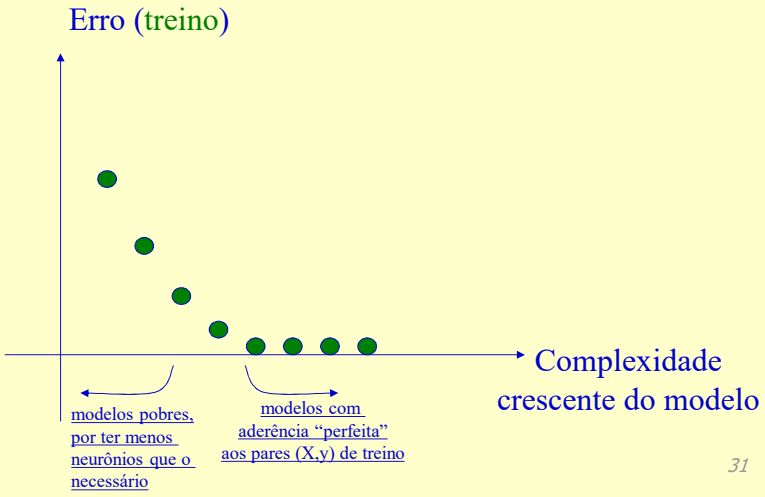
Um tema importante no curso:

<u>Sobreapredizado / Sobreajuste</u>

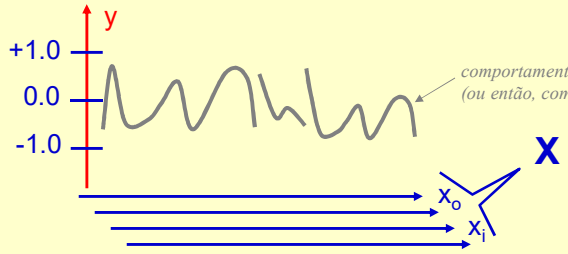
Conceito, entendimento da sua origem e formas de limitá-lo

Aumento de aderência aos dados de treino com o aumento de nós da RNA ...

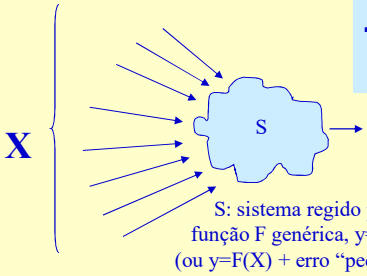
recordando



A função y(X) "a descobrir", num caso geral de função analógica y(X) ....



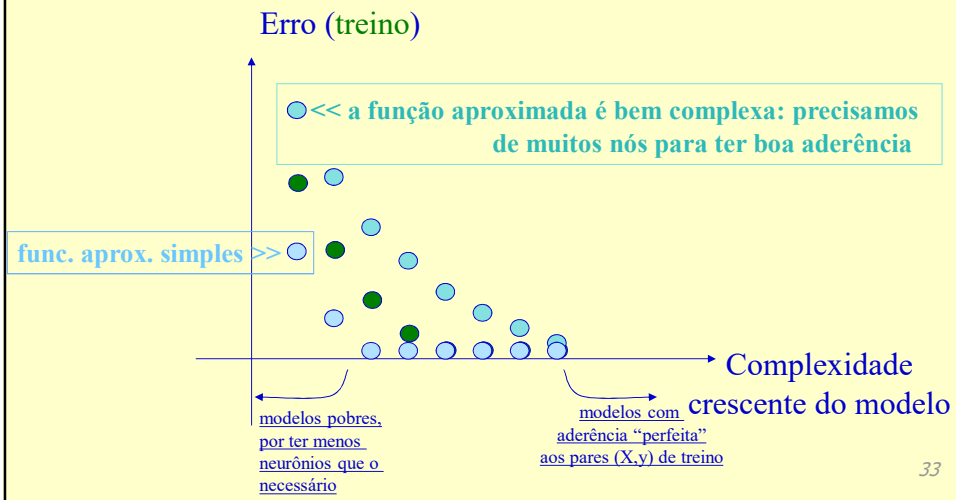
- ESTIMADOR ;
- MODELAGEM POR MAPEAMENTO X→y



(.... VALORES CONTÍNUOS!!!)

S: sistema regido por função F genérica, y=F(X) (ou y=F(X) + erro "pequeno")

Aumento de aderência aos dados de treino com o aumento de nós da RNA ...



Isto quer dizer que sempre é melhor termos um modelo com mais nós neurais que um modelo com menos nós neurais?

Afinal, da mesma maneira que a computação de um regressor polinomial de grau seis engloba a computação dos regressores polinomiais de graus menores, os modelos com mais nós neurais englobam os mais simples (em termos de capacidades de computações possíveis) correto?

Sim, correto! Mas há um limite no "lucro" em tal estratégia, dado pelo fenômeno de

Sobreaprendizado e perda de generalização ...

Sobreaprendizado:

Primeiro entendamos o conceito num contexto mais familiar (e simples), o de regressão polinomial univariada, para dados com tendência linear ou não linear de "grau" moderado, sujeitos a alguma flutuação não significativa em "y"

...

Depois, vocês pensem nos equivalentes dos nossos raciocínios, mas para as RNAs e outros tipos de modelos com número de parâmetros variável (complexidade variável) que você conheça ...

