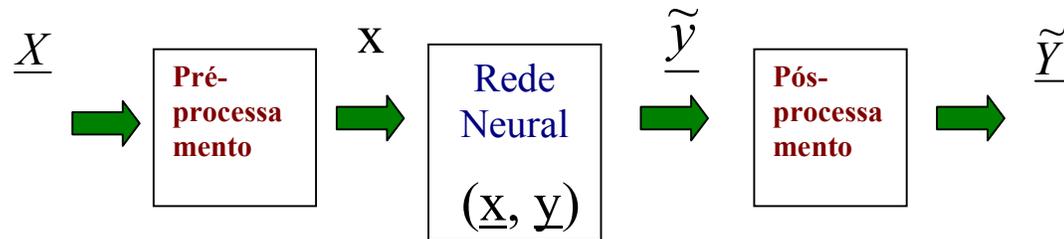


Redes Neurais Artificiais -

Pré e pós-processamento dos dados



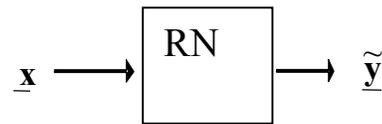
Pré-processamento dos dados

Preparação dos dados de entrada e saída

- 1 - Escolha das variáveis
- 2 – Compactação / Parametrização das variáveis
- 3 – Escalamento das variáveis
- 4 - Pares entrada – saída:

Preparação dos dados de entrada e saída

1 - Escolha das Variáveis



saída \underline{y} – as que desejamos

entrada \underline{x} – as necessárias para gerar as saídas

**“Como entrada escolha as variáveis relevantes,
todas as variáveis relevantes
e somente as variáveis relevantes”**

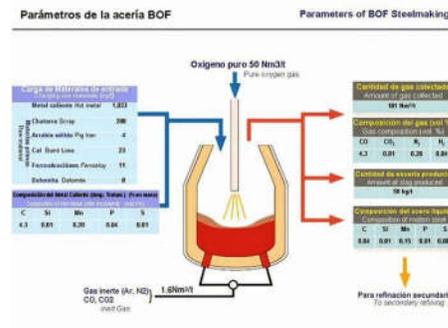
Relevância - como saber se uma entrada é relevante ?

Fenomenologia >>>> candidatas à relevantes

Correlação com as variáveis de saída

Relevância (pós-processamento)

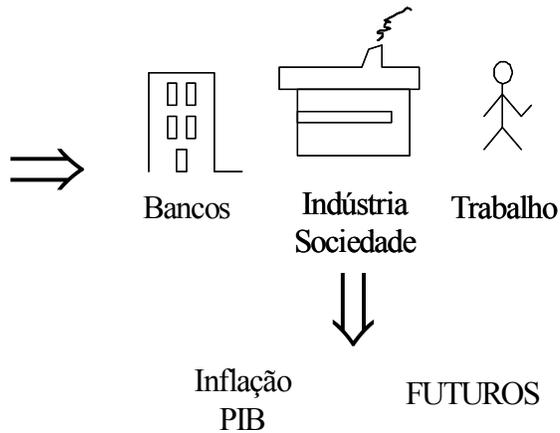
Conversor Siderúrgico



Entradas	Saídas
Peso de gusa	Houve projeção ?
Temperatura do gusa	Temperatura final
Peso de minério + sucata	% carbono
Peso de calcita	% enxofre
Volume de oxigênio	% fósforo
Temperatura da corrida anterior	
Tempo decorrido da corrida anterior	
Temperatura ambiente	
etc.	

Sistema Econômico

- Produção de bens de consumo
 - intermediários
 - capital
 - Quantidade de moeda
 - Inflação
- PASSADOS



800 séries de variáveis.

Quais usar ?

Teste de Relevância**Independência / Dependência Estatística entre Variáveis****Coefficiente de Correlação de Pearson**

Pares $(x_i, y_i) \quad i=1, \dots, P$

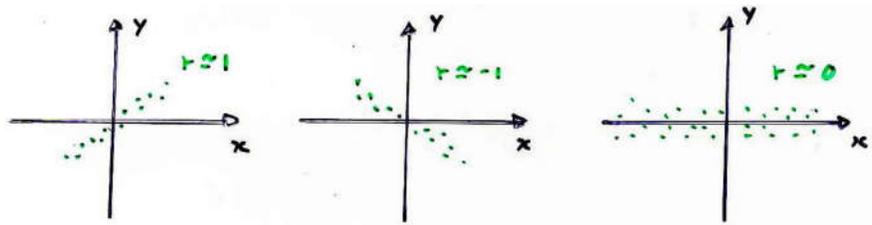
$$r(x, y) = \frac{\frac{1}{P-1} \sum_{i=1}^P (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y} \quad \text{onde}$$

$$\mu_x = \frac{1}{P} \sum_{i=1}^P x_i \quad e \quad \sigma_x = \sqrt{\frac{1}{P-1} \sum_{i=1}^P (x_i - \mu_x)^2}$$

em nosso caso faremos $\mu_x = \mu_y = 0$ e $\sigma_x = \sigma_y = 1$ e então

$$r(x, y) = \frac{1}{P-1} \sum_{i=1}^P x_i y_i$$

$$-1 \leq r \leq 1$$



Valores randômicos correlação $r = 0$

$$\mu(r) \cong 0$$

$$\sigma(r) \cong \frac{1}{\sqrt{P}}$$

95% confiança na correlação

$$|r| \geq 2\sigma(r) = \frac{2}{\sqrt{P}}$$

Matrizes de correlação

entradas – saídas

	y_1	...	y_m
x_1	r_{1v1}		r_{1vm}
x_2	r_{2v1}		r_{2vm}
x_3	r_{3v1}		r_{3vm}
...			
x_n	r_{nv1}		r_{nvm}

entre entradas (simétrica)

