Pré-processamento dos dados

Preparação dos dados de entrada e saída

1 - Escolha das variáveis

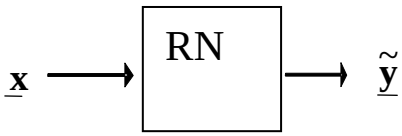
2 – Compactação / Parametrização das variáveis

3 – Escalamento das variáveis

4 - Pares entrada – saída:

Preparação dos dados de entrada e saída

1 - Escolha das Variáveis



saída y – as que desejamos

entrada x – as necessárias para gerar as saídas

"Como entrada escolha as variáveis relevantes,
todas as variáveis relevantes
e somente as variáveis relevantes"

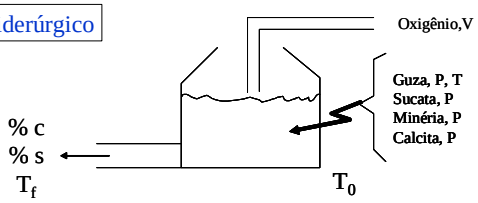
Relevância - como saber se uma entrada é relevante ?

Fenomenologia >>>> candidatas à relevantes

Correlação com as variáveis de saída >>>>

Relevância (pós-processamento)

Conversor Siderúrgico

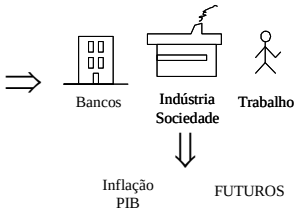


Entradas	Saídas
Peso de gusa	Houve projeção ?
Temperatura do gusa	Temperatura final
Peso de minério + sucata	% carbono
Peso de calcita	% enxofre
Volume de oxigênio	% fósforo
Temperatura da corrida anterior	
Tempo decorrido da corrida anterior	
Temperatura ambiente	
etc.	

Sistema Econômico

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

PASSADOS



800 séries de variáveis.

Quais usar ?

Independência / Dependência Estatística entre Variáveis

Coeficiente de Correlação de Pearson

$$Pares\ (x_i, y_i) \qquad i = 1,\ldots,P$$

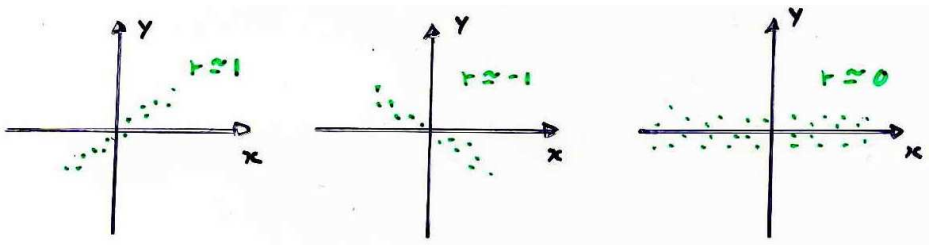
$$r(x,y) = \frac{\frac{1}{P-1}\sum_{i=1}^{P}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x\,\sigma_y} \qquad onde$$

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

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

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

$$-1 \le r \le 1$$



Valores randômicos **correlação r = 0**

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

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

95% confiança na correlação

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

Matrizes de correlação

entradas – saídas

	y_1	...	y_m
x_1	r_{1y1}		r_{1ym}
x_2	r_{2y1}		r_{2ym}
x_3	r_{3y1}		r_{3ym}
...			
x_n	r_{ny1}		r_{nym}

entre entradas (simétrica)

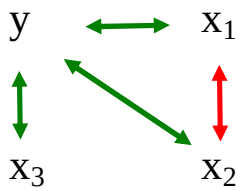
x_1	x_2	x_3	...	x_n
1	r_{12}	r_{13}		r_{1n}
r_{12}	1	r_{23}		r_{2n}
r_{13}	r_{23}	1		r_{3n}
...				
r_{1n}	r_{2n}	r_{3n}		1

completa

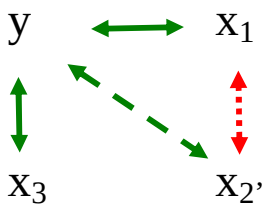
	x_1	x_2	x_3	...	x_n	y_1	...	y_m
x_1	1	r_{12}	r_{13}		r_{1n}	r_{1y1}		r_{1ym}
x_2	r_{12}	1	r_{23}		r_{2n}	r_{2y1}		r_{2ym}
x_3	r_{13}	r_{23}	1		r_{3n}	r_{3y1}		r_{3ym}
...								
x_n	r_{1n}	r_{2n}	r_{3n}		1	r_{ny1}		r_{nym}

