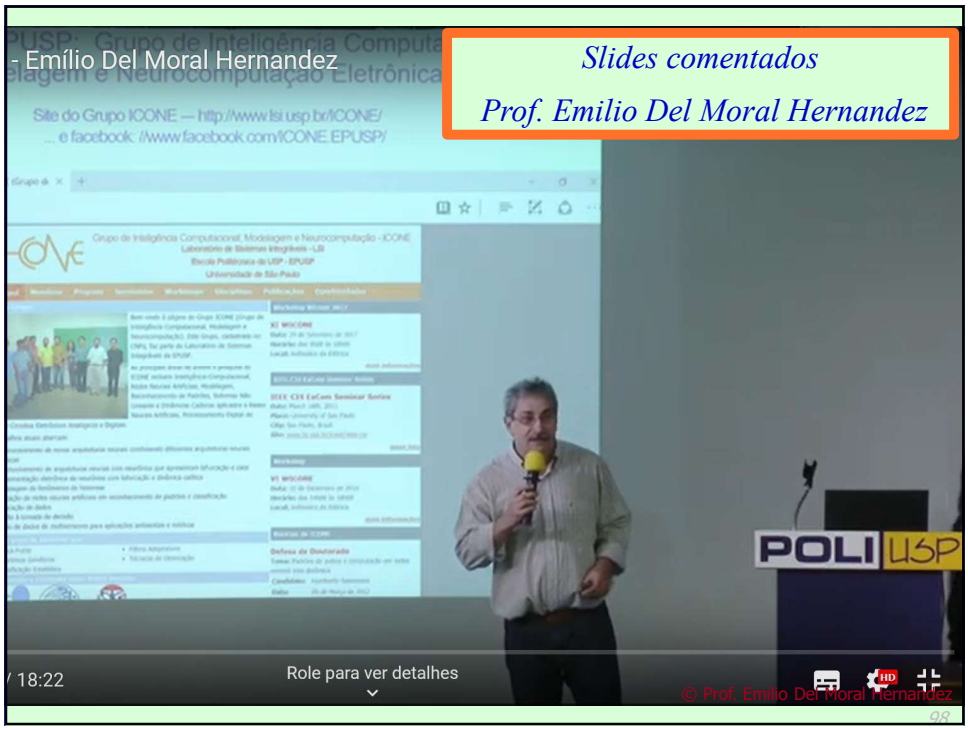
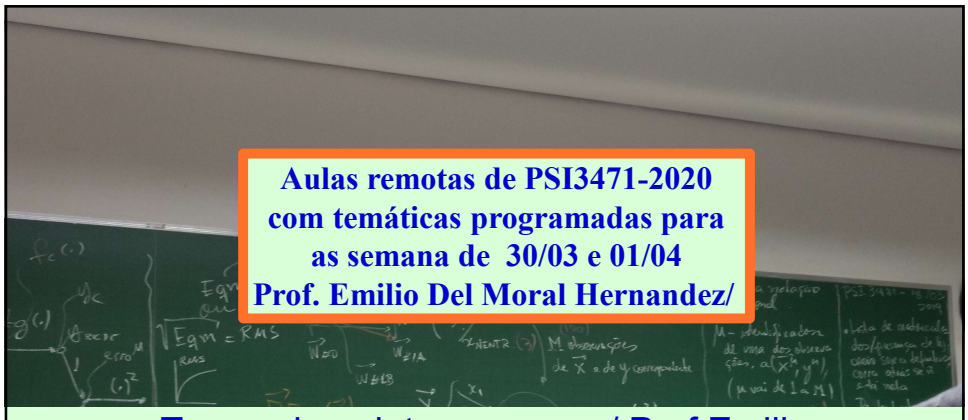


98



Temas da quinta semana c/ Prof Emilio

#9 (30/março – 2ªf) Foco da semana: Regressores - Casos simples de aproximação de funções univariadas. Teorema de Cybenko: o MLP como aproximador universal de funções multivariadas; implicações práticas do teorema para a implementação de regressores e reconhecedores de padrões não lineares multivariados genéricos.
#10 (01/abril – 4ªf) ... Medidas de qualidade diversas para regressores multivariados (distintas do erro quadrático médio); Flutuação do desempenho do modelo com as particulares amostras de treino e de teste e técnicas de reamostragem; técnica de validação cruzada, k-fold cross validation e leave one out. Sobreajuste / sobreaprendizado / perda de generalização em regressão polinomial e em redes neurais; limitação do número de nós neurais para evitar o sobreajuste e otimizar a generalização da rede neural; partição do volume de observações em conjuntos de treino, validação e teste.
06 e 08 de abril: Semana Santa – não há aula

... nestes slides: segunda destas 2 aulas (#10 -01/04)

99

Temas da quinta semana c/ Prof. Emilio

#9 (30/março – 2ªf) Foco da semana: Regressores - Casos simples de aproximação de funções univariadas. Teorema de Cybenko: o MLP como aproximador universal de funções multivariadas; implicações práticas do teorema para a implementação de regressores e reconhecedores de padrões não lineares multivariados genéricos.
#10 (01/abril – 4ªf) Medidas de qualidade diversas para regressores multivariados (distintas do erro quadrático médio): Flutuação do desempenho do modelo com as particulares amostras de treino e de teste e técnicas de reamostragem; técnica de validação cruzada, k-fold cross validation e leave one out. Sobreajuste / sobreaprendizado / perda de generalização em regressão polinomial e em redes neurais; limitação do número de nós neurais para evitar o sobreajuste e otimizar a generalização da rede neural; partição do volume de observações em conjuntos de treino, validação e teste.

100

100

Sobreajuste
=
"Sobreaprendizado"

101

... Uma pergunta que muitos de vocês devem estar se fazendo ... Mas afinal, como escolho o número de neurônios na RNA, já que a prova de Cybenko não diz nada a respeito?

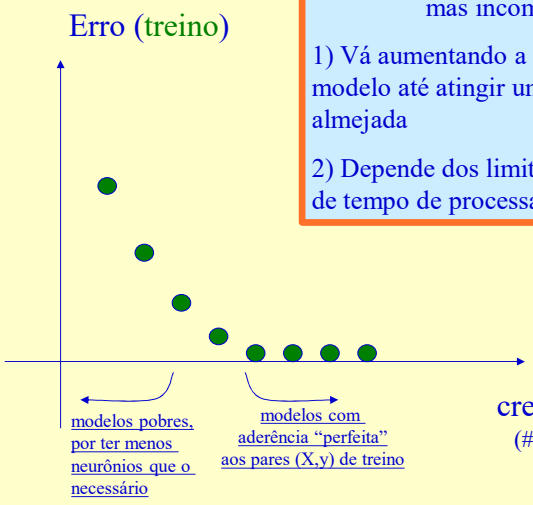
102

Quantos neurônios devo usar na RN?

Algumas respostas simples, mas incompletas ...

1) Vá aumentando a complexidade do modelo até atingir uma aderência almejada

2) Depende dos limites de computação / de tempo de processamento



103

Quantos neurônios devo usar na RN?

