

# Preços de Carros e Classificador de Cogumelos Venenosos

Custodio Moreira Brasileiro Silva

Fausto Coutinho Lourenço

Luiz Felipe França Amadio

Leilson José de Araujo

Prof. Emílio Del Moral Hernandez

São Paulo, 18 de Junho de 2014

# Componentes do Grupo



Custódio



Fausto



Leilson



Luiz Felipe

# Agenda

- Distribuição de Atividades
- Ferramentas Utilizadas
- Preços de Carros
  - Motivação
  - Conjuntos de Treino/Teste e Variáveis
  - Definição de Parâmetros
  - Definição de Topologia e Sobre-treino
- Classificador de Cogumelos Venenosos
  - Motivação
  - Conjuntos de Treino/Teste e Variáveis
  - Definição de Topologia
  - Matriz de Confusão

# Distribuição das Atividades

- Propostas de Tema – Todos
- Análise e Escolha – Todos
- Regressor
  - Pesquisa Teórica – Custódio
  - Aquisição de Dados – Custódio e Luiz Felipe
  - Protótipo Inicial/Simulação Inicial – Custódio
  - Otimização – Luiz Felipe
- Classificador
  - Pesquisa Teórica – Fausto e Leilson
  - Aquisição de Dados – Fausto
  - Protótipo Inicial/Simulação Inicial – Leilson
  - Otimização – Fausto e Leilson
- Relatório e Apresentação - Todos

# Ferramentas Utilizadas

- Ambiente Computacional Multiple Back Propagation (MBP) – Simulações das topologias
- Microsoft Office Excel – Pré processamento dos dados



Fonte: <http://mbp.sourceforge.net>



Fonte: <http://dataniitro.com>

# Regressor – Preço de Veículos

- Motivação
  - 87 milhões de veículos produzidos no mundo em 2013<sup>[1]</sup>
  - 3,7 milhões de veículos produzidos no Brasil em 2013<sup>[1]</sup>
  - 50 grandes montadoras<sup>[1]</sup>
  - Capacidade de predição



<http://hdbackgroundpoint.com/mustang-backgrounds.html>

# Regressor – Preço de Veículos

- Fonte de Dados
  - UCI / Wards Automotive Yearbook 1985
  - Utilização original: seguro de veículos



Fonte: <http://novoguscar.blogspot.com.br/2009/08/historia-uno-25-anos-no-brasil.htm>

# Regressor – Preço de Veículos

- Variáveis de entrada:

- |               |                   |                    |
|---------------|-------------------|--------------------|
| - montadora   | - largura         | - posição do motor |
| - portas      | - peso            | - com turbo        |
| - tração      | - consumo cidade  | - cilindros        |
| - entre-eixos | - consumo estrada | - tamanho do motor |
| - comprimento | - potência        |                    |
| - altura      | - RPM max         |                    |

- Variável de saída:

- preço do veículo



# Regressor – Preço de Veículos

- Etapas da otimização/definição da rede:
  - Topologia:
    - número de camadas escondidas
    - número de neurônios nas camadas escondidas
  - Space network:
  - Outras configurações:
    - Intervalo dos  $W$ 's iniciais
    - Atualização automática de parâmetros durante o treino
    - Passos de adaptação variáveis
    - Ligação direta nós de entrada e saída
    - Cross Validation
    - Sensibilidade de entradas
    - Critério de parada

# Regressor – Preço de Veículos

- Considerações sobre os ensaios a seguir:
  - Dados das tabelas referentes ao menor RMS de teste obtido após 10 diferentes re-sorteios dos  $W$ 's iniciais
  - Critério de parada:  $\text{RMS treino} < 0.01$  (a menos que indicado contrário)
- Análise importante -> compromisso entre:
  - Tempo total (proporcional a complexidade do modelo)
  - RMS teste obtido
    - Em muitos casos a seguir veremos modelos muito complexos e que não possuem um RMS satisfatoriamente mais baixo do que modelos menos complexos

# Regressor – Preço de Veículos

- Número de camadas escondidas

Topologia	RMS Teste	T(s)	obs:
16-5-1	0,032169	11	
16-5-5-1	0,029874	16	
16-5-5-5-1	0,031409	15	
16-5-5-5-5-1	0,038854	19	
16-5-5-5-5-5-1	0,03643	29	

- Neurônios na primeira camada escondida

Topologia	RMS Teste	T(s)	obs:
16-3-5-1	0,031226	92	finalizado manualmente com RMS treino = 0.0147
16-5-5-1	0,031664	16	
16-10-5-1	0,031001	9	
16-20-5-1	0,02982	12	
16-50-5-1	0,030328	21	
16-3-10-10-1	0,0260835	102	finalizado manualmente com RMS treino = 0.0151
16-10-10-10-1	0,031989	15	
16-50-10-10-1	0,034246	37	

- Poucos neurônios na entrada implicam em lentidão ou falta de convergência
- Aumento do número de camadas escondidas não melhorou RMS e aumentou o tempo
- Aumento do número de neurônios não melhorou o RMS e aumentou o tempo