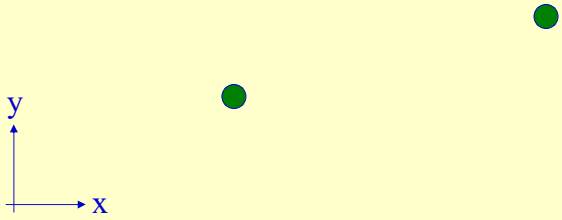
Falemos em lousa um pouco sobre a reta média para um conjunto de pares (x,y), a parábola média, a cúbica média ... etc

$$y \sim ax+b \; ; \quad y \sim ax^2 +bx +c \; ; \quad y \sim ax^3 +bx^2 +cx +d$$

falemos sobre regressão polinomial univariada, com o grau do polinômio aproximador podendo ser 1, 2, 3, ... 50, 51 etc.

$$y \sim ax^{51} +bx^{50} +cx^{49} + ...$$

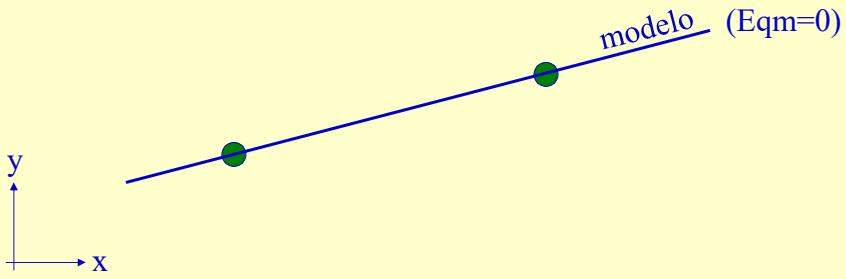
Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada



Os dados empíricos (x^{μ},y^{μ}) estão em verde;

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

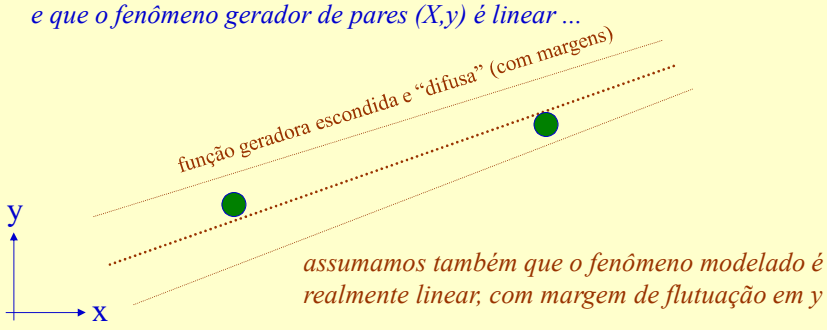
façamos uso de modelagem linear ...



Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

e que o fenômeno gerador de pares (X,y) é linear ...

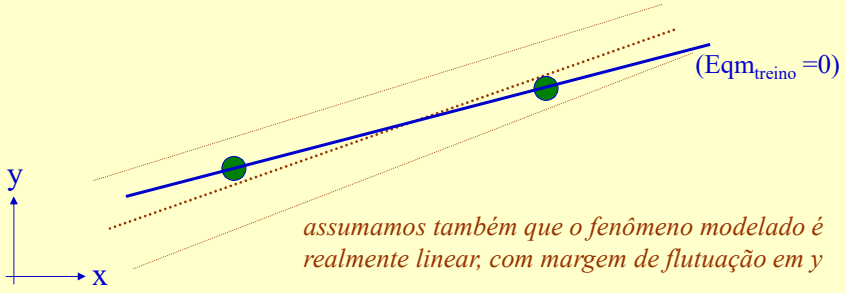


assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

façamos uso de modelagem linear ...

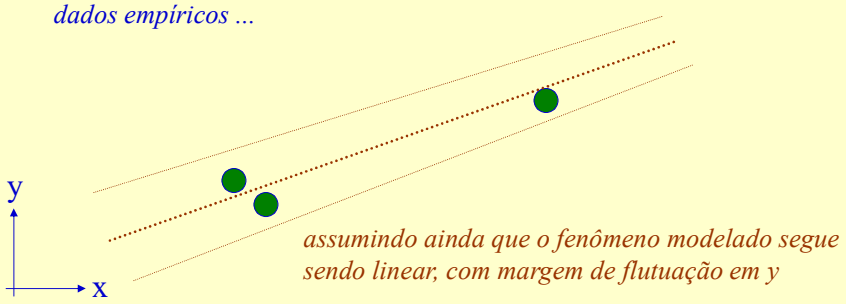


assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

Consideremos agora nova situação com mais dados empíricos ...

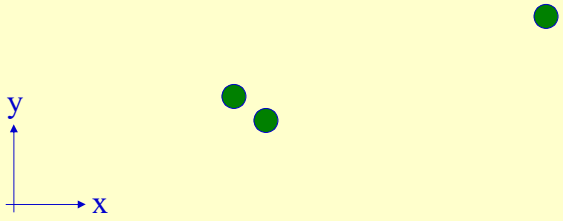


assumindo ainda que o fenômeno modelado segue sendo linear, com margem de flutuação em y

Os dados empíricos (x^{μ},y^{μ}) estão em verde;

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

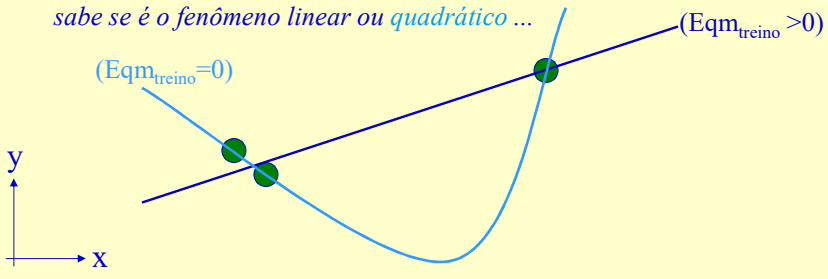
Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



Os dados empíricos (x^{μ},y^{μ}) estão em verde;

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

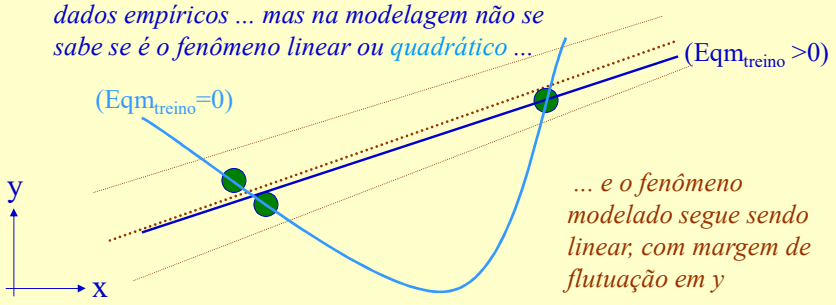
Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

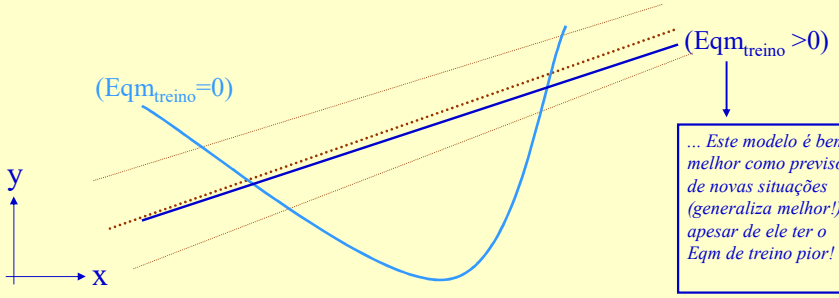
Consideremos agora nova situação com mais dados empíricos ... mas na modelagem não se sabe se é o fenômeno linear ou quadrático ...