Algumas respostas simples,
mas incompletas ...

1) Vá aumentando a complexidade do modelo até atingir uma aderência almejada

2) Depende dos limites de computação / de tempo de processamento

... um terceiro critério, mais sofisticado ...

3) O número de neurônios que maximize o desempenho da rede na generalização / que minimize a sua degradação por

Sobreajuste

04

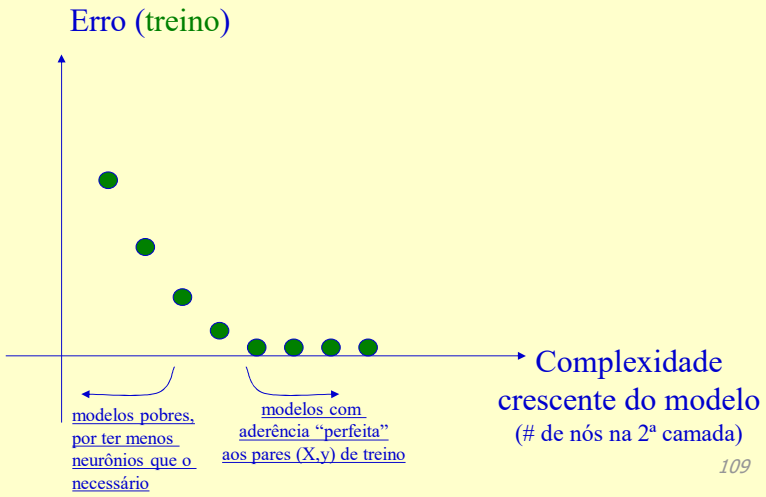
104

... Passos preliminares ao tema: resgatemos alguns aspectos d prova de Cybenko referentes ao grau de ajuste entre modelo neural e função sendo aproximada por esse modelo

108

108

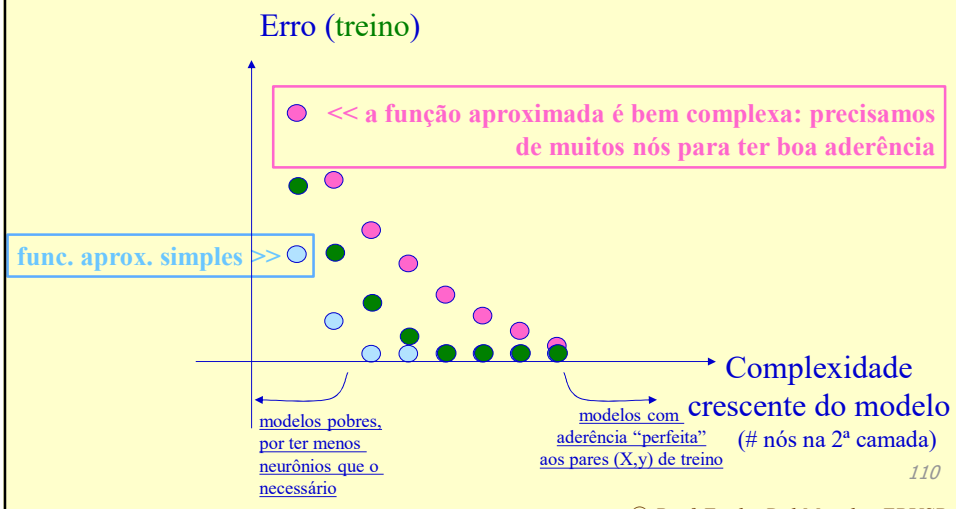
Aumento de aderência aos dados de treino com o aumento de nós da RNA ...



109

109

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



110

110

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 ...

Sobreajuste em polinômios:

Entendamos inicialmente o conceito de sobreajuste / sobreaprendizado num universo mais familiar (e simples) do que a modelagem com RNAs; trabalhemos no universo de modelagem por regressão polinomial univariada, usada para representar dados com comportamentos lineares e não lineares.

Depois, vocês mesmos podem estender os nossos raciocínios feitos aqui no universo de polinômios, para o universo de regressão por RNAs, e mesmo para outros tipos de modelos, com número de parâmetros variável (complexidade variável) que você conheça ...

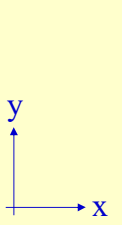
Falemos em lousa um pouco sobre a reta média para um conjunto de pares (x,y), a parábola média, a cúbica média ... etc

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

e mais além, falemos sobre regressão polinomial univariada, com o grau do polinômio aproximador podendo ser 1, 2, 3, ou mesmo graus bastante mais altos como 50, 51 etc.

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

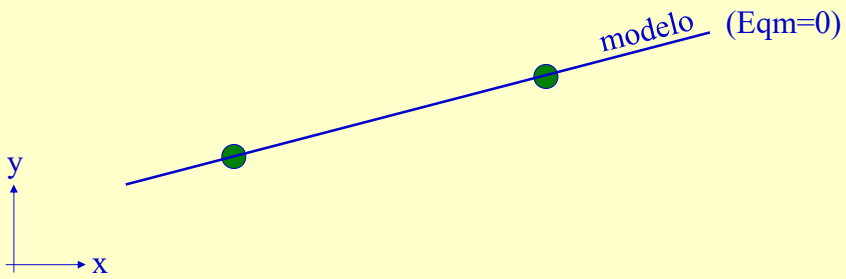
Sobreaprendizado ilustrado em sua fenomenologia, na regressão polinomial (de vários graus) univariada



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 ...

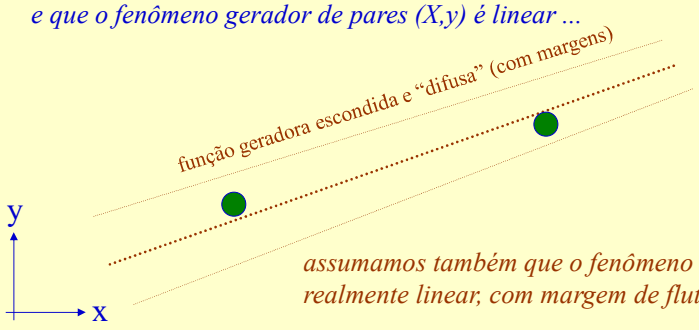


Os dados empíricos (x^{μ},y^{μ}) estão em verde;
O modelo linear gerado a partir dos dados, em azul.

