Neural Network Interest Group

Título/*Title*:
Estudo sobre a Capacidade de Aprendizagem
das Redes Neuronais

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 5 /2003

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:
Estudo sobre a Capacidade de Aprendizagem
das Redes Neuronais

Autor(es)/*Author(s)*:

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 5 /2003

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Estudo sobre a Capacidade de Aprendizagem das Redes Neuronais

J.P. Marques de Sá – jmsa@fe.up.pt
INEB – Instituto de Engenharia Biomédica

Bibliografia

Anthony M, Bartlett PL (1999) Neural Network Learning: Theoretical Foundations. Cambridge University Press.

Baum EB, Haussler D (1989) What Size Net Gives Valid Generalization? Neural Computation, 1:151-160.

Bishop CM (1995) Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the Vapnik-Chernovenkis Dimension. J Ass Comp Machinery, 36:929-965.

Cherkassky V, Mulier F (1998) Learning from Data, John Wiley & Sons, Inc.

Ehrenfeucht A, Haussler D, Kearns M, Valiant L (1989) A General Lower Bound on the Number of Examples Needed for Learning. Information and Computation, 82:247-261.

Haykin S (1999) Neural Networks. A Comprehensive Foundation. Prentice Hall Inc., New Jersey.

Kearns MJ, Vazirani UV (1997) An Introduction to Computational Learning Theory. The MIT Press.

Mirchandani G, Cao W (1989) On Hidden Neurons for Neural Nets. IEEE Tr Circ Syst, 36: 661-664.

Mitchell TM (1997) Machine Learning. McGraw Hill Book Co., New York.

Simon HU (1997) Bounds on the Number of Examples Needed for Learning Functions. SIAM J. of Computing, 26:751-763.

Vapnik VN (1998) Statistical Learning Theory. Wiley, New York.

Símbolos

x	objecto (instância, exemplo, caso)
d	número de atributos do objecto
n	número de objectos
w	número de pesos de um MLP
$\mathbf{x}$	vector (d-dimensional)
t	valor alvo
$\hat{t}$	valor estimado de t
X	conjunto de todos os objectos (instâncias)
S	amostra com n objectos escolhidos aleatoriamente
P	Probabilidade discreta
p	fdp
Pe	Probabilidade de erro
$x \in X \sim D$	x extraído de X segundo a distribuição D

1 Aprendizagem PAC

1.1 Questões Centrais da Aprendizagem

Complexidade amostral:

Qual o n = card(X_n) necessário para o algoritmo convergir (com alta probabilidade) para uma aprendizagem eficaz?

Complexidade computacional:

Que esforço computacional é requerido para o algoritmo convergir (com alta probabilidade) para uma aprendizagem eficaz?

Desempenho do algoritmo:

Quantos objectos serão erradamente classificados (erro) até o algoritmo convergir para uma aprendizagem eficaz?

1.2 Definições

X - Domínio das instâncias.

$$X = \text{Conjunto de pessoas}$$

C - Espaço de conceitos, $C \subseteq 2^X$ (conjunto de dicotomias em X)

$$C = \{\text{Caucasiano, Português, Obeso, ...}\}$$

$$c = \text{Obeso}$$

t_c - Função alvo, indicadora de um conceito

$$t_{\text{obeso}} \in T: \quad X \quad \rightarrow \quad \{0, 1\}$$

$$t_{\text{obeso}}(\text{João}) = 1$$

Frequentemente tomamos $c \equiv t_c$: obeso(João) = 1

D - Distribuição amostral, estacionária

$$D = \text{distribuição de pessoas num supermercado}$$

Nota: Quando a distribuição amostral dos objectos se rege um modelo conhecido, pode-se, em princípio, determinar uma resposta exacta às questões precedentes (classificação estatística paramétrica)

L - Conjunto de algoritmos de aprendizagem

$$L = \{l : S \rightarrow H\}$$

O *aprendiz* $l \in L$ considera:

- Um conjunto de treino *S*, gerado segundo *D*.
- Um conjunto de possíveis hipóteses *H* com vista a aprender o conceito.

Exemplo:

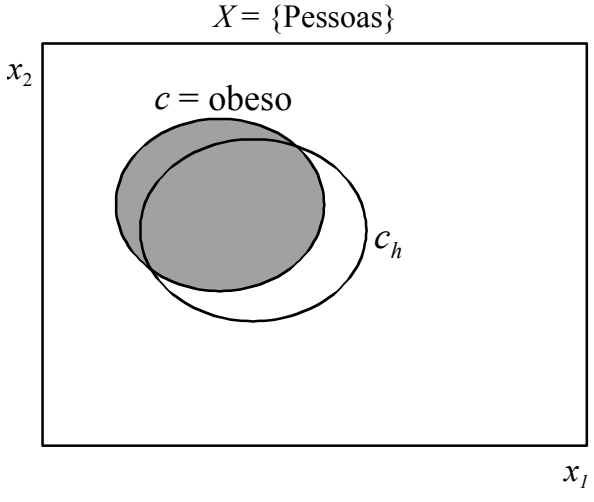
$$H = \{h : X \rightarrow \{0,1\} :$$
$$h(x) = w_2 x_2 + w_1 x_1 + w_0; \quad x_1 \equiv \text{altura}(x), x_2 \equiv \text{largura}(x), w_0, w_1, w_2 \in \Re\}$$

c_h - Conjunto induzido por *h* em *X*

$$c_h = \{x \in X; \quad h(x) = 1\}$$

Exemplo:

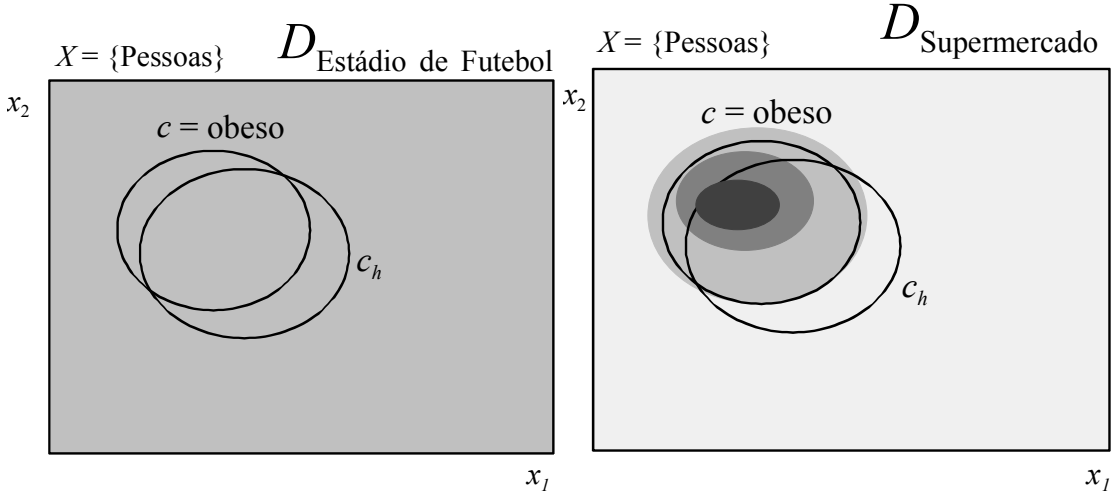
$$c_h = \{x \in X; \quad \mathbf{w}'\mathbf{x} + w_0 = 1\}; \ \mathbf{w}' = \begin{bmatrix} w_1 & w_2 \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



Pe - Erro (verdadeiro) da hipótese *h*

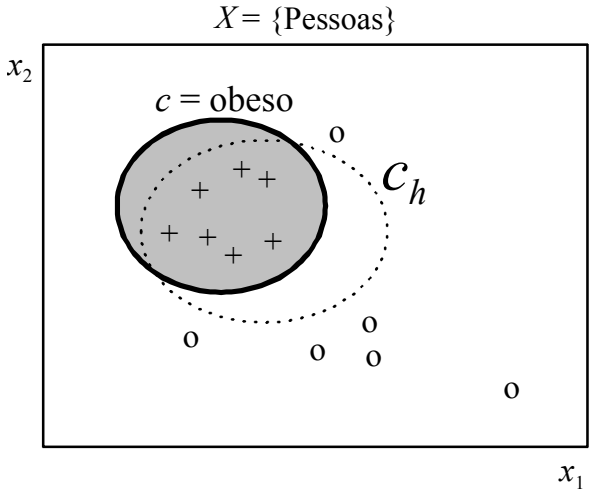
$$Pe(h) \equiv Pe_D(h) = \underset{x \in X \sim D}{P}\left(c(x) \neq h(x)\right)$$

O erro depende da distribuição *D*:



Hipótese consistente

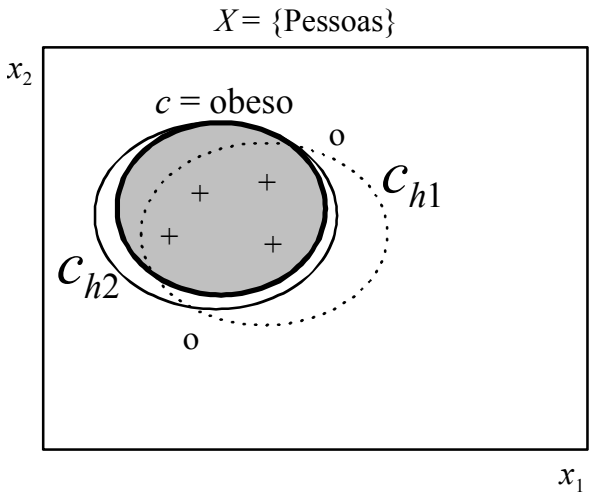
h consistente sse $\forall x \in X_n$, $c(x) = h(x)$, i.e., $Pe_{emp}(h)=0$



1.3 Conceito PAC

Dado um $l \in L$, gerando a hipótese *h*, será realista esperar *Pe*(*h*)=0?

Em geral ($X_n \neq X$), pode haver vários *h*s consistentes com o conjunto de treino e não estamos certos qual deles aprende o conceito.



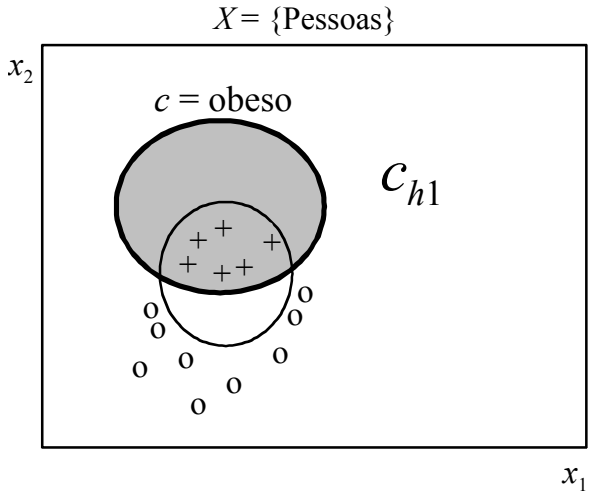
h_1 e h_2 são ambos consistentes; contudo, h_2 aprende melhor o conceito (menor erro).