... e o fenômeno modelado segue sendo linear, com margem de flutuação em y

Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

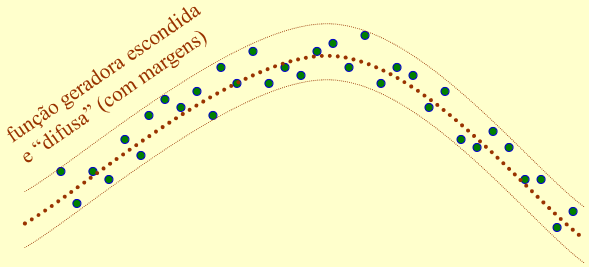


... Este modelo é bem melhor como previsor de novas situações (generaliza melhor!), apesar de ele ter o Eqm de treino pior!

Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo

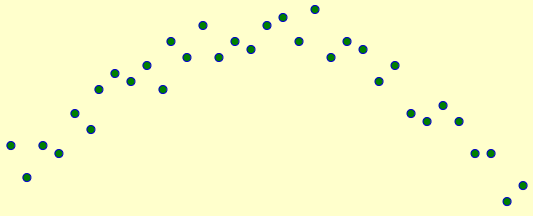


Os dados empíricos (x^{μ},y^{μ}) estão em verde;

O fenômeno gerador de pares (x,y) é quadrático em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo

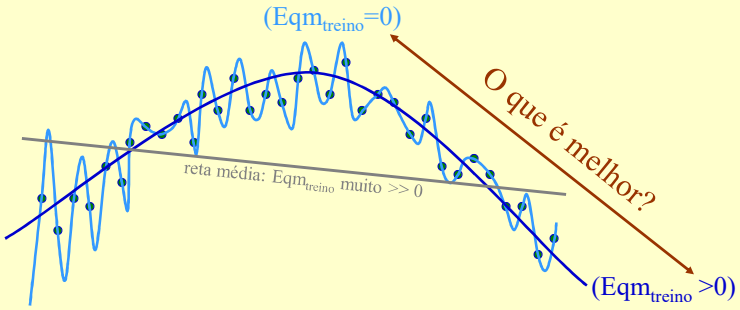


Os dados empíricos (x^{μ},y^{μ}) estão em verde;

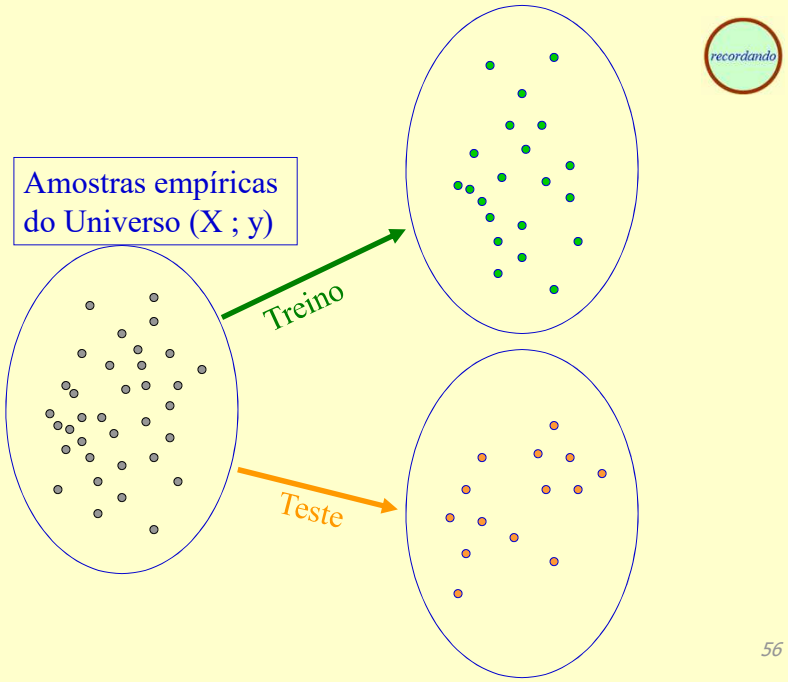
O fenômeno gerador de pares (x,y) é quadrático em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada

um novo exemplo

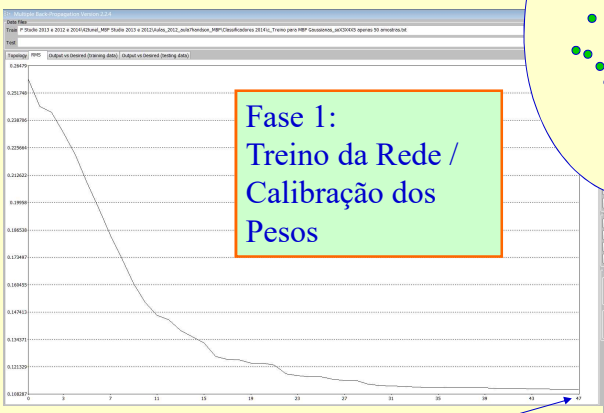


Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.

