Tópicos Especiais:

INTELIGÊNCIA DE NEGÓCIOS II

Mineração de Dados - II

Sylvio Barbon Junior
barbon@uel.br

Sumário

- Etapa II
  - Algoritmos Básicos
  - Weka: Framework para Machine Learning
  - Avaliando os Resultados
  - Estudo de Caso 2: Produção de Uvas

Inteligência de Negócios
Algoritmos Básicos

- Serão apresentados os seguintes algoritmos:
  - MLP - Multilayer Perceptron;
  - SVM - Support Vector Machine;
  - K-means;
  - PCA - Principal Component Analysis;
  - Apriori;
  - SMOTE - Sythetic Minority Oversampling Technique;
  - J48.

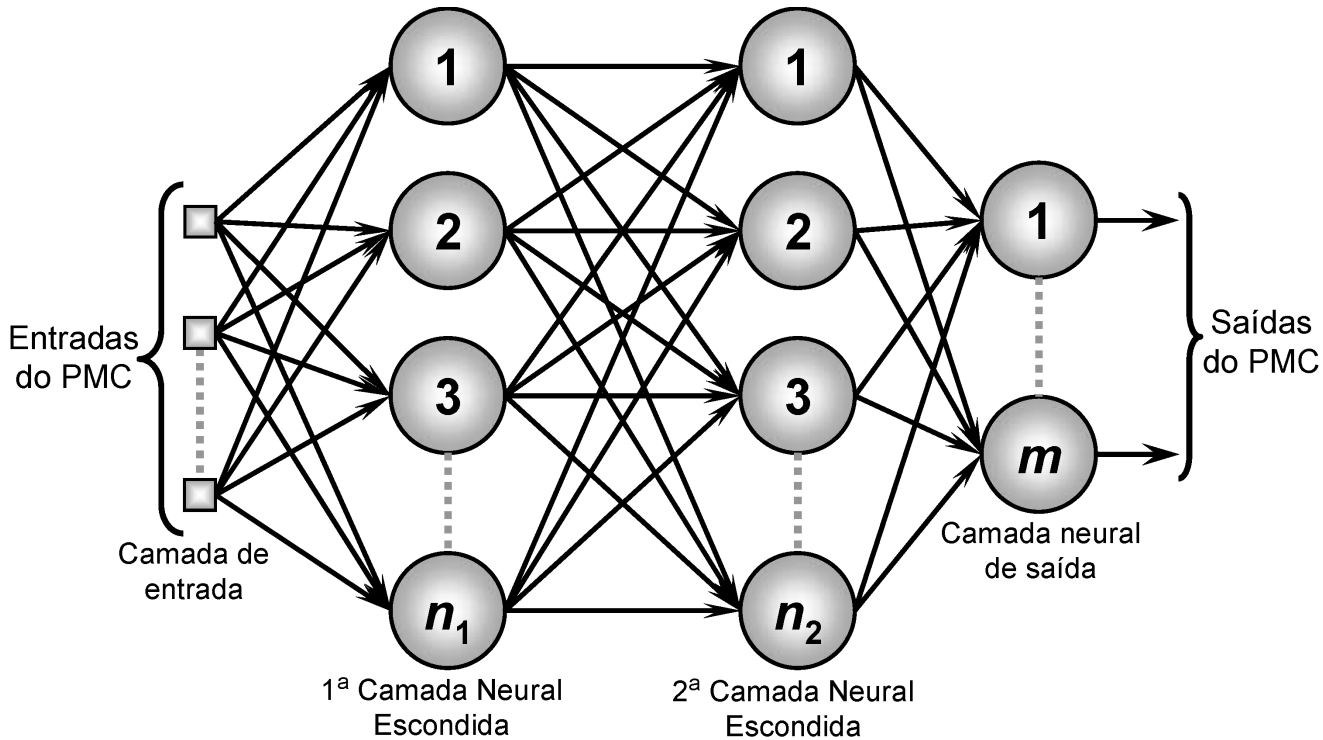
Solucionando problemas de classificação como um ser humano!

Inteligência de Negócios
MLP - MultiLayer Perceptron

- “Redes neurais artificiais são modelos computacionais inspirados no sistema nervoso de seres vivos”;
- Possuem capacidade de aquisição e manutenção do conhecimento;
- Principais funções:
  - Classificação de Padrões;
  - Predição de Comportamentos;
- Processo de estabelecimento de arquitetura é um processo empírico, tal fato implica em abordagens de tentativa e erro para reconhecimento da solução.

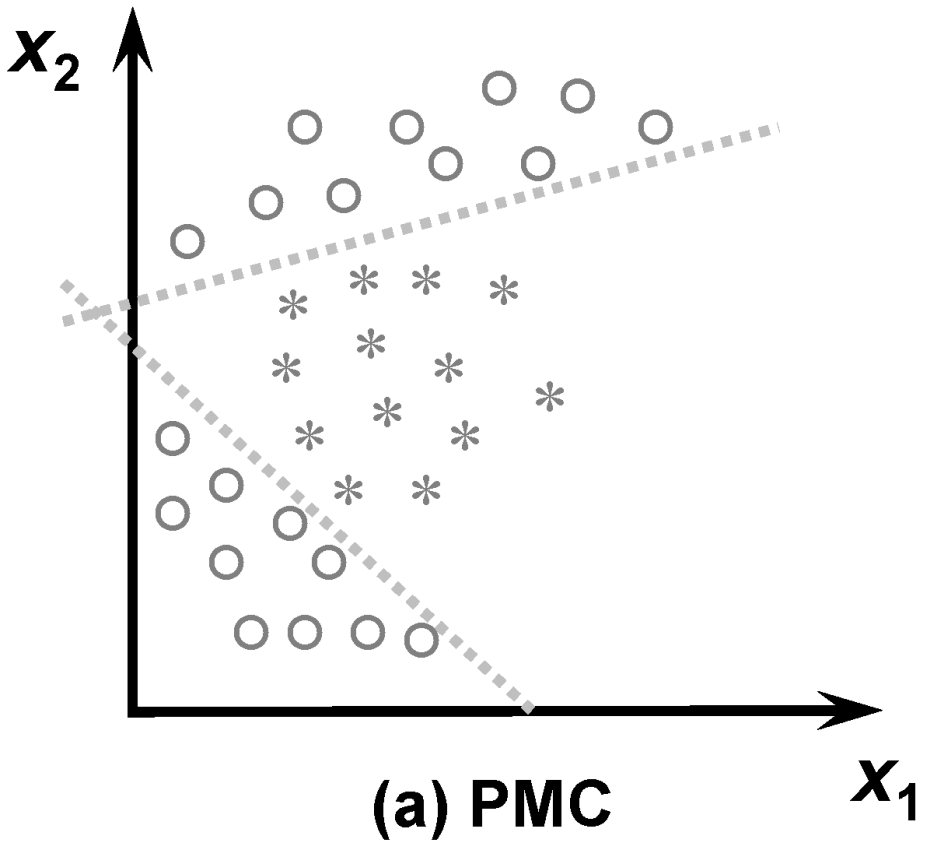
Inteligência de Negócios
MLP - MultiLayer Perceptron