Eliminando variáveis “caras”

Descorrelação entre entradas:

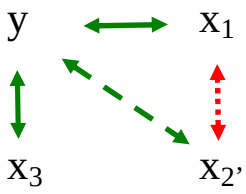


	x_1	x_2	x_3	Y
x_1	1	r_{12}	r_{13}	r_{1y}
x_2	r_{12}	1	r_{23}	r_{2y}
x_3	r_{13}	r_{23}	1	r_{3y}



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

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



	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 Silva[1], Marcio H. S. Siqueira[1], Luiz P. Calôba[2], Ivan C. da Silva[1], Antonio A. de Carvalho[1] and João Marcos A. Rebello[1].

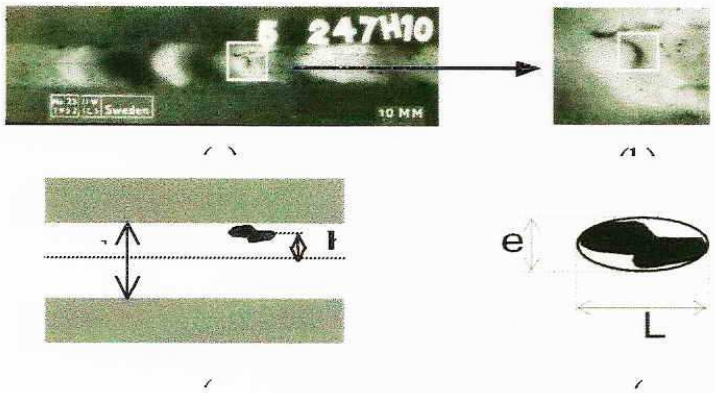


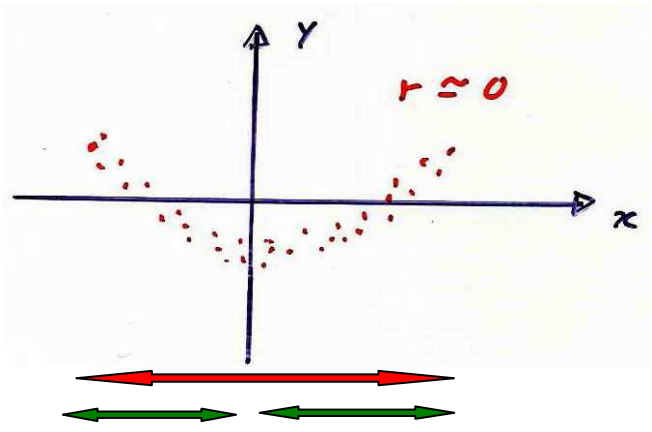
Table 1. Correlation matrix with correlated parameters.

	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

Coeficiente de Correlação de Spearman

Coeficiente 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: $\chi^2 \quad qui - quadrado$

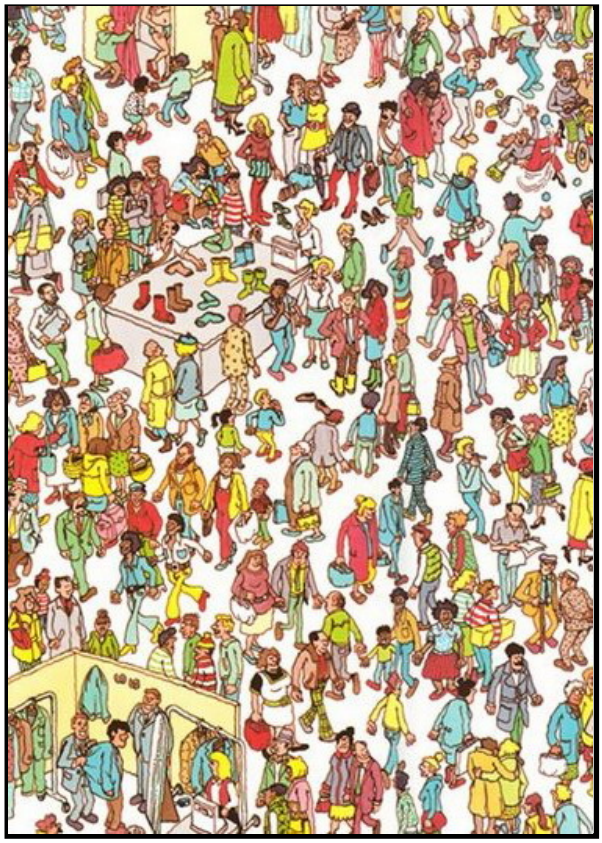
Mas o coeficiente de correlação de Pearson sempre dá alguma informação

