Pré-processamento dos dados

Preparação dos dados de entrada para treinamento não supervisionado

(parecido – mas não igual – ao das redes feedforward)

1 - Escolha das variáveis de entrada

2 – Compactação / Parametrização das variáveis

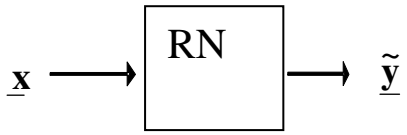
3 – Escalamento das variáveis

4 - Pares entrada – saída

5 – Relevância das entradas

1 - Escolha das Variáveis de Entrada

"Como entrada escolha as variáveis relevantes, todas as variáveis relevantes e somente as variáveis relevantes para a classificação desejada.



Usualmente a classificação depende fortemente das entradas utilizadas

Como saber se uma entrada é relevante ?

Fenomenologia >>>> candidatas à relevantes

Análise estatística (pré-processamento)

Relevância (pós-processamento)

Entradas não relevantes dificultam o treinamento (porque se comportam como ruído) e eventualmente podem gerar novas classes "artificiais", sem interesse.

1.1 - Treinamento Supervisionado

Independência / Dependência Estatística entre Variáveis

Coeficiente de Correlação de Pearson

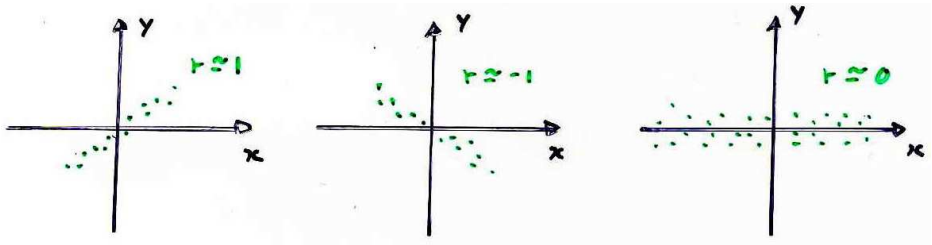
$$Pares\ (x_i, y_i) \qquad i = 1,...,P$$

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

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

embora não seja o mais recomendado, também oferece informação relevante usado com variáveis discretas.

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



Valores randômicos **correlação r = 0**

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

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

95% confiança na correlação

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

Matriz de correlações entradas – saídas

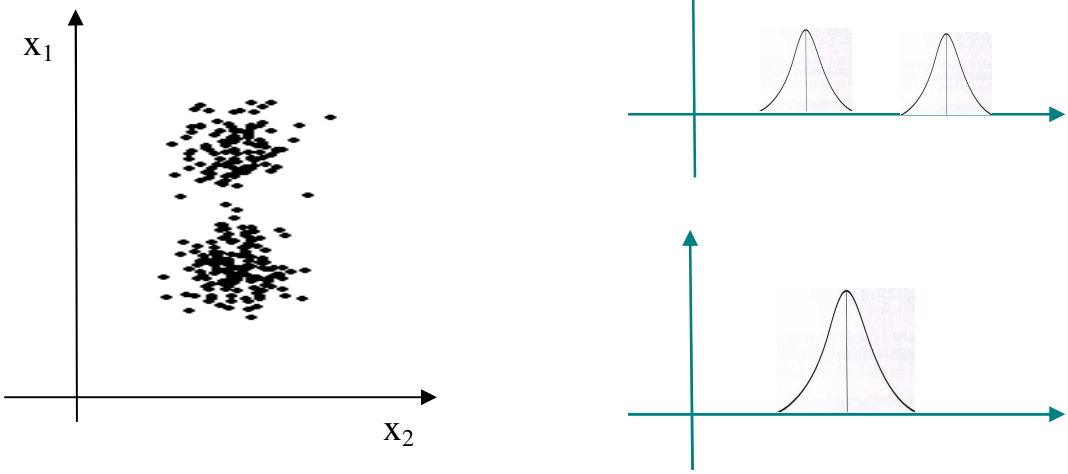
	y_1	...	y_m
x_1	r_{1y1}		r_{1ym}
x_2	r_{2y1}		r_{2ym}
x_3	r_{3y1}		r_{3ym}
...			
x_n	r_{ny1}		r_{nym}

usada para decidir se a variável será utilizada ou não

1.2 - Treinamento não supervisionado

Analisar a distribuição das variáveis em cada dimensão

Variáveis com modas bem diferenciadas no histograma de diferença de valores entre os elementos tem melhor possibilidade de gerar bons agrupamentos.