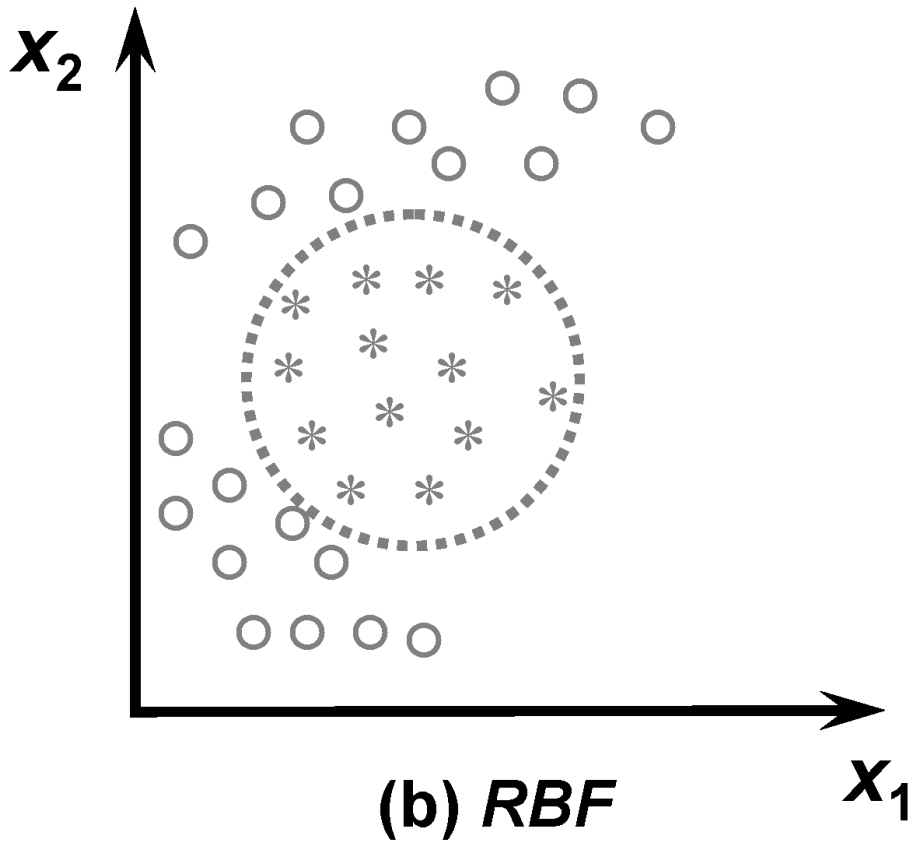
- Perceptron de múltiplas camadas (PMC);
- Apresenta no mínimo duas camadas de neurônios;
- Modelo supervisionado e *Feedfoward*;
- Publicada em 1986;



Inteligência de Negócios
MLP - MultiLayer Perceptron

- Complexidade: O(#epócas*#amostras*#características*#neurônio)

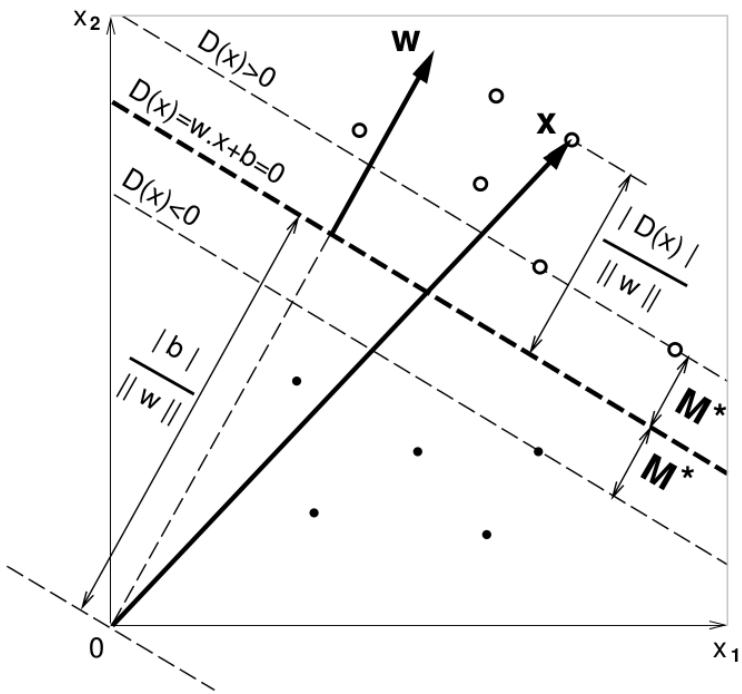


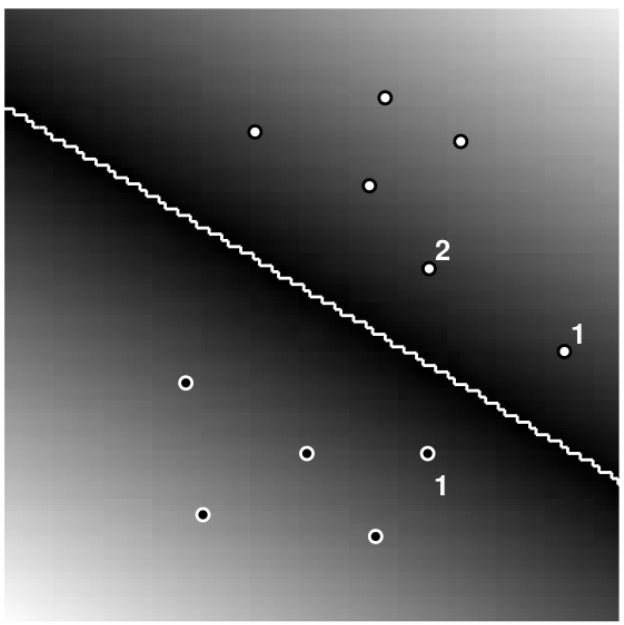


Classificador para qualquer tipo de problema, sem muitas configurações e ideal para problemas binários!