118

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

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



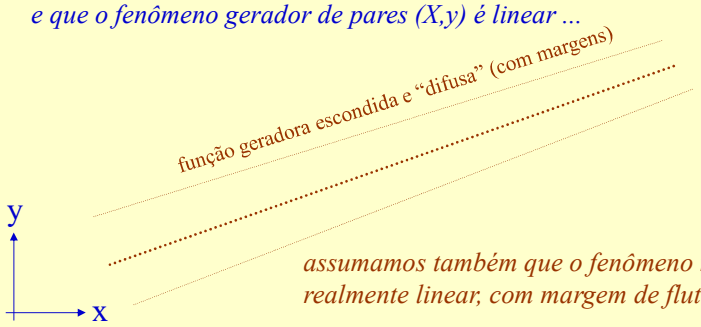
assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

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

119

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

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

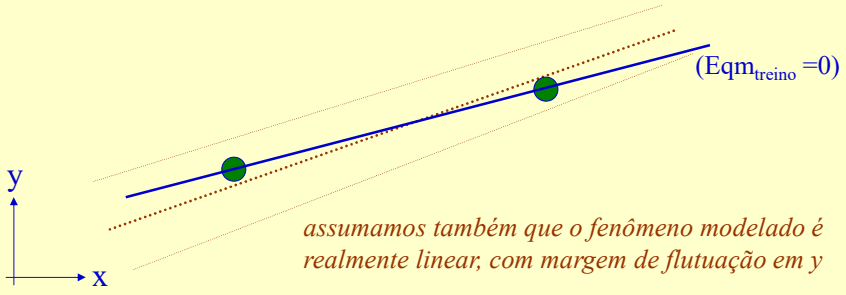


assumamos também que o fenômeno modelado é realmente linear, com margem de flutuação em y

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 ...

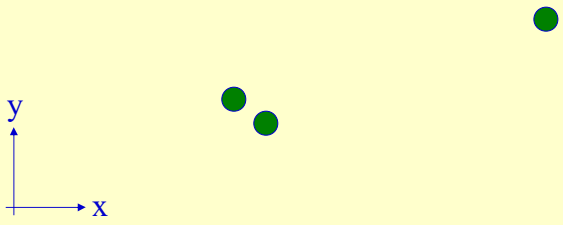
assumindo ainda que o fenômeno modelado segue sendo linear, com margem de flutuação em y

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

122

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 ...

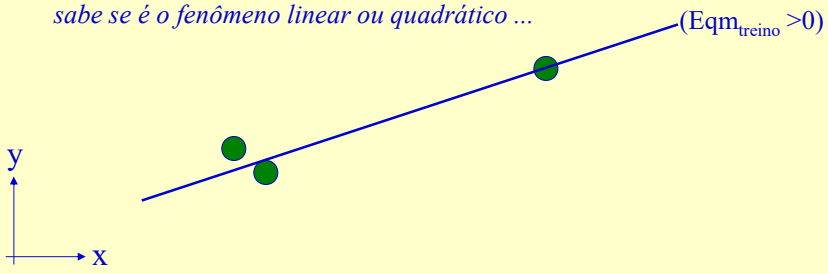


Os dados empíricos (x^{μ},y^{μ}) estão em verde;

123

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 ...



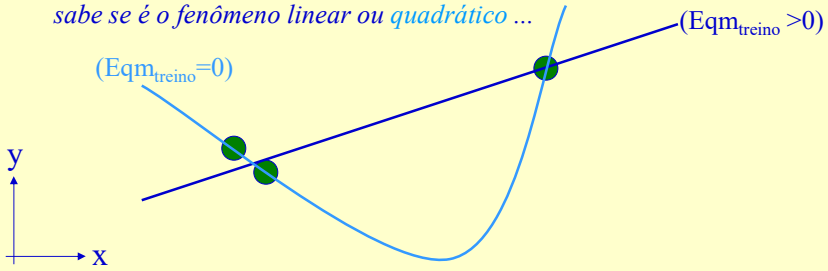
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

124

124

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 ...



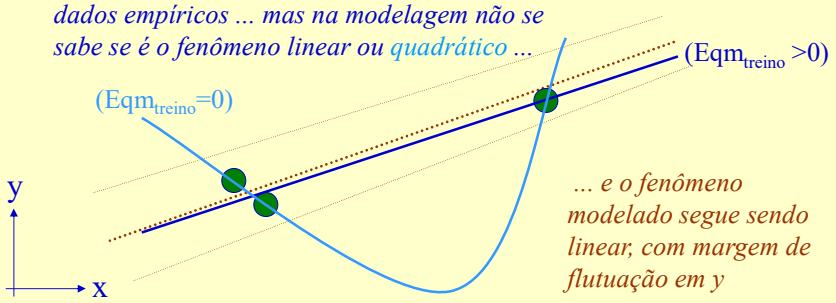
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis. O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

125

125

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 ...



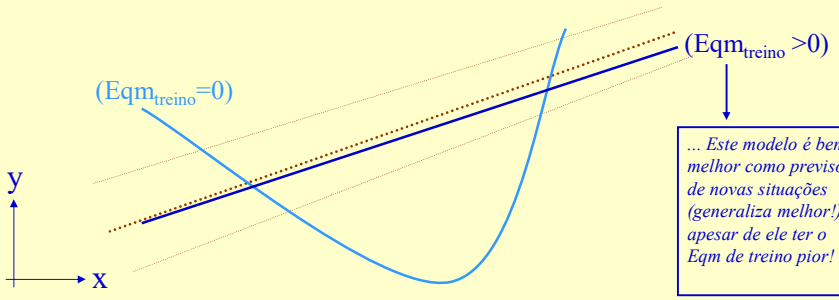
... e o fenômeno modelado segue sendo linear, com margem de flutuação em y

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

126

126

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!

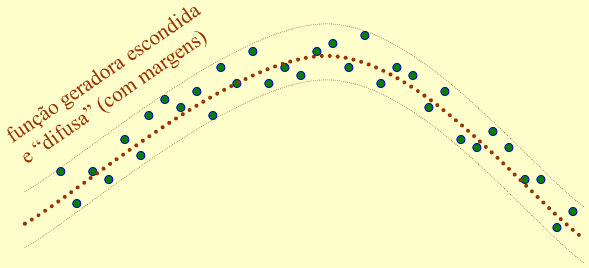
Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Dois modelos polinomiais gerados a partir dos dados, em azuis.
O fenômeno gerador de pares (x,y) é linear em essência, mas tem alguma flutuação randômica em y. A tendência e os limites da flutuação estão representados em marrom

127

127

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

um novo exemplo



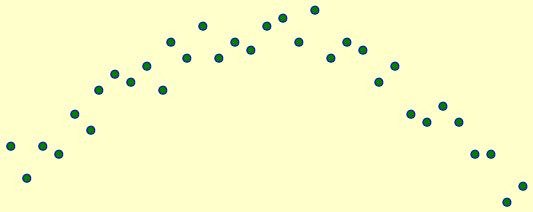
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

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

128

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

um novo exemplo



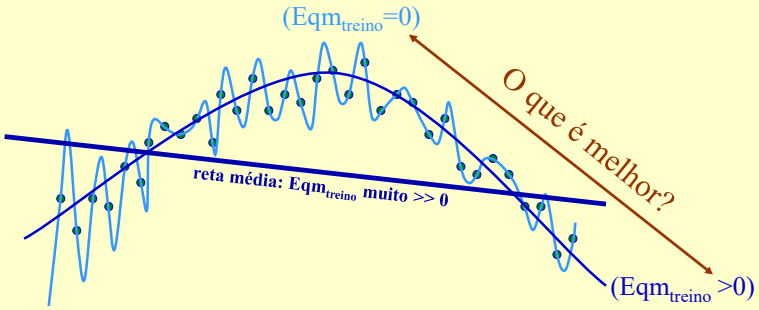
Os dados empíricos (x^{μ},y^{μ}) estão em verde;

