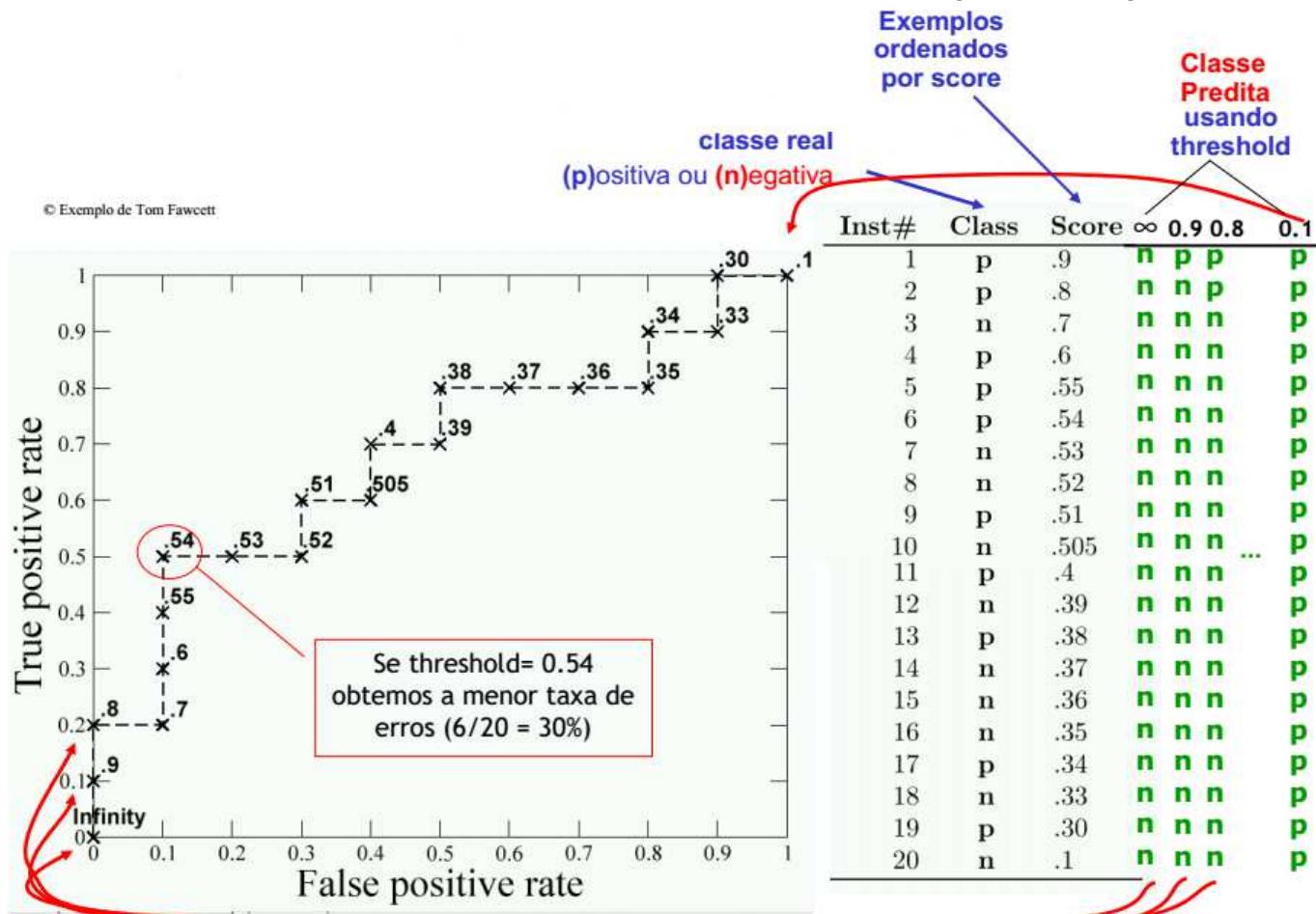
- **Curva ROC – Classificadores Contínuos (“Soft”) – Ranker**
- Se o classificador probabilístico tem um problema binário:
 $P(\text{"yes"} | x) = p \Rightarrow \text{então } P(\text{"no"} | x) = 1 - p$
⇒ só necessitamos especificar a probabilidade de uma classe
- Um **Ranker** é um classificador suave que proporciona um valor entre 0 e 1 da probabilidade de uma das classes. Este valor também se denomina “score”
- Exemplos:
 - Probabilidade de que um cliente compre um produto
 - Probabilidade de relevância de um documento
 - Probabilidade de que um correio seja spam

