RELATÓRIO

LISTA DE EXERCÍCIOS 3

INTRODUÇÃO AO APRENDIZADO DE MÁQUINA (COE782) | COPPE | UFRJ

Luis Paulo Albuquerque Guedes
Programa de Engenharia Elétrica
Universidade Federal do Rio de Janeiro
Av. Athos da Silveira Ramos, 149 - Bloco H, 3º andar
luis.albuquerque@marinha.mil.br

1 CONHECIMENTO PRÉVIO RELEVANTE

1.1 MÉTODOS DE MÍNIMOS QUADRADOS (LS)

O modelo linear tem sido uma peça fundamental da estatística nos últimos 30 anos e continua sendo uma de nossas ferramentas mais importantes. Dado um vetor de entradas $X^T = (X_1, X_2, \ldots, X_3)$, tem-se a seguinte saída Y por meio do modelo [1]:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{1}$$

O termo $\hat{\beta}_0$ é o intercepto, também conhecido como viés, em aprendizado de máquina. Muitas vezes, é conveniente incluir a variável constante 1 em X, incluir $\hat{\beta}_0$ no vetor de coeficientes $\hat{\beta}$, e, depois, escrever o modelo linear em forma de vetor como um produto interno [1]:

$$\hat{Y} = X^T \hat{\beta} \tag{2}$$

Em que X^T denota a transposição do vetor ou da matriz (sendo X um vetor coluna) [1].

Como ajustar o modelo linear a um conjunto de dados de treinamento? Existem muitos métodos diferentes, mas de longe o mais popular é o método dos mínimos quadrados. Nessa abordagem, os coeficientes β são escolhidos de forma a minimizar o resíduo[1] da soma dos quadrados:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{3}$$

[1] O método dos mínimos quadrados é uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados (tais diferenças são chamadas resíduos).

$RSS(\beta)$ é uma função quadrática, e, portanto, seu mínimo sempre existe, mas pode não ser único. A solução é mais fácil de caracterizar em notação de matriz. Assim, a expressão (3) pode ser escrita da seguinte forma:

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

Em que $\boldsymbol{X}$ é uma matriz $N \times p$ com cada linha sendo um vetor de entrada, e $\boldsymbol{y}$ é um vetor das saídas no conjunto de treinamento.

1.2 REGRESSÃO RIDGE

A regressão Ridge reduz os coeficientes de regressão impondo uma penalidade ao seu tamanho. Os coeficientes Ridge minimizam a soma dos quadrados dos resíduos penalizada,

$$\hat{\beta}^{ridge} = argmin_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad (4)$$

Aqui, $\lambda \geq 0$ é um parâmetro de complexidade que controla a quantidade de contração (tradução livre para *shrinkage*): quanto maior o valor de λ, maior a quantidade de contração. Os coeficientes são contraídos em direção a zero (e entre si). A ideia de penalizar pela soma dos quadrados dos parâmetros também é utilizada em redes neurais, onde é conhecida como decaimento de peso [1].

Uma forma equivalente de escrever o problema da regressão Ridge é:

$$\hat{\beta}^{ridge} = argmin_{\beta}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad (5)$$

Em que:

$$\sum_{j=1}^{p}{\beta_j}^2 \leq t, \quad (6)$$

o que torna explícita a restrição de tamanho nos parâmetros. Existe uma correspondência biunívoca entre os parâmetros λ em (1) e t em (2). Quando há muitas variáveis correlacionadas em um modelo de regressão linear, seus coeficientes podem se tornar mal determinados e exibir alta variância. Um coeficiente positivo excessivamente grande em uma variável pode ser anulado por um coeficiente negativo igualmente grande em sua variável correlacionada. Ao impor uma restrição de tamanho nos coeficientes, como em (2), esse problema é atenuado [1].

As soluções de Ridge não são invariantes à escala das entradas, e assim normalmente as entradas são padronizadas antes de resolver (1). Além disso, observa-se que o intercepto β_0 foi deixado de fora do termo de penalidade. A penalização do intercepto faria com que o procedimento dependesse da origem escolhida para Y; ou seja, adicionar uma constante c a cada um dos alvos y_i não resultaria simplesmente em um deslocamento das previsões pelo mesmo valor de c [1].

Os coeficientes restantes são estimados por uma regressão Ridge sem interceptação, utilizando os x_{ij} centrados. Daqui em diante, assume-se que esse centramento foi feito, de modo que a matriz de entrada $\mathbf{X}$ tem p colunas (em vez de $p + 1$). Ao escrever o critério em (4) em forma de matriz,

$$RSS(\lambda) = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) + \beta^T\beta, \quad (7)$$

as soluções da regressão ridge podem ser facilmente vistas como:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}, \tag{8}$$

Em que $\mathbf{I}$ é a matriz identidade p × p [1].

1.3 REGRESSÃO LASSO

A regressão Lasso visa identificar as variáveis e os coeficientes de regressão correspondentes que levam a um modelo que minimize o erro de predição. Isto é conseguido impondo uma restrição aos parâmetros do modelo, que "reduz" os coeficientes de regressão para zero, ou seja, forçando a soma do valor absoluto dos coeficientes de regressão a ser inferior a um valor fixo (λ). Num sentido prático, isto restringe a complexidade do modelo. Variáveis com coeficiente de regressão zero após a contração são excluídas do modelo. A escolha de λ é frequentemente feita usando uma abordagem automatizada de validação cruzada *k-fold*.