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

129

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

um novo exemplo

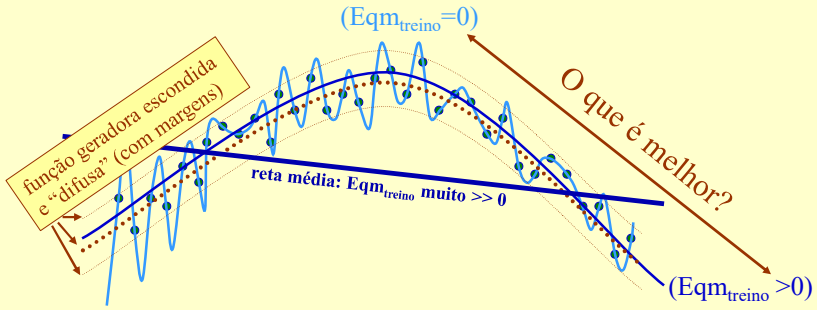


Os dados empíricos (x^{μ},y^{μ}) estão em verde;
Três modelos polinomiais gerados a partir dos dados, em azuis.

130

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;
Três modelos polinomiais gerados a partir dos dados, em azuis.

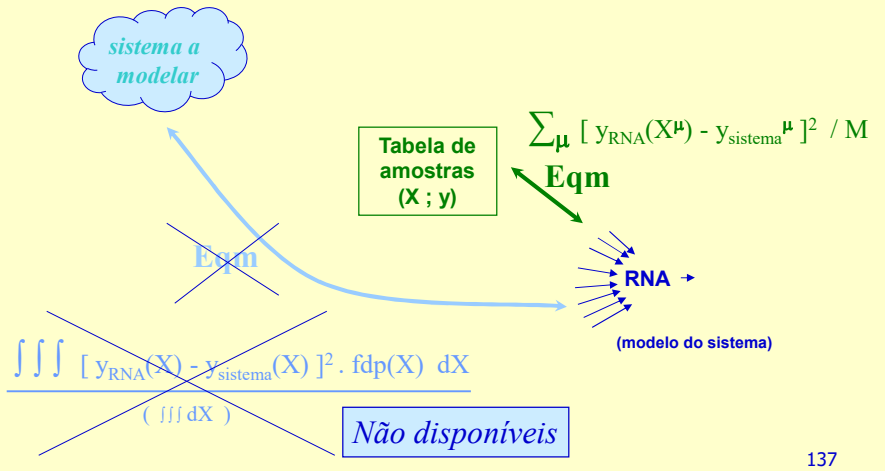
131

Como saber se os efeitos negativos do sobreajuste na degradação no modo inferência / efeitos de degradação da generalização, estão ocorrendo ou não em uma dada modelagem / para um dado modelo de regressão ???

(seja ele polinomial, seja ele neural)

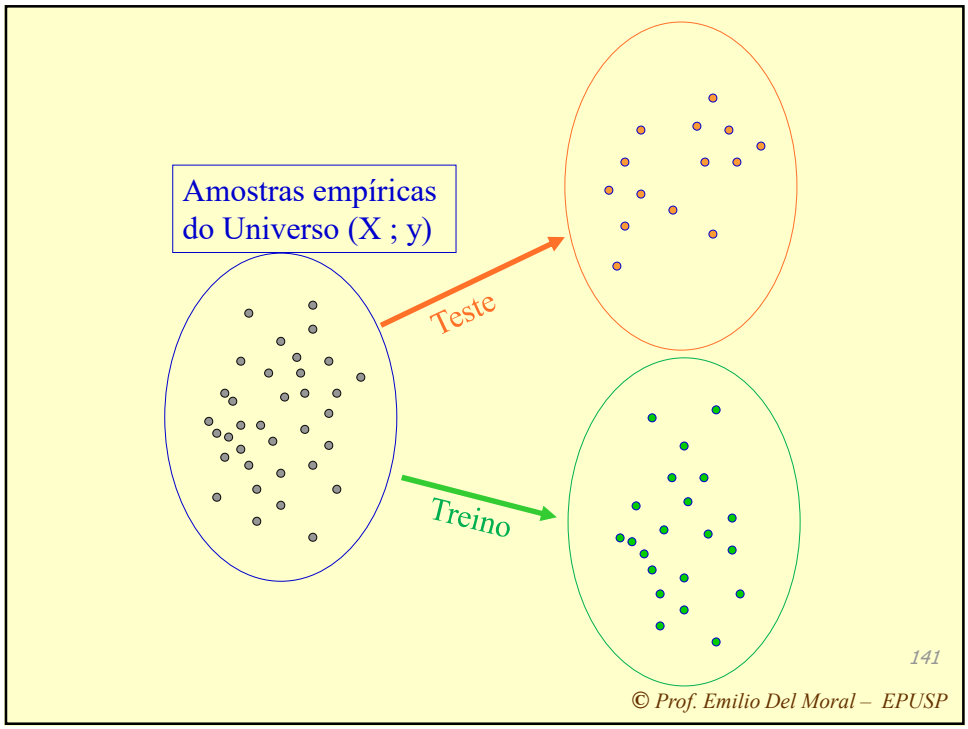
Como quantificar a qualidade de generalização de diversos modelos que queremos contrastar (treinados a partir dos mesmos dados empíricos)?

Como avaliar a qualidade de generalização da RNA?



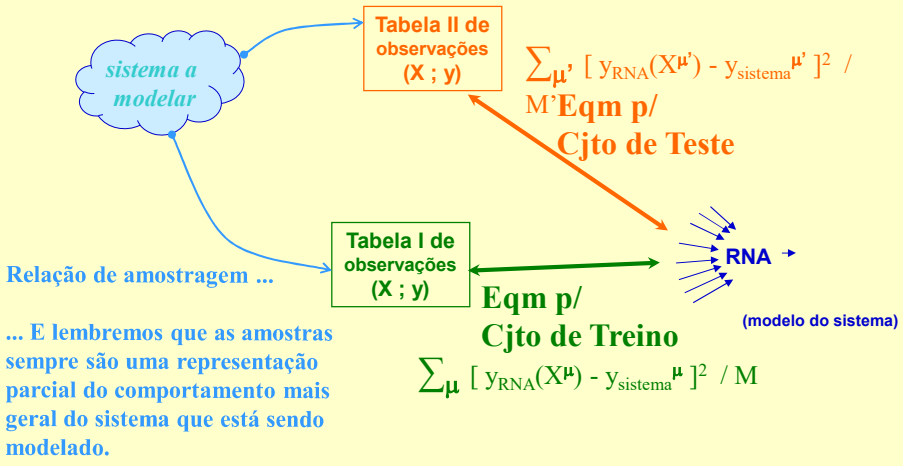
Necessidade de validação cruzada e do conceito de conjuntos empíricos de treino e de teste

140



141

O treinamento mira minimizar o Eqm das amostras (X ; y) de treino. A avaliação de qualidade de generalização é feita a partir de Eqm com novas amostras (X ; y) ... Estas são chamadas amostras de teste :



Relação de amostragem ...

