Universidade Federal de Minas Gerais

Escola de Engenharia

Redes Neurais Artificiais: Teoria e Aplicações

Turma A - 2021/2

Lista de Exercícios - Adaline

Autores:
Vítor Gabriel Reis Caitité

Email:
vcaitite@ufmg.br

2 de dezembro de 2021

Sumário

Lista de Figuras

1 Objetivo

Para a base de dados Boston Housing, será resolvido o problema de regressão (13 variáveis de entrada e variável MEDV de saída) usando a rede Adaline, conforme a seção 3.4.6 das notas de aula. Além disso, tendo como base a figura 3.7 das notas de aula, será avaliado, também, o desempenho de modelos (pelo menos 3 modelos diferentes) com menos variáveis (por exemplo: as variáveis 9 e 10 apresentam alta correlação linear, logo, um dos modelos não incluirá a variável 9).

2 Base de Dados

As principais características do *dataset* "Boston house prices" [1] são:

- Número de instâncias: 506
- Número de atributos: 13 preditivos numéricos / categóricos. O valor mediano (atributo 14) é geralmente o alvo de predição.

Na Figura 1 é possível observar as primeiras 5 instâncias dessa base de dados.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

Figura 1: Organização da base de dados "Boston house prices".

Este conjunto de dados foi obtido da biblioteca StatLib, que é mantida na Carnegie Mellon University. Além disso ele obtido pelo link: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/.

3 Desenvolvimento

3.1 Desenvolvimento e Validação do Modelo

Primeiramente o algoritmo do Adaline ser desenvolvido conforme as instruções presentes nas notas de aula. O treinamento da rede foi realizado utilizando 70% dos dados para treinamento e 30% para teste.

Na Figura 2 é possível observar a evolução do erro quadrático por época do treinamento.

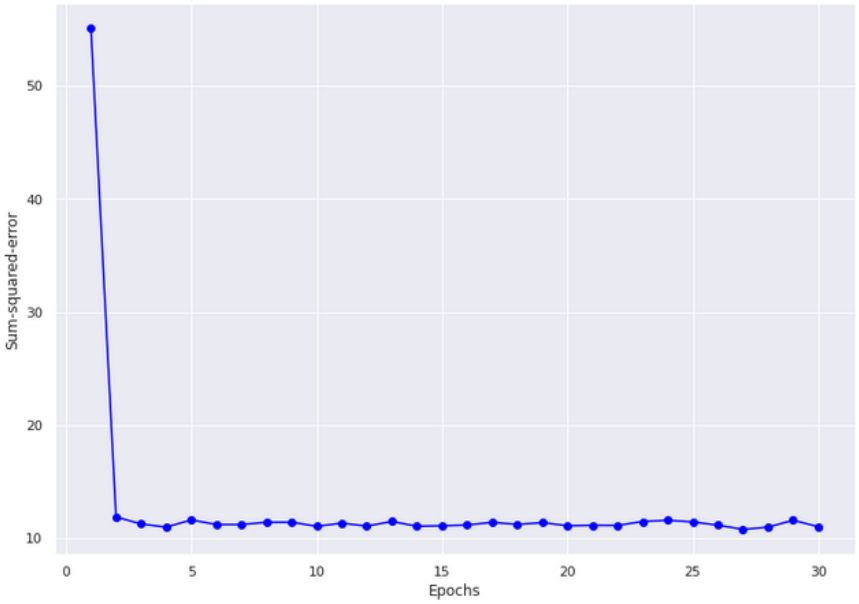

Figura 2: Evolução do erro quadrático durante o treinamento.

Para avaliação do modelo a principal medida utilizada foi a RMSE (*root mean squared error*). Essa medida calcula "a raiz quadrática média"dos erros entre valores observados (reais) e predições (hipóteses).