Para esta abordagem, o conjunto de dados é particionado aleatoriamente em k subamostras de igual tamanho. Enquanto as subamostras $k - 1$ são usadas para desenvolver um modelo de predição, a subamostra restante é usada para validar este modelo. Este procedimento é realizado k vezes, sendo cada uma das k subamostras por sua vez utilizadas para validação e as demais para desenvolvimento do modelo. Um resultado geral é produzido combinando os k resultados de validação separados para um intervalo de valores de λ e escolhendo o λ preferido, que é então usado para determinar o modelo final. Uma vantagem particular desta técnica é que ela reduz o *overfitting* sem restringir um subconjunto do conjunto de dados para uso exclusivo para validação interna [2].

Portanto, tem-se que a regressão Lasso consiste em um método de contração semelhante ao Ridge, com diferenças sutis, mas importantes. A estimativa Lasso é definida por [1]:

$$\hat{\beta}^{ridge} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \tag{9}$$

Em que:

$$\sum_{j=1}^{p} |\beta_j| \leq t, \tag{10}$$

Na literatura de processamento de sinais, o lasso também é conhecido como busca de base [3].

Também podemos escrever o problema do lasso na forma equivalente de Lagrange:

$$\hat{\beta}^{ridge} = argmin_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{11}$$

Observa-se a semelhança com o problema de regressão Ridge em (4) ou em (5): a penalidade L2 Ridge $\sum_{j=1}^{p} x_{ij} \beta_j$ é substituída pela penalidade L1 Lasso $\sum_{j=1}^{p} |\beta_j|$. Essa última restrição torna as soluções não lineares em relação aos y_i, e não há uma expressão em forma fechada como na regressão Ridge.

2 RESOLUÇÃO DOS EXERCÍCIOS PROPOSTOS

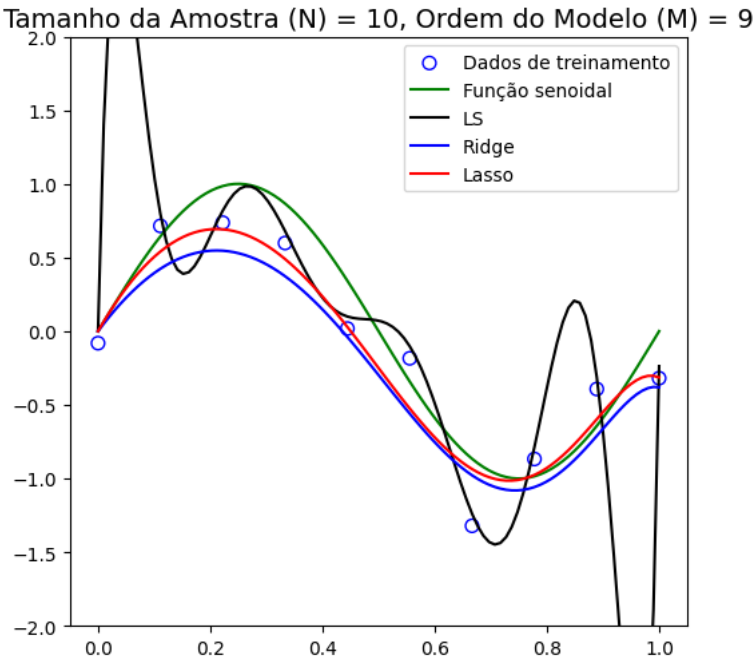
E1 Considere o experimento computacional denominado "*Polynomial Curve Fitting*", usado diversas vezes no livro texto (veja páginas 4 e 5 do livro, bem como Apêndice A), considerando a ordem do modelo sendo $M = 9$ e o tamanho da amostra sendo $N = 10$. Faça:

a) Calcule a solução de mínimos quadrados (LS) w_{LS};
b) Calcule a solução via regressão *ridge* (escolha um fator de regularização razoável) w_{ridge};
c) Calcule a solução via regressão lasso (escolha um fator de regularização razoável) w_{lasso};
d) Monte uma tabela exibindo os 10 coeficientes w para as 3 soluções obtidas nos itens acima e comente/compare os resultados;
e) Plote uma figura contendo o processo gerador em verde (a senoide), e suas estimativas y_{LS} , y_{ridge} e y_{lasso} em preto, azul e vermelho, respectivamente.
f) Repita todos os itens anteriores para $N = 20$ e $N = 50$.

Solução:

https://colab.research.google.com/drive/1DRtDIlLoZfXqLBhCD4LOMwdgegEUdN3n#scrollTo=zxX5m9v9Famc

i. Tamanho da Amostra (N) = 10



	Mínimos Quadrados (LS)	Regressão Ridge	Regressão Lasso
w_0	0.0000	0.0000	0.0000
w_1	174.4710	5.1076	6.6470
w_2	-3858.9079	-11.2848	-16.0724
w_3	34977.1201	-3.8828	0.0000
w_4	-167273.7280	2.8898	0.0000

w_5	465233.4874	5.8596	14.5649
w_6	-777903.9487	5.5076	0.0000
w_7	769690.3013	2.8354	0.0000
w_8	-414690.6455	-1.2700	0.0000
w_9	93651.6126	-6.1429	-5.4536

<u>Análise para Tamanho da Amostra (N) = 10:</u>

Mínimos Quadrados (LS):

- Coeficientes têm magnitudes extremamente altas, indicando possível overfitting.
- Coeficientes variam muito, sugerindo sensibilidade aos dados de treinamento.