R. Deep, “Probability and Statistics”, Elsewier, 2006
P.A. Barbetta, M.M. Reis, A.C. Bornia, “Estatística para Cursos de Engenharia e Informática”, Atlas, 2004.

Relevância –

efeito de
dados não relevantes

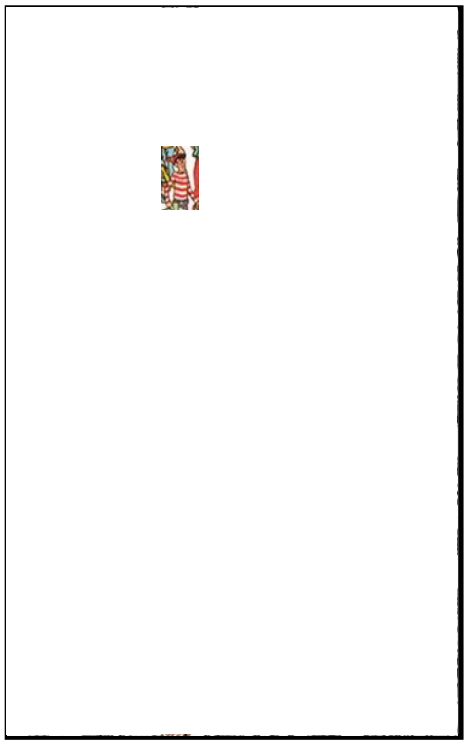
Onde está Wally ?

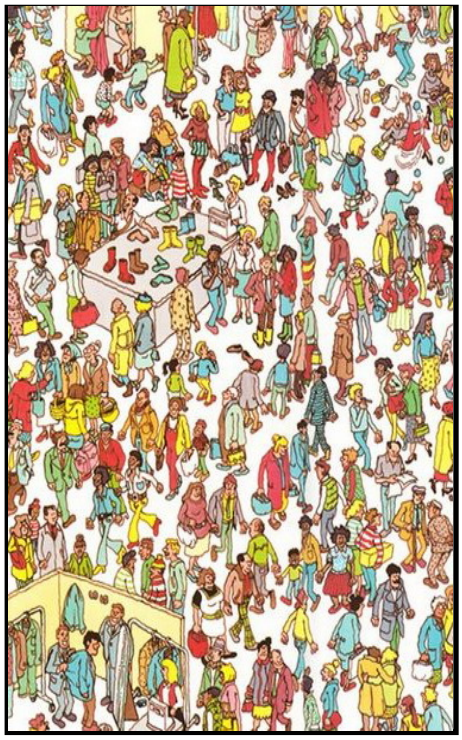


Onde está Wally?

Dados irrelevantes dificultam o treinamento e a operação.

Eventualmente prejudicam a operação com a introdução de ruído.





Lixo na entrada, lixo na saída !

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

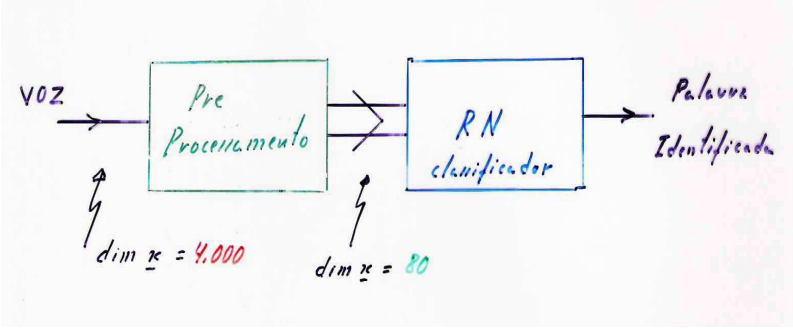
Compactar e/ou parametrizar entradas muito redundantes