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

Solução: Compactar e/ou parametrizar entradas muito redundantes.

Mas a parametrização pode alterar a classificação.

2.1 Processos de parametrização / compactação:

– Baseados na Fenomenologia

Ex: Formantes e coeficientes cepstrais para voz,
Parâmetros de taxonomia biológica, etc.

- Transformadas Matemáticas:

Ex: Fourrier, Wavelets, QV, etc.
PCA, PCA generalizadas, ICA, etc.

– Processos de Invariância, Insensibilidade

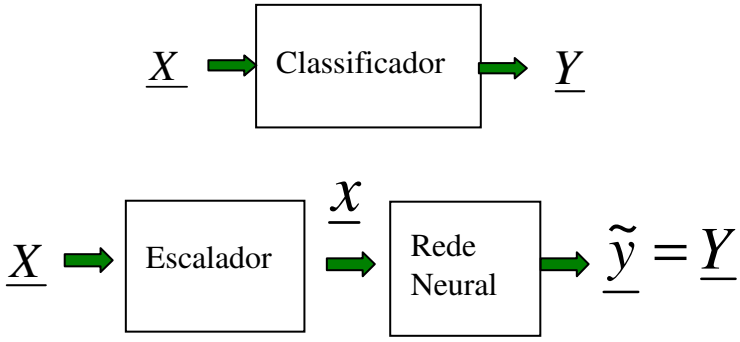
Ex: Invariância à translação, escala e rotação de imagens
Insensibilidade ao locutor na análise de conteúdo de voz, etc.

3 – Escalamento das Variáveis – Variáveis Quantitativas

fundamental para o bom condicionamento do processo numérico

X_i = variável original >>> x_i = variável escalada

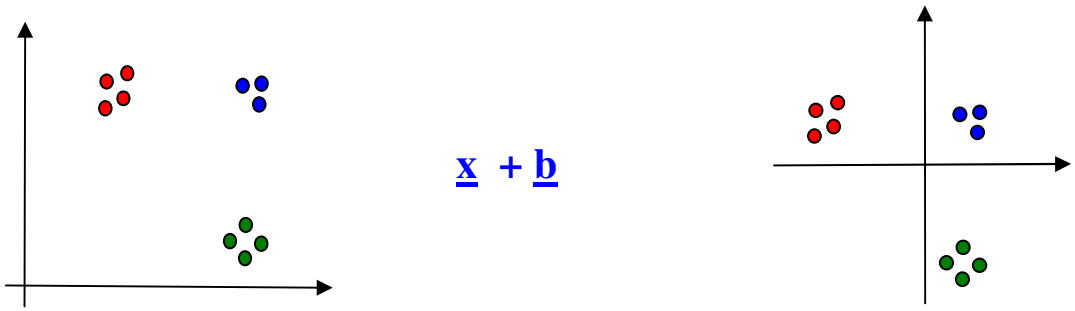
Treinamento e Operação da Rede com variáveis normalizadas



3.1 – Invariâncias e dependências na classificação

Agrupamento "natural":

Insensível à translação



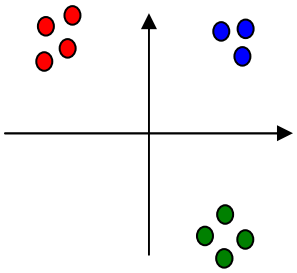
Usar variáveis com média nula

$\underline{x} \Longrightarrow \underline{x} - \underline{\mu_x}$

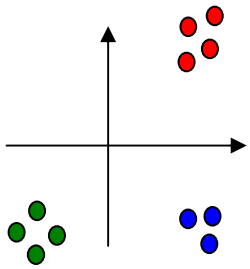
$x_i \Longrightarrow x_i - \mu_{xi}$

Agrupamento "natural":