# Regressor – Preço de Veículos

- Neurônios na última camada escondida

Topologia	RMS Teste	T(s)
16-10-3-1	0,030756	8
16-10-5-1	0,028355	6
16-10-10-1	0,030449	10
16-10-20-1	0,030518	14
16-10-50-1	0,031994	26
16-10-10-3-1	0,0295295	17
16-10-10-10-1	0,0316919	24
16-10-10-50-1	0,0335459	37

- Aumento do número de neurônios não melhorou RMS e aumentou o tempo

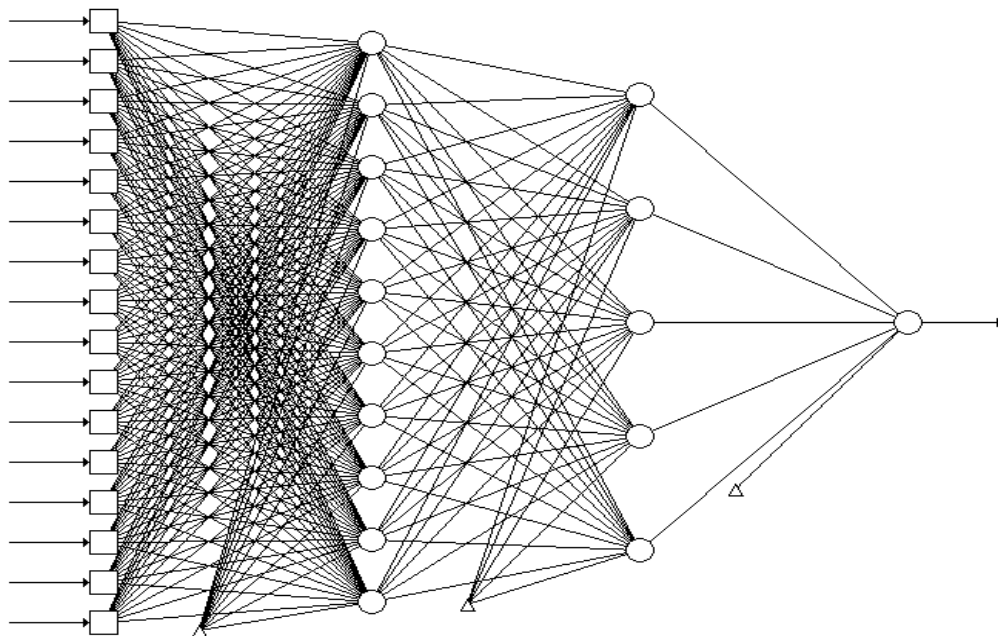
- Neurônios na camada escondida intermediária

Topologia	RMS Teste	T(s)
16-10-3-10-1	0,0306721	15
16-10-10-10-1	0,0294712	16
16-10-50-10-1	0,0290703	45

- Aumento do número de neurônios não melhorou o RMS e aumentou o tempo

# Regressor – Preço de Veículos

- Topologia Escolhida (16-10-5-1)



# Regressor – Preço de Veículos

- Space Network

Topologia	RMS Teste	T(s)
padrão	0,0315473	17
5	0,0302515	15
10	0,0312256	16,6
10-5	0,0330973	21,8

- A utilização de Space Network não deixou a rede mais rápida e nem com RMS mais baixo
- A atualização automática dos parâmetros se viu se não necessária, ao menos muito útil na melhoria de velocidade e precisão da RNA

- Atualização automática de parâmetros

Topologia	RMS Teste	T(s)	Obs
16-10-5-1	0,0327890	108	Sem atualização de parâmetros / Encerrado manualmente com RMS treino = 0.013
16-10-5-1	0,0291654	9	Com Atualização de momento
16-10-5-1	0,0381569	66	Com Atualização de aprendizagem / Encerrado manualmente com RMS treino = 0.0137
16-10-5-1	0,0283133	65	Com Atualização de momento e aprendizagem / Encerrado manualmente com RMS treino = 0.0148
16-10-5-1	0,0189453	7	Ajuste automático de pesos adaptativos, aprendizagem e momento (padrão)

# Regressor – Preço de Veículos

- Intervalo de sorteio dos W's iniciais

Topologia	RMS Teste	T(s)	Obs
16-10-5-1	0,0306713	13,9	-01, até 0,1
16-10-5-1	0,0310365	11,4	-0,3 até 0,3
16-10-5-1	0,0309999	10,9	-0,5 até 0,5
16-10-5-1	0,0308683	9,2	-0,8 até 0,8
16-10-5-1	0,0292790	10	-1 até 1
16-10-5-1	0,0369451	21,1	-5 até 5
16-10-5-1	0,0562929	44	-10 até 10

- Ligação direta de entrada e saída

Topologia	RMS Teste	T(s)	Obs
16-10-5-1	0,0299353	14,9	Com ligação direta entre entrada e saída
16-10-5-1	0,0292790	10	Sem ligação direta entre entrada e saída

- Intervalo pequeno no sorteio de W's iniciais deixou o modelo mais lento e igualmente preciso
- Intervalo grande no sorteio de W's iniciais deixou o sistema saturado, lerdo e impreciso
- Ligação direta entre nós de entrada e saída possuiu pouca melhora no RMS com um gasto de tempo ligeiramente maior