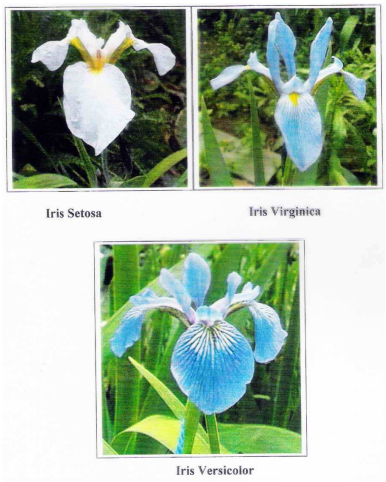
Processos de parametrização:

1 – Baseados na Fenomenologia

EX: Voz >> Formantes, Taxonomia biológica

Cepstrum, etc.





Processos de Compressão:

2 - Transformadas Matemáticas:

Fourrier, Wavelets, QV, etc.

PCA, PCA generalizadas, ICA, etc.

Sonar (PCA)

Imagens (QV – 1x30)

Invariância, Insensibilidade

Imagens Translação
Escala
Rotação

Insensibilidade - conteúdo da voz vs. locutor, etc.

3 – Escalamento das Variáveis

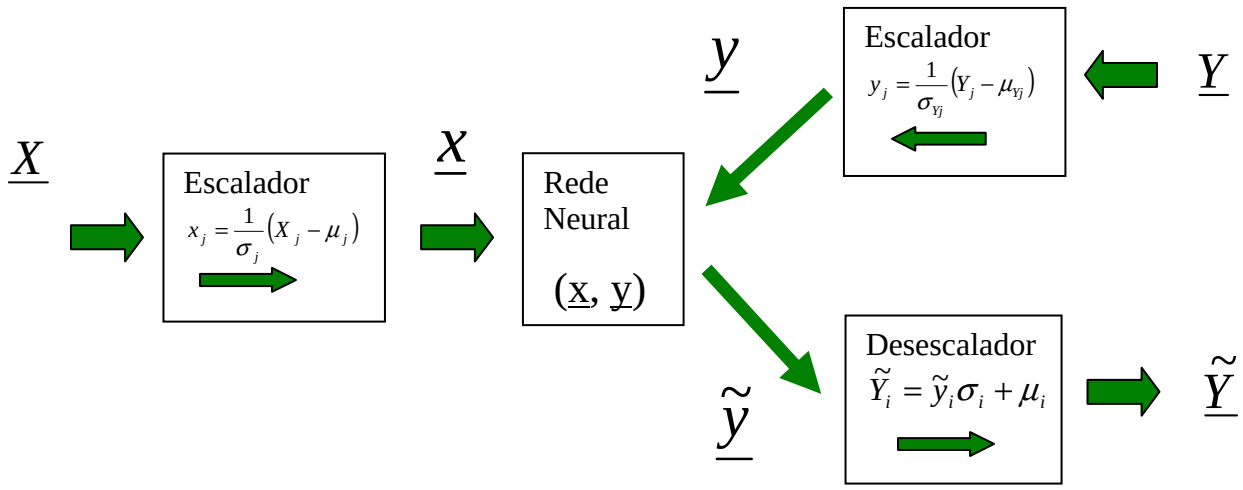
fundamental para o bom condicionamento
do processo numérico de otimização

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

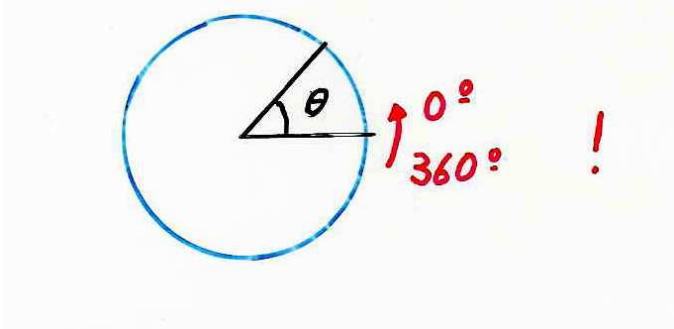
usar escalamento estatístico:

$$x_i = \frac{1}{\sigma_{X_i}} \left(X_i - \mu_{X_i} \right) \quad >>> \quad \mu_x = 0 \;\; e \;\; \sigma_x = 1$$