Avaliação de Modelos

- **Curva ROC – Classificadores Contínuos (“Soft”)**
- Um classificador “soft” pode converter-se num classificador “crisp” se utilizarmos um *threshold*
 - Exemplo: se score > 0.7 então classe “+” caso contrário classe “-”
- Para diferentes *thresholds* obtemos diferentes classificadores “crisp”
- Assim, fixando um *threshold*, obtemos um classificador “crisp” e
 - podemos desenhar-lo como um ponto no gráfico ROC
 - gera-se uma curva escalonada (a curva ROC)
 - quanto maior o número de instâncias a curva fica mais contínua

Avaliação de Modelos

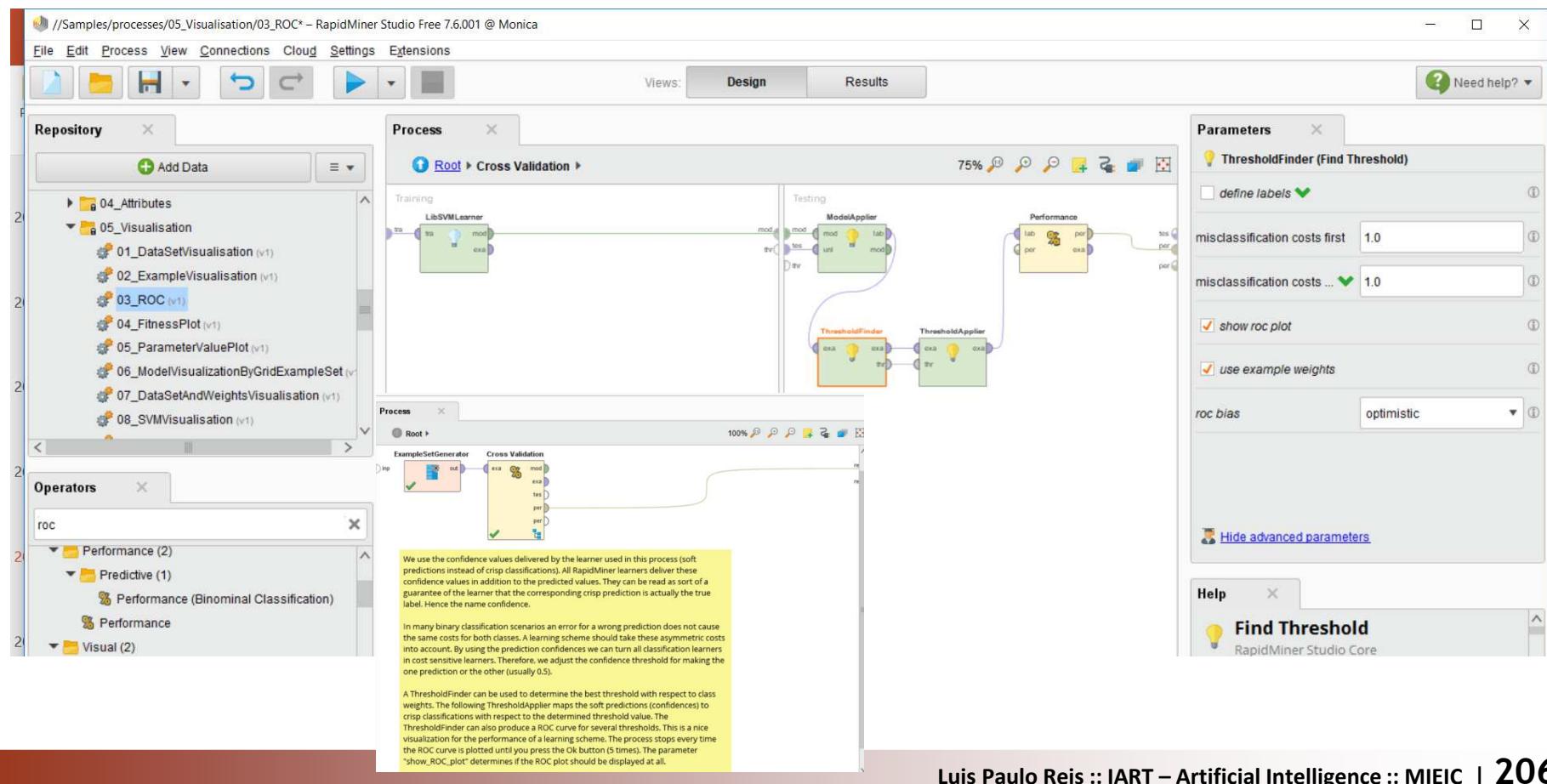
- Curva ROC – Classificadores Contínuos (“Soft”)