# Regressor – Preço de Veículos

- Sensibilidade

Ruído de:	Alteração no RMS do conjunto de:			
	Treino		Teste	
	14ª entrada	16ª entrada	14ª entrada	16ª entrada
0%	0,009998092	0,009998092	0,016022281	0,016022281
2%	0,010053116	0,010035943	0,016033124	0,01603441
5%	0,010337129	0,010232421	0,016090356	0,016097736
10%	0,01105849	0,010739176	0,016245433	0,016263759
Sempre no Máximo	0,019325625	0,01895848	0,022914016	0,018521042
Sempre na média	0,016482571	0,014182876	0,015462838	0,01905083
Sempre no Mínimo	0,038831827	0,026957902	0,022909578	0,024993152

<b>RMS obtido:</b>	<b>0,01622208</b>
--------------------	-------------------

- Utilizamos a função da sensibilidade do MBP
- No destaque, as entradas que provocaram menores alterações no RMS.
- Mesmo as entradas com menores sensibilidades provocam alterações da ordem de grandeza do RMS obtido, com ruído de apenas 2%
- Entrada 14 = RPM Max
- Entrada 16 = Consumo na estrada
- Por fim, os menores valores de RMS obtidos utilizando diferentes combinações de variáveis de entrada



# Regressor – Preço de Veículos

- Sensibilidade

	Melhor resultado		2º melhor resultado	
	Iterações	RMS teste	Iterações	RMS teste
Sem entrada 14	2248	0,029174	2048	0,029391
Sem entrada 16	1547	0,027223	1620	0,028503
Sem entradas 14 e 16	2232	0,027443	2264	0,028655
Com todas entradas	1642	0,015786	1540	0,01858

- Podemos observar, de fato, que houve pouca alteração de RMS retirando uma ou duas variáveis, como era de se esperar, de acordo com a tabela anterior

- Utilizamos a função da sensibilidade do MBP
- No destaque, as entradas que provocaram menores alterações no RMS.
- Mesmo as entradas com menores sensibilidades provocam alterações da ordem de grandeza do RMS obtido, com ruído de apenas 2%
- Entrada 14 = RPM Max
- Entrada 16 = Consumo na estrada
- Por fim, os menores valores de RMS obtidos utilizando diferentes combinações de variáveis de entrada

# Regressor – Preço de Veículos

- Cross Validation

Conjunto 1		
Iteração	RMS Treino	RMS Teste
768	0,009992	0,027999
1024	0,009628	0,037795

Conjunto 2		
Iteração	RMS Treino	RMS Teste
1664	0,009938	0,027988
1408	0,009934	0,022871

Conjunto 3		
Iteração	RMS Treino	RMS Teste
1408	0,009896	0,021175
1408	0,00996	0,022846

Conjunto 4		
Iteração	RMS Treino	RMS Teste
1536	0,009961	0,036506
1408	0,009861	0,038141

Conjunto 5		
Iteração	RMS Treino	RMS Teste
1280	0,009783	0,023165
1408	0,009961	0,028695

Media	Pior caso	
RMS Teste	RMS teste	Conjunto
0,0287181	0,043141	4

- Estes dados são referentes a 2 diferentes simulações realizadas com 5 grupos de dados
- Ao final, temos um valor da média de todas as medidas, bem como a situação de pior caso
- Os conjuntos que apresentaram piores resultados são o 1 e o 4
- Optamos por utilizar o conjunto 1 pois ele possui exatamente um dado de cada uma das 19 montadoras de carros analisadas (o conjunto de testes 4 foi obtido aleatoriamente)

# Regressor – Preço de Veículos

- Critério de parada + Cross Validation

Conjunto 1			
	Iteração	RMS Treino	RMS Teste
Final	49920	0,005221	0,030992
RMS teste MIN	23936	0,005743	0,025386
Parada Padrão	768	0,009992	0,027999

Final	49920	0,004526	0,049412
RMS teste MIN	128	0,018626	0,027852
Parada Padrão	1024	0,009628	0,037795

Conjunto 4		
Iteração	RMS Treino	RMS Teste
49920	0,005981	0,048066
256	0,017545	0,023298
1536	0,009961	0,036506

49920	0,005945	0,049371
256	0,015818	0,024615
1408	0,009861	0,038141

- Inicialmente:
  - Parada com RMS treino  $< 0.01$
- Porém, este valor pode não ser suficientemente pequeno para que uma análise completa e precisa possa ser feita
- Nesse caso, foram realizados testes com os 2 piores casos obtidos no Cross Validation, e com o critério de parada de 50.000 iterações. Duas situações para cada conjunto estão mostradas acima.
- Resultados mostram que houve pouca alteração entre o menor RMS obtido durante todo o processo de treino, com 50.000 iterações, e com o que foi obtido pela parada padrão (RMS treino  $< 0.01$  / aprox. 1.000- 1.500 iterações)