Tudo o que podemos esperar do aprendiz é que:

$Pe_D(h) \leq \varepsilon$,
ε : *parâmetro de erro*.

aprendiz aproximadamente correcto...

Como o conjunto de treino é escolhido aleatoriamente existe sempre uma probabilidade não nula de ser escolhido um conjunto que contém instâncias que induzem em erro.



Logo só podemos esperar que:

$$P(Pe_D(h) \leq \varepsilon) \geq 1 - \delta$$

δ: *parâmetro de confiança*.

aprendiz aproximadamente correcto, provavelmente...

Definição de aprendizagem PAC - *Provavelmente Aproximadamente Correcta*:

Seja C um conjunto (classe) de conceitos definidos em X e l um aprendiz usando $X_n \subseteq X$ e um espaço de hipóteses H.

C é *PAC-apreensível* por l (l é um algoritmo de aprendizagem PAC para C), se:

$$\forall\, c \in C, \quad \forall\, D \text{ (em } X), \quad \forall\, \varepsilon, \delta, \quad 0 < \varepsilon, \delta < 0.5,$$

$$l \in L \quad \text{determina} \quad h \in H, \quad P(Pe(h) \le \varepsilon) \ge 1 - \delta\,,$$

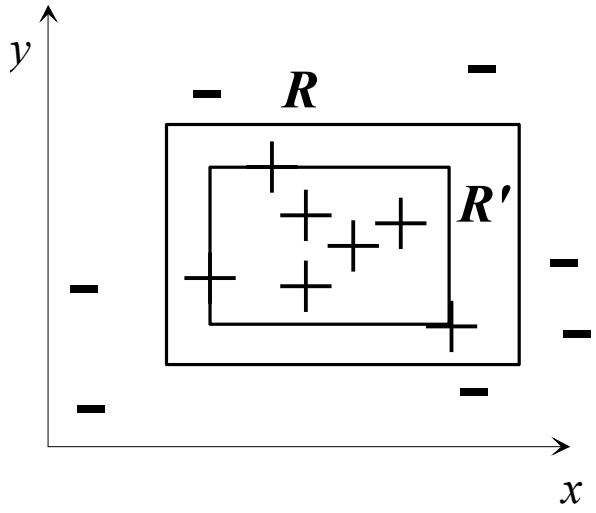
em tempo polinomial em $1/\varepsilon$, $1/\delta$, n e size(c).

size(c) - Número de elementos independentes que se utiliza na representação dos conceitos.

Representação	**size(c)**
Expressão Booleana canónica conjuntiva	Nº de literais Booleanos
Árvore de decisão	Nº de nodos da árvore
Perceptrão Multicamada (MLP)	Nº de pesos do MLP

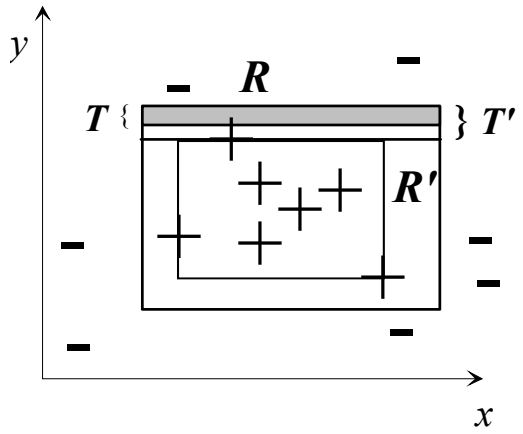
1.4 Exemplo

1 - A classe de conceitos correspondente a rectângulos alinhados com os eixos em $\Re^2$, é PAC-apreensível (Kearns, Vazirani, 1997).



- R : conceito a aprender
- O aprendiz l gera a hipótese R' : rectângulo com ajuste perfeito aos exemplos positivos

Então: $R' \subset R$ e R'-R é a reunião de 4 tiras rectangulares (e.g. T')



Dado ε seja T a tira que (para um certo D) corresponde a: $P(\mathbf{x} \in T) = \frac{\varepsilon}{4}$.

Qual a probabilidade de $P(\mathbf{x} \in T') > \frac{\varepsilon}{4}$?

$P(\mathbf{x} \in T') > \frac{\varepsilon}{4} \quad \Rightarrow \quad T' \supset T \quad \Rightarrow \quad T$ não contém nenhum ponto de X_n.

Probabilidade de T não conter nenhum ponto de X_n:

$$\left(1 - \frac{\varepsilon}{4}\right)^n$$

Logo:

$$P\left(P\left(\mathbf{x} \in T'\right) > \frac{\varepsilon}{4}\right) = \left(1 - \frac{\varepsilon}{4}\right)^n \quad \Rightarrow$$

$$P\left(P\left(\mathbf{x} \in R' - R\right) > \varepsilon\right) \le \delta \qquad \text{com} \qquad \frac{\delta}{4} = \left(1 - \frac{\varepsilon}{4}\right)^n \quad \Rightarrow$$

$$P(Pe(h) \le \varepsilon) \ge 1 - \delta$$

Logo, dado ε e δ o conceito é PAC-apreensível desde que disponhamos de n tal que:

$$\left(1 - \frac{\varepsilon}{4}\right)^n \le \frac{\delta}{4}$$

$$\left(1 - x\right) \le e^{-x} \quad \Rightarrow \quad 4e^{-n\varepsilon/4} \le \delta \quad \Rightarrow \quad n \ge \left(\frac{4}{\varepsilon}\right) \ln\left(\frac{4}{\delta}\right)$$

n é polinomial em $1/\varepsilon$ e $1/\delta$. P. ex. para $\varepsilon = \delta = 0.05$: $n > 351$

2 Complexidade Amostral em Espaços de Hipóteses Finitos

Será que é possível obter um limite inferior da complexidade amostral para todas as situações?

i.e.

