

- Emilio Del Moral Hernandez

Slides comentados
Prof. Emilio Del Moral Hernandez

18:22 Role para ver detalhes

© Prof. Emilio Del Moral Hernandez

240

Temas da quinta semana c/ Prof. Emilio

241

#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 – 3ªf) 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.

Aprofundando / Completando alguns dos temas planejados para a 5ª semana ...

© Prof. Emilio Del Moral Hernandez

241

Aprofundando / Completando alguns dos temas planejados para a 5ª semana ...

242

- O papel de componentes randômicas em regressores e em reconhecedores de padrões no sobreajuste
- Os mínimos locais no aprendizado por EBP e como eles podem dificultar o controle de sobreajuste
- Outros usos do conceito de conjunto de validação além da seleção de complexidade de modelo, aprendido anteriormente
- O conceito de validação cruzada, a técnica de k-fold cross validation, e o "leave one out"

© Prof. Emilio Del Moral Hernandez

242

242

Explicitando alguns dos elementos que dão margem ao sobreaprendizado

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

243

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

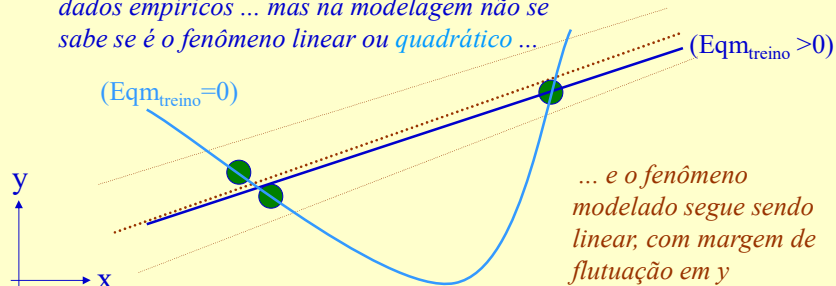
244

© Prof. Emilio Del Moral – EPUSP

244

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^i, y^i) 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

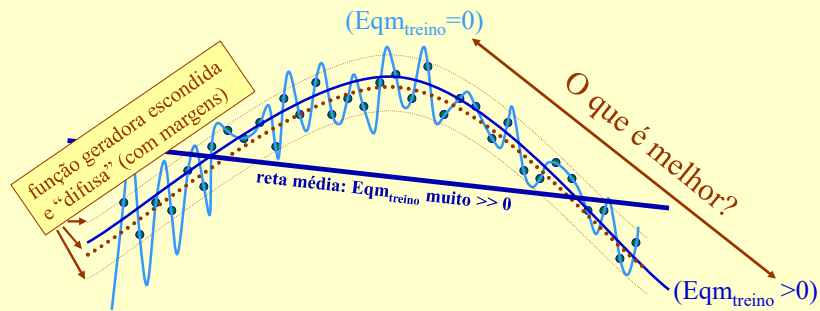
245

© Prof. Emilio Del Moral – EPUSP

245

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

um novo exemplo



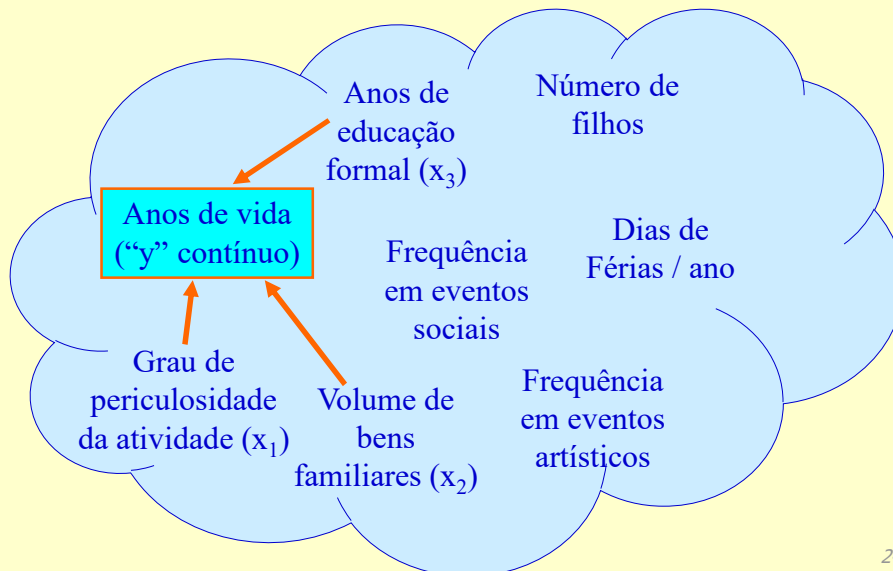
Os dados empíricos (x^i, y^i) estão em verde;
Três modelos polinomiais gerados a partir dos dados, em azuis.