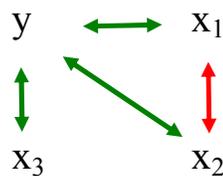
x_1	x_2	x_3	...	x_n
1	r_{12}	r_{13}		r_{1n}
r_{12}	1	r_{23}		r_{2n}
r_{13}	r_{23}	1		r_{3n}
...				
r_{1n}	r_{2n}	r_{3n}		1

completa

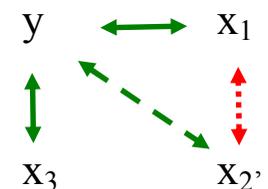
	x_1	x_2	x_3	...	x_n	y_1	...	y_m
x_1	1	r_{12}	r_{13}		r_{1n}	r_{1v1}		r_{1vm}
x_2	r_{12}	1	r_{23}		r_{2n}	r_{2v1}		r_{2vm}
x_3	r_{13}	r_{23}	1		r_{3n}	r_{3v1}		r_{3vm}
...								
x_n	r_{1n}	r_{2n}	r_{3n}		1	r_{nv1}		r_{nvm}

Eliminando variáveis “caras”

Descorrelação entre entradas:

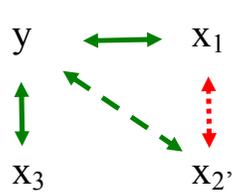


	x_1	x_2	x_3	Y
x_1	1	r_{12}	r_{13}	r_{1v}
x_2	r_{12}	1	r_{23}	r_{2v}
x_3	r_{13}	r_{23}	1	r_{3v}



$$\mu(x_1) = \mu(x_2) = \mu(x_3) = \mu(y) = 0$$

$$x_2 = ax_1 + x_2', \quad x_2' = x_2 - ax_1 \quad a = \arg \text{Min } E[(x_2')^2] = \frac{\sigma_2}{\sigma_1} r_{12}$$



	x_1	x_2	x_3	y
x_1	1	0	r_{13}	r_{1y}
x_2'	0	1	r_{23}	$r_{2'y}$
x_3	r_{13}	r_{23}	1	r_{3y}

$$r_{2'y} = \frac{r_{2y} - r_{1y}r_{12}}{\sqrt{1 - r_{12}^2}}$$

Poda de entradas – pós-processamento

Exemplo:

CONTRIBUTION TO THE DEVELOPMENT OF A RADIOGRAPHIC INSPECTION AUTOMATED SYSTEM

Romeu Ricardo da Silva¹, Marcio H. S. Siqueira¹, Luiz P. Calôba², Ivan C. da Silva¹, Antonio A. de Carvalho¹ and João Marcos A. Rebello¹.

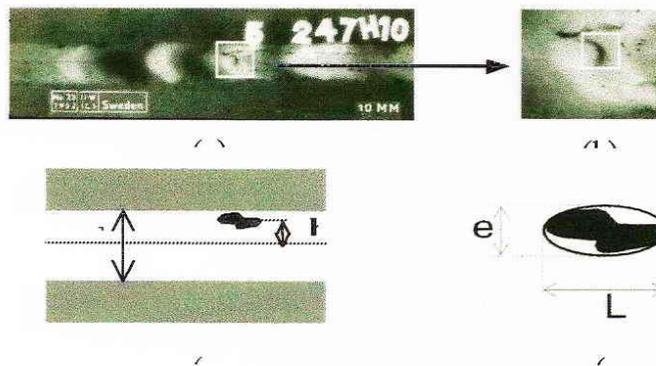


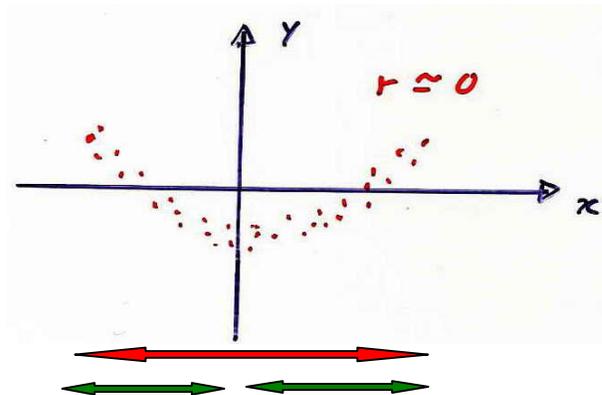
Table 1. Correlation matrix with correlated parameters.

Characteristic Parameters							Defects			
$2/\sqrt{N}$	0.20						0.28	0.48	0.53	0.51
	C	a	L/A	e/A	R	P	IE	PO	FP	MO
C	1.00						0.19	-0.30	-0.02	0.08
a	0.15	1.00					-0.07	-0.38	0.50	0.02
L/A	-0.19	0.03	1.00				0.28	0.06	-0.02	-0.44
e/A	-0.23	-0.53	0.60	1.00			0.06	0.56	-0.34	-0.33
R	0.13	0.78	0.14	-0.50	1.00		0.02	-0.42	0.46	-0.04
P	-0.06	-0.16	-0.39	-0.14	-0.24	1.00	-0.34	0.11	-0.48	0.82

Table 2. Correlation matrix with de-correlated parameters.

Characteristic Parameters							Defects			
$2/\sqrt{N}$	0.20						0.28	0.48	0.53	0.51
	C	a	L/A	e/A	R	P	IE	PO	FP	MO
C	1.00						0.20	-0.18	-0.10	0.01
a	0.01	1.00					-0.15	-0.06	0.24	0.03
L/A	-0.10	0.36	1.00				0.16	0.11	-0.22	-0.12
e/A	0.00	0.00	0.60	1.00			0.06	0.56	-0.34	-0.33
R	-0.02	0.66	0.46	0.00	1.00		-0.07	-0.11	0.16	0.05
P	-0.10	0.00	0.00	-0.14	0.00	1.00	-0.34	0.11	-0.48	0.82

Problema: correlações de ordens mais elevadas



Correlação em partes do domínio

Coefficiente de Correlação de Spearman

Coefficiente de Correlação de Kendall

Independência estatística

Relevância (pós-processamento)

Variáveis discretas

discreta – discreta

discreta - contínua

Testes de independência estatística

condição: $p(x \cap y) = p(x)p(y)$

teste: χ^2 qui – quadrado

Mas o coeficiente de correlação de Pearson sempre dá alguma informação

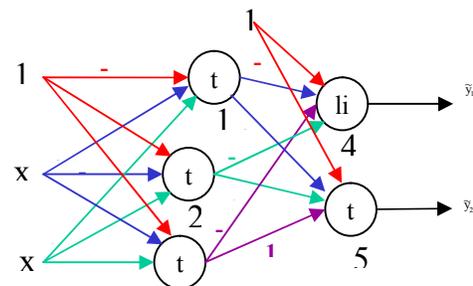
R. Deep, “Probability and Statistics”, Elsevier, 2006

P.A. Barbeta, M.M. Reis, A.C. Bornia, “Estatística para Cursos de Engenharia e Informática”, Atlas, 2004.

