Universidade Federal do Rio de Janeiro (UFRJ)
Programa de Engenharia Elétrica (COPPE/UFRJ)
Disciplina: Introdução ao aprendizado de máquina (COE 782)
Professor: Markus Vinicius Santos Lima

Aluna: Vivian de Carvalho Rodrigues
Programa de Engenharia Civil da COPPE/UFRJ
DRE:125228569

Objetivo: Fazer um relatório da Lista 3. Os exercício foram realizados na linguagem de programação Python v3.8.18, através do Notebook Jupyter v.6.5.2 (Gerenciador Anaconda).

1) Experimento "Polynomial Curve Fitting"

Toda a resolução computacional se encontra no anexo E.1.

2) Descrição dos Dados — Prostate Dataset

Número de observações: 97 indivíduos
Número de preditores (colunas 1–8):

- lcavol – log do volume do tumor no exame de cavidade
- lweight – log do peso da próstata
- age – idade
- lbph – log do tamanho do aumento da próstata benigna
- svi – invasão de vesícula seminal (0 ou 1)
- lcp – log do tumor da cápsula prostática
- gleason – escore de Gleason (agressividade do câncer)
- pgg45 – percentual de células com nota Gleason 4 ou 5
- Variável alvo (coluna 9):
- lpsa – log do antígeno prostático específico (PSA), o alvo da regressão

- Indicador de treino/teste (coluna 10):

Mostra quais das 97 observações pertencem ao conjunto de treinamento (67) e ao de teste (30).

3) Correção Importante

O valor de "lweight" na observação 32 estava errado: era 6.1, o que implicava um peso de próstata de 449g.

Valor correto: log(44.9) ≈ 3.80

4) Observação

Este exercício foi resolvido por etapas e dividido em três arquivos:

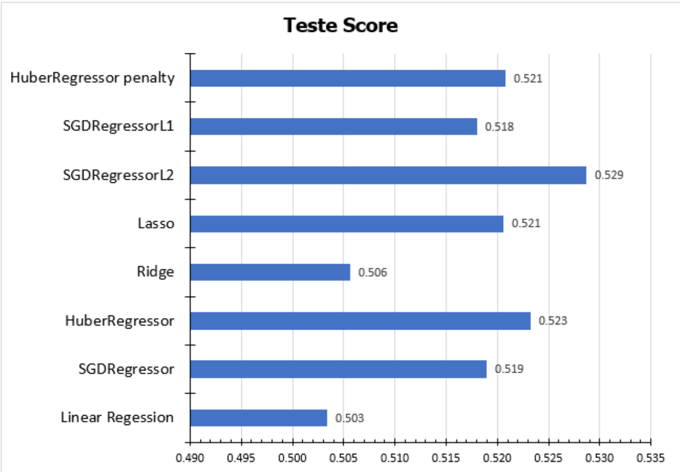
- COE782_ML_Lista3_E2_Vivian_parte1.ipynb (Anexo E.2_1)
  Contém a resolução dos itens a), b), c) , d) e f)
- COE782_ML_lista3_E2_Vivian_iteme_cross_v2. Ipynb (Anexo E.2_2)
  Resolução do item e)
- Bonus.ipynb (Anexo E.2_3)

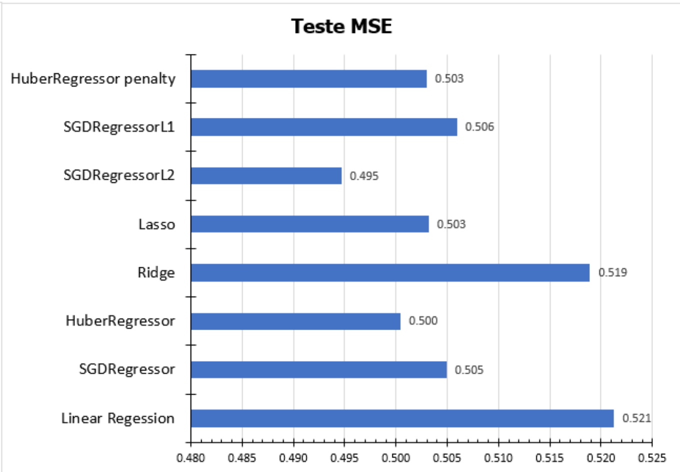
5) Modelos a serem utilizados - itens c) , d) e f)