Obs 1:

– **Variáveis cíclicas**

Não introduzir descontinuidades abruptas em variáveis originalmente contínuas



X: 0 – 360º >>>>> **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.

$$x_i = \frac{1}{\sigma_{\ln X_i}} \left(\ln X_i - \mu_{\ln X_i} \right) \quad >>> \quad \mu_x = 0 \;\; e \;\; \sigma_x = 1$$

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. $X_i \in$ {frio, quente} ou $X_i \in$ {feio, bonito})

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

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

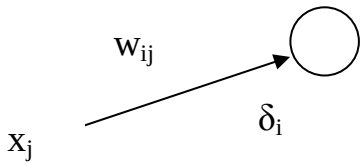
x_i em notação binária
maximamente esparsa

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

Variáveis binárias - Uso na entrada

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

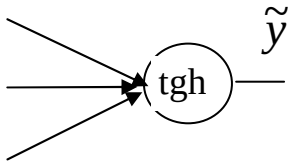
x_j=0 >> w_{ij} não treina !



Variáveis binárias – Interpretação da saída

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

$$\tilde{y}_{logico} = sign(\tilde{y})$$



Erro médio quadrático

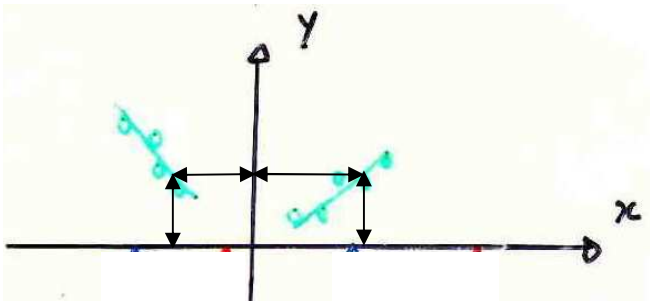
Erro de classificação $\frac{1}{2}\left| y - \tilde{y}_{logico} \right|$

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



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

Mapeamento unívoco

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

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

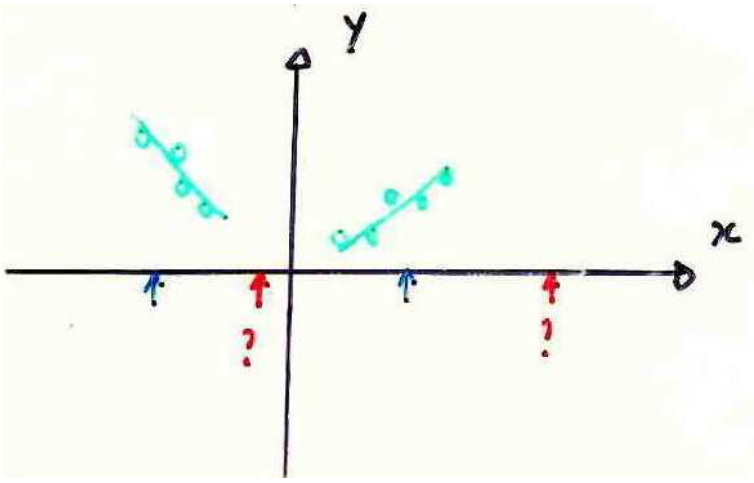
Mapeamento bi-unívoco ?

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 !

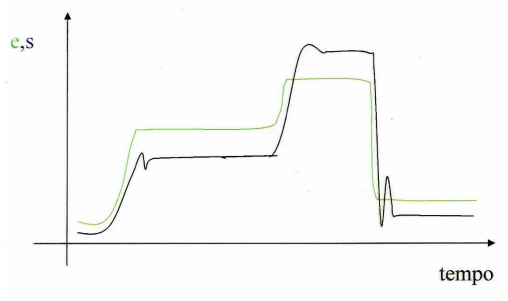
População local

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 !)

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
a probabilidade $p(x_i)$ de ocorrência de valores nos intervalos é:

Faixa	$0 < \|x_i\| < 1$	$1 < \|x_i\| < 1.5$	$1.5 < \|x_i\| < 2$	$2 < \|x_i\| < 2.5$	$2.5 < \|x_i\| < 3$	$3 < \|x_i\|$
$p(x_i)$	**68 %**	**18 %**	**9 %**	**3 %**	**1 %**	**0,3 %**

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

(numero real no intervalo) >>
número previsto = p.(número total eventos)

então provavelmente existem intrusos no intervalo.