Inteligência de Negócios
SVM - Support Vector Machine

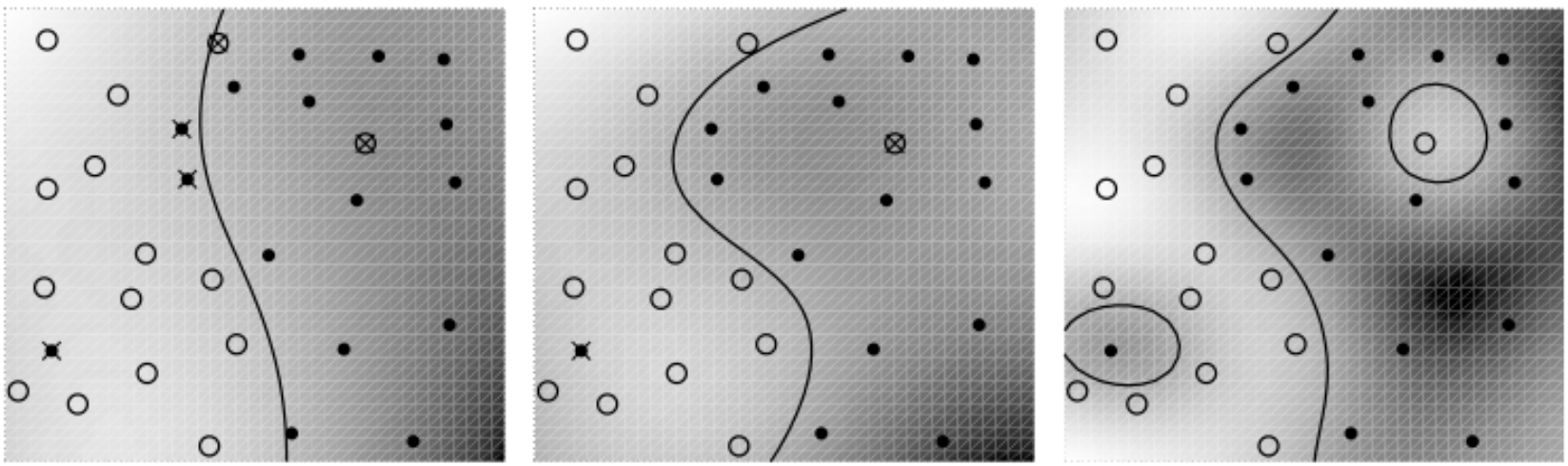
- Máquina de Vetor de Suporte - MVS;
- Foi publicada em 1992;
- Tem a característica da teoria do aprendizado estatístico da década de 60;
- Complexidade: $O(n^3)$





Inteligência de Negócios
SVM - Support Vector Machine

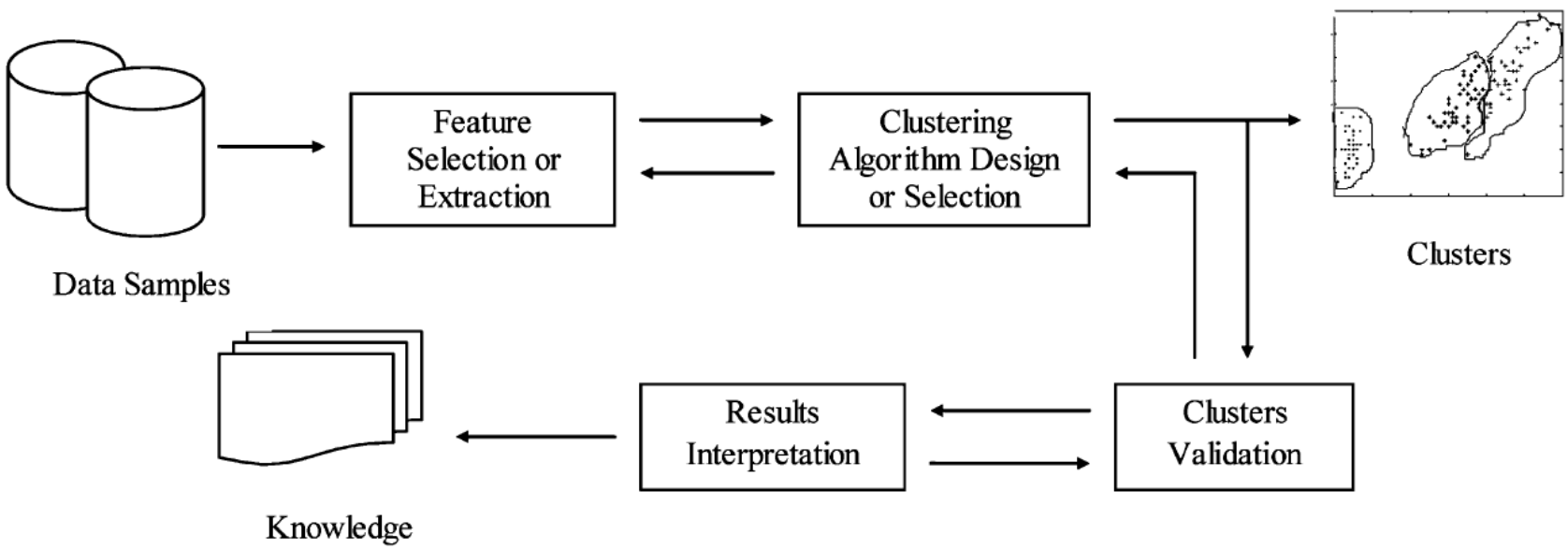
- Ideal para aplicações de classificação **binária**;
- O kernel da SVM define o comportamento do hiperplano criado;
- O kernel mais robusto é o SMO (*Sequencial Minimal Optimization*)



Dada uma base de dados, quais itens se assemelham?