Quantos elementos deve ter no mínimo um conjunto de treino para com alta probabilidade podermos determinar uma hipótese eficaz?

2.1 Espaço de Versões

Definições:

Espaço de versões:

$$VS_{H,S} = \{\, h \in H; \quad \forall\,(x, c(x)) \in S, \quad h(x) = c(x) \,\}$$

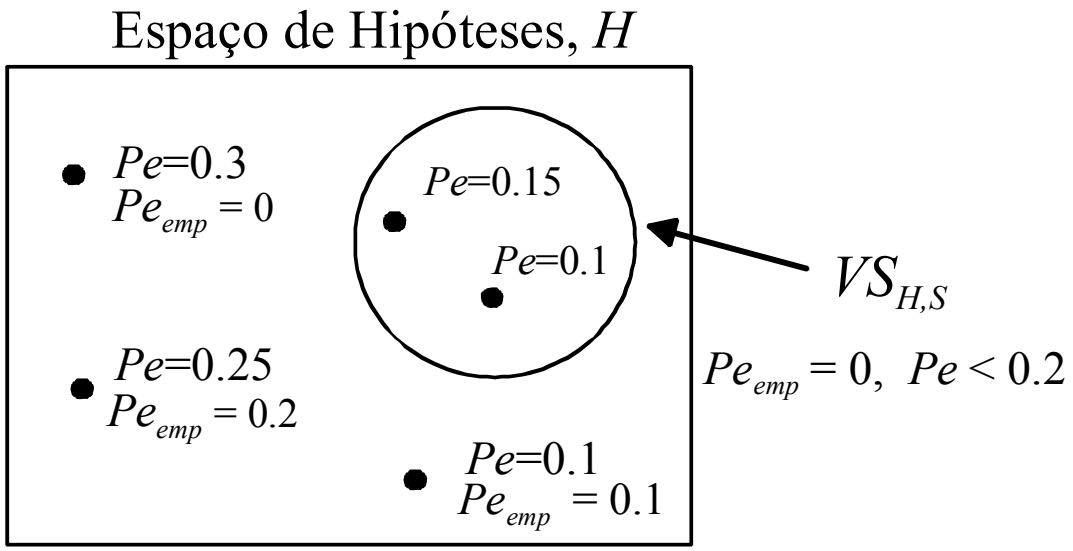
Conjunto das hipóteses consistentes,
com *erro de treino* $Pe_{emp}(h) = 0$.

Espaço de versões ε-exausto:

Seja c um conceito. O espaço de versões diz-se ε-exausto com respeito a c e D se toda a hipótese de $VS_{H,S}$ tem erro menor que ε com respeito a c e D.

$$\forall\ h \in VS_{H,S}, \quad Pe(h) < \varepsilon$$

Exemplo de espaço de versões 0.2-exausto:



2.2 Generalização de Hipóteses de Treino

Teorema:

Para H finito com $|H|$ hipóteses distintas e S uma amostra com $n \geq 1$ exemplos escolhidos aleatoriamente de um conceito alvo c, então para $0 \leq \varepsilon \leq 1$ a probabilidade de o espaço de versões $VS_{H,S}$ não ser ε-exausto (com respeito a c) é menor ou igual a:

$$\left|H\right| e^{-\varepsilon n}$$

Formulação informal:
A probabilidade de encontrarmos uma boa hipótese de treino (consistente com o conjunto de treino) mas, de facto, má hipótese em geral (com erro verdadeiro superior a ε) é inferior a $\left|H\right| e^{-\varepsilon n}$, sendo n o número de exemplos de treino.

Demonstração:

1. Sejam $h_1, h_2, ..., h_k$ todas as hipóteses com $Pe \geq \varepsilon$.

2. $VS_{H,S}$ não é ε-exausto se $\exists\, h_i \in VS_{H,S} \quad i = 1, \ldots, k$

3. $Pe(h_i) \geq \varepsilon \quad \Rightarrow \quad P(h_i(x) = c(x)) = 1 - \varepsilon, \quad \forall x \in X_n$

4. $P(h_i \text{ consistente}) = P(h_i(x_1) = c(x_1) \wedge \ldots \wedge h_i(x_n) = c(x_n)) = (1-\varepsilon)^n$

5. $P(h_1 \text{ consistente} \vee \ldots \vee h_k \text{ consistente}) =$
$k(1-\varepsilon)^n \leq |H|(1-\varepsilon)^n \leq |H| e^{-\varepsilon n}$

O número de exemplos de treino requeridos para tornar a probabilidade menor que um certo valor δ é:

$$|H|\, e^{-\varepsilon n} \le \delta \qquad \Rightarrow \qquad n \ge \frac{1}{\varepsilon}(\ln|H| + \ln(\frac{1}{\delta}))$$

Notas:

1. Notar a semelhança da expressão obtida com as anteriores.

2. Notar que este limite de *n* pode ser demasiado pessimista. De facto o Teorema fornece um limite de probabilidade que cresce com |*H*| (pode assim ultrapassar 1!).

3. Notar que o Teorema não considera o caso de |*H*| infinito. Precisamos para isso de uma outra medida da complexidade de *H*.