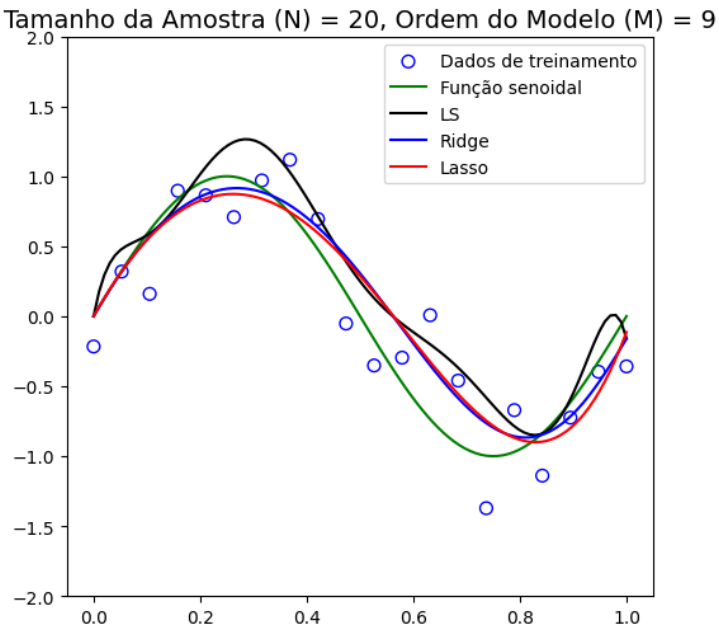
Regressão ridge:

- Coeficientes são reduzidos em comparação com Mínimos Quadrados, indicando regularização.
- Menos sensíveis aos dados de treinamento, o que ajuda a evitar overfitting.
- Para a regressão ridge foi escolhido o valor de $alpha\_ridge = 0.001$. Percebemos, também, que os valores de w estão bem mais moderados se comparados com os de LS.

Regressão lasso:

- Parece ser mais robusto a overfitting comparado a Mínimos Quadrados.
- Para a regressão lasso foi escolhido o valor de $alpha\_lasso = 0.0005$. Percebemos, também, que os valores de w estão bem mais moderados até mesmo se comparados com os da regressão ridge. Ainda, podemos evidenciar que há coeficientes de w iguais a zero, o que contribui para a redução da complexidade do modelo, a partir da redução da dimensionalidade.

ii. Tamanho da Amostra (N) = 20



	Mínimos Quadrados (LS)	Regressão Ridge	Regressão Lasso
w_0	0.0000	0.0000	0.0000
w_1	20.1664	6.6245	6.7673
w_2	-325.1092	-10.9064	-13.2312
w_3	2718.8536	-4.8623	0.0000
w_4	-11275.1975	1.6735	0.0000
w_5	24951.1049	4.5781	6.8182
w_6	-30345.1419	4.4855	0.0000
w_7	19244.4125	2.5520	0.0000
w_8	-5066.1360	-0.4030	0.0000
w_9	76.8880	-3.9033	-0.4679

Análise para Tamanho da Amostra (N) = 20:

Mínimos Quadrados (LS):

- Coeficientes ainda têm magnitudes relativamente altas.
- Ainda existe sensibilidade aos dados de treinamento.

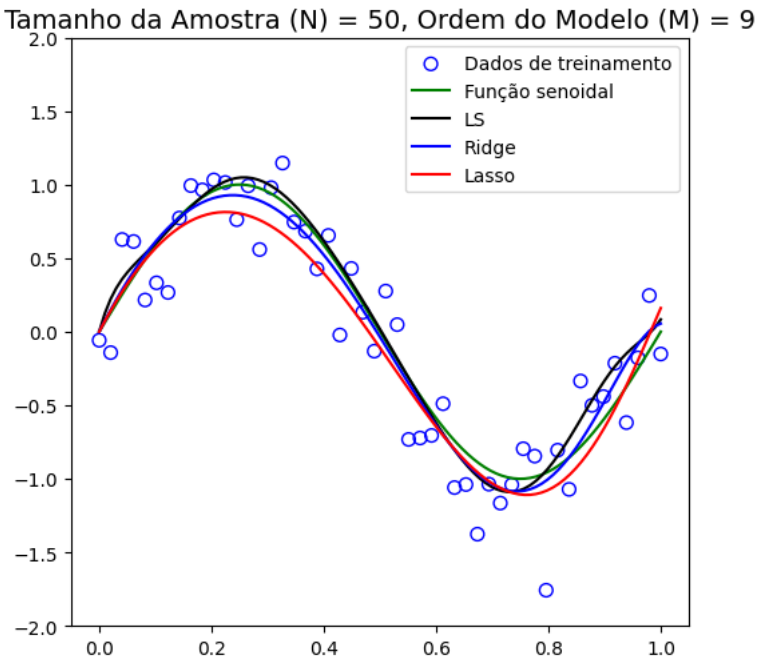
Regressão ridge:

- Coeficientes continuam a ser reduzidos em comparação com Mínimos Quadrados.
- Regularização é evidente e ajuda na estabilização dos coeficientes.
- Para a regressão ridge foi escolhido o valor de $alpha\_ridge = 0.001$. Percebemos, também, que os valores de w estão bem mais moderados se comparados com os de LS.

Regressão lasso:

- Coeficientes são mais esparsos em comparação com Ridge.
- Para a regressão lasso foi escolhido o valor de $alpha\_lasso = 0.0005$. Percebemos, também, que os valores de w estão bem mais moderados até mesmo se comparados com os da regressão ridge. Ainda, podemos evidenciar que há coeficientes de w iguais a zero, o que contribui para a redução da complexidade do modelo, a partir da redução da dimensionalidade.

iii. Tamanho da Amostra (N) = 50



	Mínimos Quadrados (LS)	Regressão Ridge	Regressão Lasso
w_0	0.0000	0.0000	0.0000
w_1	14.1518	7.5839	7.3279
w_2	-190.5701	-14.0968	-16.6494
w_3	1753.8674	-7.1305	0.0000
w_4	-8714.7782	3.0054	0.0000
w_5	24415.1417	8.2673	12.6420
w_6	-40571.5305	8.2538	0.0000
w_7	39720.5753	4.4603	0.0000
w_8	21121.0404	-1.5786	0.0000
w_9	4694.2663	-8.7097	-3.1599

Análise para Tamanho da Amostra (N) = 50:

Mínimos Quadrados (LS):

- Magnitudes dos coeficientes diminuem, indicando menor sensibilidade aos dados.