Avaliação de Modelos

- Curva ROC – No RapidMiner - Curvas ROC

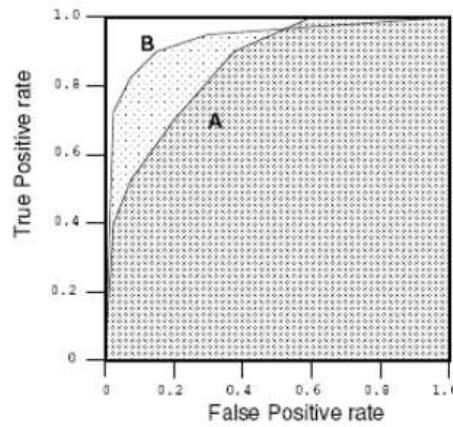
Path: Repository -> Samples -> processes -> 05_Visualization -> 03_ROC



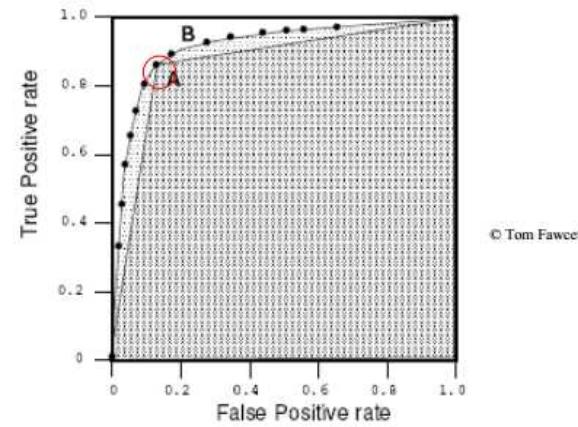
Avaliação de Modelos

- **Curva ROC – Comparação de Classificadores**
- Reduzir a curva ROC de cada classificador a um valor escalar:
 - a área AUC (*area under the ROC curve*)
- AUC varia entre 0 e 1 (1 = área de um quadrado unitário)
- 0.5 é a área de um classificador aleatório
⇒ nenhum classificador real deve ter um valor AUC < 0.5