# Regressor – Preço de Veículos

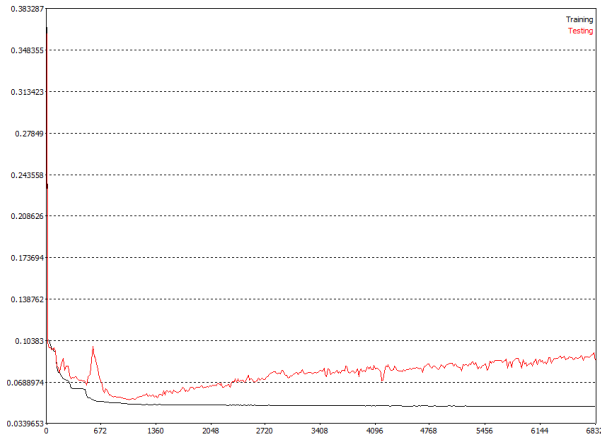
- Critério de parada + Complexidade

	3 camadas escondidas			4 camadas escondidas			5 camadas escondidas		
<b>Final</b>	9984	0,009483	0,046041	99840	0,008432	0,019431	99920	0,00642	0,070487
<b>RMS teste MIN</b>	160	0,02062	0,02203	86016	0,008588	0,018596	2048	0,009494	0,023386
<b>Parada Padrão</b>	5504	0,009984	0,04808	4864	0,009998	0,02319	1664	0,009903	0,024057

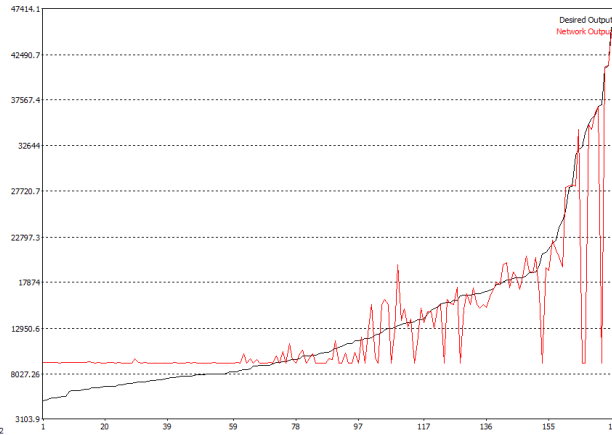
- Analisando, agora, a alteração no critério de parada em relação ao aumento de complexidade do sistema, realizamos alguns testes com 3, 4 e 5 camadas internas, com critério de parada de 100.000 iterações
- Mais uma vez, devemos comparar o RMS mínimo obtido em toda a simulação com aquele que seria obtido com o critério de parada padrão (RMS treino < 0.01)
- Resultados mostram que realmente o aumento da complexidade não melhora de maneira satisfatória no desempenho da rede, além de evidenciar que o critério de parada padrão, para o nosso caso, é um bom critério.

# Regressor – Preço de Veículos

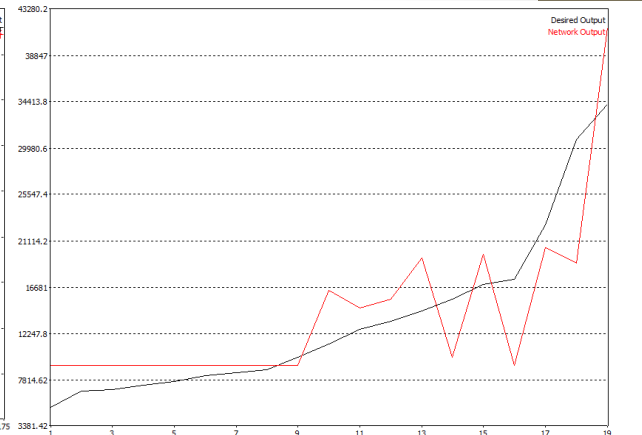
- Definição de parâmetros



RMS x iterações



Saída x exemplar (treino)

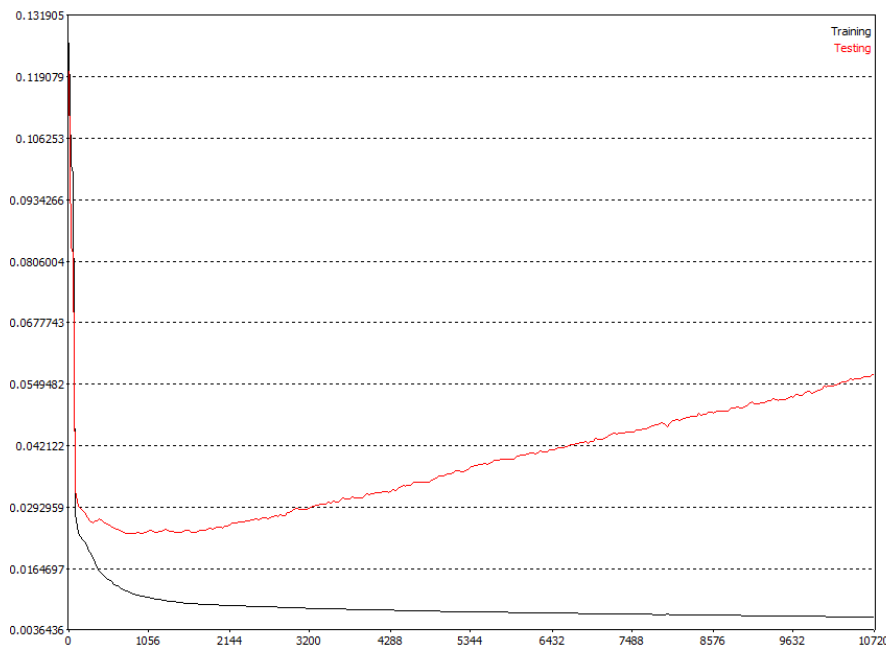


Saída x exemplar (teste)