3 Complexidade Amostral em Espaços de Hipóteses Infinitos

3.1 Limites da Aprendizagem PAC

Definição:

Seja C uma classe de conceitos, $C \subseteq 2^X$. A dimensão de Vapnik-Chervonenkis de C, $d_{VC}(C)$, é a cardinalidade do maior conjunto finito de pontos $X_n \subseteq X$ que é estilhaçado por C.
Se conjuntos arbitrariamente grandes de pontos podem ser estilhaçados por C, $d_{VC}(C)$ é infinita.

Teorema, Blumer *et al.* (1989):

Seja C uma classe de conceitos. Então:

i. C é PAC-apreensível sse $d_{VC}(C)$ é finita.

ii. Se $d_{VC}(C)$ é finita, então:

(a) Para $0 < \varepsilon < 1$ e dimensão amostral de pelo menos

$$n_u = \max\left[\frac{4}{\varepsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d_{VC}(C)}{\varepsilon}\log_2\left(\frac{13}{\varepsilon}\right)\right], \qquad (3)$$

qualquer algoritmo consistente é de aprendizagem PAC para C.

(b) Para $0 < \varepsilon < 1/2$ e dimensão amostral menor que

$$n_l = \max\left[\frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right), d_{VC}(C)\left(1-2\left(\varepsilon\left(1-\delta\right)+\delta\right)\right)\right], \quad (4)$$

nenhum algoritmo de L, para qualquer espaço de hipóteses H, é de aprendizagem PAC para C.

Aplicação aos MLPs:

Limite inferior, n_l: ε : *Pe* aceitável
Usar fórmula (4) com fórmula (1).

Limite superior, n_u:
(pouco realista) Usar a fórmula (2) e a fórmula (3)

Baum e Haussler (1989) mostraram que um MLP complexo com erro de treino ε terá um erro de teste de no máximo 2ε para (limite mais realista):

$$n_u = \frac{w}{\varepsilon}\ln\left(\frac{u}{\varepsilon}\right), \qquad (5)$$

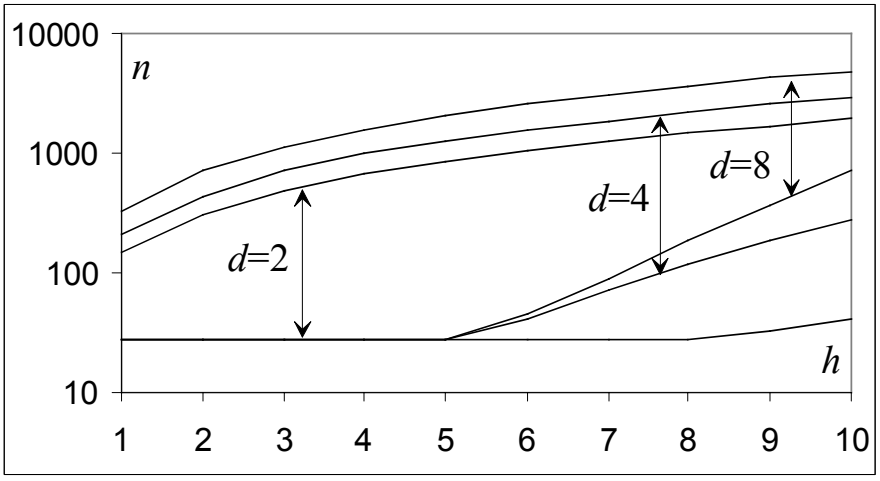
com

$$\delta = 8\left(2uen_u / w\right)^w e^{-\varepsilon n_u/16}.$$

δ é muito baixo (δ<0.005) mesmo para baixos valores de *d* e *m*.

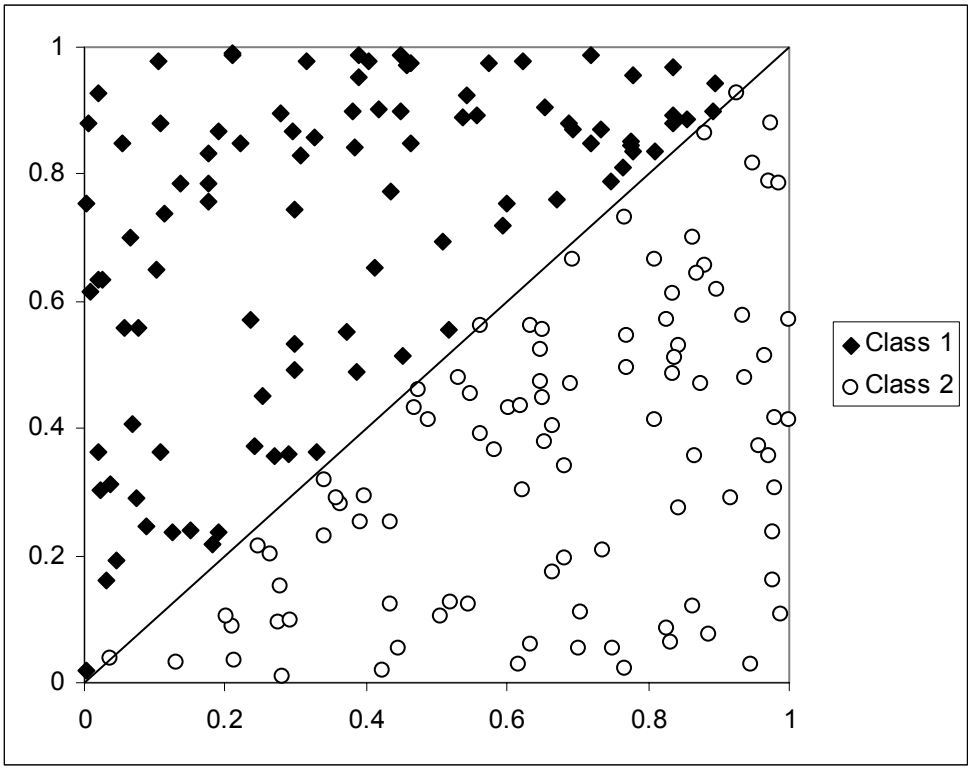
Regra prática baseada neste limite: $n_u = w/\varepsilon$