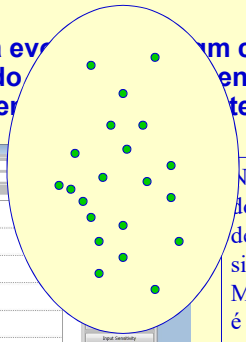


recordando

Gráfico fornecido pelo ambiente MBP da ev… m com o número de repetidos usos da "bússola do … endente": isto conecta o MBP com o gráfico apresen… terior



Fase 1:
Treino da Rede / Calibração dos Pesos



Nota: o RMS do eixo vertical deste gráfico significa Root Mean Square, e é a raiz quadrada do nosso conhecido Eqm

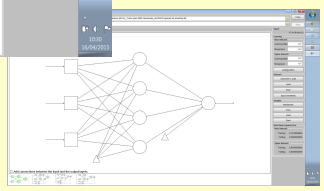
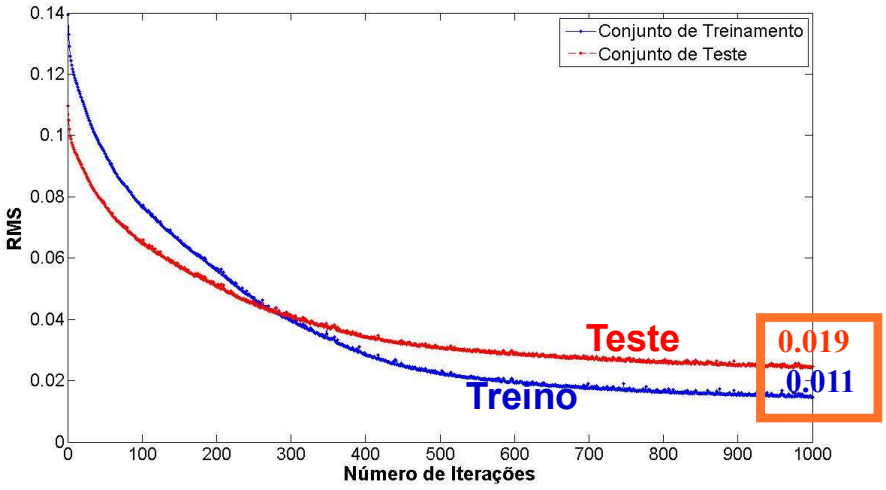


Gráfico mostrando as primeiras 47 iterações do processo de refinamentos sucessivos do modelo ...

Raiz do Erro Quadrático Médio (RMS) p/ conjuntos de treino e teste ... $RMS_{teste} > RMS_{treino}$

recordando



recordando

O Ciclo completo da modelagem:

0) Formalização do problema, mapeamento quantitativo em um modelo neural inicial e ... 0b) coleta de pares empíricos (X,y)

1) Fase de TREINO da RNA (MLP): com conhecimento dos X e dos y, que são ambos usados na calibração do modelo

2) Fase de TESTE / Caracterização da qualidade da RNA para generalizar: temos novos pares X e y, com y guardado "na gaveta", usado apenas para avaliação, não para re-calibração. É como um ensaio de uso final do modelo, com possibilidade de medir a sua qualidade com o y que foi guardado na gaveta.

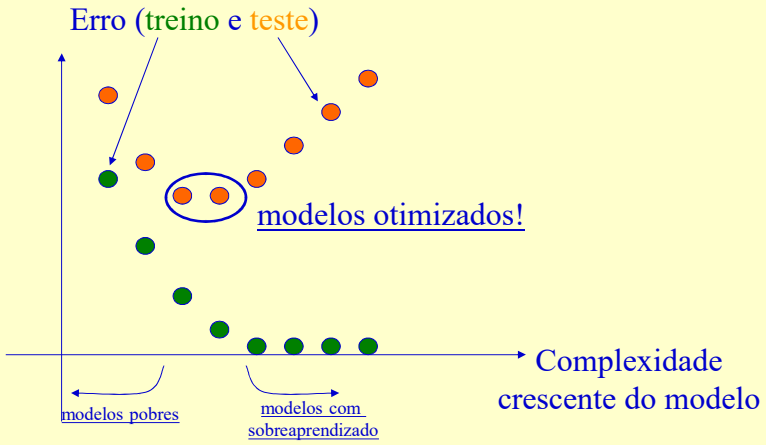
[Fase de refinamentos da RNA, dados e modelo, em ciclos, desde 0]

3) Fase de USO FINAL da RNA, com y efetivamente não conhecido, e estimado com conhecimento dos X + uso do modelo calibrado.

.... Diferenças e semelhanças entre 1, 2 e 3

59

Sobreaprendizado em "sumário executivo"



60

Mas cuidado ... você experimentou rodar mais de uma vez o MBP sobre os mesmos dados de treino e ver se o resultado final é o mesmo?

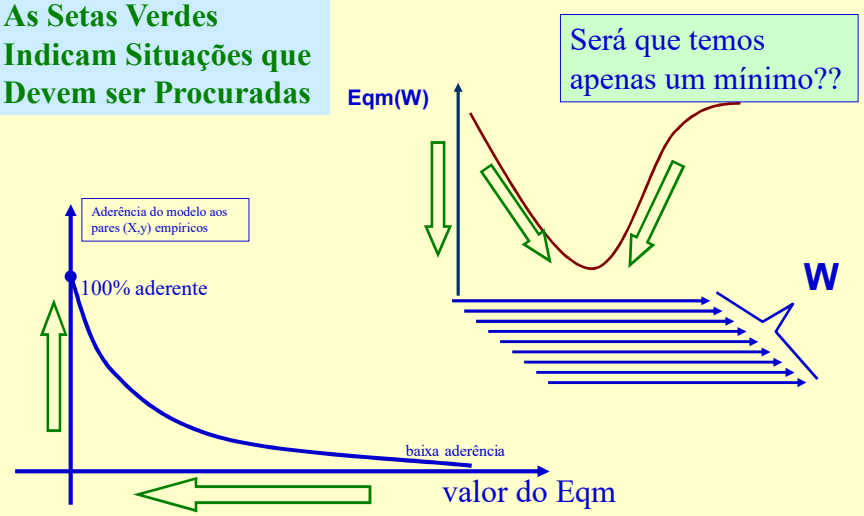
- **Não há garantia de que o W randômico, seguido de Gradiente Descendente leve ao mínimo global quando há vários mínimos, só há garantia de que levará a um dos mínimo locais, que pode não ser o mínimo global**
- **Isto faz necessário rodar várias vezes a otimização de pesos e selecionar a configuração com melhor Eqm**
- **É legítimo ficarmos com a rede otimizada de menos Eqm final? Ou deveríamos ter um ataque de avaliar o Eqm médio de várias tentativas?**