246

© Prof. Emilio Del Moral – EPUSP

246

Um hipotético universo de variáveis interdependentes, passível de modelagem/ens

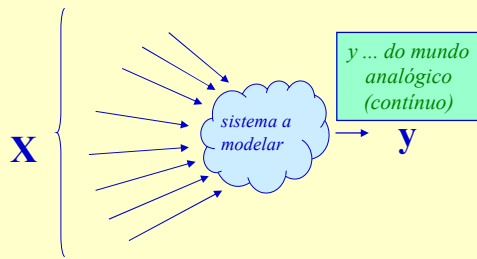


247

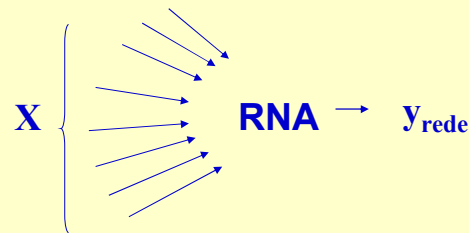
© Prof. Emilio Del Moral – EPUSP

247

**Modelagem de um sistema por função de mapeamento $X \rightarrow y$
(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



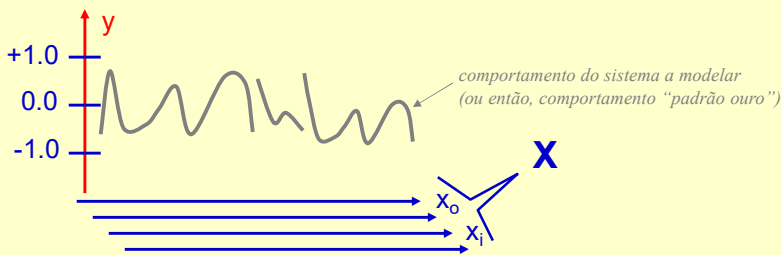
A RNA, para ser um bom modelo do sistema, deve reproduzir essa relação entre X e y , tão bem quanto possível

248

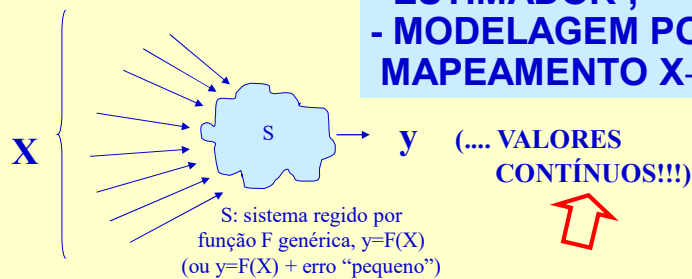
© Prof. Emilio Del Moral – EPUSP

248

A função $y(X)$ “a descobrir”, num caso geral de função analógica $y(X)$



**- ESTIMADOR ;
- MODELAGEM POR
MAPEAMENTO $X \rightarrow y$**

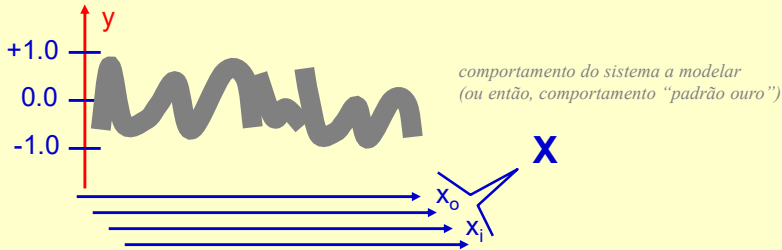


249

© Prof. Emilio Del Moral – EPUSP

249

Cenário mais real: a “função” $y(X)$ do sistema modelado é “difusa”: $y = F_{\text{médio}}(X) + \text{flutuação} \dots$



.... 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 : $y = F_{\text{médio}}(X) + \text{flutuação}$.