Regressão ridge:

- Coeficientes continuam a ser reduzidos, mas menos esparsos em comparação com Lasso.
- Para a regressão ridge foi escolhido o valor de $alpha\_ridge = 0.001$. Percebemos, também, que os valores de w estão bem mais moderados se comparados com os de LS.

Regressão lasso:

- Para a regressão lasso foi escolhido o valor de $alpha\_lasso = 0.0005$. Percebemos, também, que os valores de w estão bem mais moderados até mesmo se comparados com os da regressão ridge. Ainda, podemos evidenciar que há coeficientes de w iguais a zero, o que contribui para a redução da complexidade do modelo, a partir da redução da dimensionalidade.

E2 (Dados de câncer de próstata)

Data Source:

- Info:
  https://web.stanford.edu/~ hastie/ElemStatLearn/datasets/prostate.info.txt
- *Database*:
  https://web.stanford.edu/~ hastie/ElemStatLearn/datasets/prostate.data

"The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen (lpsa) and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45)."

Considere a variável (lpsa) como 'target' e as variáveis (lcavol), (lweight), (age), (lbph), (svi), (lcp), (gleason) e (pgg45) como 'entradas'.

Siga o roteiro abaixo:

a) Padronize os atributos de entrada para que eles tenham média 0 e variância 1;

b) Divida o *dataset* em dois conjuntos, treinamento e teste, conforme indicado nos índices da última coluna (T = treinamento, F = teste);

c) Encontre o modelo linear de regressão ótimo no critério de mínimos quadrados (solução LS);

d) Implemente modelos lineares regularizados pelos métodos 'Ridge' e 'Lasso' que minimizam a função objetivo L(w) = 1/(2N) RSS(w) + \lambda ||w||^q. Apresente resultados para um \lambda = 0.25;

e) Aplicando as regressões 'Ridge' e 'Lasso' e utilizando *k-fold cross-validation*, é possível selecionar um valor para \lambda que resulta em um modelo com melhor capacidade de generalização. Isso é feito selecionando o \lambda relativo à menor estimativa do erro de predição quadrático médio (usualmente chamado de *validation score*) ao longo dos *k-folds*. Também é possível selecionar um valor de \lambda que seleciona o modelo mais simples dentro de uma tolerância da estimativa do erro de predição quadrático médio. Isso é particularmente útil quando se deseja encontrar soluções esparsas (no caso do Lasso) ou de menor norma L2 (no caso do Ridge).
Para tal, um critério comumente adotado é a 'Regra de 1 desvio padrão', onde escolhe-se o maior \lambda cujo validation score seja igual ou pouco menor do que o 'score mínimo'+ '1 desvio padrão do score mínimo' .

- Monte as curvas de validation score de *k-fold cross validation* em função de \lambda para os modelos regularizados por 'ridge' e 'lasso'; (Sugestão: use k = 10, e procure \lambda em um intervalo [0, 0.5]).
- Calcule o desvio padrão do 'score' mínimo em cada respectiva curva e desenhe-o como barra de erro em torno daquele ponto.
- Determine o \lambda que resulta no modelo mais simples de acordo com a 'Regra de 1 desvio padrão'.
- Treine o modelo final 'ridge' e 'lasso' utilizando todos os dados (de treinamento) e o respectivo \lambda encontrado e apresente os resultados.

f) Utilizando o conjunto de teste construído no item (b), calcule a estimativa do erro de predição quadrático médio do conjunto de teste para cada modelo (mínimos quadrados, 'ridge' e 'lasso').
Disserte sobre os resultados obtidos.

(Bônus) Estime o desvio padrão dos coeficientes do modelo obtido pelo método de bootstrap dos resíduos.
Dica: Veja os slides 9 e 10 de:
http://www.est.ufmg.br/~cristianocs/MetComput/Aula8.pdf

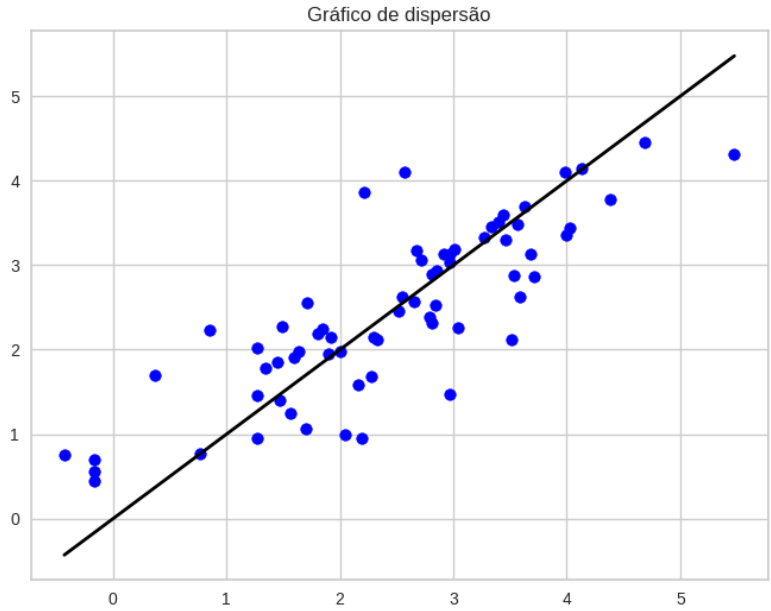
Solução:

https://colab.research.google.com/drive/1KjHurE-aLxJQJdX4Yvfn99mEGfgXPZhT#scrollTo=_yiemUWGeSSf

a/b) Comandos no código acima.