- ***Outro assunto ... e quanto à variação associada ao k-fold Cross Validation? Média ou melhor Eqm?***

O que devemos mirar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

Devemos buscar Maximização da aderência = Minimo Eqm possível

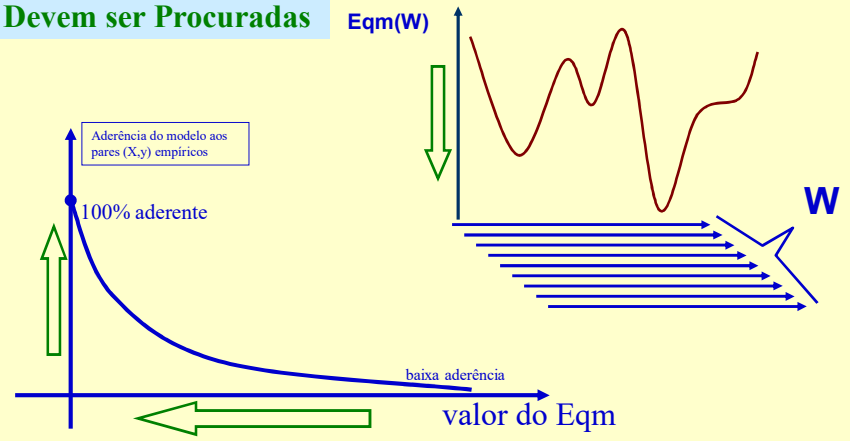
As Setas Verdes Indicam Situações que Devem ser Procuradas



O que devemos mirar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

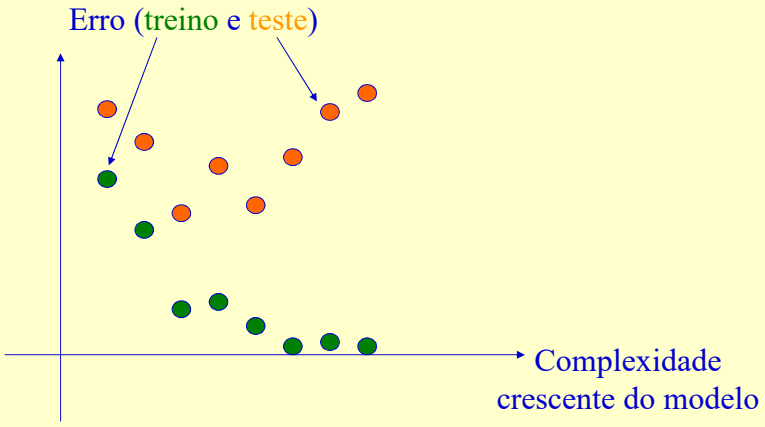
Devemos buscar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas



64

Atenção para componentes randômicas que impactam muito quando se faz um único ensaio de medida de erro, para cada tamanho de rede específico (um ensaio apenas, para cada grau de complexidade) ...



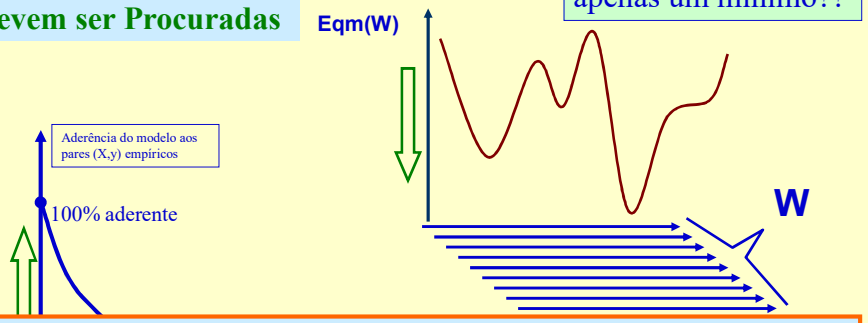
65

O que devemos mirar quando exploramos o espaço de pesos W buscando que a RNA seja um bom modelo?

Devemos buscar Maximização da aderência = Minimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas

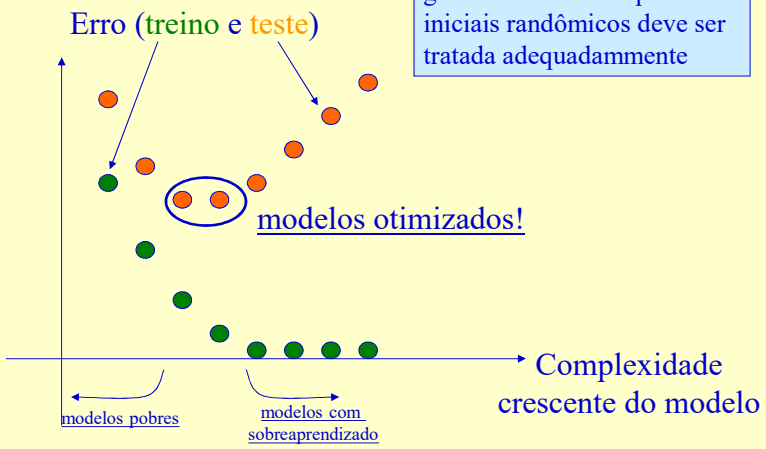
Será que temos apenas um mínimo??



... Para não sermos reféns de mínimos locais com alto Eqm, podemos aplicar o gradiente descendente repetidamente na mesma RNA, com novos pesos iniciais randômicos em cada rodada, mantendo para o modelo final apenas os valores de pesos associados ao ensaio com o melhor dos resultados finais no Eqm!