Topologia: 16-10-5-1
6844 iterações em 1m38s
RMS treino = 0.00484725547
RMS teste = 0.0893262065
intervalo W -10 até 10

# Regressor – Preço de Veículos

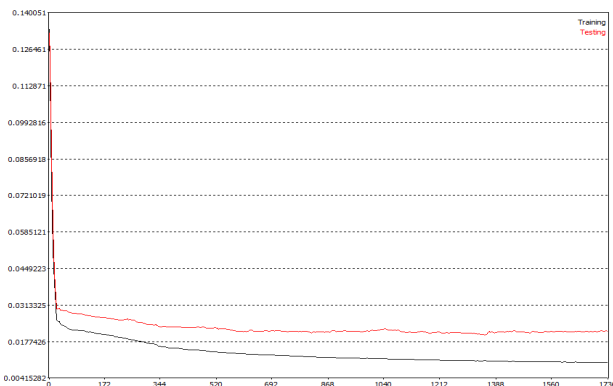
- Sobretreino



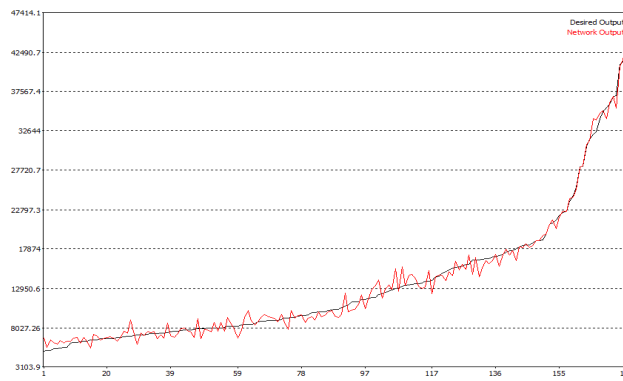
Topologia: 16-30-30-30-30-30-1
10729 iterações em 33m 42s
RMS treino = 0,0064020097
RMS teste = 0,056755096

# Regressor – Preço de Veículos

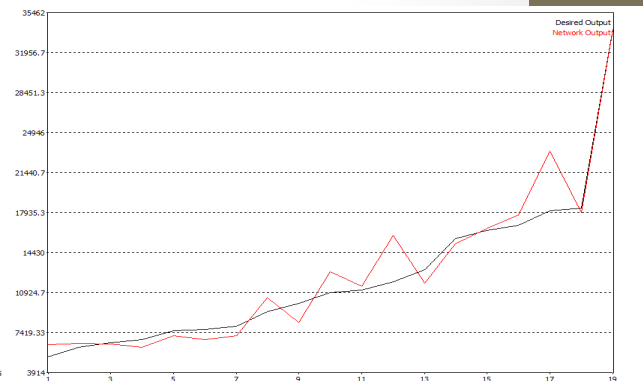
- Topologia Escolhida (16-10-5-1)



RMS x iterações



Saída x exemplar (treino)



Saída x exemplar (teste)

Topologia: 16-10-5-1
1736 iterações em 16s
RMS treino = 0.0099976953
RMS teste = 0.0214601043

# Classificador de Cogumelos

- **Motivação**

- Grande Utilidade Alimentícia e Farmacêutica<sup>[2]</sup>
- Aproximadamente 14000 espécies diferentes<sup>[2]</sup>
- Muito sensíveis a características morfológicas

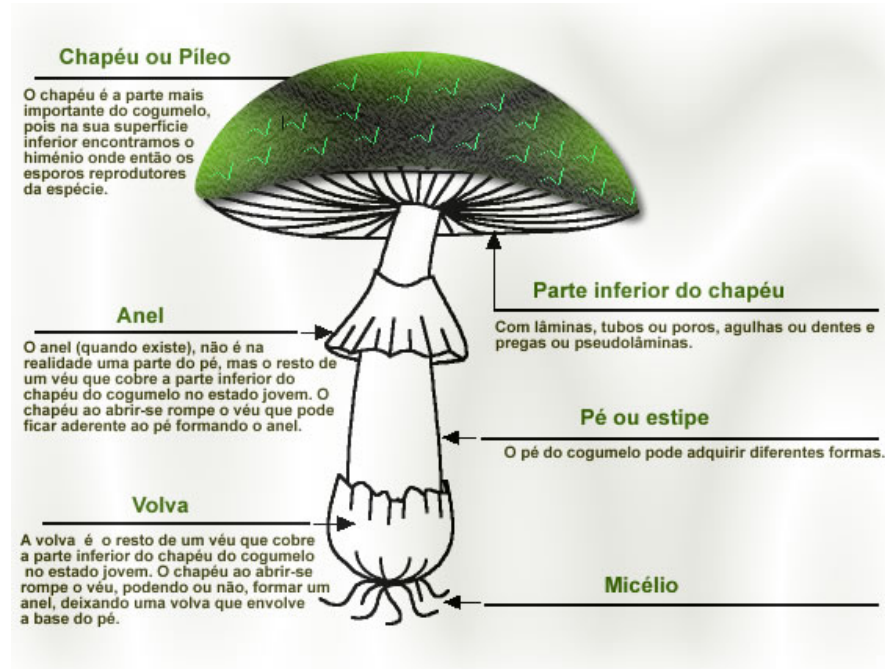


Fonte: <http://flores.culturamix.com/diccas/cultivo-de-cogumelos-comestiveis-a-arte-de-frutificar-fungos>