1.1 Entradas relevantes faltantes

Tabela de pares entrada-saída

Par	x1	x2	x3	x4	y
1	7,73	-1,57	1,38	4,56	4,62
2	9,19	-1,80	2,84	4,58	2,20
3	9,32	-1,25	2,63	6,63	3,49
4	7,40	-0,64	1,47	6,19	5,14
5	12,49	-0,07	1,09	2,93	3,24
6	8,18	-2,01	2,47	5,87	4,21
7	7,40	-0,64	1,47	5,20	4,23
8	9,11	-2,91	2,93	5,99	2,92
9	9,72	-1,77	1,33	3,37	3,59
10
11



$$y = f(x_1, x_2, x_3, x_4)$$

Rede Neural

entradas: x_1, x_2, x_3, x_4

$$\tilde{y} = f(x_1, x_2, x_3)$$

pares 4 e 7 Saída: y ?

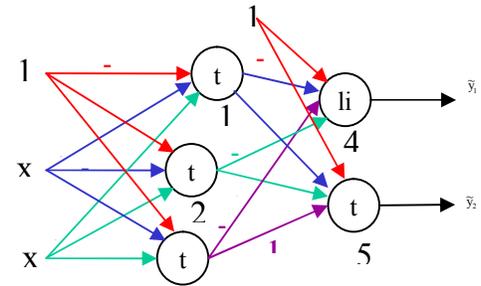
1.2 Entradas não relevantes presentes

$$\Delta_p w_{ij} = \alpha v_j \delta_i$$

$$E(\Delta w_{ij}) = \frac{E(\Delta_p w_{ij})}{p} = \alpha \frac{E(v_j \delta_i)}{p} = 0$$

porque $r(v_j \delta_i) = 0$

então $E(w_{ij}) = w_{ij \text{ inicial}}$ não treinam

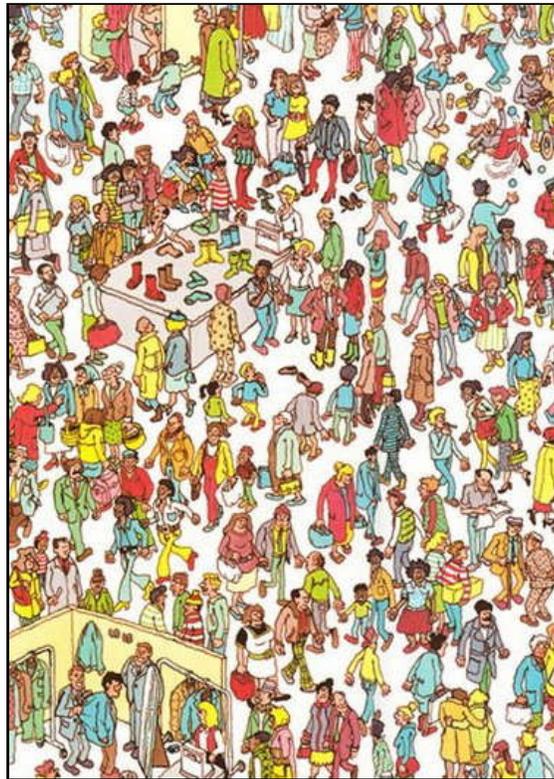


atuam como ruído, degradam a aproximação

Relevância –

efeito de dados não relevantes

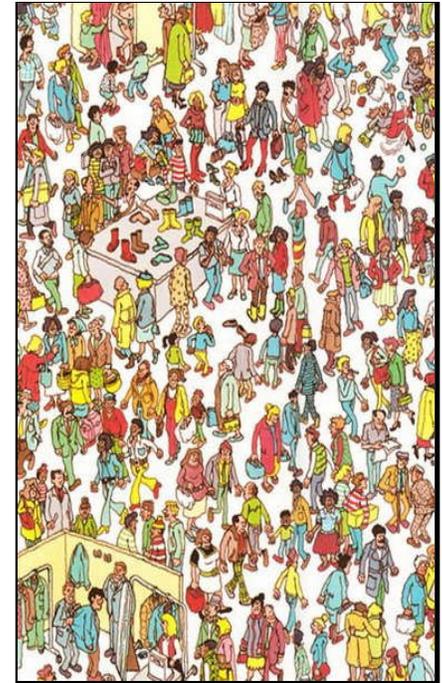
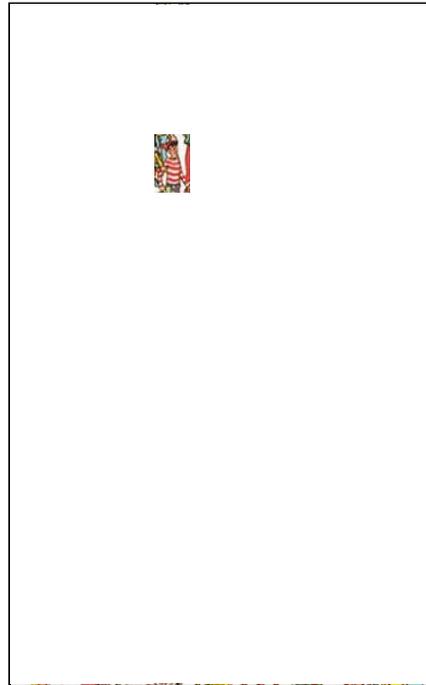
Onde está Wally ?



Onde está Wally?

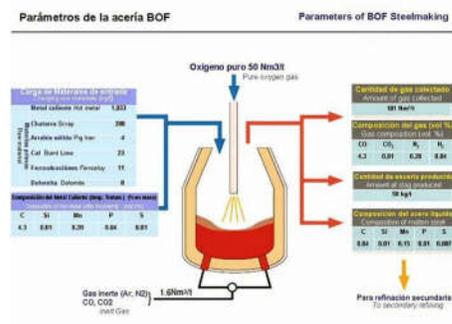
Dados irrelevantes dificultam o treinamento e a operação.

Eventualmente prejudicam a operação com a introdução de ruído.



Lixo na entrada, lixo na saída !