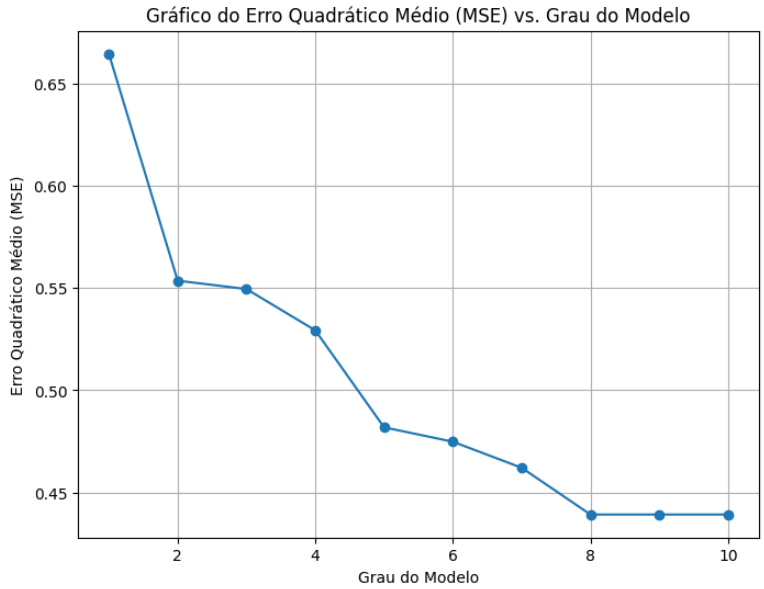
- Quantidade de dados marcados com "F"= 30 dados.
- Quantidade de dados marcados com "T"= 67 dados.

c)

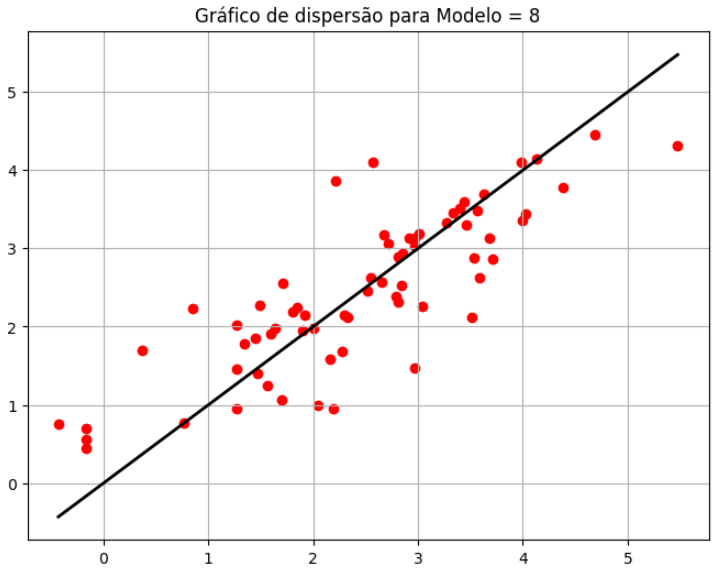


Atributos	Coeficientes
intercepto	2.4523450850746262
lcavol	0.71104059
lweight	0.29045029
age	-0.14148182
lbph	0.21041951
svi	0.30730025
lcp	-0.28684075
gleason	-0.02075686
pgg45	0.27526843
MSE	0.43919976805833433

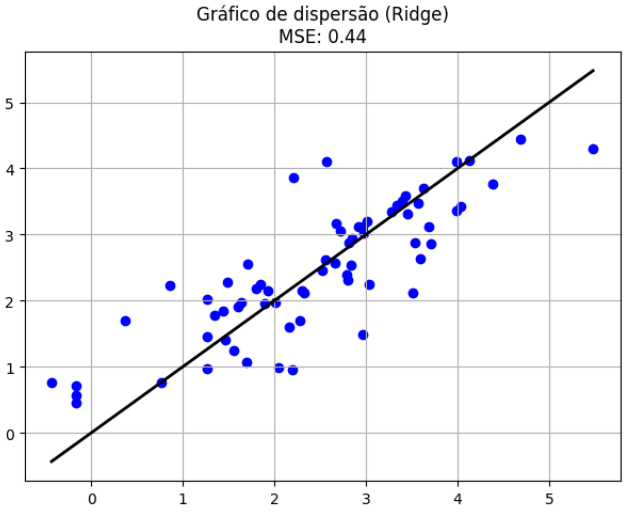
Matriz de correlação							
	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.30023						
age	0.28632	0.31672					
lbph	0.06316	0.43704	0.28734				
svi	0.59294	0.18105	0.12890	-0.13914			
lcp	0.69204	0.15682	0.17295	-0.08853	0.67124		
gleason	0.42641	0.02355	0.36591	0.03299	0.30687	0.47643	
pgg45	0.48316	0.07416	0.27580	-0.03040	0.48135	0.66253	0.75705