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

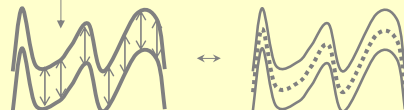
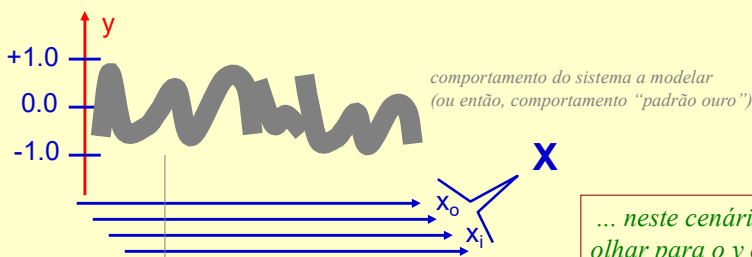
$$\dots y_{\text{rede}} \sim y_{\text{médio}} \text{ esperado para um dado } X$$

250

© Prof. Emilio Del Moral – EPUSP

250

Cenário mais real: a “função” $y(X)$ do sistema modelado é “difusa”: $y = F_{\text{médio}}(X) + \text{flutuação} \dots$



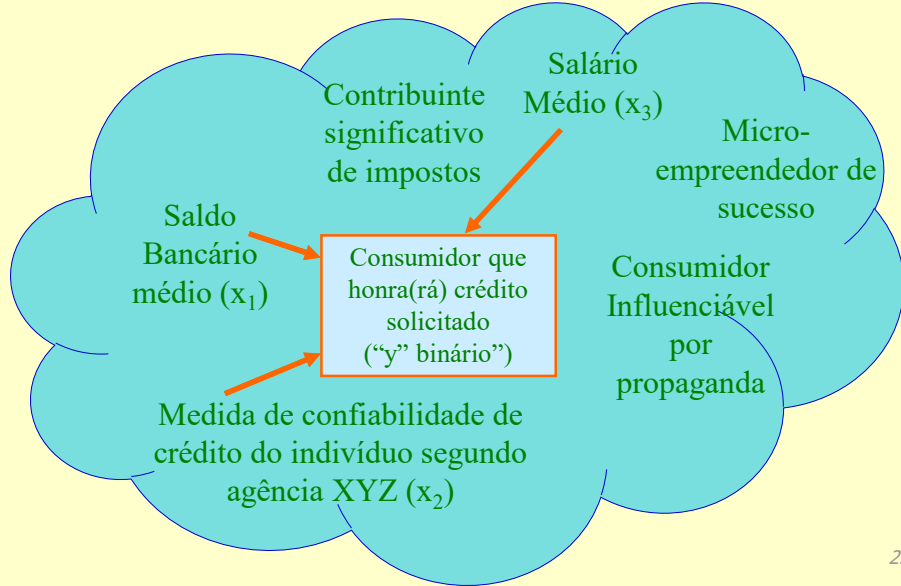
... neste cenário, podemos olhar para o y observado no sistema que se deseja modelar não mais como um valor específico bem definido, mas como um valor médio esperado (dado valor de X) e uma faixa de valores em torno desse valor médio esperado.

251

© Prof. Emilio Del Moral – EPUSP

251

Um hipotético universo de variáveis interdependentes, passível de modelagem/ens



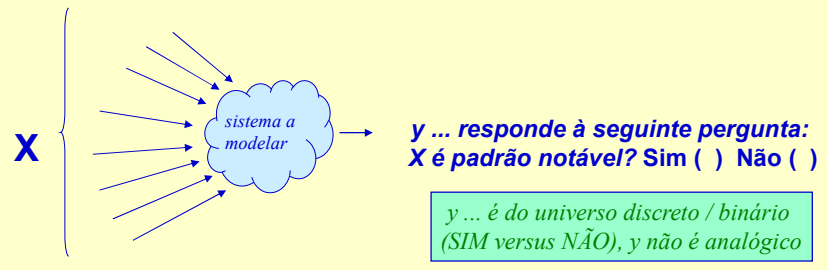
252

© Prof. Emilio Del Moral – EPUSP

252

RNAs como reconhecedor / detetor de padrões

...