Conversor Siderúrgico



Entradas	Saídas
Peso de gusa	Houve projeção ?
Temperatura do gusa	Temperatura final
Peso de minério + sucata	% carbono
Peso de calcita	% enxofre
Volume de oxigênio	% fósforo
Temperatura da corrida anterior	
Tempo decorrido da corrida anterior	
Temperatura ambiente	
etc.	

2 – Compactação / Parametrização das variáveis

Informação redundante:

Ex: Voz, Imagens, Sonar, etc.

pode reduzir o ruído mas

dificulta e torna lento o treinamento e a operação

Compactar e/ou parametrizar entradas muito redundantes

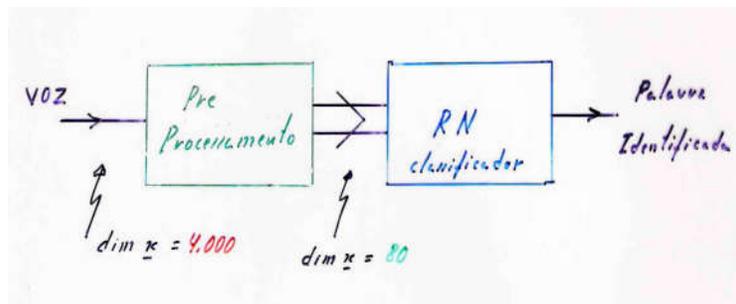
Processos de parametrização:

1 – Baseados na Fenomenologia

EX: Voz >> Formantes,
biológica

Cepstrum, etc.

Taxonomia



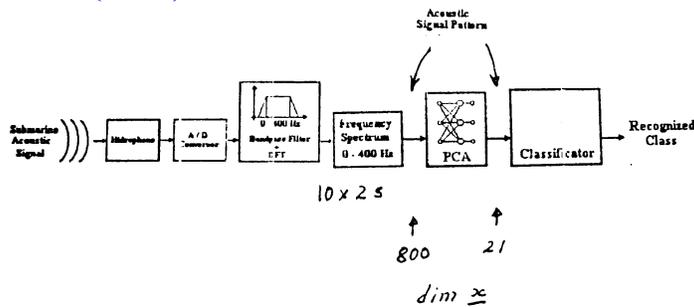
Processos de Compressão:

2 - Transformadas Matemáticas:

Fourrier, Wavelets, QV, etc.

PCA, PCA generalizadas, ICA, etc.

Sonar (PCA)



Imagens (QV – 1x30)



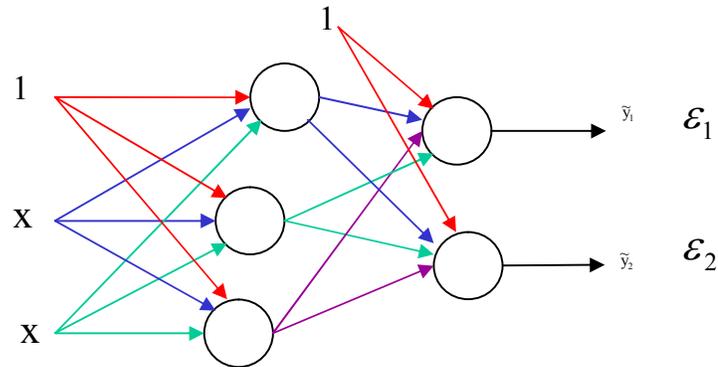
2.1 Invariância, Insensibilidade

Imagens Translação
 Escala
 Rotação

Insensibilidade - conteúdo da voz vs. locutor, etc.

3 – Escalamento das Variáveis

**fundamental para o bom condicionamento
do processo numérico de otimização**



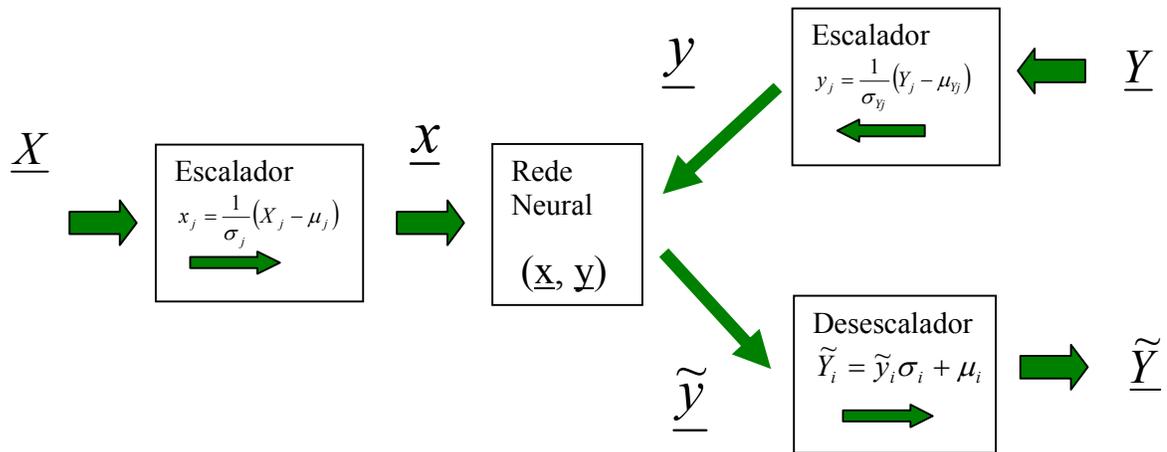
$$\Delta_p w_{ij} = \alpha v_j \delta_i \qquad \Delta w_{ij} = E_p(\Delta_p w_{ij})$$

$X_i =$ variável original >>> $x_i =$ variável escalada

critérios: média nula

maioria dos valores no intervalo (-1,+1)

Treinamento e Operação da Rede com variáveis normalizadas



3.1 - Variáveis quantitativas:

contínuas (e.g. temperatura, comprimento)

discretas (e.g. número de filhos) – representável por variável contínua

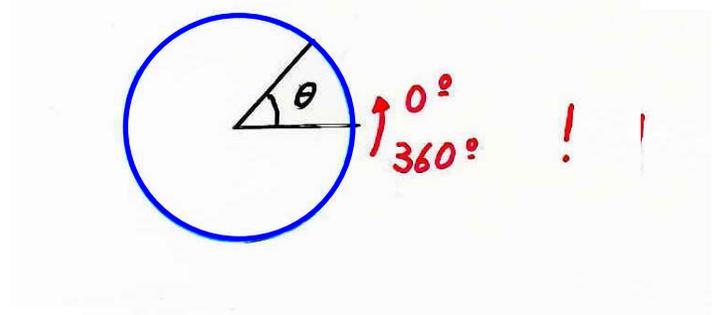
usar escalamento estatístico:

$$x_i = \frac{1}{\sigma_{x_i}} (X_i - \mu_{x_i}) \quad \gg \gg \quad \mu_x = 0 \text{ e } \sigma_x = 1$$