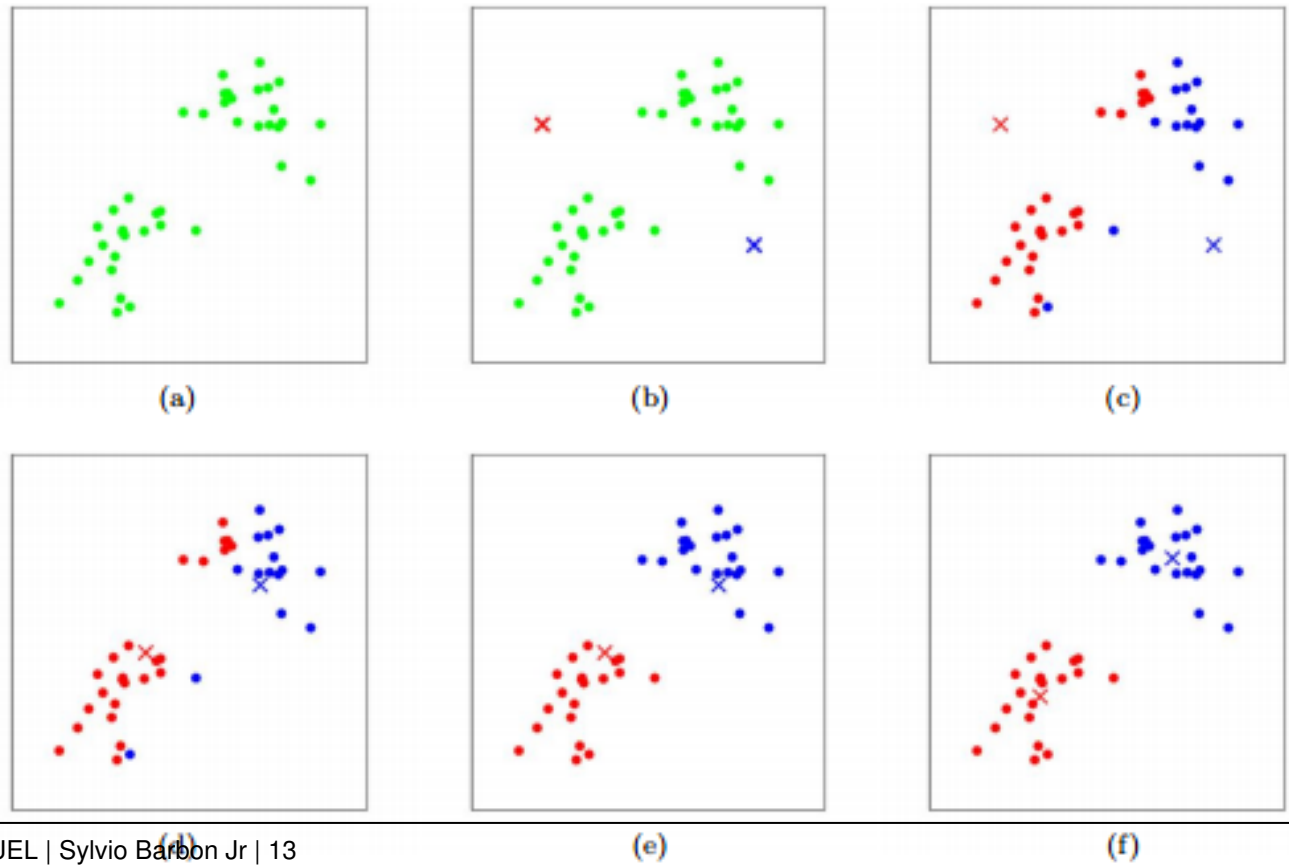
Inteligência de Negócios
K-means

- K-médias
- Principal propósito é a divisão das amostras em subgrupos (clusters, subsets ou categorias) que compartilham características;
- Não oferece um "modelo"com os resultados esperados, ele rotula as instâncias baseado na distância entre as características;
- Aprendizado não supervisionado;



Inteligência de Negócios
K-means

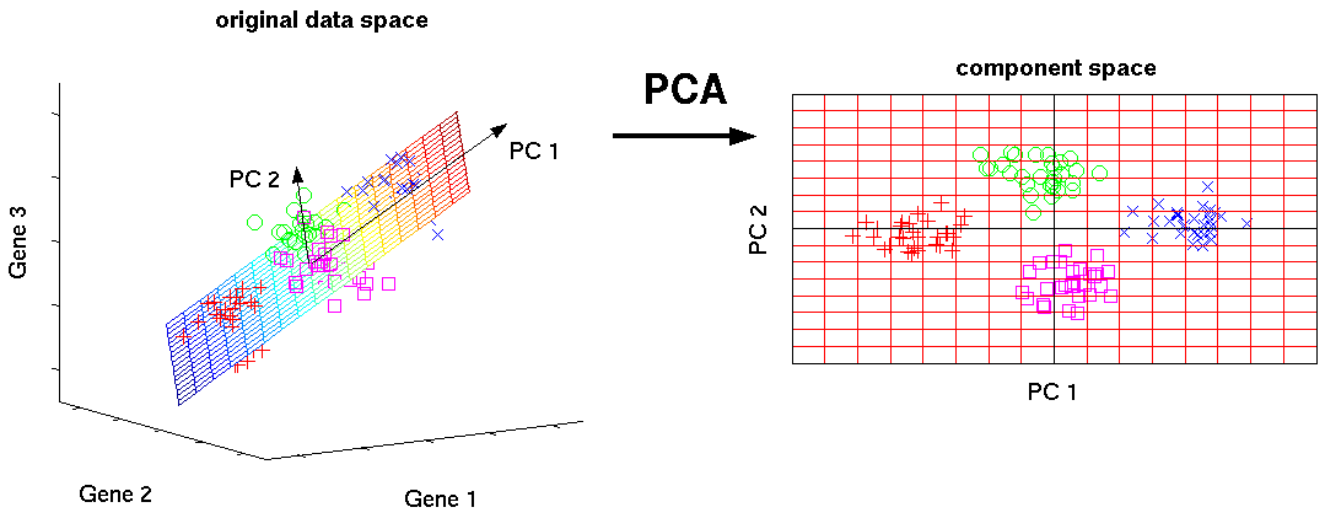
- A quantidade de subgrupos a serem posicionados devem ser passados como parâmetro ao algoritmo.
- Complexidade: O(amostras*centróides(k)*características)



Muitas características, como reduzi-las para encontrar quais são mais adequadas e auxiliar na compreensão do problema?