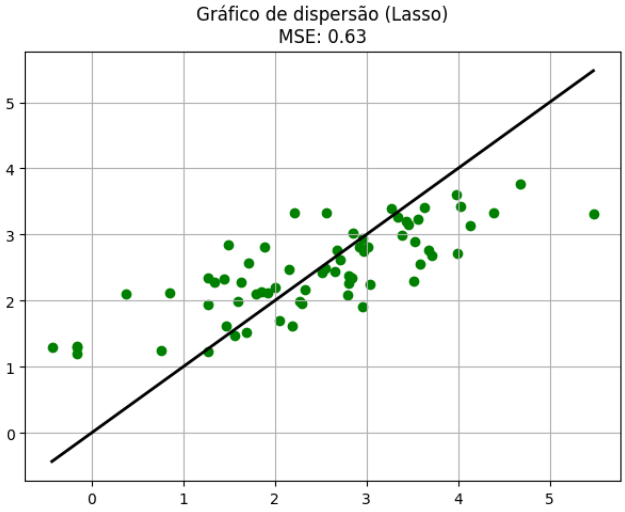


Portanto, temos que o melhor grau do modelo é 8, com MSE mínimo de 0.44. Para este modelo, temos o seguinte:



d)





Regressão ridge ($\lambda = 0.25$):

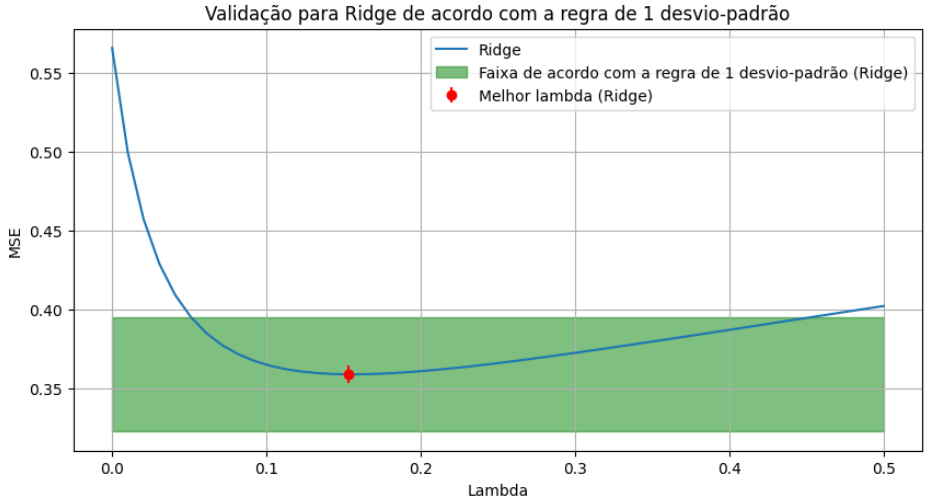
Atributos	Coeficientes
intercepto	2.4523450850746262
lcavol	0.70434465
lweight	0.29026491
age	-0.13964242
lbph	0.20992007
svi	0.30583538
lcp	-0.27831864
gleason	-0.01818066
pgg45	0.27009569
MSE	0.44

Regressão lasso (λ = 0.25):

Atributos	Coeficientes
intercepto	2.4523450850746267
lcavol	0.54140921
lweight	0.15677364
age	0
lbph	0
svi	0.06816303
lcp	0
gleason	0
pgg45	0
MSE	0.63

e) Análise para "Regra de 1 desvio padrão":

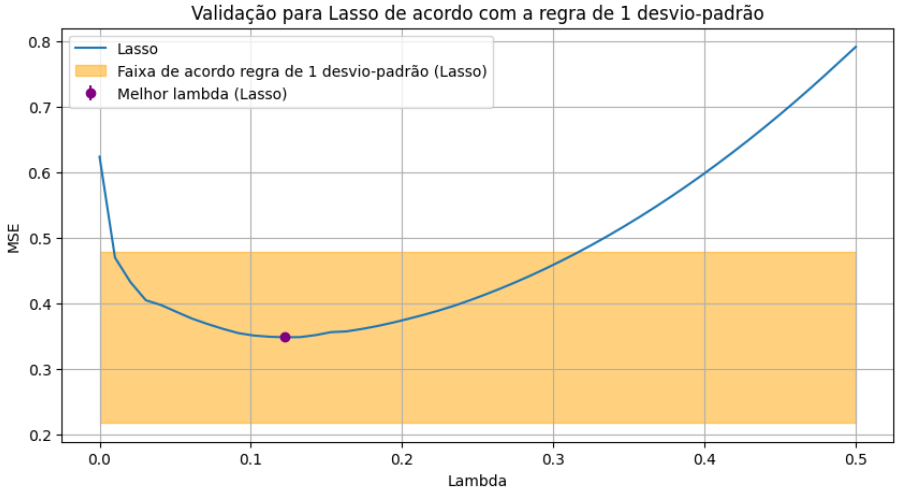
Regressão ridge:



Resultados:

- Melhor λ para ridge (MSE mínimo) = 0.1530612244897959.
- R^2 para Ridge com λ=0.01020408163265306 = 0.6943619175831537.
-