Modelo	**Penalização**	**Robustez**	**Seleção de variáveis**	**Recomendado para...**
LinearRegression	Nenhuma	Não	Não	Casos simples, sem regularização
Ridge	L2	Parcial	Não	Multicolinearidade
Lasso	L1	Parcial	Sim	Seleção de variáveis
SGDRegressor	L1/L2/EN	Não	Sim (com L1)	Grandes volumes de dados
HuberRegressor	Nenhuma (com função de perda robusta)	Sim	Não	Dados com outliers

6) Resultados do item c) , d) e f)

	Penalização			Coef.Determinacao - R^2		MSE	
Modelo	**Ridge - l2**	**Lasso - L1**	**Lambda**	**Treino**	**Teste**	**Treino**	**Teste**
Linear Regession	-	-	N.A	0.694	0.503	0.439	0.521
SGDRegressor	-	-	N.A	0.684	0.519	0.454	0.505
HuberRegressor	-	-	N.A	0.678	0.523	0.463	0.500
Ridge	Sim	Não	0.25	0.694	0.506	0.439	0.519
Lasso	Não	Sim	0.25	0.568	0.521	0.621	0.503
SGDRegressorL2	Sim	Não	0.25	0.664	0.529	0.482	0.495
SGDRegressorL1	Não	Sim	0.25	0.567	0.518	0.622	0.506
HuberRegressor penalty	-	-	0.25	0.679	0.521	0.462	0.503





Análise de discussões

1. <u>Regressões sem regularização</u> (lambda = 0):

Modelo Score de teste (R^2)

LinearRegression 0.503

SGDRegressor 0.519

HuberRegressor 0.523

Todos os modelos estão explicando cerca de 50–55% da variância dos dados de teste, o que indica que estes modelos têm algum poder preditivo, mas podem haver variáveis importantes faltando.

2. <u>Modelos com regularização</u> (lambda = 0.25):

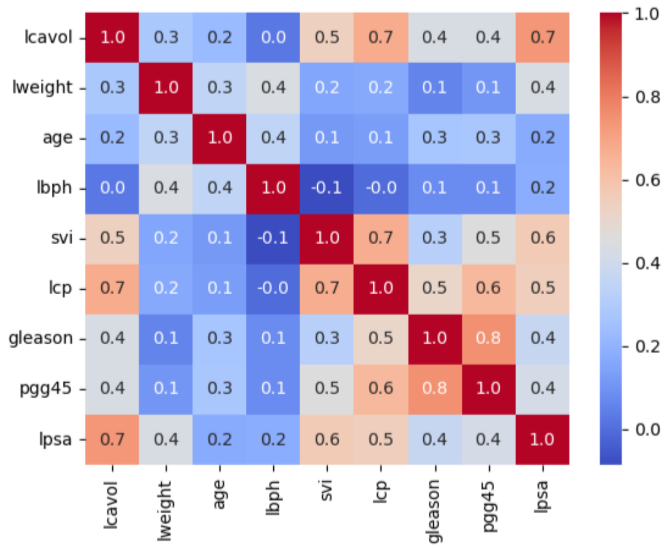
Modelo	Score de teste (R^2)
Ridge (L2)	0.506
Lasso (L1)	0.521
SGDRegressor (penalty='l2')	0.529
SGDRegressor (penalty='l1')	0.518
HuberRegressor (com penalidade)	0.521

SGD com L2 manteve bom desempenho, quase igual ao caso sem penalidade — sugerindo que o modelo está estável. SGD com L1 e Huber com penalidade não trazem melhorias significativas.

Nenhum modelo superou significativamente os outros. A diferença entre 0.503 e 0.529 é relativamente pequena e pode estar dentro da flutuação esperada dependendo do split treino/teste. Modelos com penalidade (especialmente Ridge) podem trazer leve ganho em estabilidade e generalização.

Os modelos estão aprendendo padrões reais dos dados (score de treino acima de 0.60 mostra isso). Mas **não estão generalizando perfeitamente** para os dados de teste (perda de ~0.1 em R^2). Diferença moderada entre treino e teste é comum quando

- O *dataset* é pequeno (97 amostras).
- As relações são fracas ou parcialmente não lineares
- Há variáveis correlacionadas. Neste experimento foi observado que existem algumas variáveis correlacionadas.