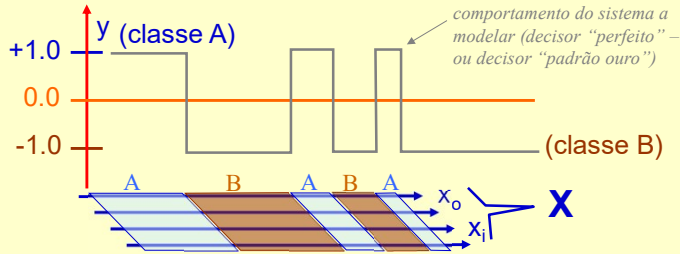


253

© Prof. Emilio Del Moral – EPUSP

253

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



**CLASSIFICADOR;
RECONHECEDOR DE
PADRÕES NOTÁVEIS**



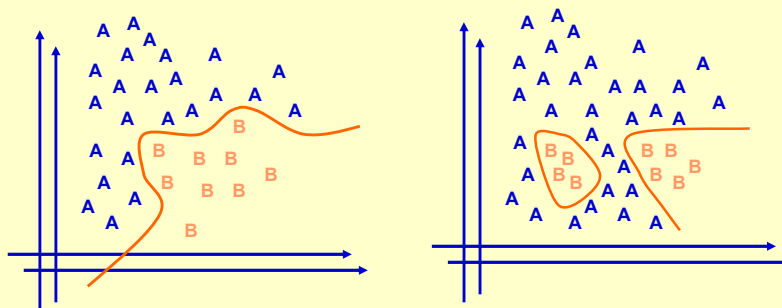
254

© Prof. Emilio Del Moral – EPUSP

254

Capacidade de reconhecimento de padrões em casos complexos NÃO LINEARES

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

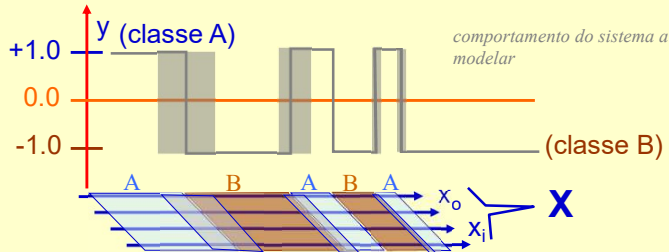


255

© Prof. Emilio Del Moral – EPUSP

255

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 :

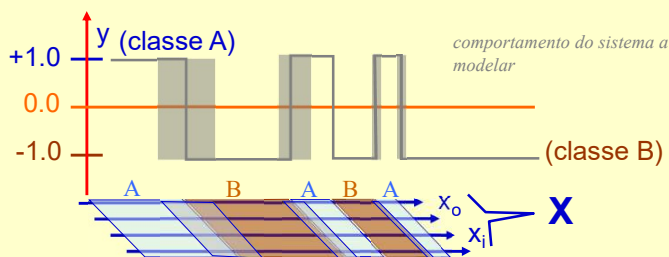
... $y_{rede} \sim$ classe 'mais esperada' para um dado X

256

© Prof. Emilio Del Moral – EPUSP

256

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



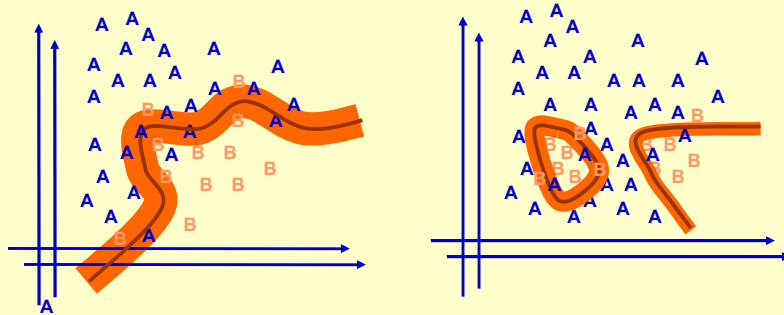
... podemos olhar para o y (classe A ou B) observado no sistema que se deseja modelar não mais como uma classe sempre bem definida e com fronteiras de separação entre A e B bem definidas no espaço de valores de X , mas como sendo delineadas na modelagem através de fronteiras com eventuais faixas de tolerância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