... E lembremos que as amostras sempre são uma representação parcial do comportamento mais geral do sistema que está sendo modelado.

142

142

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 <u>novos</u> 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]

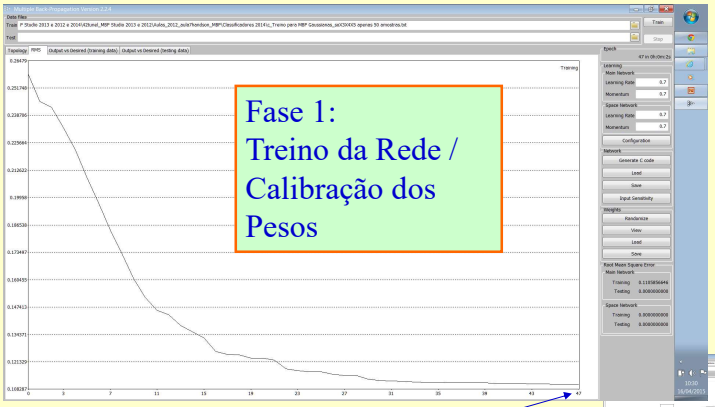
3) Fase de USO FINAL da RNA, com y efetivamente não conhecido, e estimado com conhecimento dos X + uso do modelo calibrado.

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

143

143

Gráfico fornecido pelo ambiente MBP da evolução do Eqm com o número de repetidos usos da "bússola do gradiente descendente": isto conecta o MBP com o gráfico apresentado no slide anterior



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

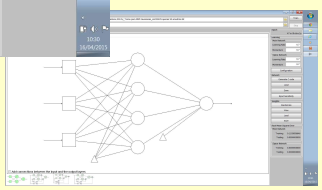
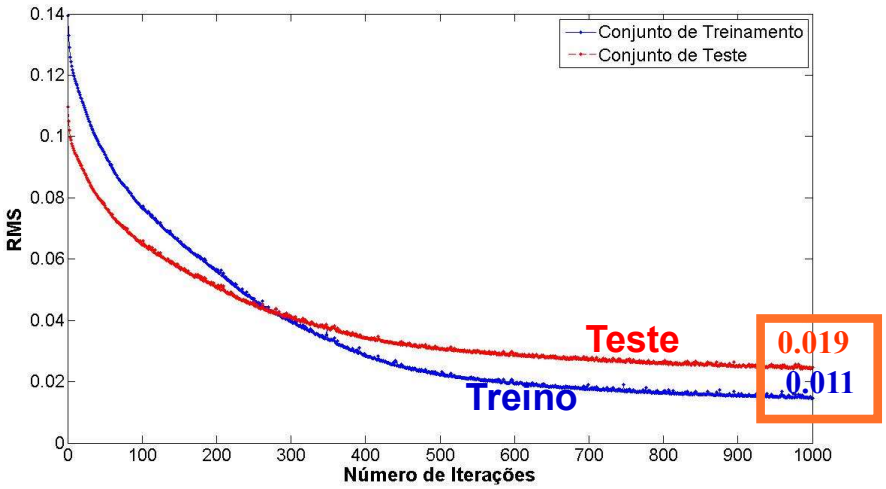


Gráfico mostrando as primeiras 47 iterações do processo de refinamentos sucessivos do modelo ...

144

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



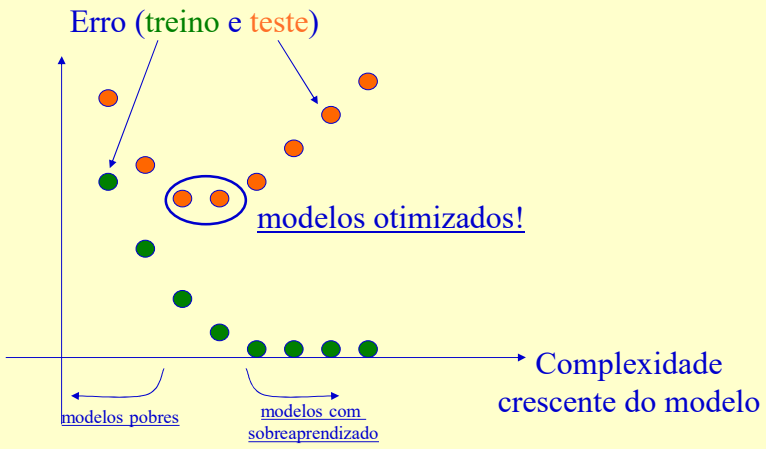
145

145

Limitando o sobreajuste pela limitação da complexidado da Rede Neural (~número de neurônios)

154

Sobreaprendizado em "sumário executivo"



155

155

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

157

O conceito de sobreaprendizado nos dá critérios adicionais para a definição do número de neurônios / grau de complexidade de uma rede neural, critérios esses que vão bem além de simples economia computacional; esses critérios miram o aumento de precisão na generalização

165

Quantos neurônios devo usar na RN?

Algumas respostas simples,
mas incompletas ...

1) Vá aumentando a complexidade do modelo até atingir uma aderência almejada

2) Depende dos limites de computação / de tempo de processamento

... um terceiro critério, mais sofisticado ...

3) O número de neurônios que maximize o desempenho da rede na generalização / que minimize a sua degradação por

Sobreajuste

66

166

Revisitando os Conjuntos de Dados Empíricos ...

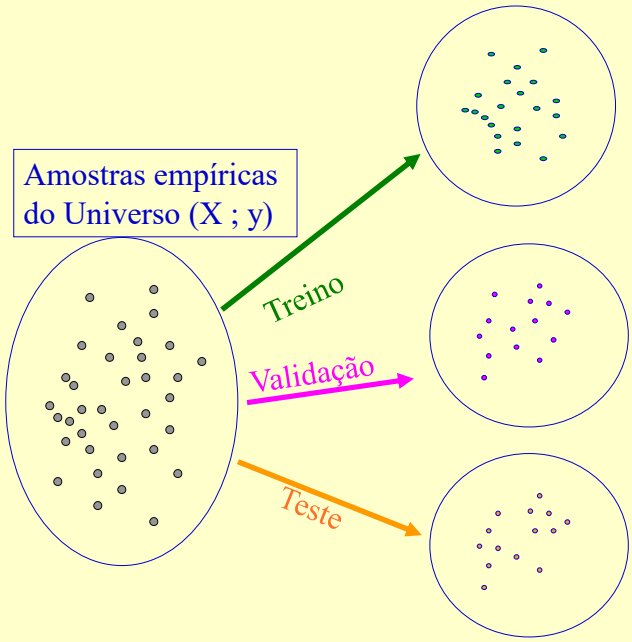
Treino + Teste ....