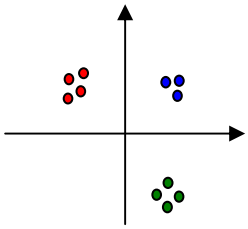
Insensível à



Rotação
R x

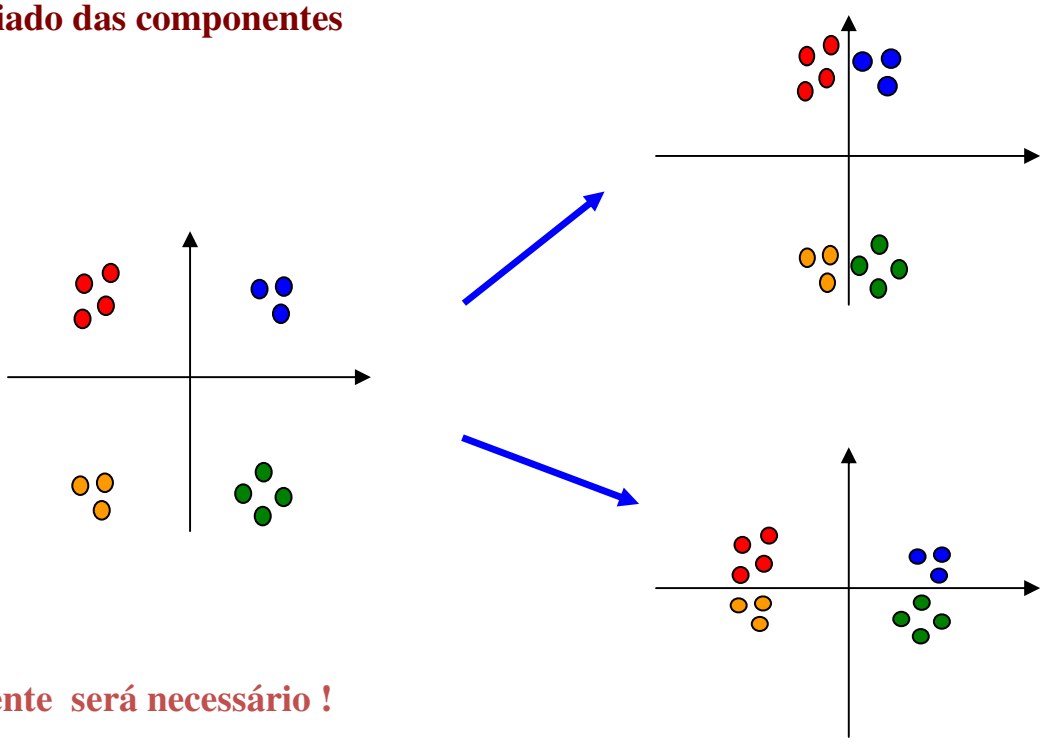


Escala total
k x



Sensível ao escalamento

diferenciado das componentes



Usualmente será necessário !

3.2 - Variáveis quantitativas (entradas), escalamento inicial:

contínuas (e.g. temperatura, comprimento)

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

usar **média nula e dispersão adequada** por variável:

$$x_i = k_i \left(X_i - \mu_{X_i} \right) \qquad onde \quad \mu_{X_i} = \underset{\forall \vec{x}}{E} X_i$$

x_i tem média nula e k_i controla sua dispersão.

Para termos uma visão global inicialmente aplicamos

$$k_i = \frac{1}{\sigma_{X_i}} \qquad onde \quad \sigma_{X_i} = \left[\underset{\forall \vec{x}}{E} \left(X_i - \mu_{X_i} \right)^2 \right]^{\frac{1}{2}}$$

Mas o valor de k_i poderá ser modificado posteriormente no processo de branqueamento do ruído.

3.2.1 Reescalamento - Treinamento Supervisionado

Como as classes são conhecidas

$$k_i = \min_{\forall C_j}[\frac{1}{\sigma_{ji}}] \quad \text{sendo} \quad \sigma_{ji}^2 = \underset{\forall \vec{x} \in C_j}{E}(x_i - \mu_{ji})^2 \quad \text{onde } \mu_{ji} = \underset{\forall \vec{x} \in C_j}{E} x_i$$

onde

$\mu_{ji} =$ **média da componente (direção) i dos elementos da Classe j**

$\sigma_{ji} =$ **desvio padrão da componente (direção) i dos elementos da Classe j**

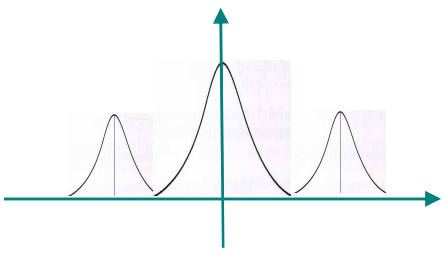
3.2.2 – Reescalamento - Treinamento não Supervisionado