257

© Prof. Emilio Del Moral – EPUSP

257

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



259

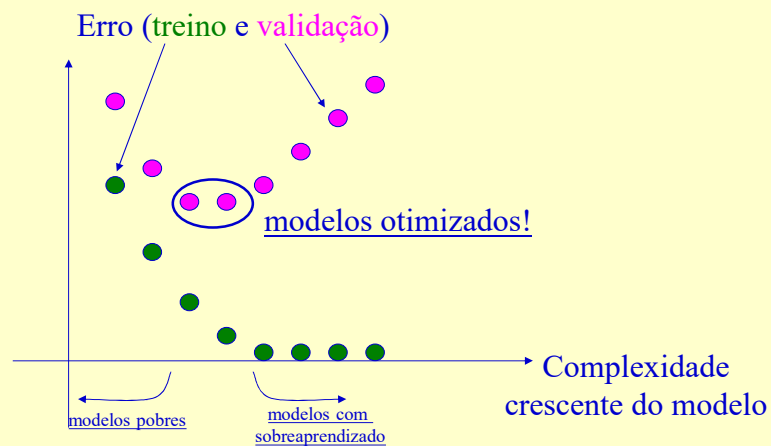
© Prof. Emilio Del Moral – EPUSP

259

Dificuldades para escolha empírica da complexidade da rede neural para controle do sobreajuste, trazidas pelos mínimos locais do Eqm, intrínsecos ao Error Back Propagation / Método do Gradiente, e forma prática de contornar tal problema

261

Sobreaprendizado em “sumário executivo”



262

© Prof. Emilio Del Moral Hernandez

262

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 mínimos 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?*

264

© Prof. Emilio Del Moral Hernandez

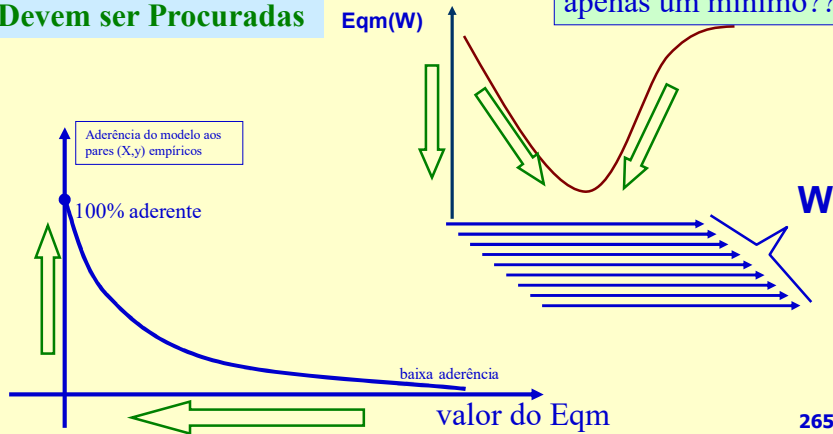
264

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

Devemos buscar Maximização da aderência = Mínimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas

Será que temos apenas um mínimo??