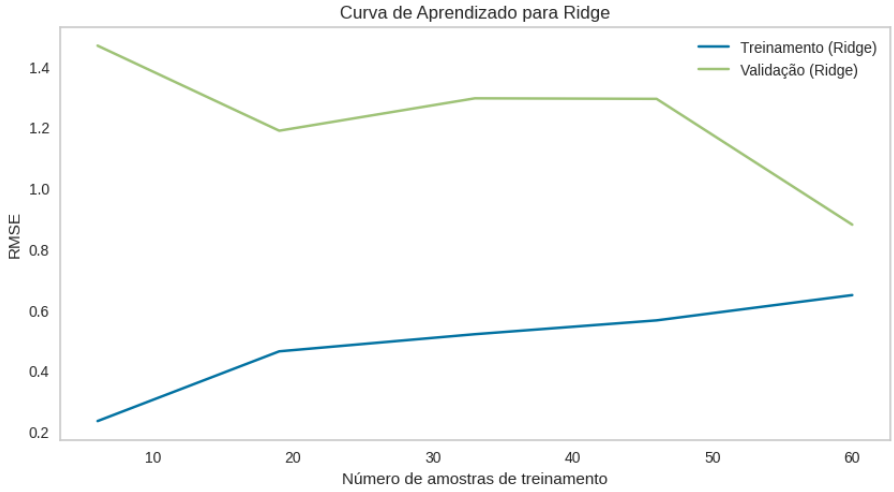
Regressão lasso:

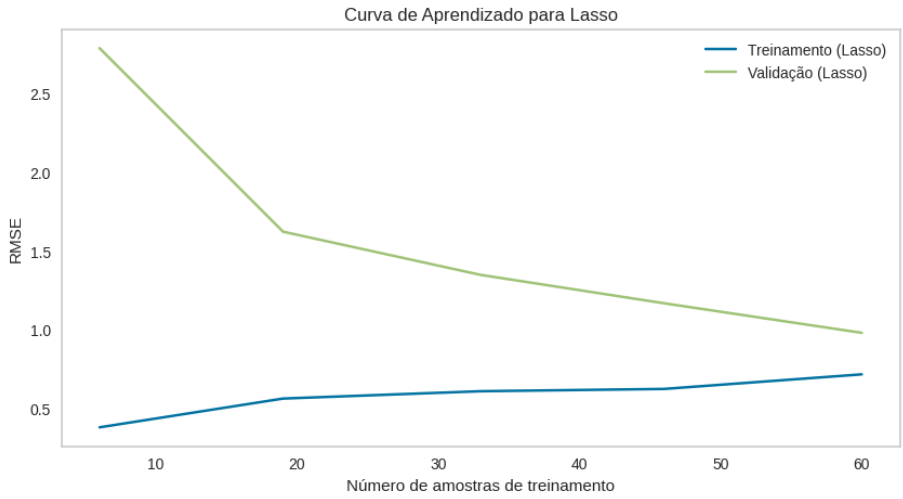


Resultados:

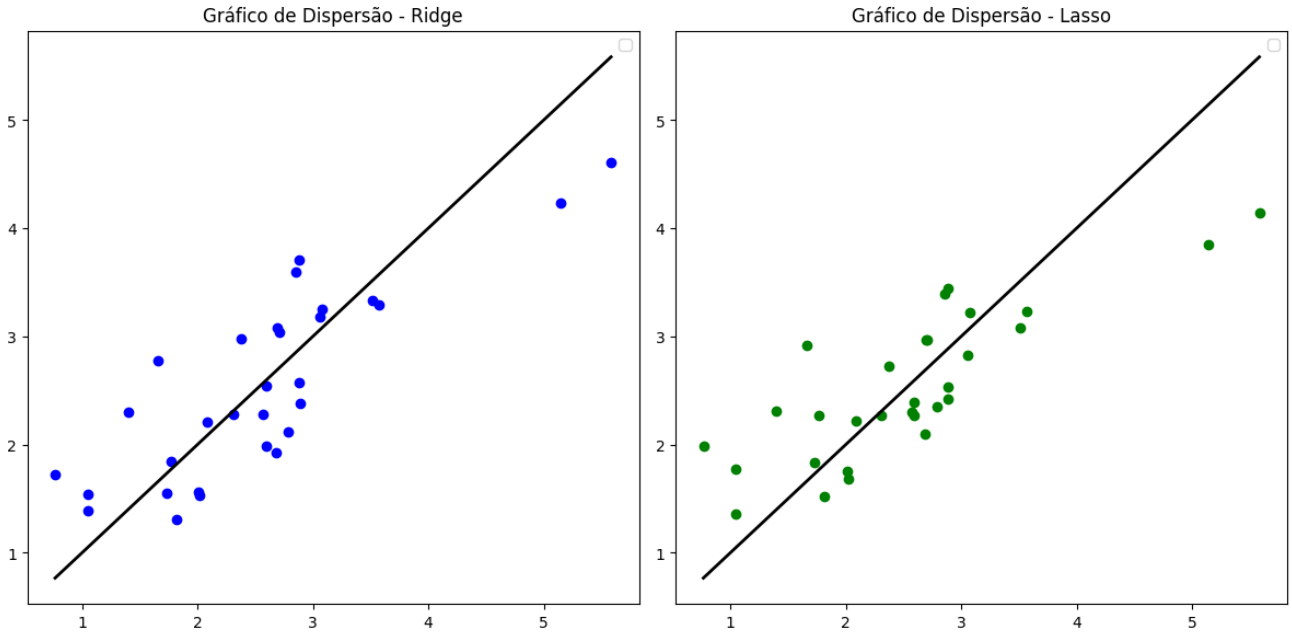
- Melhor λ para lasso (MSE mínimo) = 0.12244897959183673.
- R^2 para lasso com λ=0.11224489795918366 = 0.6382516987578413.