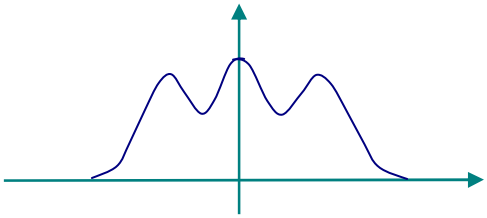
Para simplificar a análise gráfica realizamos inicialmente o escalamento estatístico

$$x_i = \frac{1}{\sigma_i}\left(X_i - \mu_{X_i}\right) \quad \text{onde} \quad \mu_{X_i} = \underset{\forall X_i}{E} X_i \quad \text{e} \quad \sigma_i^2 = \underset{\forall X_i}{E}(X_i - \mu_{X_i})^2$$

e em seguida estimar o valor de $k_{ji} = \frac{1}{\sigma_{ji}}$

para cada dimensão i a partir da análise gráfica de $p(x_i)$ na dimensão i

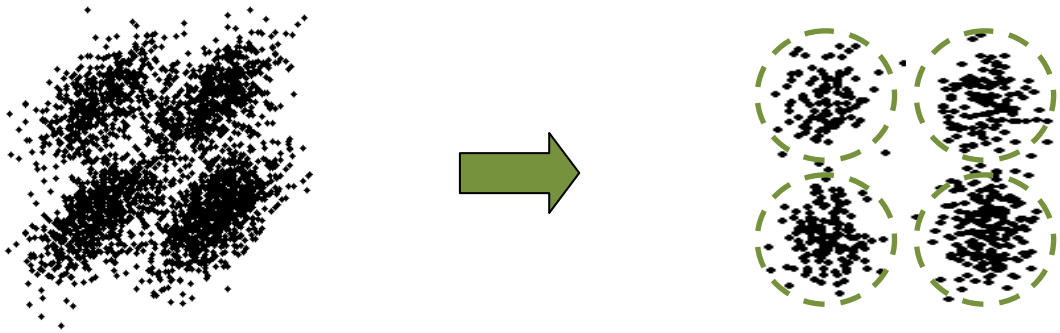




3.3 Branqueamento do Ruído – gerando classes esféricas

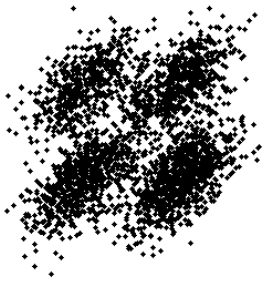
classes sphereing

É comum o ruído nas várias dimensões de x apresentar correlação entre si e ter potências diferentes. Em uma entrada x é dita ter ruído branco quando o ruído em todas as dimensões não apresenta correlação entre si e tem potência igual. O processo de branqueamento do ruído leva a classes esféricas, que permitirá o uso de um classificador muito mais simples. O branqueamento do ruído das entradas é obtido em duas etapas, o descorrelacionamento e a normalização da potência.



3.3.1 - Descorrelação do Ruído

Se os ruídos das componentes forem correlatos os domínios das classes serão elipsoides e estarão contidos por esferas maiores, que incluem espaços da não classe.



A descorrelação dos ruídos r_i pode ser obtida representando os dados x (padrão da classe + ruído) em uma base B do espaço das componentes principais (PCA) do ruído r de cada dado de entrada (PCA dos ruídos, e não dos sinais!). Para tanto necessitamos calcular as PCA de r e a matriz de mapeamento B do ruído r em sua PCA p

$$\vec{x} \in C_j \qquad \vec{x} = \vec{m}_j + \vec{r}_j \qquad \therefore \qquad \vec{r} = \vec{x} - \vec{m}_j$$

Utilizando B determinamos o mapeamento de cada ruído r em sua PCA p

$$\vec{p} = B\,\vec{r} \qquad \text{onde} \qquad [B] = [\vec{b}_1^{\,t}\ \vec{b}_2^{\,t}\ \vec{b}_3^{\,t}\ ...\vec{b}_m^{\,t}] \qquad \vec{x} \Rightarrow \underline{B}\,\vec{x}$$

A multiplicação das entradas x pela matriz B corresponde à uma rotação no espaço das entradas e não influi na disposição das classes entre si, isto é, na classificação.

3.3.2 Normalização das Potências de Ruído nas Diversas Dimensões

A segunda etapa deve ser a equalização da potência do ruído nas diversas dimensões de x. Calculamos a variância de cada componente p_i do ruído