# Classificador de Cogumelos

- Motivação



# Classificador de Cogumelos

- Fonte de Dados

- UCI / Mushroom (1987)
- 22 atributos e aproximadamente 8200 amostras

- Forma do chapéu
- Superfície do chapéu
- Cor do chapéu
- Odor
- Espaçamento das lâminas
- Tamanho das lâminas
- Cor das lâminas

- Forma da estipe
- Superfície da estipe acima do anel
- Superfície da estipe abaixo do anel
- Cor da estipe acima do anel
- Cor da estipe abaixo do anel
- Cor do Veu

- Número de anéis
- Morfologia dos anéis
- Cor dos esporos
- População
- Habitat
- Manchas
- Venenoso ou não
- Forma do micélio (Excluída)

# Classificador de Cogumelos

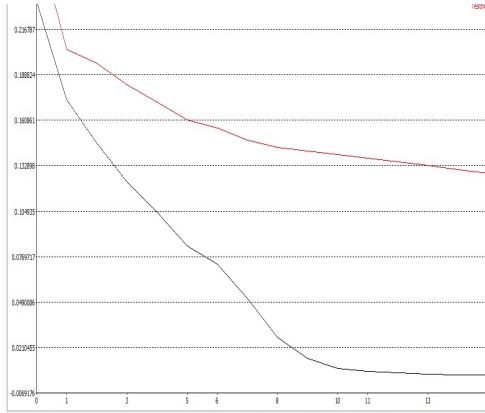
- Pré-processamento
  - Atribuição de valores numéricos às características morfológicas utilizando o Excel
  - Exclusão de atributos com dados faltantes
  - Ordenação lógica quanto à curvatura (variável forma do chapéu) e quanto à rugosidade ( variáveis superfície do chapéu e da estipe)
  - Separação em variáveis binárias de características não ordenáveis logicamente (Odor, Cor, Morfologia,Habitat)
- 84 variáveis de entrada e 1 de saída

# Classificador de Cogumelos

- Pré-processamento
  - Condicionamento: Mesmo número de exemplares venenosos e não-venenosos nos conjuntos de treino e teste, de modo a garantir que a rede seja treinada corretamente e testada a maioria de casos possíveis
  - Ordenação : Treinar primeiramente com todos os casos não-venenosos para em seguida treinar os venenosos, de modo a evitar uma adaptação muito brusca dos pesos sinápticos.

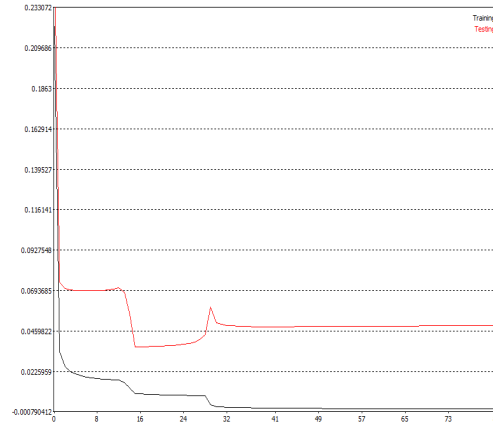
# Classificador de Cogumelos

- Comparação de pré-processamento (Topologia 84-2-1)