Obs 1:

– Variáveis cíclicas

**Não introduzir descontinuidades
abruptas em variáveis
originalmente
contínuas**



X: 0 – 360° >>>>> **x: (sen 2πX/360 ; cos 2πX/360)**

X: 0 – 24 h. >>>>> **x: (sen 2πX/24 ; cos 2πX/24)**

X: 0 – 12 meses >>>>> **x: (sen 2πX/12 ; cos 2πX/12)**

Obs 2:

- Variáveis quantitativas contínuas cobrindo faixas muito extensas (várias décadas)

Escalamento não linear: as variáveis podem ser comprimidas em uma escala logarítmica antes da normalização.

$$x_i = \frac{1}{\sigma_{\ln X_i}} (\ln X_i - \mu_{\ln X_i}) \quad \gg \gg \quad \mu_x = 0 \text{ e } \sigma_x = 1$$

Note que utilizar log transforma a minimização do erro médio quadrático na minimização do erro relativo médio quadrático.

3.2 Variáveis categóricas

binárias (e.g. $X_i \in \{\text{frio, quente}\}$ ou $X_i \in \{\text{feio, bonito}\}$)

$$x_i \in \{-1, +1\}$$

nominais (e.g. $X_i \in \{\text{solteiro, casado, separado, viuvo}\}$)

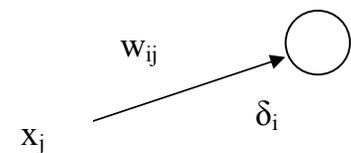
x_i em notação binária
maximamente esparsa

$$\text{e.g. } \underline{x}_i \in \left\{ \begin{array}{l} \begin{bmatrix} +1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ +1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ +1 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ -1 \\ +1 \end{bmatrix} \end{array} \right\}$$

Variáveis binárias - Uso na entrada

$$\Delta w = 2 \alpha x_j \delta_i$$

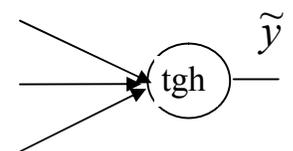
$x_j=0 \gg w_{ij}$ não treina !



Variáveis binárias – Interpretação da saída

$$y \in \{-1, +1\} \quad \tilde{y} = \text{tgh}(u) \in (-1, +1)$$

$$\tilde{y}_{\text{lógico}} = \text{sign}(\tilde{y})$$



Erro médio quadrático

Erro de classificação $\frac{1}{2} |y - \tilde{y}_{\text{lógico}}|$

Obs: Erro médio quadrático x Erro de classificação

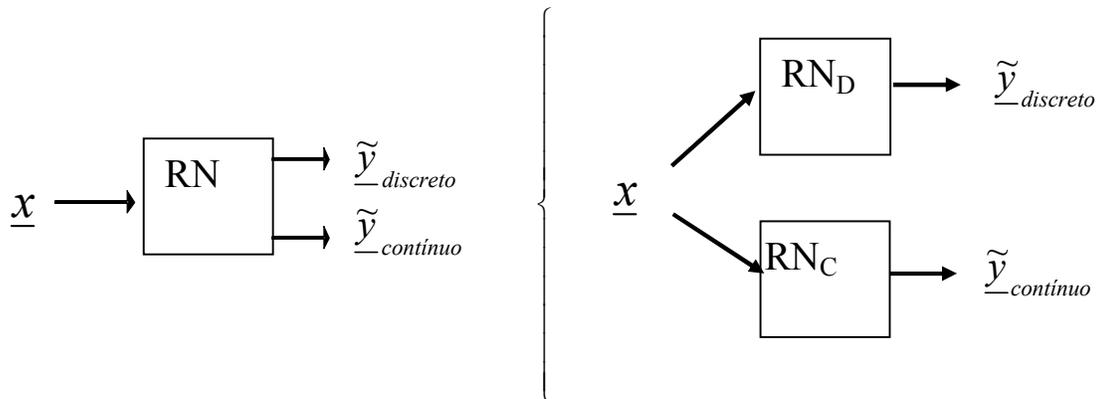
Em casos críticos

não utilizar saídas contínuas e discretas na mesma rede,

devido aos conceitos de erro que realmente importam

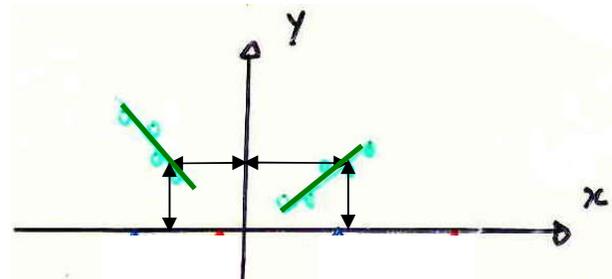
(emq para saídas contínuas e erro de classificação para saídas binárias)

serem muito distintos e serem minimizados por um único critério (emq) !

**4 - Pares entrada – saída:****4.1 Tipo de mapeamento**

$$\underline{x}_1 \rightarrow \underline{y}_1 \quad \underline{x}_2 \rightarrow \underline{y}_2$$

Mapeamento unívoco



se $\underline{x}_2 = \underline{x}_1$ então $\underline{y}_2 = \underline{y}_1$

à menos do ruído. Na prática: $\underline{y}_2 = \underline{y}_1 + \underline{r}$

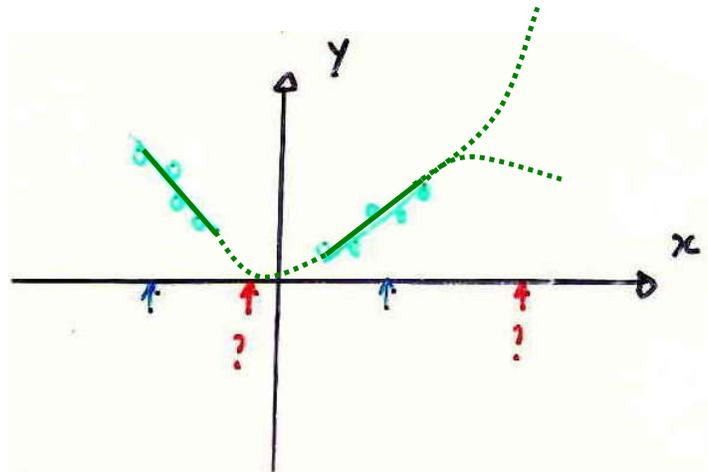
Mapeamento bi-unívoco ?

se $\underline{y}_2 = \underline{y}_1$ então $\underline{x}_2 = \underline{x}_1$? Não é necessário.