Limites de *n*, (4) e (5), para ε= 0.05 e δ= 0.01:



3.2 Caso de Estudo

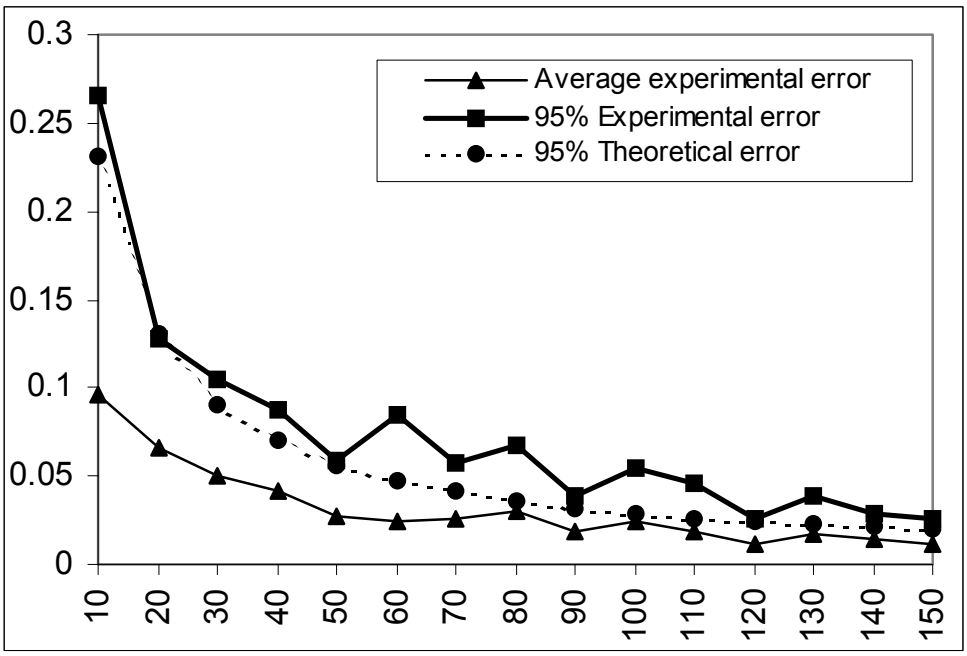
Duas classes de pontos distribuídos em $[0, 1]^2$, linearmente separáveis.



- Hipótese ideal: $x_2 = x_1$
- Distribuição amostral D : distribuição uniforme

Ensaios com perceptrão simples (MLP2:1)

Para cada n=10, 20,...,150 são gerados 25 X_n, e determinados os respectivos MLP2:1



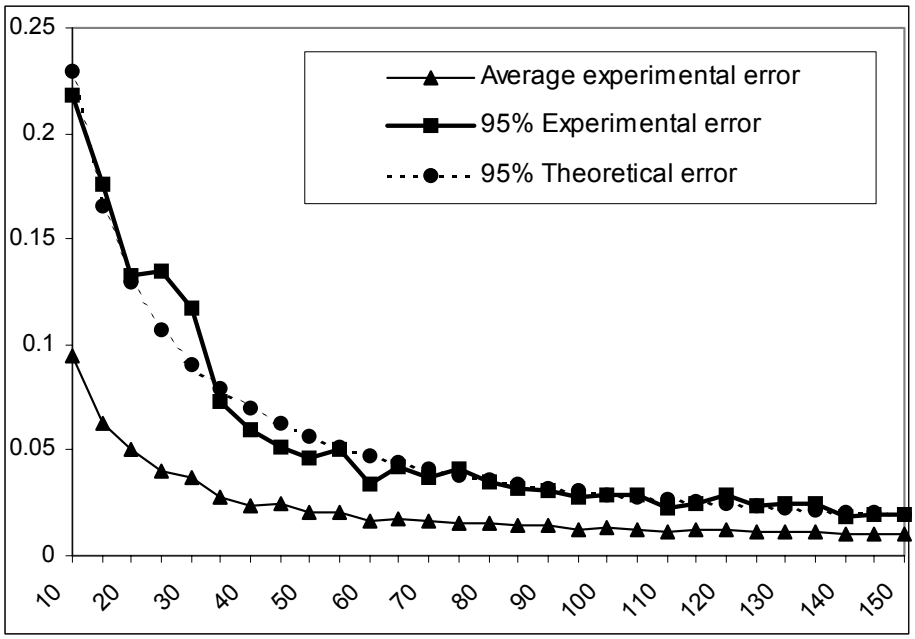
Erro médio em 25 experiências

Percentil 95% dos erros em 25 experiências

Erro, ε, correspondente a δ=95% para $n_l = n$ e d_{VC}=3 (fórmula de Blumer *et al.*)