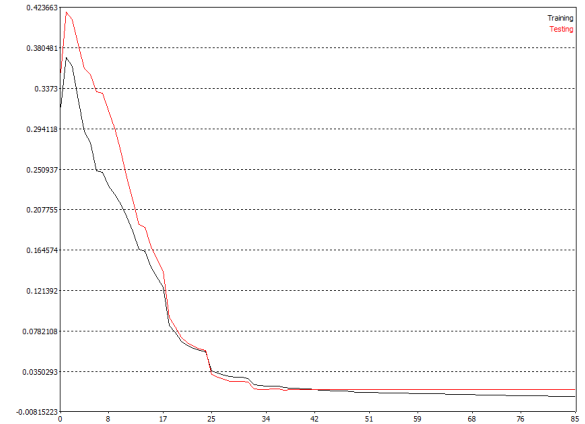
Sem condicionamento e sem ordenação

RMS Teste = 0,1689



Condicionado e sem ordenação

RMS Teste = 0,0493

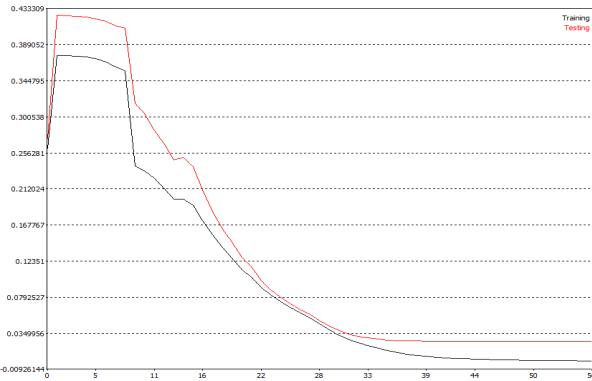


Condicionado e ordenado

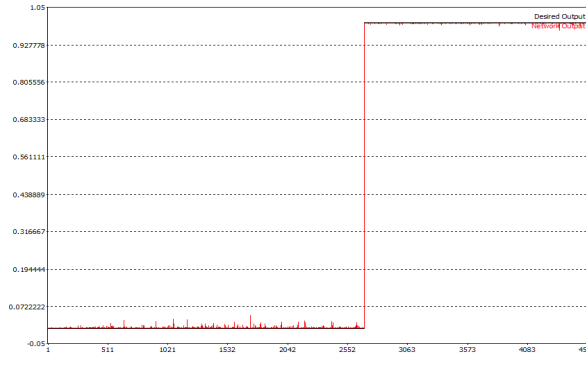
RMS Teste = 0,0159  
(melhor desempenho)

# Classificador de Cogumelos

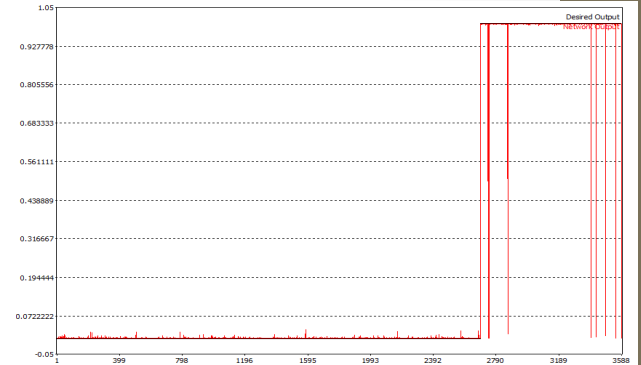
- Comparação de Topologias - Topologias 84-30-10-5-1



RMS x iterações



Saída x exemplar (treino)

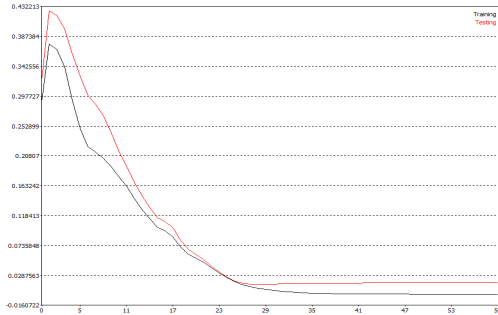


Saída x exemplar (teste)

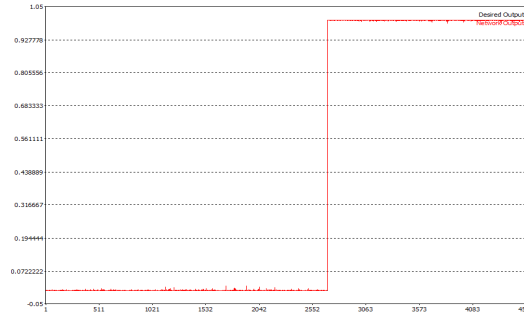
56 iterações em 2m43s
RMS treino = 0,0011684900
RMS teste = 0,0249287901

# Classificador de Cogumelos

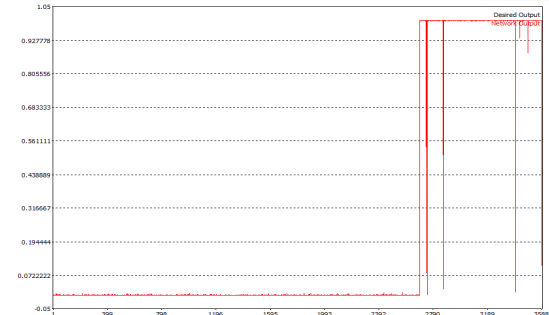
- Comparação de Topologias - Topologias 84-30-10-1



RMS x iterações



Saída x exemplar (treino)



Saída x exemplar (teste)