4.2 Domínio dos pares – População local

Número de pares entrada-saída

deve caracterizar estatisticamente bem o domínio de operação



A rede só aprende o que treinou !

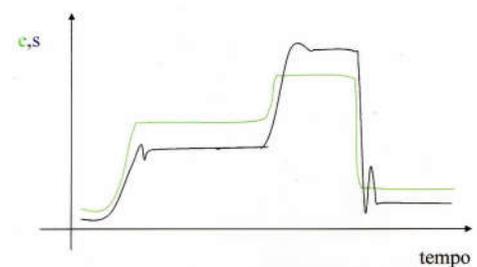
População local

Efeito de população local reduzida:

Ex: Cartão de Crédito

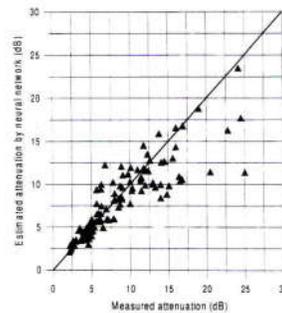
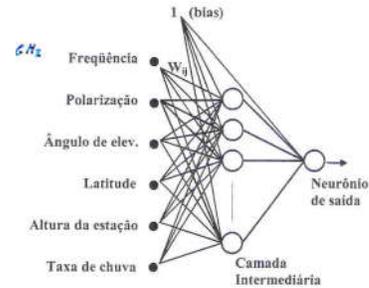
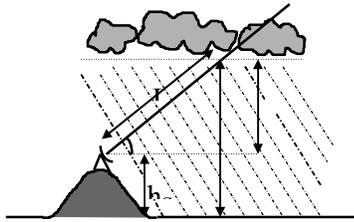
Confiáveis 99 %
Não confiáveis 1 %

Planta de produção



- Efeito ? alto erro na região de baixa população
- Correção ? replicar população das classes (ou regiões) de baixa população (mas isto não aumenta a generalização !)

Ex: Atenuação de sinal em enlace terra-satélite

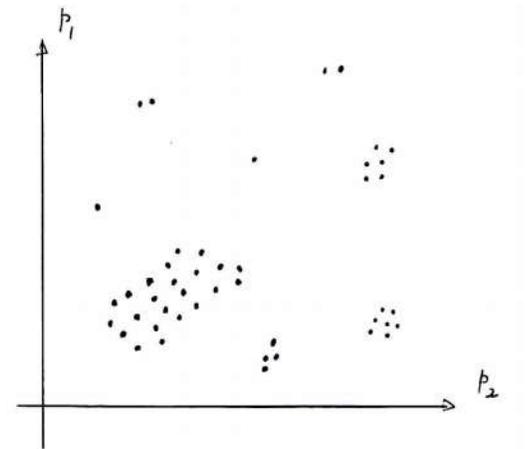


Clusterizar o espaço \underline{e} = entrada + saída

$$\underline{e} = \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix}$$

Balançar a população em cada cluster e

Montar os conjuntos de treinamento e teste usando todos os clusters



Modelos	\sqrt{F}	%
UIT-R	35.1	
RN 1 (pop. estocástica)	29.6	
RN 2 (pop. balanceada)	20	

4.3 Intrusos (outlayers, outsiders) - Pontos errôneos

Como detetar ?

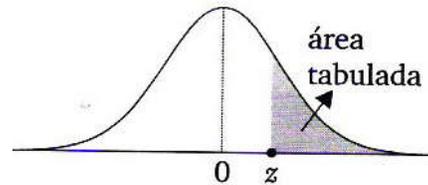
Primeira análise: valor das variáveis

Se a distribuição de uma variável x_i normalizada $\mu_i = 0$ e $\sigma_i = 1$ é normal com P elementos

o número n esperado de eventos com valores com $x > z^*$ (ou $x < z^*$) é:

$$n/P = \text{erf}(z^*)$$

para $n = 1$ $z^* = \text{erf}^{-1}(1/P)$



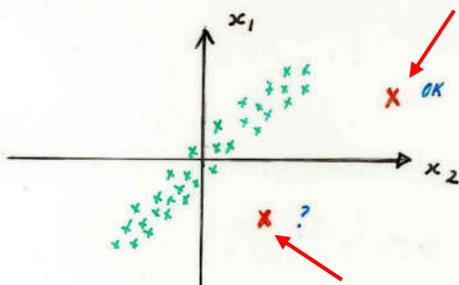
Se o número real de ocorrências no intervalo for significativamente maior que o número previsto

então provavelmente existem intrusos no intervalo. →

Intrusos - deteção

Inspeção em cada variável isoladamente

$$x_i > 3 \sigma_{xi} \rightarrow \text{provável intruso}$$



Melhor Análise:

- Clusterização no espaço

\underline{e} = entrada + saída

$$\underline{e} = \begin{bmatrix} x \\ y \end{bmatrix}$$

descorrelacionar as componentes de \underline{e}

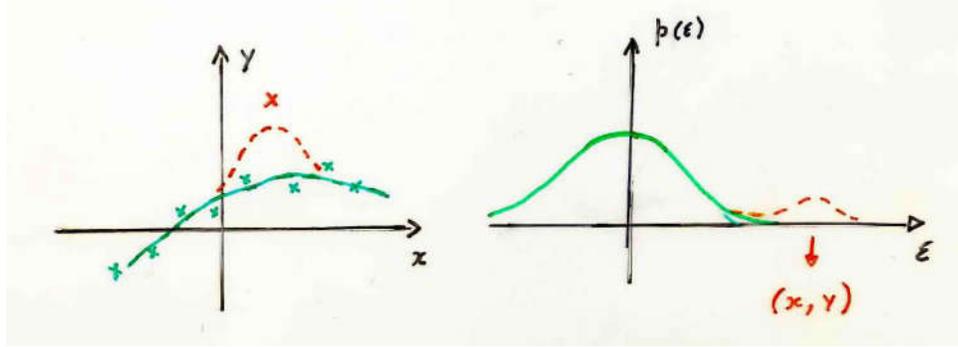
e.g. PCA \underline{e} \underline{z}

$$z_i > 3 \sigma_{zi} \rightarrow \text{provável intruso}$$

Intrusos – possíveis efeitos no erro

$$(\underline{x}, \underline{y}) \longrightarrow \underline{\tilde{y}} \longrightarrow \varepsilon = |\underline{y} - \underline{\tilde{y}}|$$