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|} \tag{1}$$

O resultado médio obtido após 10 execuções foi:

```
The model performance for testing set

--------------------------------------

RMSE is 5.270774704797643
R2 score is 0.7001715444096551
```

Nas Figuras 3 e 4 é possível observar a distribuição da variável de predição MEDV para os valores reais e para os previstos pelo modelo. Essas figuras foram geradas

para os 30% de dados separados para teste.

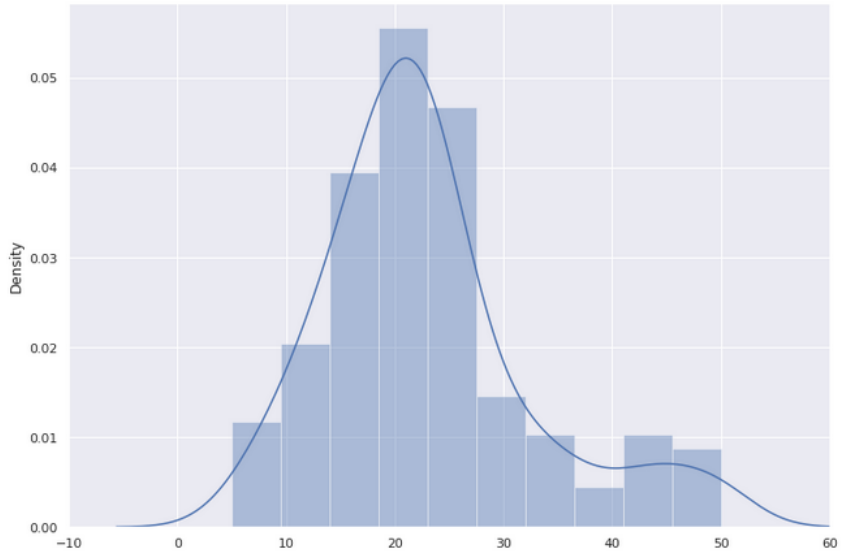

Figura 3: Distribuição da variável MEDV para os dados de teste.

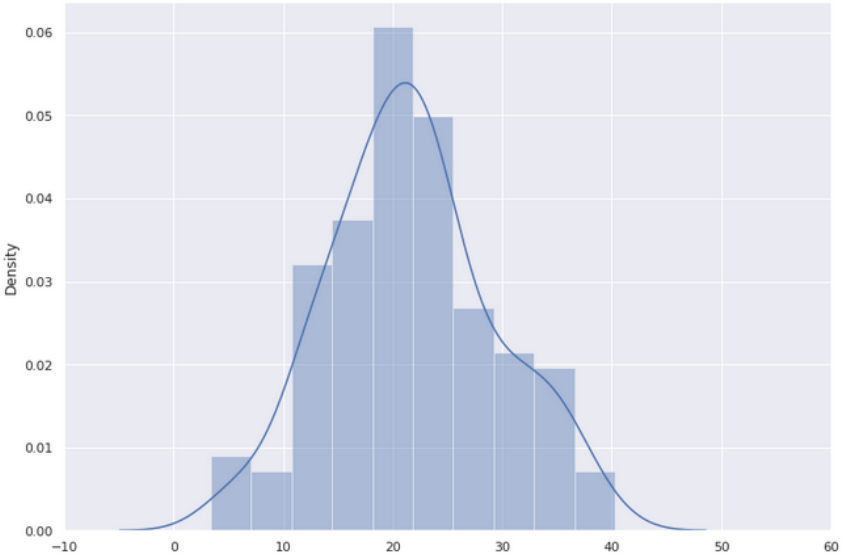

Figura 4: Distribuição encontrada pelo modelo para a variável MEDV para os dados de teste.

Através desses resultados foi possível observar que o modelo proposto conseguiu aproximar o comportamento da variável MEDV com base nas *features* de entrada.

3.2 Geração de Modelos com Menos Variáveis

Primeiramente foi plotado a matriz de correlação como pode ser vista na Figura 5.

Figura 5: Correlação entre as variáveis do dataset.