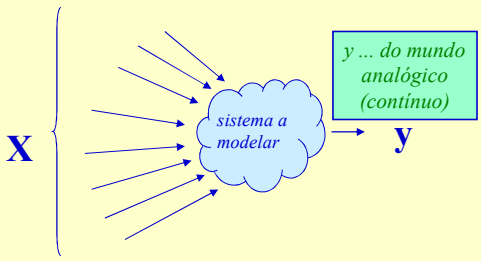
Com repetidos ensaios em cada grau de complexidade os mínimos locais são evitados e detectamos adequadamente o sobreaprendizado

Ou seja, a flutuação nos gráficos devida aos pesos iniciais randômicos deve ser tratada adequadammente

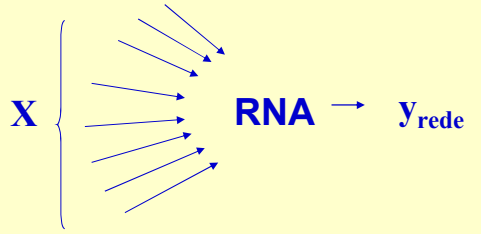


Identificando os ingredientes para o risco de sobreaprendizado nos contextos de regressão multivariada e de reconhecimento de padrões multivariado

Modelagem de um sistema por função de mapeamento X -> y (a RNA como regressor analógico não linear multivariável)

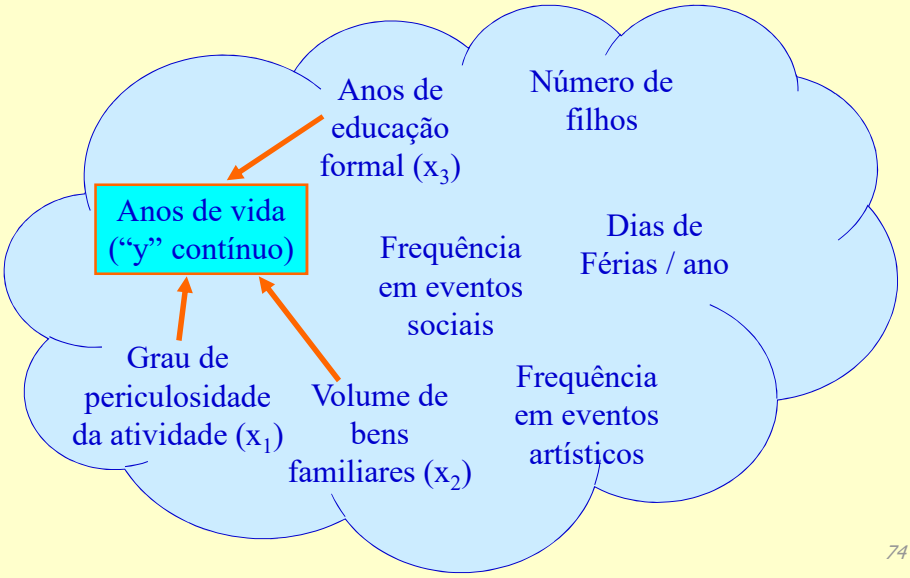


Assumimos que a variável y do sistema a modelar é uma função (normalmente desconhecida e possivelmente não linear) de diversas outras variáveis desse mesmo sistema

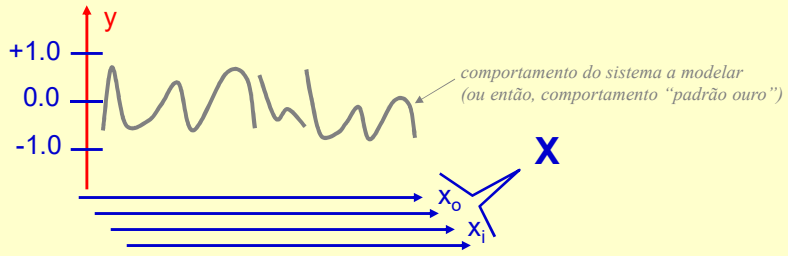


A RNA , para ser um bom modelo do sistema, deve reproduzir essa relação entre X e y, tão bem quanto possível

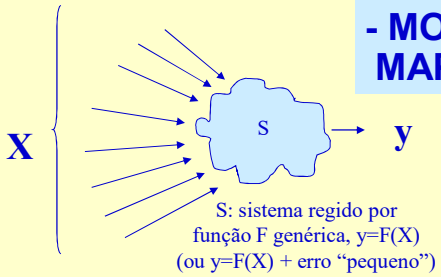
Um hipotético universo de variáveis inter-dependentes, passível de modelagem/ens



A função y(X) "a descobrir", num caso geral de função analógica y(X) ....



- ESTIMADOR ;
- MODELAGEM POR MAPEAMENTO X→y



Cenário mais real: a "função" y(X) do sistema modelado é "difusa": $y=F_{médio}(X)+$ *flutuação* ....

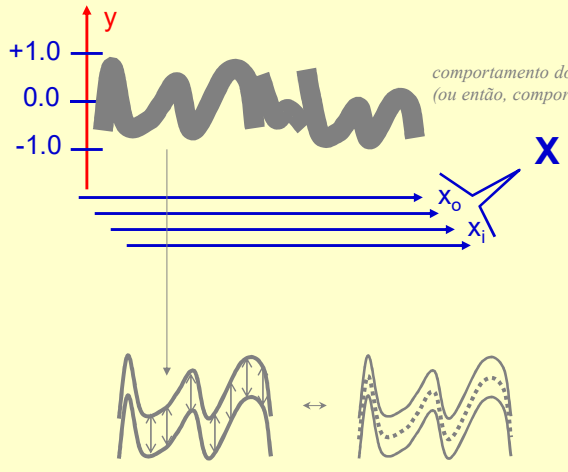
comportamento do sistema a modelar
(ou então, comportamento "padrão ouro")

.... em problemas concretos / reais, há sempre alguma ambiguidade no mapeamento que leva valores de X a valores de y. Para decepção de Cybenko, não temos uma função y=F(X) no sentido matemático exato, pois para uma dada ênupla de valores X fixados, temos tipicamente uma faixa de valores que podem ser observados para a variável y: $y=F_{médio}(X)+$ flutuação.

Neste cenário, buscamos que o modelo capture o comportamento médio das relações observadas entre X e y:

... y_{rede} ~ ymédio esperado para um dado X

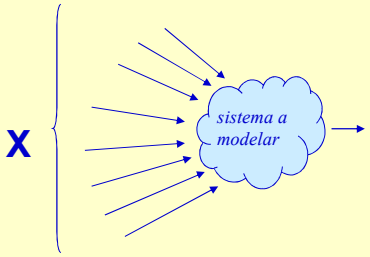
Cenário mais real: a "função" y(X) do sistema modelado é "difusa": $y=F_{médio}(X)+$ *flutuação* ....



comportamento do sistema a modelar
(ou então, comportamento "padrão ouro")

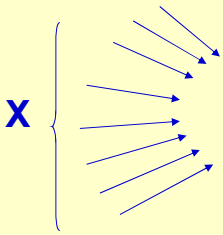
... neste cenário, podemos olhar para o y observado no sistema que se deseja modelar não mais como um valor específico bem definido, mas como um valor médio esperado (dado valor de X) e uma faixa de valores em torno desse valor médio esperado.