Inteligência de Negócios
PCA - Principal Component Analysis

- Análise de Componentes Principais - ACP
- Exitem diversas variações.
- É um método simples e **não paramétrico** usado para extrair informação relevante de uma base redundante e ruidosa;
- PCA é uma transformação linear que minimiza a redundância (covariância) e maximiza a informação (variância).
- Complexidade: O(*atributos*^{2}*exemplos + *atributos*^{3})



Como reconhecer um padrão frequente e quais itens influenciam nas combinações?

Inteligência de Negócios
Apriori

- Utilizado na Mineração de Padrões Frequentes;
- A partir de conjuntos frequentes, é possível derivar as regras de associação.
- O espaço de busca de todos os possíveis conjuntos de itens para um conjunto A é de exatamente $2^{|A|}$ itemsets diferentes.
- A representação tradicional é um reticulado que apresenta em suas extremidades um conjunto vazio e um conjunto com todos os itens na base.
- Se $|A|$ é grande o suficiente, então uma proposta simples de gerar e contar os suportes de todos os itemsets não é viável.
- A proposta de nível de confiança do algoritmo *Apriori* implica em diversas varreduras sobre o banco de dados para calcular o suporte dos itemsets frequentes candidados.

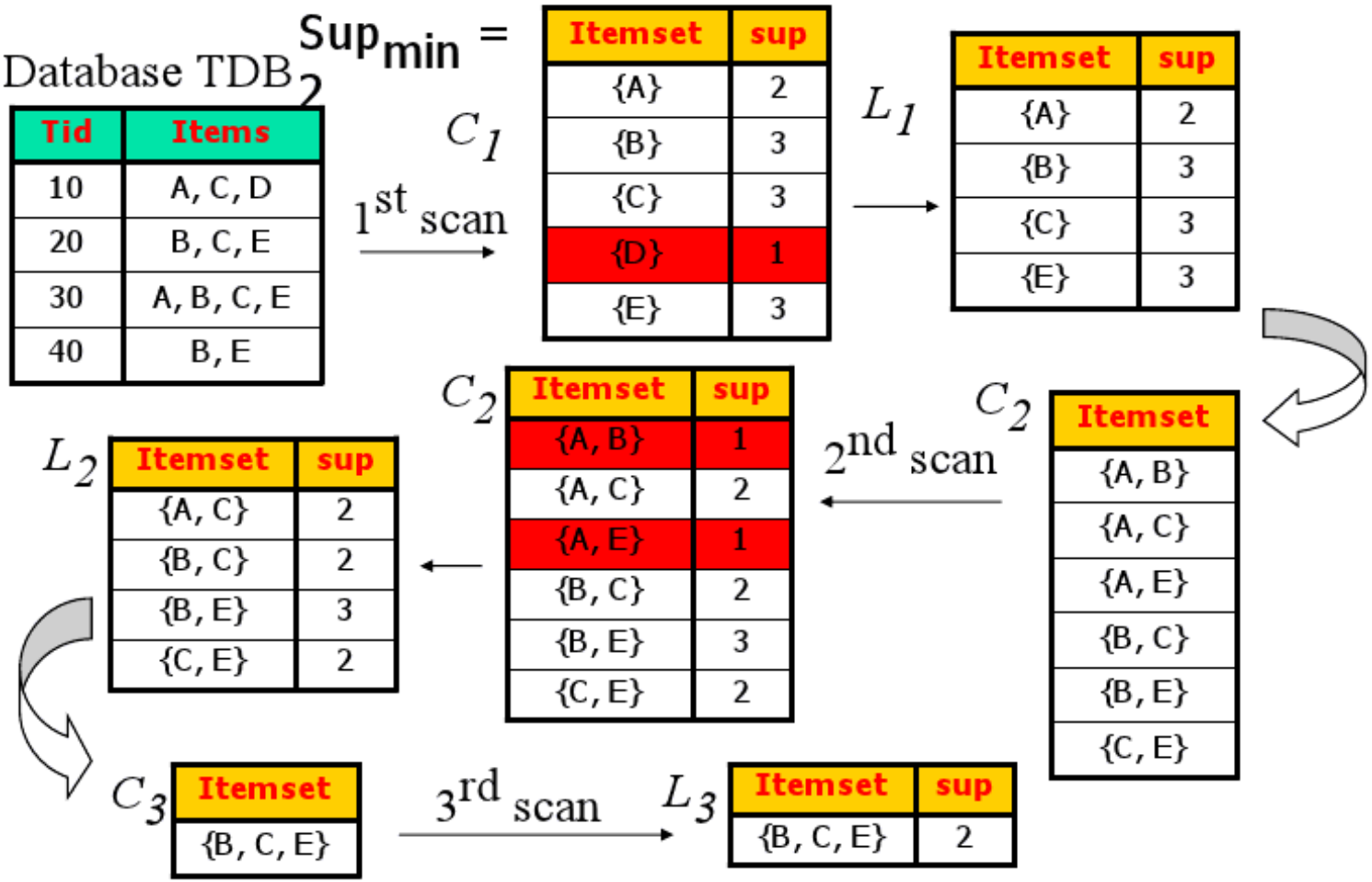
Inteligência de Negócios
Apriori

- Utilizado na Mineração de Padrões Frequentes;
- O suporte está relacionado à frequencia mínima;

TID	Itens
1	{a,d,e}
2	{b,c,d}
3	{a,c,e}
4	{a,c,d,e}
5	{a,e}
6	{a,c,d}
7	{b,c}
8	{a,c,d,e}
9	{b,c,e}
10	{a,d,e}

0 itens	1 item	2 itens	3 itens
Ø: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Inteligência de Negócios
Apriori