Análise comparativa das curvas de aprendizado para dados de treinamento:





f)



Regressão ridge:

Atributos	Coeficientes
intercepto	2.5365468833333336
lcavol	0.46083965
lweight	0.16291644
age	-0.06049845
lbph	-0.11487751
svi	0.26243087
lcp	0.23704707
gleason	-0.01141195
pgg45	0.01696037
MSE	0.3231996421134671

Regressão lasso:

Atributos	Coeficientes
intercepto	2.5365468833333336
lcavol	0.41748456
lweight	0
age	0
lbph	0
svi	0.1674072
lcp	0.2015806
gleason	0
pgg45	0
MSE	0.3729181259106822

Análise comparativa de resultados:

Regressão ridge

Atributos	Coeficientes para $\lambda = 0.25$	Coeficientes para $\lambda = 0.153061$
intercepto	2.4523450850746262	2.5365468833333336
lcavol	0.70434465	0.46083965
lweight	0.29026491	0.16291644
age	-0.13964242	-0.06049845
lbph	0.20992007	-0.11487751
svi	0.30583538	0.26243087
lcp	-0.27831864	0.23704707
gleason	-0.01818066	-0.01141195
pgg45	0.27009569	0.01696037
MSE	0.44	0.3231996421134671

Regressão lasso:

Atributos	Coeficientes para $\lambda = 0.25$	Coeficientes para $\lambda = 0.122448$
intercepto	2.4523450850746267	2.5365468833333336
lcavol	0.54140921	0.41748456
lweight	0.15677364	0
age	0	0
lbph	0	0
svi	0.06816303	0.1674072
lcp	0	0.2015806
gleason	0	0
pgg45	0	0
MSE	0.63	0.3729181259106822

Efeito da Regularização:

Ao comparar os dois conjuntos de coeficientes, pode-se observar que, em geral, os coeficientes tendem a ser ligeiramente menores (mais próximos de zero) para o maior valor de λ (0.25). Isso é esperado, já que um valor maior de λ impõe uma penalidade maior sobre a magnitude dos coeficientes na tentativa de reduzir overfitting.

MSE:

O MSE é uma medida da qualidade do modelo, com valores menores indicando um melhor ajuste. Os modelos com $\lambda = 0.153061$ para a regressão ridge e com $\lambda = 0.122448$ para a regressão lasso, possuem MSE menores (0.3231996421134671 e 0.3729181259106822, respectivamente) em comparação com o modelo com $\lambda = 0.25$ (0.44 e 0.63, respectivamente), sugerindo que o modelo com menor regularização se ajusta melhor aos dados no caso em análise.

Curva de aprendizado:

Cabe mencionar, ainda, que a curvas de aprendizado mostram como os erros de treino e generalização evoluem conforme utilizamos mais dados de treinamento. Nesse sentido, a curva de aprendizado é uma ferramenta visual para avaliar como o desempenho do modelo muda à medida que o número de amostras de treinamento aumenta [4]. Assim, podemos extrair as seguintes conclusões a partir da análise gráfica das curvas de aprendizado:

Desempenho no Treinamento:

- Em ambos os modelos, a linha azul representa o RMSE do conjunto de treinamento. Ela começa baixa e permanece relativamente constante à medida que o número de amostras de treinamento aumenta. Isso sugere que o modelo não está significativamente melhorando seu desempenho no conjunto de treinamento com a adição de mais dados. Isso pode ser um

indicativo de que os modelos já aprenderam o que podiam das amostras de treinamento disponíveis ou que mais dados de treinamento são necessários para melhorar as performances.

Desempenho na Validação:

- Em ambos os modelos, a linha verde representa o RMSE do conjunto de validação. Podemos observar que o erro diminui à medida que mais dados de treinamento são usados. Isso indica que os modelos estão generalizando melhor e aprendendo a fazer previsões mais precisas à medida que recebem mais dados de treinamento.

Convergência das Curvas:

- Em ambos os modelos, as duas linhas parecem estar convergindo, o que é um sinal positivo. Em um modelo ideal, esperamos que o desempenho no conjunto de treinamento e no conjunto de validação sejam semelhantes, o que significa que os modelos estão bem ajustados e generalizam bem para novos dados.

***Overfitting* e *Underfitting*:**

- Em ambos os modelos, não parece haver um problema significativo de *overfitting,* já que a linha de treinamento não tem um RMSE muito baixo em comparação com a de validação. Também não parece haver *underfitting* significativo, já que a performance no conjunto de validação está melhorando com mais dados.

Bônus

Regressão ridge:

Desvio padrão dos coeficientes obtidos pelo método de *bootstrap* dos resíduos		
Atributos	**Coeficientes *Bootstrap***	**desvio padrão**
intercept	-0.09453	0.093871
lcavol	0.532221	0.124273
lweight	0.248539	0.119341
age	-0.162538	0.099765
lbph	0.157043	0.11242
svi	0.2793	0.135156
lcp	-0.070522	0.158699
gleason	0.062012	0.146296
pgg45	0.018656	0.169442

Regressão lasso:

Desvio padrão dos coeficientes obtidos pelo método de *bootstrap* dos resíduos		
Atributos	**Coeficientes *Bootstrap***	**desvio padrão**
Intercept	-0.053043	0.104225
lcavol	0.623143	0.159437
lweight	0.125313	0.098099
age	0.0	0.042059
lbph	0.0	0.048692
svi	0.142413	0.110345
lcp	0.0	0.047877
gleason	0.0	0.035823
pgg45	0.0	0.060935

REFERÊNCIA:

[1] T. Hastie, R. Tibshirani e J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition", 2009.

[2] J. Ranstam and J. A. Cook, "LASSO regression," British Journal of Surgery, vol. 105, no. 10, p. 1348, 2018.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit", SIAM Journal on Scientific Computing, pp. 33–61, 1998.

[4] M. Facure, "Avaliando um Modelo de Aprendizado de Máquina", Disponível em: https://matheusfacure.github.io/2017/03/04/aval-modelo-am/. Acesso em: 19/12/2023.