RNAs como reconhecedor / detetor de padrões ...



y ... responde à seguinte pergunta:
X é padrão notável? **Sim ( ) Não ( )**

y ... é do universo discreto / binário (SIM versus NÃO), y não é analógico

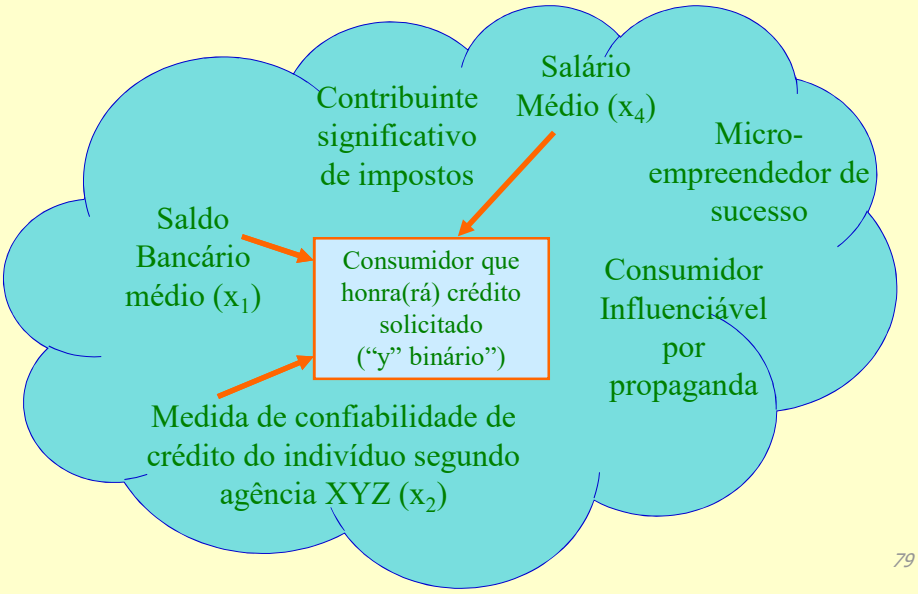


RNA → y_{rede} ... **+1 ( )** *versus* **- 1 ( )**

... ou alternativamente: y_{rede} ... **0.5 ( )** *versus* **- 0.5 ( )**
... ou alternativamente: ... y_{rede} **> 0 ( )** *versus* y_{rede} **< 0 ( )**

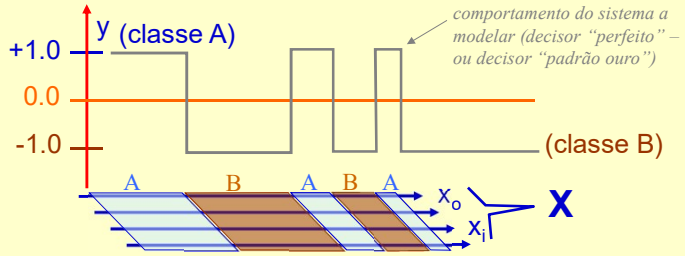
78

Um hipotético universo de variáveis inter-dependentes, passível de modelagem/ens

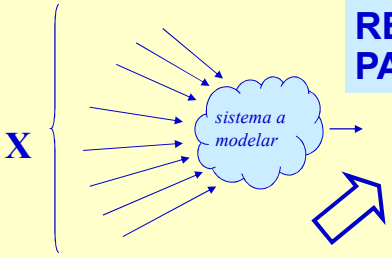


79

Caso de classificação binária / reconhecimento de padrões, será do tipo ...



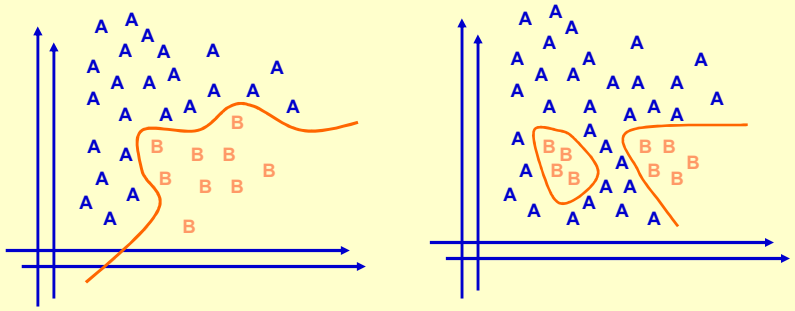
CLASSIFICADOR; RECONHECEDOR DE PADRÕES NOTÁVEIS



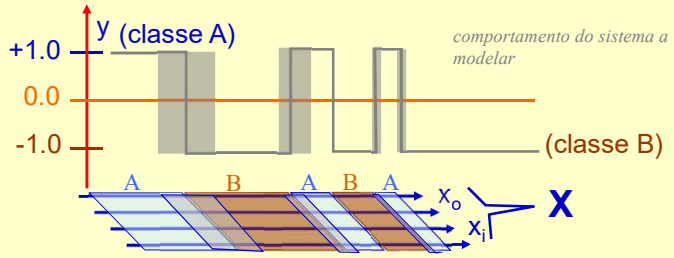
y (... restrita a valores +1 e –1)

Capacidade de reconhecimento de padrões em casos complexos NÃO LINEARES

Com as RNAs, a hipersuperfície de separação entre classes vai muito além dos hiperplanos



Cenário mais real: a separação entre regiões do espaço de X não é perfeitamente definida ....