versus

Treino + Validação + Teste

169

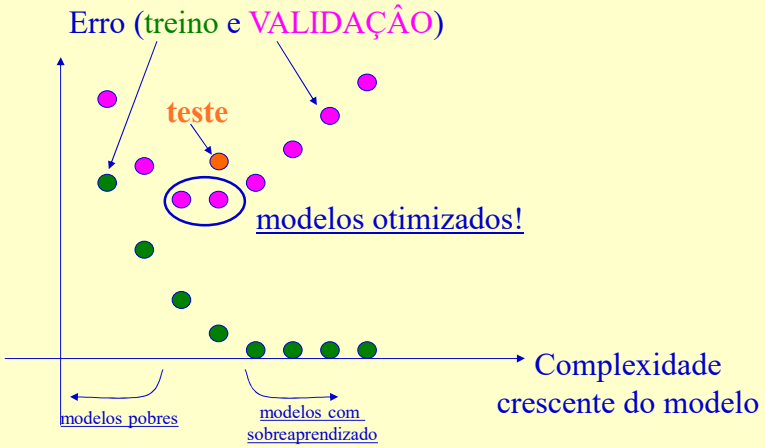
169



170

170

Sobreaprendizado em "sumário executivo"

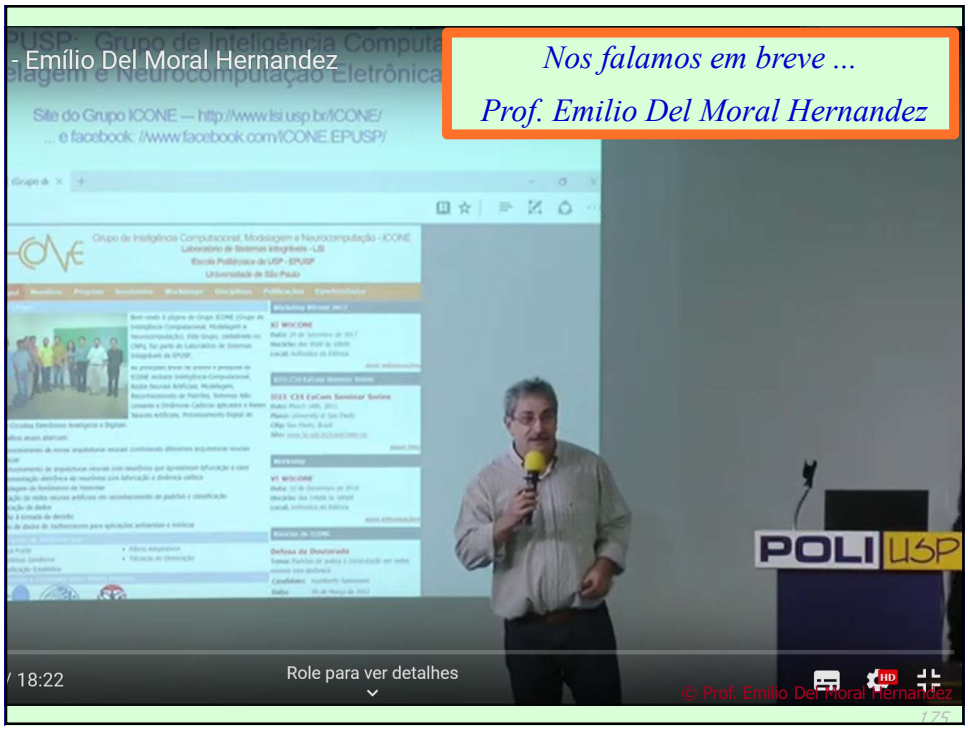


171

171

Já finalizando por hoje ...

174



175

Slides extras seguem ...

179

Explicitando alguns dos elementos que dão margem ao sobreaprendizado

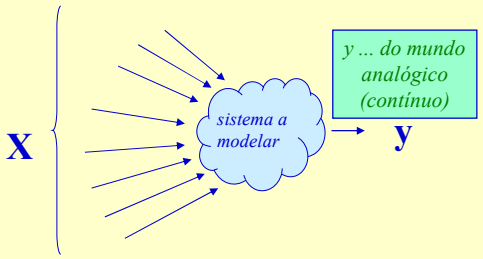
(além da complexidade alta da RNA / grande número de parâmetros no modelo)

180

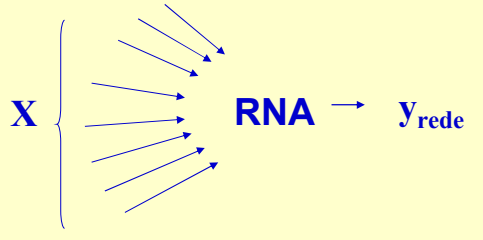
Identificando os ingredientes para o risco de sobreaprendizado nos contextos de regressão multivariada e de reconhecimento de padrões multivariado

181

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



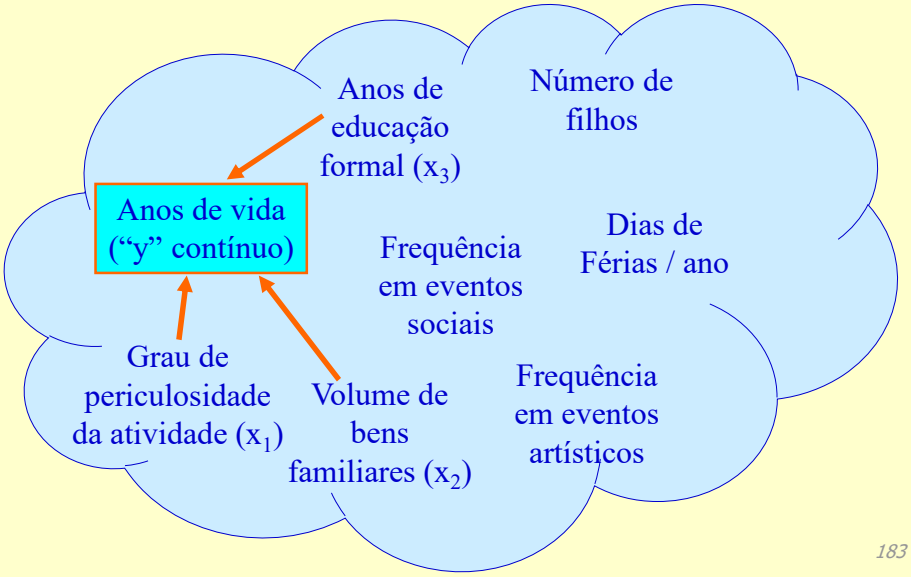
Assumimos que a variável y do sistema a modelar é uma função (normalmente desconhecida e possivelmente não linear) de diversas outras variáveis desse mesmo sistema



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

182

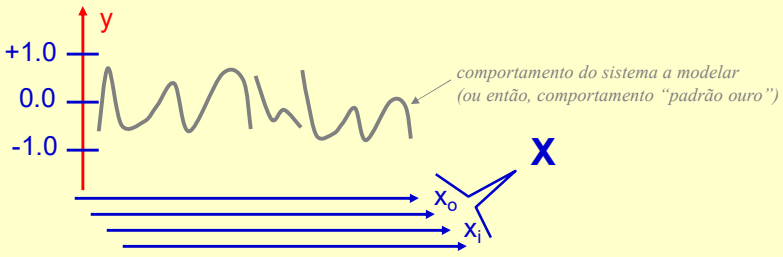
Um hipotético universo de variáveis inter-dependentes, passível de modelagem/ens