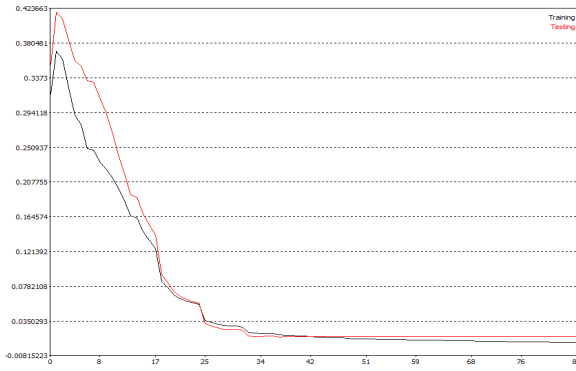
59 iterações em 2m39s

RMS treino = 0,000470065

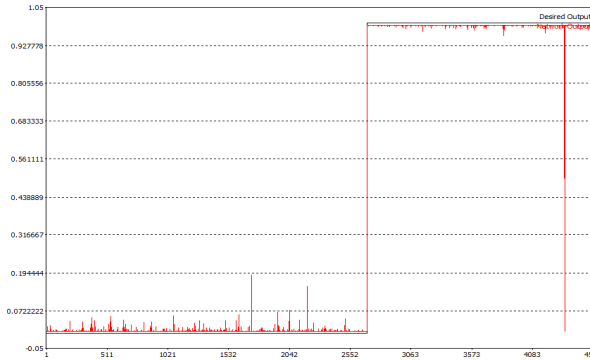
RMS teste = 0,019752192

# Classificador de Cogumelos

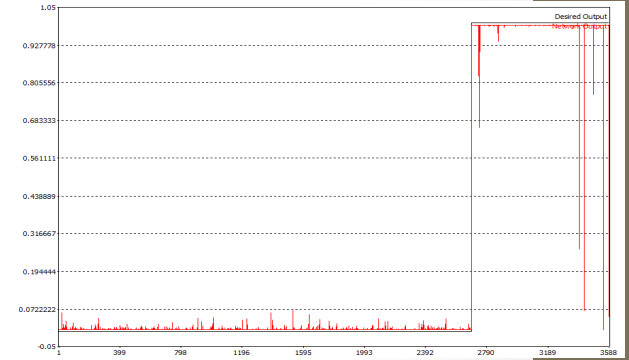
- Comparação de Topologias – Topologia 84-2-1



RMS x iterações



Saída x exemplar (treino)



Saída x exemplar (teste)

87 iteracoes em 17s
RMS treino = 0,00083199996
RMS teste = 0,0159950749

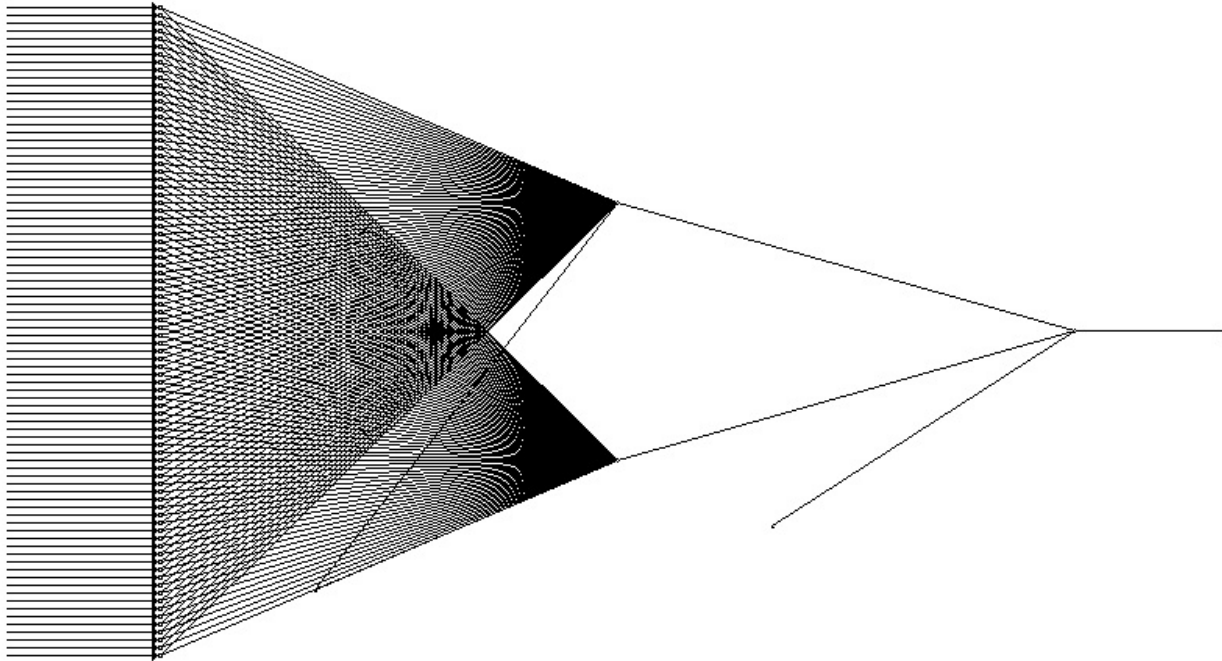


# Classificador de Cogumelos

- Comparação de Topologias
  - Topologia 84-2-1 apresentou melhor desempenho em tempo de processamento do que as redes mais complexas, aliado a um melhor desempenho de valor de RMS
  - Pré-processamento contribui para esse desempenho porque já separa características distintas na entrada

# Classificador de Cogumelos

- Topologia Escolhida (84-2-1)



# Classificador de Cogumelos

- Matriz de Confusão

Matriz de confusão		Predito	
		Venenoso	Não-Venenoso
Real	Venenoso	99.12%	0.88%
	Não-Venenoso	0.00%	100.00%

# Classificador de Cogumelos

- Matriz de Confusão

- Se a rede estimar que o cogumelo é venenoso, há boas chances do cogumelo realmente ser venenoso haja visto que todos os cogumelos classificados como venenosos eram venenosos
- Se a rede estimar que o cogumelo não é venenoso, a melhor alternativa seria realizar um teste de laboratório dada a criticidade da aplicação, porque a rede classificou alguns cogumelos venenosos como não venenosos

# Referências

[1]Oica.net – 2013 Production Statistics

[2]Miles PG, Chang ST -2004 - *Mushrooms: Cultivation, Nutritional Value, Medicinal Effectm and Environmental Impact*