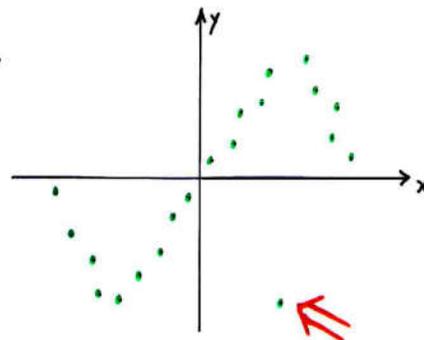
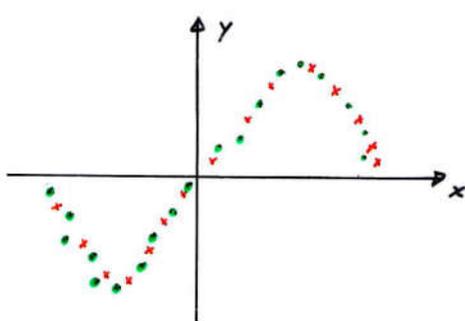
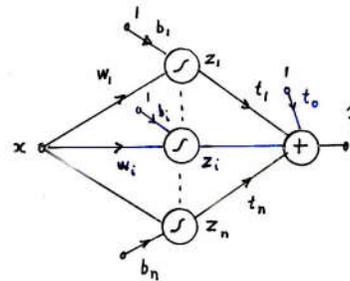
$$\varepsilon_k > 3\sigma_\varepsilon \begin{matrix} \xrightarrow{\text{green}} \\ \xleftarrow{\text{red}} \end{matrix} \begin{matrix} \text{Possível} \\ \text{Intruso} \end{matrix}$$

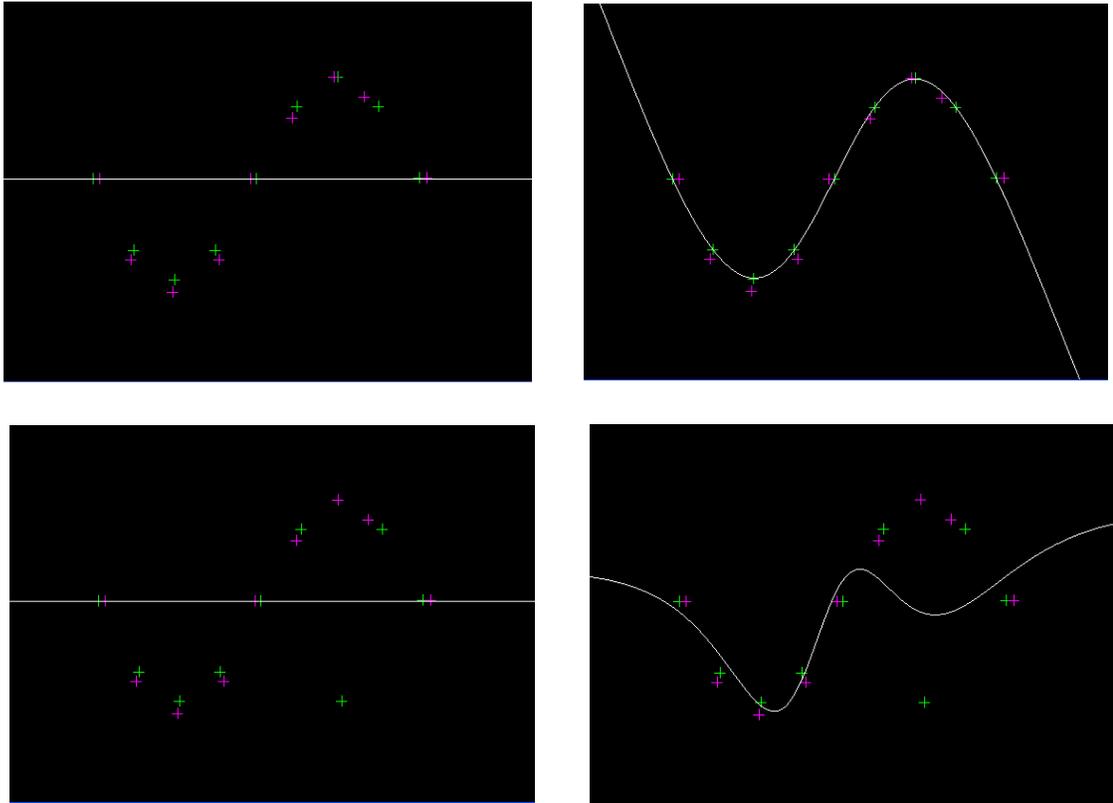


Intrusos versus Regiões de baixa população

Demo: efeito dos intrusos

$$\begin{matrix} \underline{y} = \varphi(\underline{x}) & \underline{\tilde{y}} = \varphi(\underline{x}) \\ y = \varphi(x) & \tilde{y} = \varphi(x) \end{matrix}$$





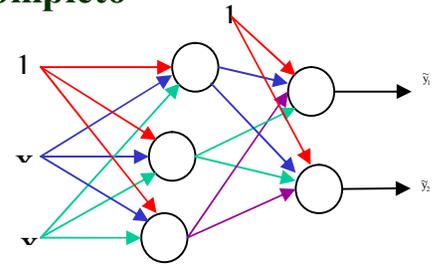
4.4 Pares incompletos, com componentes faltando