- Há ruído natural nos dados

Análise Geral dos Resultados

1. Métrica R^2 (Score)

Melhor resultado no teste: SGDRegressorL2 ($R^2 = 0.529$)

Pior resultado no teste: *Linear Regression* ($R^2 = 0.503$), apesar de ter sido um dos melhores no treino.

2. MSE

Mede o erro quadrático médio entre os valores previstos e os reais.

Melhor resultado no teste: SGDRegressorL2 (MSE = 0.495)

Pior resultado no teste: *Linear Regression* (MSE = 0.521)

Modelos com Melhor Desempenho no Teste

- SGDRegressorL2

R^2 Teste = 0.529, MSE Teste = 0.495

Melhor equilíbrio entre erro e capacidade de generalização.

- *HuberRegressor*

R^2 Teste = 0.523, MSE Teste = 0.500

Muito robusto a outliers, bom desempenho geral.

- Lasso

Curiosamente, R^2 Teste = 0.521, com MSE = 0.503

Apesar do R^2 bom no teste, tem MSE alto no treino, o que pode indicar *underfitting*.

Os modelos *Linear Regression* e *Ridge* apresentaram boa performance no treino (R^2 = 0.694), mas queda significativa no teste (R^2 ~0.50), que é um indício de *overfitting*. Já os modelos Lasso e SGDRegressorL1 apresentaram baixo R^2 no treino (~0.56), mas relativamente melhor no teste (~0.52). Isto é um potencial *underfitting*, talvez os hiper parâmetros estejam muito restritivos (muita regularização).

Conclusão

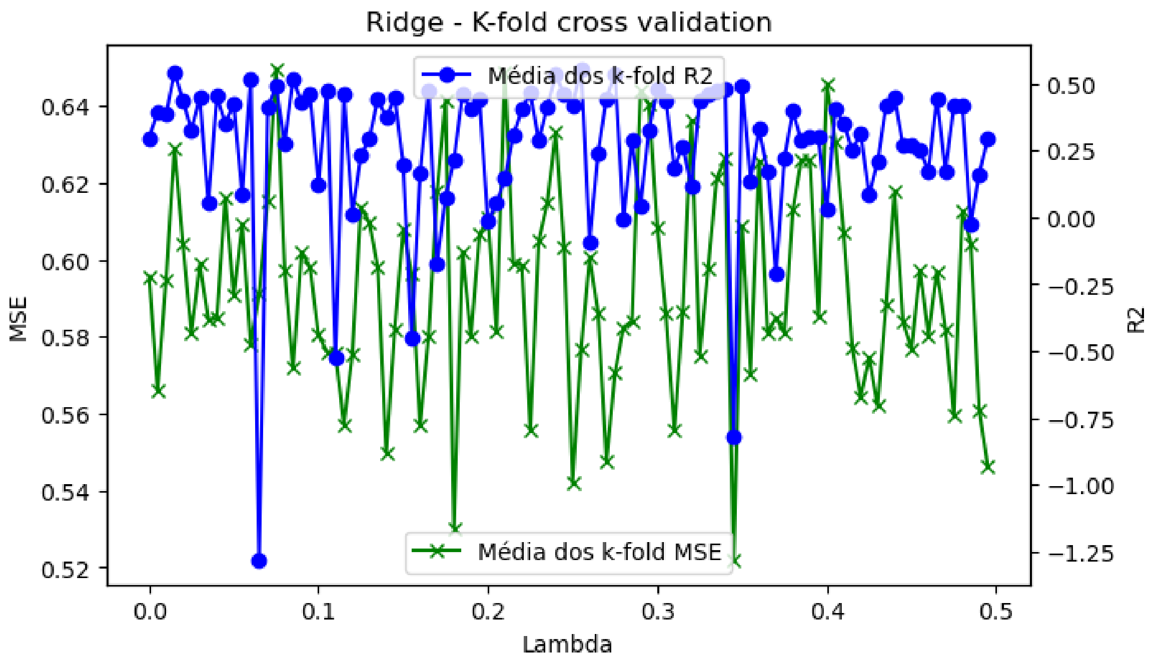
O *SGDRegressor* com regularização L2 é o mais promissor entre os testados, com melhor R^2 e menor MSE no conjunto de teste. Modelos robustos como o *HuberRegressor* também apresentam razoável capacidade de generalização.