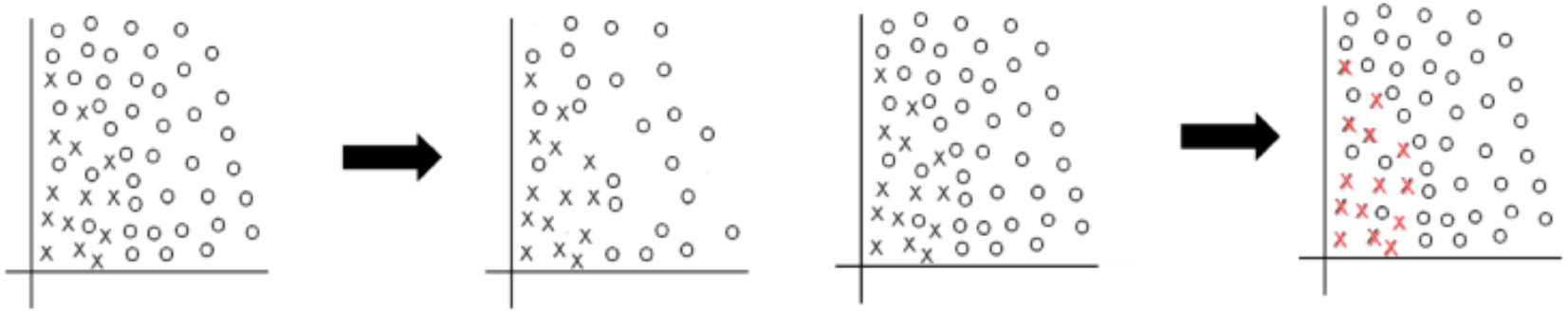


Base de Dados com quantidade de amostras discrepantes entre as classes, como corrigir?

Inteligência de Negócios
SMOTE

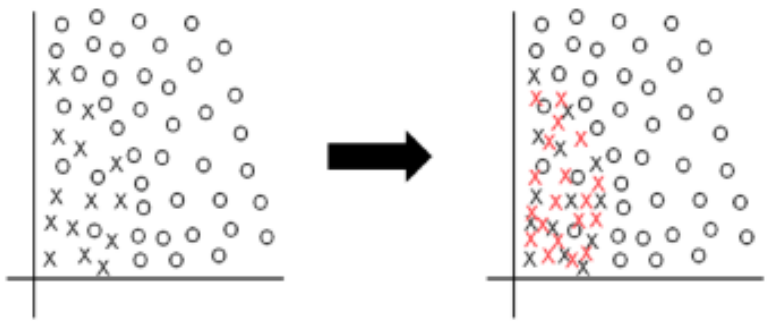
- SMOTE - Sythetic Minority Oversampling technique
- Lidar com distribuição desbalanceada (imbalanced) nas amostras é um problema para o reconhecimento de um determinado padrão ou classe;
- Isso ocorre quando o número de exemplos que representam uma classe é muito menor que outra;
- Tal problema está presente em problemas do mundo real: casos raros ou complexos, mas necessários para descrição completa de um problema;
- Esta característica ocorre em problemas de uma ou mais classes.
- A maioria dos sistemas de aprendizado de máquina não estão preparados para tratar classes desbalanceadas;
- É uma técnica de **oversampling**, criando novas amostras baseadas na interpolação de instâncias da classes minoritárias.
- Baseado em kNN (k nearest neighbors), aleatóriamente selecionam amostras das classes minoritárias e geram as novas.

Inteligência de Negócios
SMOTE



Undersampling

Oversampling



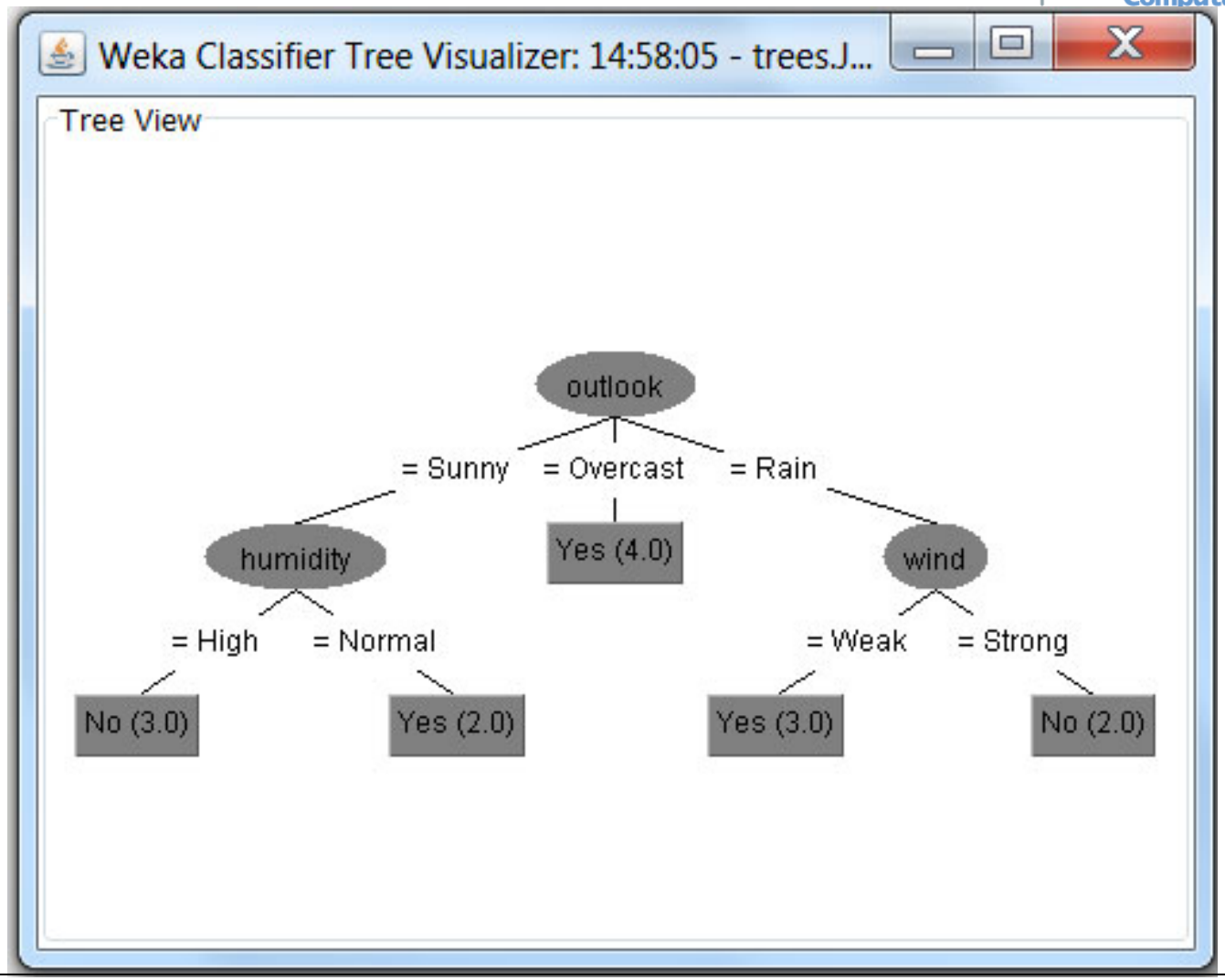
Smote[1]

Classificação, compreensão e visualização do modelo gerado!