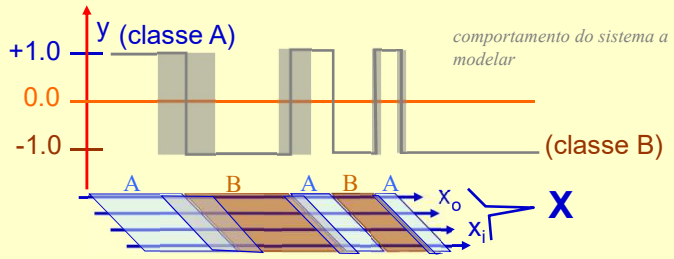


.... em problemas concretos / reais, há sempre alguma ambiguidade no mapeamento que leva valores de X aos valores discretos de y. Não temos uma função y=F(X) no sentido matemático exato, pois para uma dada ênupla de valores X fixado temos em alguns casos de fronteira a possibilidade de observar no y empírico tanto a classe A quanto a classe B: y=A ou B, com maior ou menor probabilidade para cada classe de acordo com o X. Neste desejamos que o modelo capture o comportamento médio das relações observadas entre X e y:

... y_{rede} ~ classe 'mais esperada' para um dado X

82

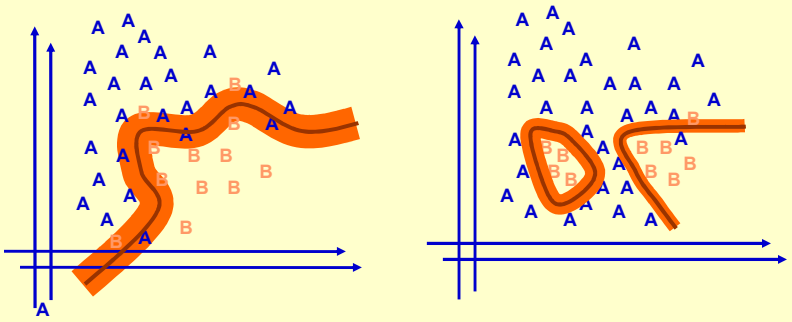
Cenário mais real: a separação entre regiões do espaço de X não é perfeitamente definida ....



... podemos olhar para o y (classe A ou B) observado no sistema que se deseja modelar não mais como uma classe sempre bem definida e com fronteiras de separação entre A e B bem definidas no espaço de valores de X, mas como sendo delineadas na modelagem através de fronteiras com eventuais faixas de tolerância e com sobreposição parcial das classes no espaço de X

83

Situações de classes com sobreposição parcial no espaço de atributos X ; situações de fronteiras de separação difusas ...



84

Ciclos de refinamento num projeto de modelagem para reconhecimento e/ou regressão, com duração de médio / longo prazo e com interesse de otimização de desempenho máxima

O Ciclo completo da modelagem:

0) Formalização do problema, mapeamento quantitativo em um modelo neural inicial e ... 0b) coleta de pares empíricos (X,y)

1) Fase de TREINO da RNA (MLP): com conhecimento dos X e dos y, que são ambos usados na calibração do modelo

2) Fase de TESTE / Caracterização da qualidade da RNA para generalizar: temos novos pares X e y, com y guardado "na gaveta", usado apenas para avaliação, não para re-calibração. É como um ensaio de uso final do modelo, com possibilidade de medir a sua qualidade com o y que foi guardado na gaveta.

[Fase de refinamentos sucessivos da RNA e/ou dos dados e/ou do modelo, em ciclos diversos, recomeçando desde o passo 0 ou do passo 1]

3) Fase de USO FINAL da RNA, com y efetivamente não conhecido, e estimado com conhecimento dos X + uso do modelo calibrado.

.... Diferenças e semelhanças entre 1, 2 e 3

... Aspectos de refinamentos nos projetos ...

Diferentes aspectos estudados no treinamento / otimização / caracterização da RNA ...

- Vários Delta W sequenciados (gradiente descendente)
- Re-sorteios de pesos iniciais (fugindo de mínimos locais)
- k-fold cross validation (avaliando sensibilidade aos dados empíricos)
- Diferentes graus de complexidade do modelo neural (evitando sobreaprendizado)
- Ensaios com vários Pré-Processamentos alternativos (aumentando desempenho)
- Descarte de algumas variáveis de menor relevância, p/ melhora do desempenho
- Aumento de M+M' c/ as mesmas variáveis (há custo extra com novas coletas X;y)
- Acréscimo de variáveis x incluídas no modelo (há custo extra com novas coletas)
- ... Outros ...

... Multi-ciclos de refinamentos diversos que são aplicados em projetos mirem a otimização de desempenho ao máximo ...

Alguns dos loops de treinamento e/ou otimização e/ou caracterização:

Loop dos Delta W sequenciados (já é intrínseco ao gradiente descendente)

Loop dos re-sorteios de pesos iniciais (fugindo de mínimos locais)

Loop do k-fold cross validation (avaliando e medindo a sensibilidade aos dados empíricos)

Loop de aumento gradual de complexidade do modelo neural (evitando sobreaprendizado)

Ensaios com vários Pré-Processamentos alternativos (aumentando desempenho)

Descarte de algumas variáveis de menor relevância (limitando ruído e sobreaprendizado)

Aumento de M+M' c/ as mesmas variáveis (há custo extra com novas coletas X;y)

Acréscimo de variáveis x incluídas no modelo (há custo extra com novas coletas)

... Multi-ciclos de refinamentos nos projetos ...

Discutamos os loops de treinamento e/ou otimização e/ou caracterização ...

1- Loop dos Delta W sequenciados (intrínseco ao gradiente descendente)

2- Loop dos re-sorteios de pesos iniciais (fugindo de mínimos locais)

3- Loop do k-fold cross validation (avaliando sensibilidade aos dados empíricos)

4- Loop de aumento gradual de complexidade do modelo neural (evitando sobreaprendizado)

5- Ensaios com vários Pré-Processamentos alternativos (aumentando desempenho)

6- Descarte de algumas variáveis de menor relevância, p/ melhora do desempenho

7- Aumento de M+M' c/ as mesmas variáveis (há custo extra com novas coletas X;y)

8- Acréscimo de variáveis x incluídas no modelo (há custo extra com novas coletas)

Essa ordem de aninhamentos se aplica bem ao seu projeto?
Ou há outra/s ordem/ns que faz/em mais sentido para as SUAS circunstâncias específicas?