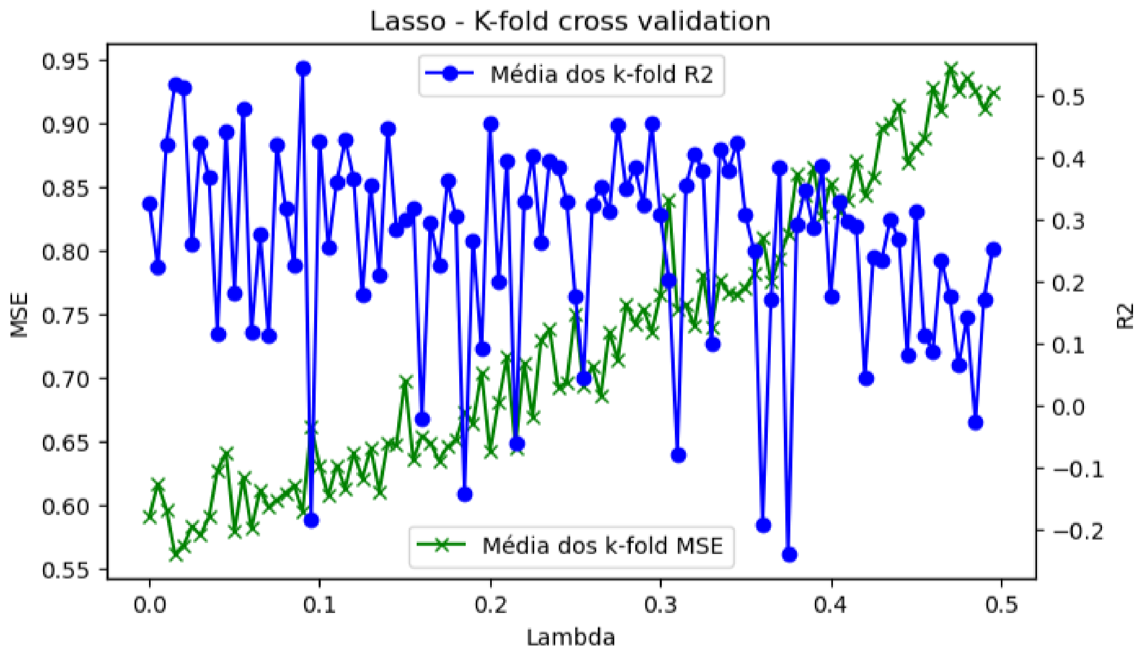
Nesta etapa do exercício é possível observar que a regularização pode estar ajudando os outros modelos a generalizar melhor.

7) Resultados do item e) – Validação Cruzada

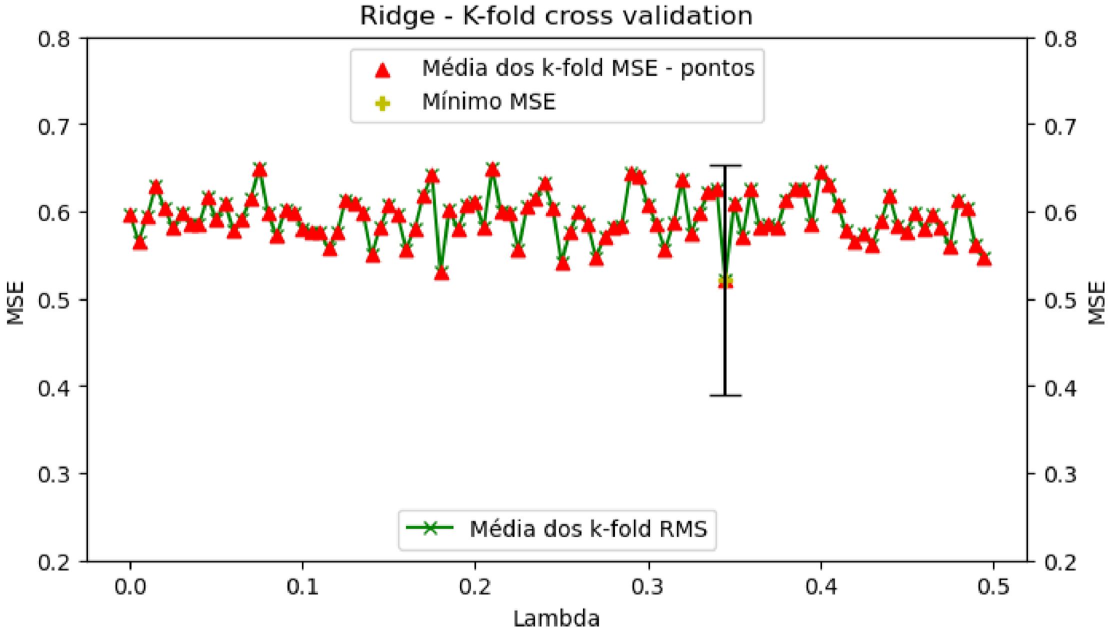
Vale observar que *score,* no *Pyhton,* se refere ao coeficiente de determinação *(R^2).* Portanto, neste relatório foi considerado o erro quadrático médio (MSE)