3.2.1 Modelo 1

O modelo 1 foi gerado desconsiderando variáveis altamente correlacionadas. Como as variáveis 9 e 10 apresentam alta correlação linear, então a variável 9 (RAD) será excluída desse modelo. O mesmo ocorre para as *features* 5, 7 e 8 (por isso as variáveis 7 (AGE) e 8 (DIS) também não serão consideradas).

Utilizando as mesmas métricas apresentadas anteriormente obteve-se o seguinte resultado:

```
The model performance for testing set

--------------------------------------
```

```
RMSE is 5.618691785743369
R2 score is 0.6357149989232859
```

3.2.2 Modelo 2

O modelo 2 foi gerado considerando apenas as três variáveis mais correlacionadas com a variável de predição MEDV. Como pôde-se observar na matriz de correlações acima, estas são: "LSTAT", "RM", "PTRATIO". O comportamento de cada uma dessas variáveis com relação a variável MEDV pode ser visto na Figura 6.

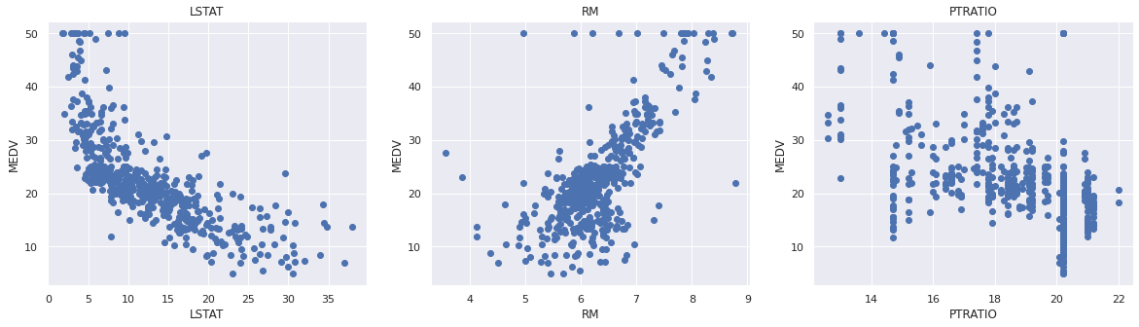

Figura 6: Comportamento das variáveis "LSTAT", "RM", "PTRATIO"com relação a variável MEDV.

Pode-se observar um comportamento aproximadamente linear principalmente das variáveis "LSTAT", "RM"com relação a variável de predição.

O modelo 2, utilizando apenas as 3 variáveis citadas, obteve-se o seguinte resultado:

```
The model performance for testing set

--------------------------------------

RMSE is 5.350940177286979
R2 score is 0.6626069312939527
```

3.2.3 Modelo 3

Por fim, para o modelo 3 foi utilizado o algoritmo de PCA para extração de *features*. O modelo foi testado para diferentes números de *features* selecionadas pelo PCA e o RMSE médio de 20 execuções para cada um dos casos pode ser visto abaixo.

```
The model performance for testing set

--------------------------------------

RMSE médio para modelo considerando 1 features: 7.392 +/- 0.539

RMSE médio para modelo considerando 2 features: 6.859 +/- 0.649

RMSE médio para modelo considerando 3 features: 5.674 +/- 0.491

RMSE médio para modelo considerando 4 features: 5.480 +/- 0.671

RMSE médio para modelo considerando 5 features: 5.447 +/- 0.626

RMSE médio para modelo considerando 6 features: 5.252 +/- 0.580

RMSE médio para modelo considerando 7 features: 5.306 +/- 0.522

RMSE médio para modelo considerando 8 features: 5.267 +/- 0.590

RMSE médio para modelo considerando 9 features: 5.266 +/- 0.529

RMSE médio para modelo considerando 10 features: 5.520 +/- 0.471

RMSE médio para modelo considerando 11 features: 5.323 +/- 0.460

RMSE médio para modelo considerando 12 features: 5.309 +/- 0.564

RMSE médio para modelo considerando 13 features: 5.179 +/- 0.507
```

Referências

[1] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.