Pares $(\underline{x}^k, \underline{y}^k)$ com componentes faltando

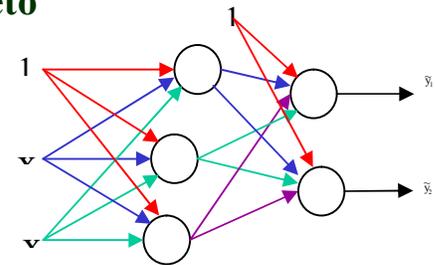
falta x_i^k ou y_i^k em alguns pares

No treinamento:**se possível não utilize o par incompleto**

caso contrário

se falta x_i^k substitua x_i^k por $\mu(x_i) = 0$ (não treina as sinapses conectadas a x_i e não influi nas demais)**se falta y_i^k faça $\varepsilon_i^k = 0$** (não treina as sinapses conectadas ao i -ésimo neurônio da camada de saída e não influi nas demais)**Na operação:****se possível não utilize o par incompleto**

caso contrário

se falta x_i^k substitua x_i^k por $\mu(x_i) = 0$ **Se a rede irá operar com dados faltantes, então****treine com dados faltantes**

Dados Faltantes - exemplos

Caso 1 – Classificação de Vinhos

3 tipos de vinhos maduros da região de Gênova

População	
Classe 1	59
Classe 2	71
Classe 3	48



Atributos (entradas)

Atributo/ Característica	Descrição – Tipo de análise armazenada
1	Álcool
2	Ácido Málico
3	Cinza
4	Alcalinidade de cinza
5	Magnésio
6	Fenóis totais
7	Flavonóides
8	Fenóis não-flavonóides
9	Proanthocyanins
10	Intensidade de cor
11	Tonalidade
12	Diluição do vinho em OD280/OD315
13	Prolina (aminoácido)

Correlação entre atributos

	Saída	Ent 1	Ent 2	Ent 3	Ent 4	Ent 5	Ent 6	Ent 7	Ent 8	Ent 9	Ent10	Ent11	Ent12	Ent13
Saída	1.0000	-0.3282	0.4378	-0.0496	0.5179	-0.2092	-0.7192	-0.8475	0.4891	-0.4991	0.2657	-0.6174	-0.7882	-0.6337
Ent1	-0.3282	1.0000	0.0944	0.2115	-0.3102	0.2708	0.2891	0.2368	-0.1559	0.1367	0.5464	-0.0717	0.0723	0.6437
Ent2	0.4378	0.0944	1.0000	0.1640	0.2885	-0.0546	-0.3352	-0.4110	0.2930	-0.2207	0.2490	-0.5613	-0.3687	-0.1920
Ent3	-0.0496	0.2115	0.1640	1.0000	0.4434	0.2866	0.1290	0.1151	0.1862	0.0097	0.2589	-0.0747	0.0039	0.2236
Ent4	0.5179	-0.3102	0.2885	0.4434	1.0000	-0.0833	-0.3211	-0.3514	0.3619	-0.1973	0.0187	-0.2740	-0.2768	-0.4406
Ent5	-0.2092	0.2708	-0.0546	0.2866	-0.0833	1.0000	0.2144	0.1958	-0.2563	0.2364	0.2000	0.0554	0.0660	0.3934
Ent6	-0.7192	0.2891	-0.3352	0.1290	-0.3211	0.2144	1.0000	0.8646	-0.4499	0.6124	-0.0551	0.4337	0.6999	0.4981
Ent7	-0.8475	0.2368	-0.4110	0.1151	-0.3514	0.1958	0.8646	1.0000	-0.5379	0.6527	-0.1724	0.5435	0.7872	0.4942
Ent8	0.4891	-0.1559	0.2930	0.1862	0.3619	-0.2563	-0.4499	-0.5379	1.0000	-0.3658	0.1391	-0.2626	-0.5033	-0.3114
Ent9	-0.4991	0.1367	-0.2207	0.0097	-0.1973	0.2364	0.6124	0.6527	-0.3658	1.0000	-0.0252	0.2955	0.5191	0.3304
Ent10	0.2657	0.5464	0.2490	0.2589	0.0187	0.2000	-0.0551	-0.1724	0.1391	-0.0252	1.0000	-0.5218	-0.4288	0.3161
Ent11	-0.6174	-0.0717	-0.5613	-0.0747	-0.2740	0.0554	0.4337	0.5435	-0.2626	0.2955	-0.5218	1.0000	0.5655	0.2362
Ent12	-0.7882	0.0723	-0.3687	0.0039	-0.2768	0.0660	0.6999	0.7872	-0.5033	0.5191	-0.4288	0.5655	1.0000	0.3128
Ent13	-0.6337	0.6437	-0.1920	0.2236	-0.4406	0.3934	0.4981	0.4942	-0.3114	0.3304	0.3161	0.2362	0.3128	1.0000

Resultados obtidos com o Banco de Vinhos.

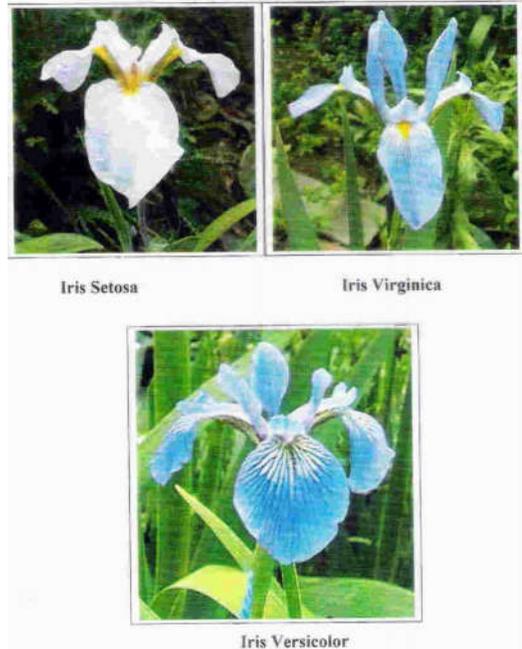


Dados faltantes no treinamento %	Dados faltantes no teste %	Percentual de acerto %
0	0	100
50	50	92
0	50	87

Caso 2 – Iris

Correlação entre atributos:

	sepal length	sepal width	petal length	petal width
sepal length	1.000			
sepal width	-0.118	1.000		
petal length	0.872	-0.428	1.000	
petal width	0.818	-0.366	0.963	1.000



Resultados obtidos com o Banco da Íris

Dados faltantes no treinamento %	Dados faltantes no teste %	Percentual de acerto %
0	0	95
25	25	86
50	50	83
0	50	74