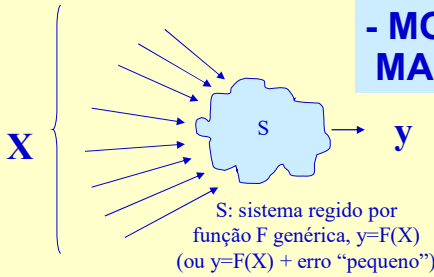
183

183

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")

184

184

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")

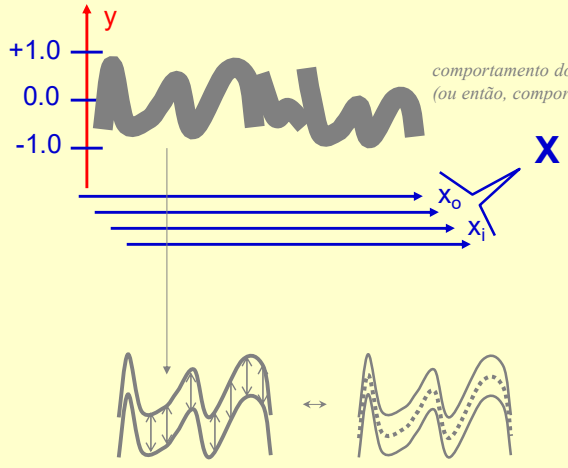
.... em problemas concretos / reais, há sempre alguma ambiguidade no mapeamento que leva valores de X a valores de y. Para decepção de Cybenko, não temos uma função y=F(X) no sentido matemático exato, pois para uma dada ênupla de valores X fixados, temos tipicamente uma faixa de valores que podem ser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

185

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

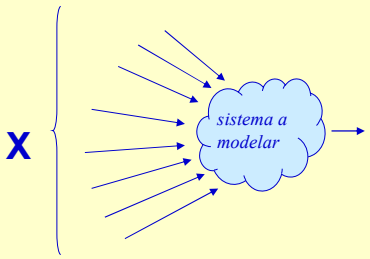


comportamento do sistema a modelar
(ou então, comportamento "padrão ouro")

... 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.

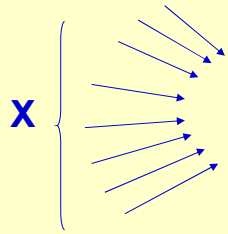
186

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 ( )

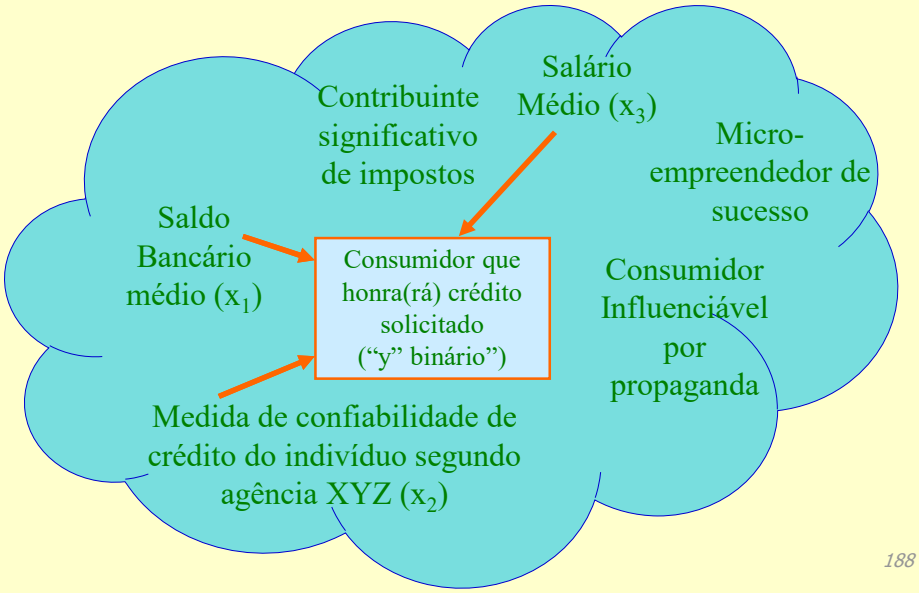
... ou alternativamente: y_{rede} ... 0.5 ( ) *versus* - 0.5 ( )
... ou alternativamente: ... $y_{rede} > 0$ ( ) *versus* $y_{rede} < 0$ ( )

187

© Prof. Emilio Del Moral – EPUSP

187

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

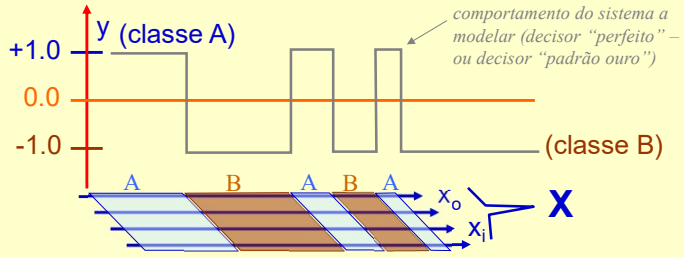


188

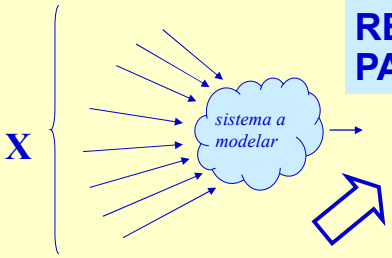
© Prof. Emilio Del Moral – EPUSP

188

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



CLASSIFICADOR; RECONHECEDOR DE PADRÕES NOTÁVEIS



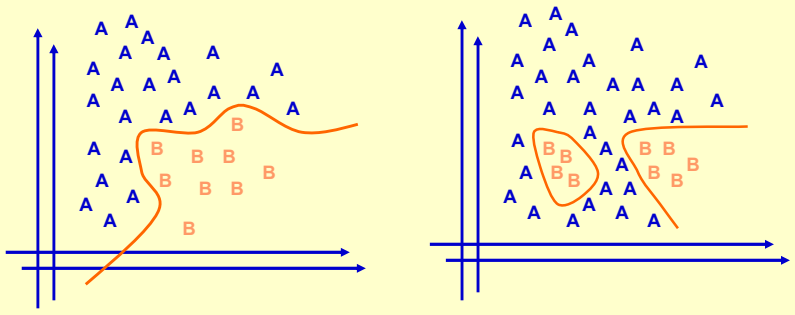
y (... restrita a valores +1 e –1)