Inteligência de Negócios
J48

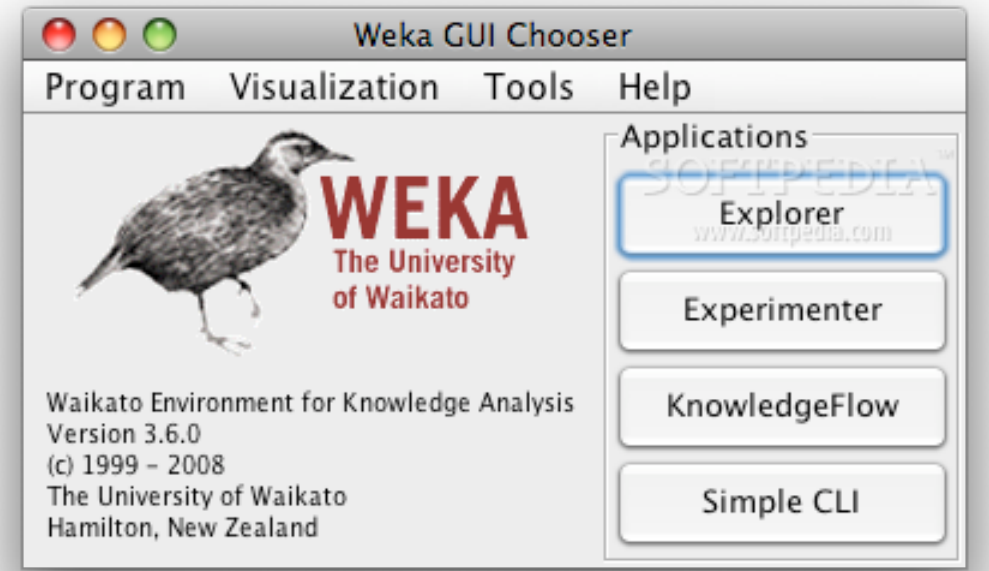
- É uma implementação do algoritmo C4.5
- Baseado em Árvore de Decisão, é referenciado como uma aboragem estatística;
- Em 2008 foi escolhido o melhor algoritmo para Mineração de Dados.
- É calculado com base na entropia e no ganho de informação.
- Ideal para classificação de padrões;
- Fortemente vinculado a uma base de treinamento;
- Exibe os atributos mais significativos hierarquizados em uma árvore;

Inteligência de Negócios
J48



Inteligência de Negócios
Weka

- É uma biblioteca com uma grande coleção de algoritmos de Aprendizado de Máquina implementado em Java;
- Permite as tarefas de Classificação, Regressão, Seleção de Atributos e Agrupamento;
- É compatível com linguagens como Python e R; aplicativos como Octave;

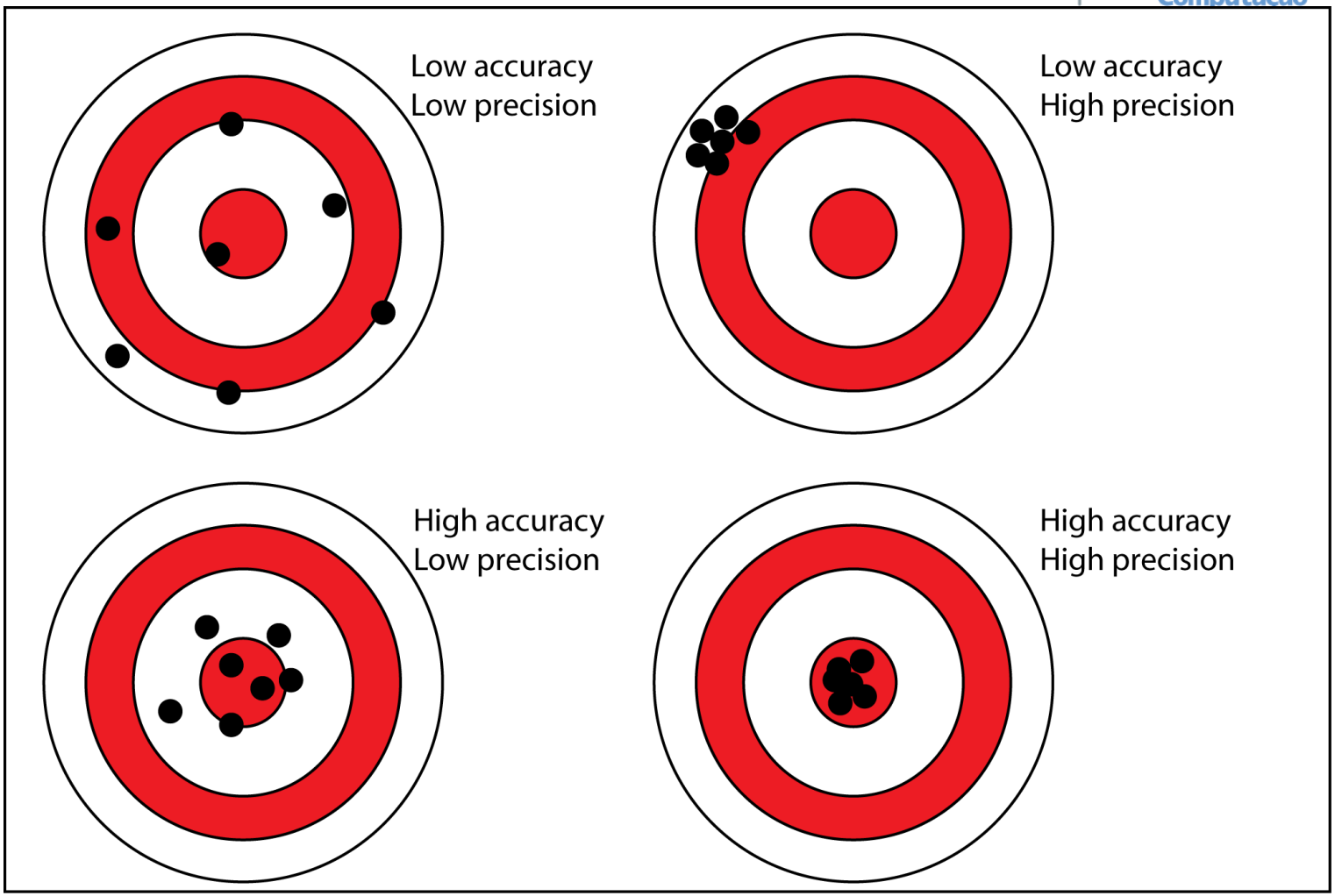


Inteligência de Negócios
Weka

- Demonstração com a base tradicional Iris.arff (problema de classificação de flores).

```
@ATTRIBUTE sepallength REAL
@ATTRIBUTE sepalwidth REAL
@ATTRIBUTE petallength REAL
@ATTRIBUTE petalwidth REAL
@ATTRIBUTE class Iris-setosa,Iris-versicolor,Iris-virginica
```