O melhor classificador é aquele que tem maior AUC



A e B - duas curvas ROC



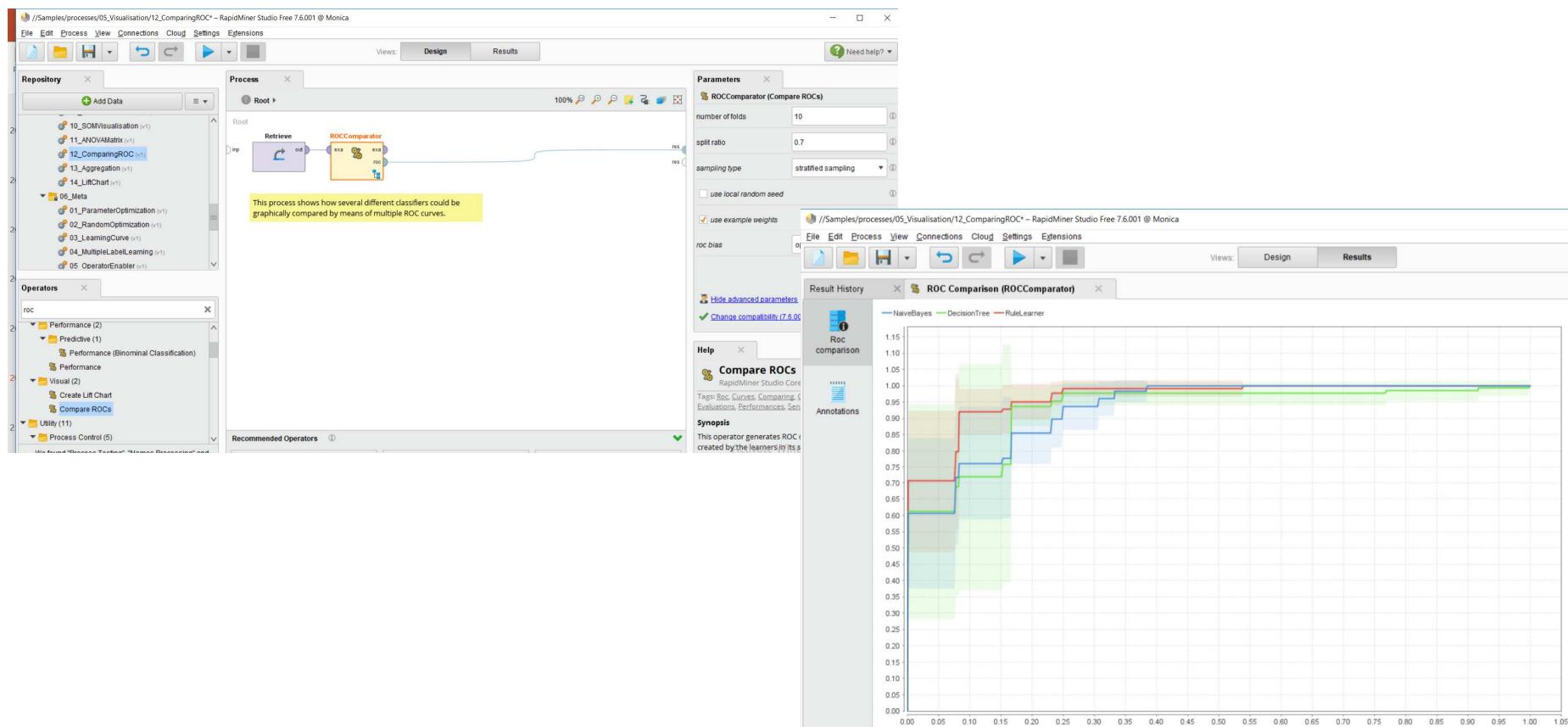
A - classificador discreto, B - classificador contínuo

© Tom Fawcett

Avaliação de Modelos

- Curva ROC – No RapidMiner (Comparar Curvas ROC)

Path: Repository -> Samples -> processes -> 05_Visualization -> 12_Comparing ROC



Bibliografia

- Tan, P., Steinbach, M. & Kumar, V. (2006). Introduction to Data Mining. Pearson Addison-Wesley.
- Adaptação de slides de “Introduction to Data Mining”, Pang-Ning Tan, Michael Steinbach, Vipin Kumar: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
- Adaptação de slides de: Gladys Castillo, Aprendizagem Computacional (Machine Learning), Universidade de Aveiro, 2008
- Adaptação de slides de: B. Mónica Faria, Extração de Conhecimento, Politécnico do Porto, 2018
- Adaptação de slides de: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
- Bergeron, B. (2003). Bioinformatics computing: the complete, practical guide to bioinformatics for life scientists. New Jersey: Prentice Hall.
- Santos, M. F. & Azevedo, C. (2005). Data mining: descoberta de conhecimento em bases de dados. Lisboa: FCA
- Hill M., Hill A. (2007) Investigação por Questionário, Edições Sílabo, 2^a Edição
- Maroco, J., Análise Estatística – com utilização do SPSS, Ed. Sílabo, Lda, Abril, 2003.
- Dawson-Saunders B, Trapp G (2004) Basic and Clinical Biostatistics, 4a Ed. Prentice-Hall Int. Inc
- RapidMiner: Data Science Platform, 2017, <https://rapidminer.com/>

Artificial Intelligence/ Inteligência Artificial

Lecture 9: Supervised Learning/Aprendizagem Supervisionada (adaptado de Faria, 2018 e Castillo 2011)

Luís Paulo Reis

lpreas@fe.up.pt

Director of LIACC – Artificial Intelligence and Computer Science Lab.
Associate Professor at DEI/FEUP – Informatics Engineering Department,
Faculty of Engineering of the University of Porto, Portugal
President of APPIA – Portuguese Association for Artificial Intelligence