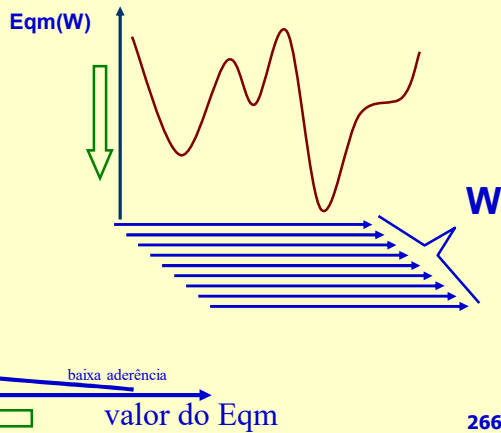
© Prof. Emilio Del Moral Hernandez

265

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

Devemos buscar Maximização da aderência = Mínimo Eqm possível

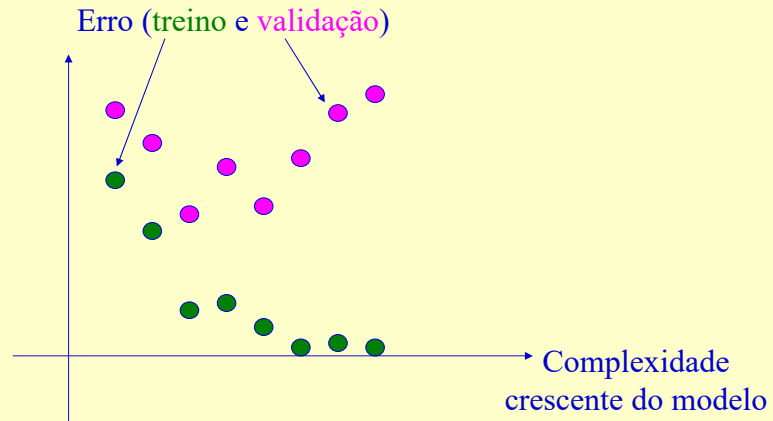
As Setas Verdes Indicam Situações que Devem ser Procuradas



© Prof. Emilio Del Moral Hernandez

266

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) ...



267

© Prof. Emilio Del Moral Hernandez

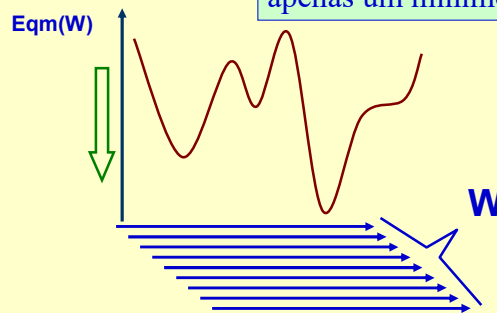
267

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

Devemos buscar Maximização da aderência = Mínimo Eqm possível

As Setas Verdes Indicam Situações que Devem ser Procuradas

Será que temos apenas um mínimo??

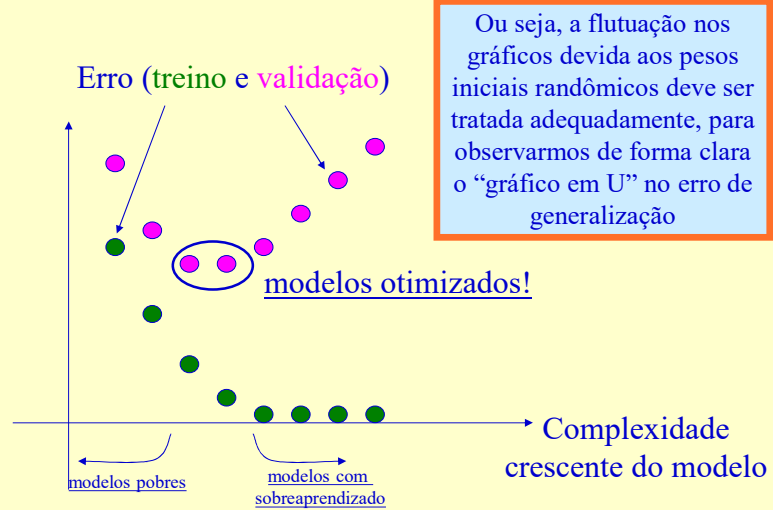


... Para não sermos reféns de mínimos locais com alto Eqm, podemos aplicar o gradiente descendente repetidamente na mesma RNA, com novos pesos iniciais randômicos em cada rodada, mantendo para o modelo final apenas os valores de pesos associados ao ensaio com o melhor dos resultados finais no Eqm!