Ensaios com SVM lineares

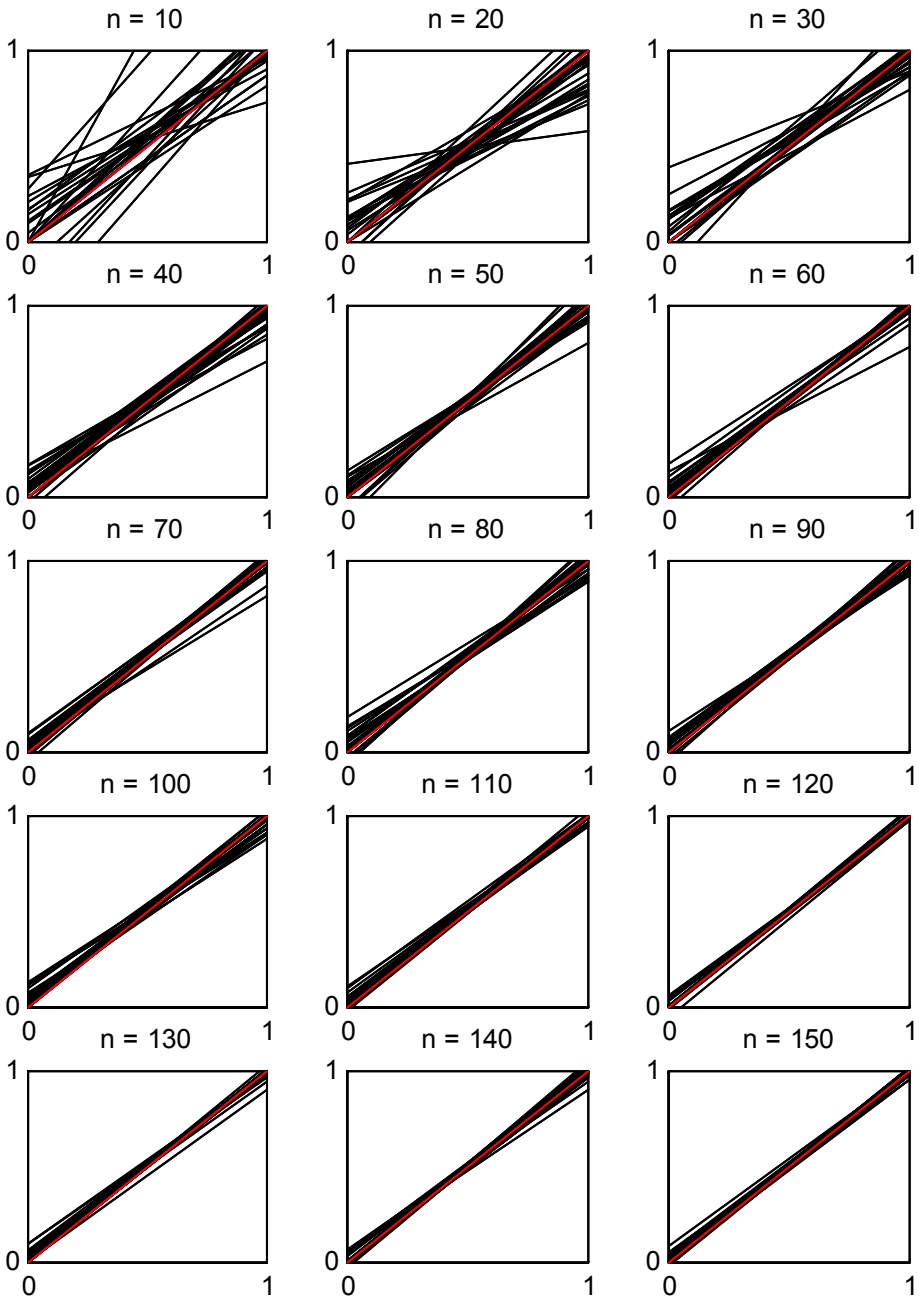
Para cada n=10, 15, ...,150 são gerados 200 X_n, e determinados os respectivos SVM.