Visto de outra forma,

Se

a distribuição de uma variável x com P elementos é normal e

$\mu_i = 0$ e $\sigma_i = 1$

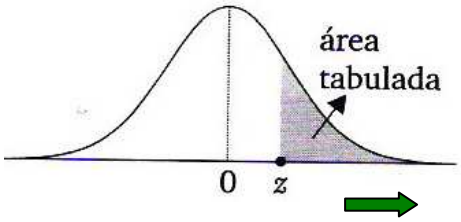
então

o valor z* tal que n ocorrências são esperadas com

$|x| > z^*$

é dado por

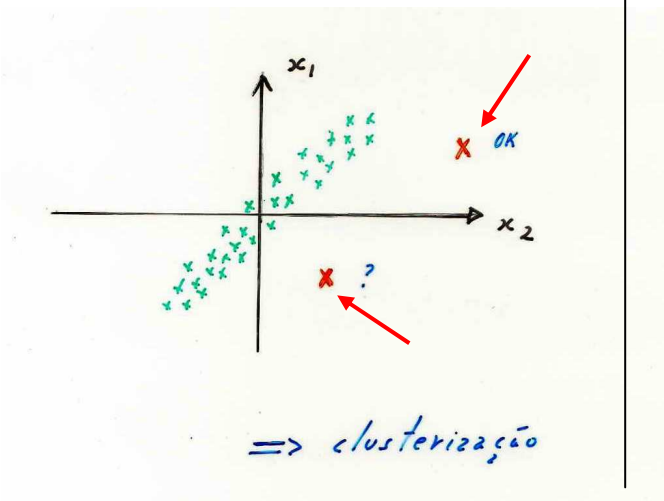
$n/2P = erf(z^*)$



Intrusos - deteção

Inspeção em cada variável isoladamente

$x_i > 3\,\sigma_{xi}$ ➡ **provável intruso**



Melhor Análise:

- Clusterização no espaço

e = entrada + saída

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

descorrelacionar as componentes de e

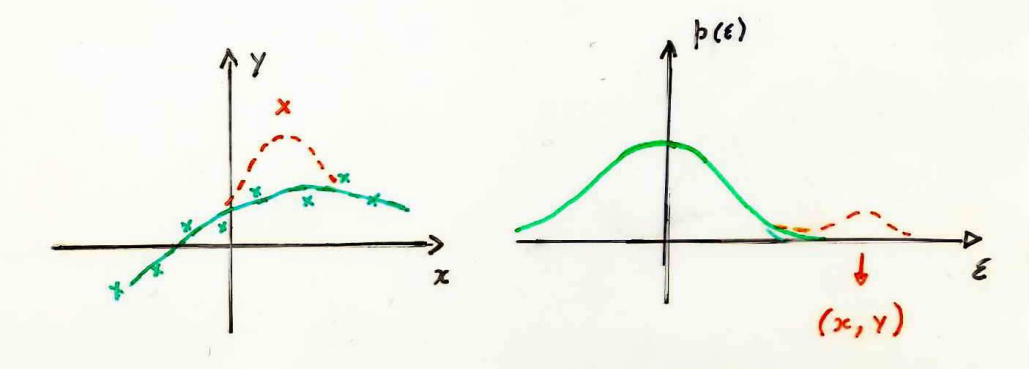
e.g. PCA e z

$z_i > 3\,\sigma_{zi}$ ➡ **provável intruso**

Intrusos – possíveis efeitos no erro

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

$$\varepsilon_k > 3\sigma_\varepsilon \quad \rightleftharpoons \quad \text{Possível Intruso}$$



Intrusos versus Regiões de baixa população

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

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 $\mu(x_i) = 0$

(mas neste caso a rede funcionará melhor se tiver sido treinada com pares incompletos, com componentes faltando)

5 – Obs complementares:

Regra para histogramas - valores tentativos para k

Tamanho da amostra (n)	Número de Classes (k)
$n<100$	$\sqrt{n}$
$n>100$	$5 \log_{10} n$