© Prof. Emilio Del Moral Hernandez

269

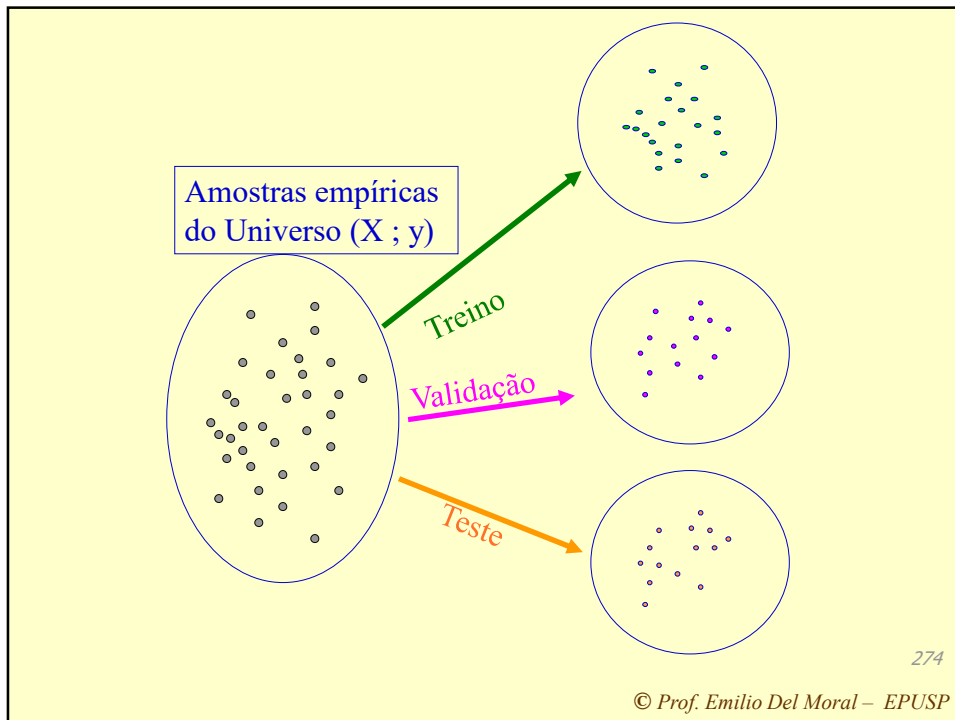
Com repetidos ensaios em cada grau de complexidade os mínimos locais são evitados e detectamos adequadamente o sobreaprendizado



270

Outros usos do conceito de conjunto de dados empíricos de validação

273



274

Alguns dos usos do conceito de conjunto de validação (adicional aos conjuntos de treino e teste):

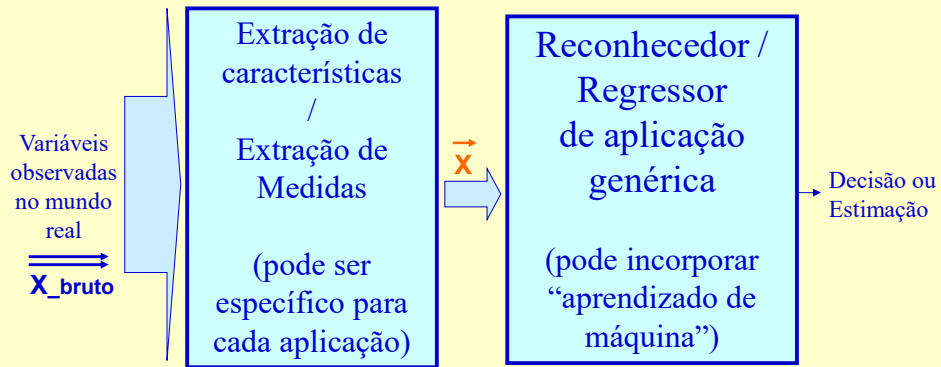
- *Seleção de complexidade do modelo neural para limitação de sobreaprendizado*
- *Ativação do early stop no aprendizado (Matlab): critério de parada adicional no processo de refinamento de pesos sinápticos*
- *Balizador no processo de seleção de estratégias de pré-processamento / extração de medidas X alternativos*
- *... etc*

275

© Prof. Emilio Del Moral Hernandez

275

Elaborando uma Solução em dois estágios



276

© Prof. Emilio Del Moral Hernandez

276

Flutuação com amostras e k-fold cross validation

278

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

© Prof. Emilio Del Moral – EPUSP

279

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?

280

© Prof. Emilio Del Moral – EPUSP

280

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!!!