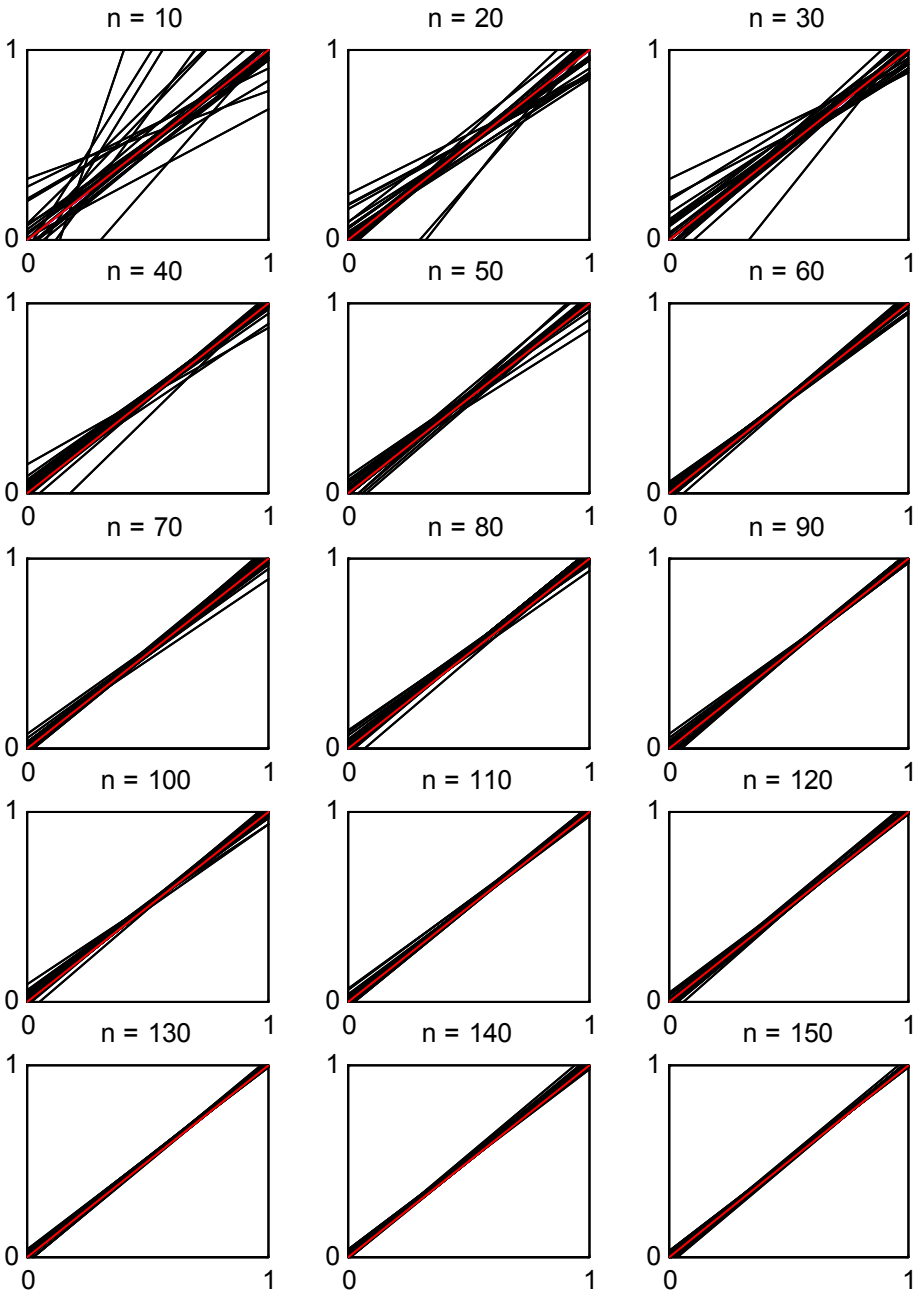
Erro médio em 200 experiências

Percentil 95% dos erros em 200 experiências

Erro, ε, correspondente a δ=95% para $n_l = n$ e d_{VC}=3 (fórmula de Blumer *et al.*)

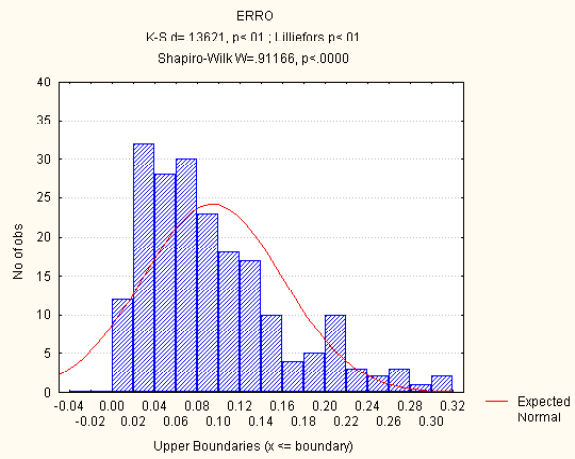


Linear discriminants produced by a Perceptron

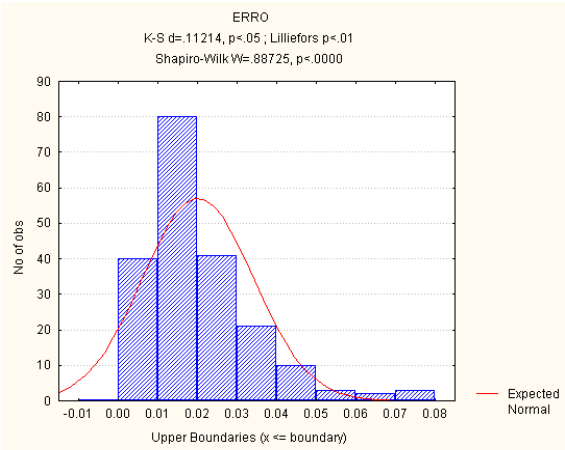


Linear discriminants produced by a Support Vector Machine

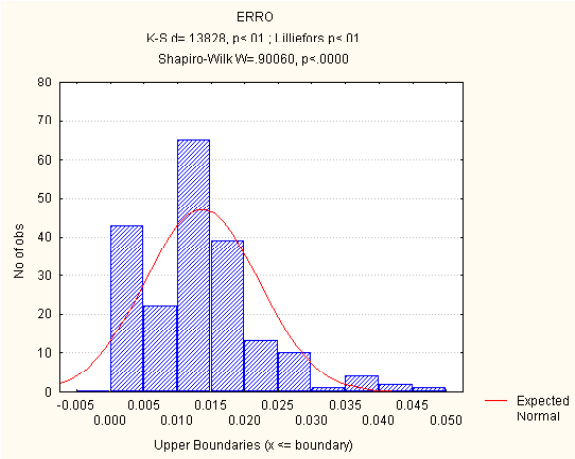
Histogramas dos erros (SVM):



Erros para n=10



Erros para n=50



Erros para n=100

Plano de Trabalho:

1. Desenvolver um programa em MATLAB que permita efectuar de forma confortável os ensaios anteriores com MLPs e SVMs

2. Repetir os ensaios para os dados do quadrado, usando as fórmulas baseadas na VCD

3. Usar a fórmula de Takashi.

4. Repetir 2 e 3 usando dados reais: 2126 casos de cardiotocogramas classificados em N e P.