$$\sigma_i^2 = E[p_i]^2 = E[\vec{b}_i^{\,t}\vec{r}]^2$$

A matriz de pesos P para normalização das potências dos ruídos será

$$P = diag\left(\frac{1}{\sigma_i}\right)$$

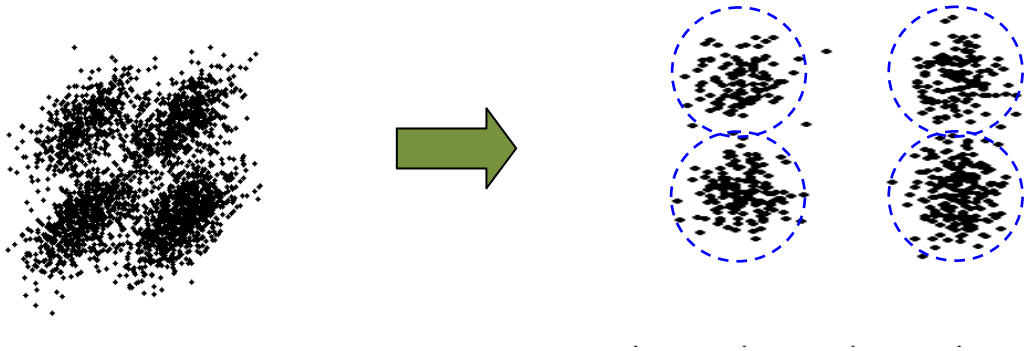
3.3.3 - Novas entradas

As novas entradas à classificar serão

$$\vec{z} = P\,B\,\vec{x} = P\,B\,\vec{m}_j + P\,B\,\vec{r} = \vec{m}_{j\,novo} + \vec{r}_{novo}$$

As classes passam a ter um novo padrão $\vec{m}_{j\,novo}$ **adicionado de um ruído branco** $\vec{r}_{novo}$ **com desvio padrão unitário.**

$$\vec{m}_{j\,novo} = P\,B\,\vec{m}_j \qquad \sigma(\vec{r}_{novo}) = \sigma(P\,B\,\vec{r}) = \vec{1} = [1\ 1\ 1 \ldots\ 1]^t$$



3.3.4 Algumas observações:

- **A multiplicação por B representa uma rotação e por isto não altera a disposição relativa das classes**
- **O uso de PCA sugere a eliminação de componentes pouco relevantes, i.e., com pequenos**

$$\sigma_i^2 = E[p_i]^2 = E[\vec{b}_i^{\,t}\vec{r}]^2$$

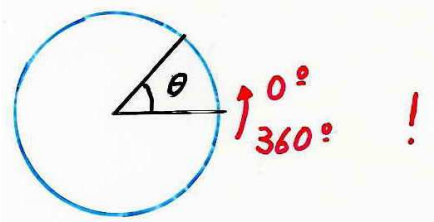
e a consequente redução da dimensionalidade e simplificação do classificador. Entretanto, a eliminação de uma componente principal para $\vec{r}_j$ **somente pode ser feita se for pouco relevante para** $\vec{x}$**, isto é, se**

$$\sigma^2(z_i) = E[\frac{1}{\sigma_i}\vec{b}_i^{\,t}\vec{x}]^2$$ **for muito pequeno comparado com seus pares.**

A relevância para $\vec{x}$ é que determina a eliminação ou não da componente !

- **A multiplicação por P realiza um escalamento diferente por componente e pode alterar a classificação, principalmente se houver grande dispersão entre os σ_i**

3.5 Variáveis cíclicas



Como para o MLP
Não introduzir descontinuidades abruptas em variáveis originalmente contínuas

$$\theta \Rightarrow sen\,\theta,\ \cos\theta$$

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

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

3.6 Variáveis cobrindo faixas muito extensas (várias décadas)

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

Note que utilizar log pode modificar os agrupamentos.

3.7 Variáveis categóricas (entradas e saídas)

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

$x_i \in \{0,1\}$

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

x_i em notação binária maximamente esparsa

e.g. $\underline{x}_i \in \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \right\}$

Ao contrário das redes MLP, aqui não há vantagens em se trabalhar com {-1,+1}

Variáveis categóricas podem influir fortemente na classificação.

4 - Pares entrada – saída:

População por classe

Usualmente pouco crítico, mas para alguns processos classes com pouca população podem não ser aprendidas corretamente

Intrusos (outlayers, outsiders)

Usualmente muito menos críticos que na backpropagation. Para a maioria dos processos simplesmente aparecem como uma classe com um único elemento.

Pares com componentes faltando

Evitar o uso. Caso imprescindível, fazer o treinamento / operação substituindo o componente pelo valor médio (zero), e eventualmente afrouxando a condição de similaridade mínima (aumentando o raio de similaridade).

5 – Conjuntos de Treinamento, Validação e Teste

Validação Cruzada – teste de generalização

trocar treinamento e teste – mesma classificação ?

Processos supervisionados – treinamento, validação e teste

Processos não supervisionados – treinamento e teste