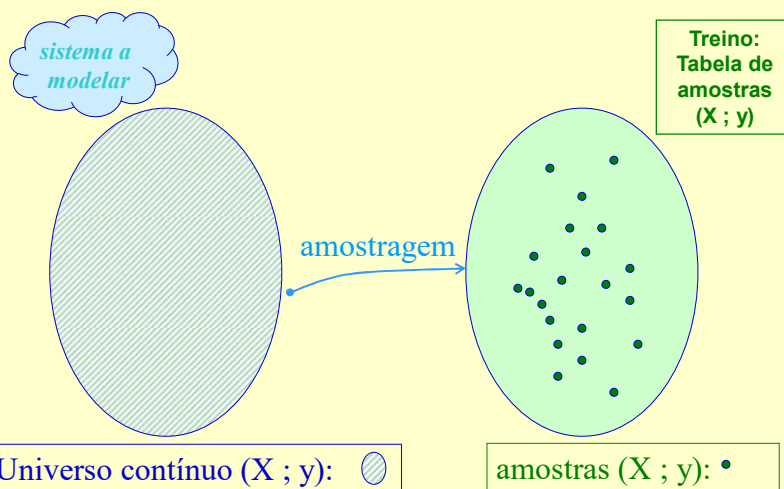
Com técnicas como a validação cruzada podemos avaliar a extensão da flutuação dessas medidas com as mudanças no conjunto de observações de treino e/ou teste (amostra estatística de dados empíricos para o treino e amostra estatística de dados empíricos para teste de generalização).

281

© Prof. Emilio Del Moral – EPUSP

281

Conjunto de treino ... Composto de amostragens discretas do universo contínuo (X ; y)



285

© Prof. Emilio Del Moral Hernandez

285

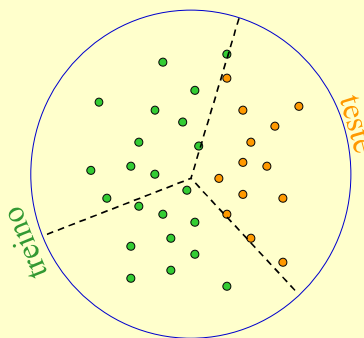
k-fold Cross Validation:
O conjunto total é “retalhado” em k partes, uma dessas partes apenas é reservada para o teste, sendo as demais k-1 partes usadas para compor o conjunto de treino; com essa estratégia, podemos ter com k ensaios distintos de treino e teste, simplesmente mudando de um ensaio para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

294

© Prof. Emilio Del Moral Hernandez

294

Cross validation / Validação cruzada – k fold cross validation



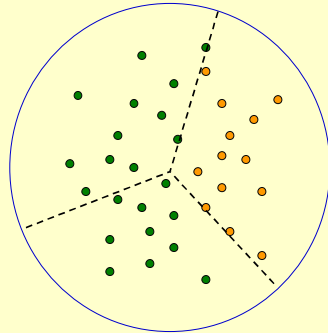
3 fold cross validation:
66% treino e
33% teste

296

© Prof. Emilio Del Moral Hernandez

296

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



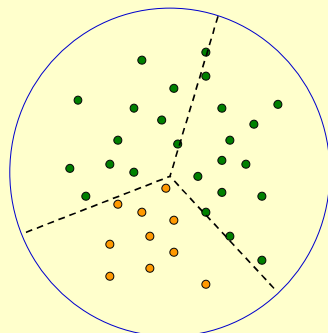
3 fold cross validation:
66% treino e
33% teste

297

© Prof. Emilio Del Moral Hernandez

297

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



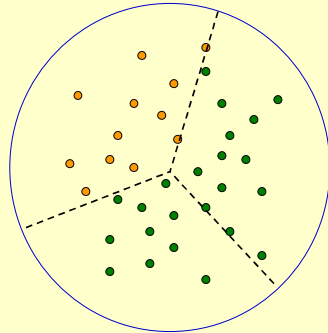
3 fold cross validation:
66% treino e
33% teste

298

© Prof. Emilio Del Moral Hernandez

298

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



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

299

© Prof. Emilio Del Moral Hernandez

299

Caso extremo do k fold cross validation ...

$$k = M_{\text{empíricos}}$$

(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 $(M_{\text{empíricos}} - 1)$ observações para treino do modelo e apenas UM exemplar empírico é deixado de fora do treino, para ser usado para o teste

300

© Prof. Emilio Del Moral Hernandez

300

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?

301

© Prof. Emilio Del Moral Hernandez

301

Próximos tópicos PSI3471-2020 ...

- Soluções em reconhecimento e em regressão em dois estágios
- 1º estágio como extrator de características / extrator de medidas
- 1º estágio como redutor de dimensionalidade de X após sua reescrita
- Técnicas de extração de características / redução de dimensionalidade (PCA e outros)

303

© Prof. Emilio Del Moral Hernandez

303

- Emilio Del Moral Hernandez

Nos falamos em breve ...
Prof. Emilio Del Moral Hernandez

304