Inteligência de Negócios
Avaliando os Resultados



Inteligência de Negócios
Avaliando os Resultados

- Quão "preditivo"é o modelo encontrado?
- Somente o "erro"não é um bom indicador de desempenho;
- Medir o taxa de sucesso om o **training set** é uma visão extremamente otimista da solução.
- Quando existe taxa de erro em avaliações baseadas no **training set**, chamamos de **resubstitution error**.
- Por isso usamos um **test set** "separado"para calcular o erro real.
- O **test set** deve ser independente do **training set**.
- Também usamos um **validation set** para aprimorar a técnica de classificação.

Inteligência de Negócios
Avaliando os Resultados

- Técnicas de Avaliação do Modelo:
  - **Holdout**: é o processo de se isolar uma parte do dataset para treinamento e outro para teste (não usado no treinamento).
  - **Crossvalidation**: Validação Cruzada, onde o dataset é divido em dobras (folds) de subamostras, onde o processo é avaliado para cada dobra, ao final é contabilizada a média de acurácia.
  - **Leave-one-out**: Uma instância é escolhida para teste e o restante para treinamento. A vantagem é que o training set é grande e a desvantagem é o custo computacional e problemas de estratificação para futuras comparações.

Inteligência de Negócios
Avaliando os Resultados

- **Matriz de Confusão**: A de uma hipótese h oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de exemplos T.

		prediction outcome		
		p	***n***	**total**
actual value	***p'***	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	total	P	N	

Inteligência de Negócios
Avaliando os Resultados

Exemplo de Matriz de Confusão para avaliação de humanos e bots.

		Human Detection		
		Positive	Negative	Total
Data set	Human	23	7	30
	Bot	5	65	70
	Total	28	72	100

Inteligência de Negócios
Avaliando os Resultados

- **Métricas**:
  - **Mean-squared error** (Erro quadrático médio): É a principal e mais comum medida. É uma medida fácil de se calculada e interpretada.
  - **Mean-absolute error** (Erro quadrático absoluto): É uma alternativa ao anterior, indicando erros individuais sem considerar o sinal.
  - **Relative squared error** (Erro quadrático relativo): Esta métrica é relativa a um classificador, considerando um classificador padrão.
  - **Relative absolute error**: É como o anterior, sem considerar sinal.
  - **Coeficiente de Correlação**: Avalia a correlação entre dois modelos, 1 é a correlação perfeita e 0 é ausência de correlação.

Inteligência de Negócios
Estudo de Caso de Produção de Uvas

- Estudo de Caso de Produção de Uvas.

19-09-2014 as 08:43

Relatório de Autorizações/PC : Módulo Cooperado cc

COOPERAD 6 **Período:** 01/01/2011 **a** 19/09/2014

EMISSAO	AUT	Ped.	Vencimento	Produto	Descrição	V. Unit.	Qtde	Val. Bruto	Funrural	Tx. Com.	Tx. Prov.	Desconto	Comp.	Val.Liquido
					TOTAIS:		2	10,08	0,23	0,40	0,05	0,00	0,00	9,40
02/07/2011	872	575	12/08/2011	58	BANANA PRATA 22KG 2	8,40	222	1.864,80	42,89	74,59	9,32	0,00	0,00	1.738,00
					TOTAIS:		222	1.864,80	42,89	74,59	9,32	0,00	0,00	1.738,00
02/07/2011	872	575	12/08/2011	57	BANANA PRATA 22KG 1	14,00	26	364,00	8,37	14,56	1,82	0,00	0,00	339,25
					TOTAIS:		26	364,00	8,37	14,56	1,82	0,00	0,00	339,25
05/07/2011	879	578	05/08/2011	9	UVA RUBI FL	18,00	102	1.836,00	42,23	73,44	9,18	0,00	0,00	1.711,15
					TOTAIS:		102	1.836,00	42,23	73,44	9,18	0,00	0,00	1.711,15
05/07/2011	879	580	05/08/2011	39	UVA RUBI FL 8KG	28,80	125	3.600,00	82,80	144,00	18,00	0,00	0,00	3.355,20
					TOTAIS:		125	3.600,00	82,80	144,00	18,00	0,00	0,00	3.355,20
07/07/2011	884	584	07/08/2011	15	UVA ITALIA FL	18,00	10	180,00	4,14	7,20	0,90	0,00	0,00	167,76
					TOTAIS:		10	180,00	4,14	7,20	0,90	0,00	0,00	167,76
07/07/2011	884	584	07/08/2011	9	UVA RUBI FL	18,00	126	2.268,00	52,16	90,72	11,34	0,00	0,00	2.113,78
					TOTAIS:		126	2.268,00	52,16	90,72	11,34	0,00	0,00	2.113,78
07/07/2011	884	585	07/08/2011	35	UVA ITALIA FL 8KG	28,80	9	259,20	5,96	10,37	1,30	0,00	0,00	241,57
					TOTAIS:		9	259,20	5,96	10,37	1,30	0,00	0,00	241,57

Referências

- **Livros:**
  - Konar, A. “Computational Intelligence: Principles, Techniques and Applications” (2005)
  - Jensen, R. Shen, Q. “Computational Intelligence and Feature Selection” (2008)
  - Witten, Ian H., and Eibe Frank. "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann (2011)
  - Silva, IN da, Danilo Hernane Spatti, and Rogério Andrade Flauzino. "Redes neurais artificiais para engenharia e ciências aplicadas."São Paulo: Artliber (2010).