189

189

Capacidade de reconhecimento de padrões em casos complexos NÃO LINEARES

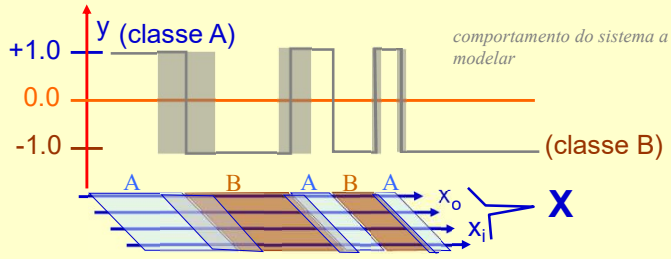
Com as RNAs, a hipersuperfície de separação entre classes vai muito além dos hiperplanos



190

190

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



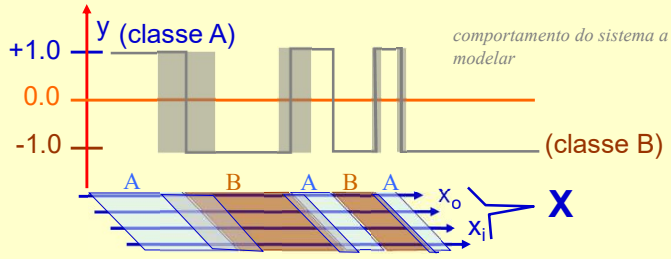
.... em problemas concretos / reais, há sempre alguma ambiguidade no mapeamento que leva valores de X aos valores discretos de y. Não temos uma função y=F(X) no sentido matemático exato, pois para uma dada ênupla de valores X fixado temos em alguns casos de fronteira a possibilidade de observar no y 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

191

191

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

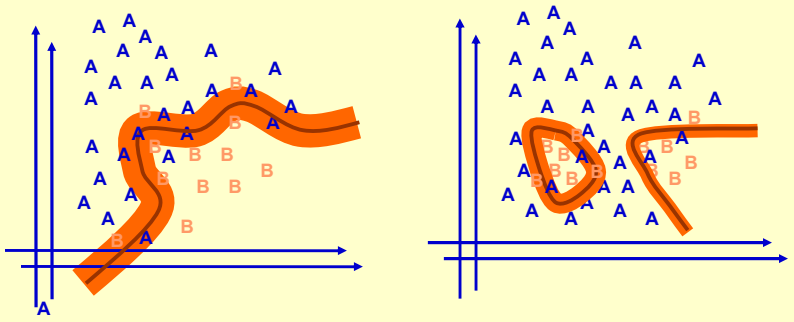


... podemos olhar para o y (classe A ou B) observado no sistema que se deseja modelar não mais como uma classe sempre bem definida e com fronteiras de separação entre A e B bem definidas no espaço de valores de X, mas como sendo delineadas na modelagem através de fronteiras com eventuais faixas de tolerância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

192

192

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



194

Flutuação com amostras e k-fold cross validation

213

Como podemos lidar de alguma forma com a relação entre variabilidade da qualidade do modelo gerado e a flutuação estatística intrinseca dos conjuntos de dados usados para treino e para teste?

214

Pergunta perturbadora ... e as medidas de aderência do modelo aos dados empíricos (para conjunto de treino) ou de qualidade de extrapolação do modelo (para o conjunto de teste) não são dependentes das específicas amostras estatísticas / dos dados empíricos coletados?

Se meus dados coletados fossem algo distintos dos que obtivemos na coleta particular realizada, os valores das medidas não seriam algo diferentes? Quanto?

215

A resposta é SIM! Sim, se os dados empíricos coletados forem algo distintos, mesmo que vindos de observações do mesmo fenômeno, os valores das medidas de qualidade do regressor seriam diferentes!!!

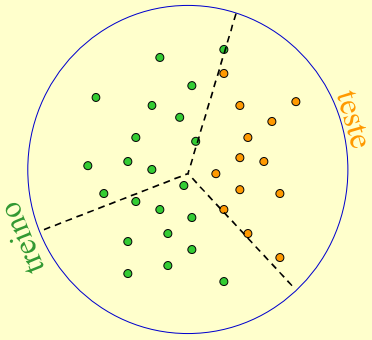
Com técnicas como a validação cruzada podemos avaliar a extensão da flutuação dessas medidas com as mudanças no conjunto de observações de treino e/ou teste (amostra estatística de dados empíricos para o treino e amostra estatística de dados empíricos para teste de generalização).

216

k-fold Cross Validation:
O conjunto total é "retalhado" em k partes, uma dessas partes apenas é reservada para o teste, sendo as demais k-1 partes usadas para compor o conjunto de treino; com essa estratégia, podemos ter com k ensaios distintos de treino e teste, simplesmente mudando de um ensaio para o outro "quem" será usado (que retalho será usado) para teste e ficará portanto "de fora" no treino feito com os demais retalhos

229

Cross validation / Validação cruzada – k fold cross validation ....

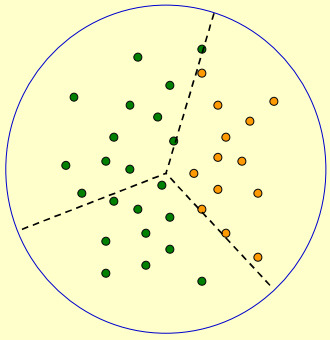


3 fold cross validation:
66% treino e
33% teste

231

231

Cross validation / Validação cruzada e flutuação estatística na partição de amostras

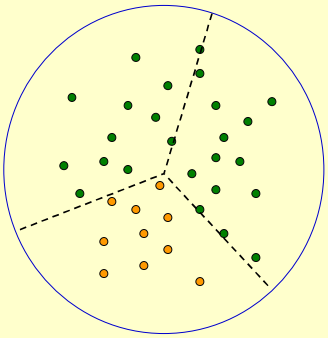


3 fold cross validation:
66% treino e
33% teste

232

232

Cross validation / Validação cruzada e flutuação estatística na partição de amostras

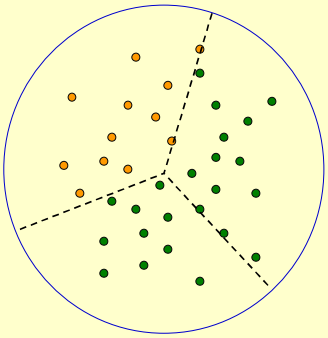


3 fold cross validation:
66% treino e
33% teste

233

233

Cross validation / Validação cruzada e flutuação estatística na partição de amostras



3 fold cross validation:
66% treino e
33% teste

234

234

Caso extremo do k fold cross validation ...

k = M_empíricos

(número de "partes da pizza" = número M de observações empíricas disponíveis, ou seja, cada parte da pizza tem um único exemplar / uma única observação)

Nesse caso, chamamos o método de "Leave one Out" – ou seja, usamos (M_empíricos - 1) observações para treino do modelo e apenas UM exemplar empírico é deixado de fora do treino, para ser usado para o teste

235

235

Conceito geral que engloba as discussões anteriores

...

Reamostragem / Data Resampling

Pergunta ... Que impacto isto tem nas medidas de qualidade de regressores? Que impacto isto tem nas medidas de classificadores? Que informação ao cliente / usuário podemos fornecer com base neste conceito?

236

236