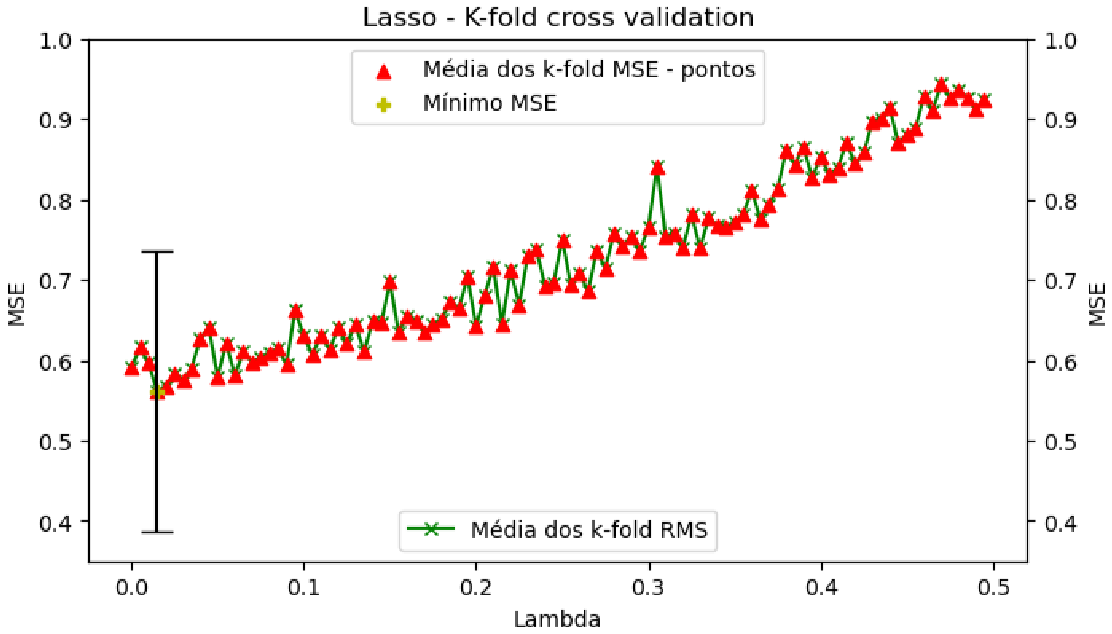
(i) Curvas de *validation* score de k-fold cross e MSE (mean square error)





(ii) Score mínimo (MSE mínimo) e o desvio padrão do score mínimo como barra de erro





No modelo Ridge, os valores de erro quadrático médio (média dos 10 folds para cada lambda) oscilou em torno de uma média, indicando que a regularização não impactou no desempenho do modelo. Foram considerados 100 lambdas na faixa [0,0.5]. Por isso considerou-se o lambda como a média dos 5 lambdas centrados no lambda com menor MSE.

No modelo Lasso, ao considerar MSE mínimo + 1 desvio padrão o valor ficou muito alto, e assim todos os fatores de regularização seriam selecionados.

Foi testada uma segunda abordagem, onde se utilizaria 'Fold MSE mínimo' + 1 desvio padrão. No entanto o valor ficou mais baixo de todos os lambdas calculados e nenhum seria selecionado.

Portanto, no modelo Lasso considerou-se MSE mínimo + 0.5 desvio padrão.

Então

Lambda (Ridge) = 0.345

MSE treino: 0.4392581936694039

MSE teste: 0.5180657802293058

Lambda (Lasso) = 0.37

MSE treino: 0.7343645001723449

MSE teste: 0.5833164758076694

8) Bonus

Estimar o desvio padrão dos coeficientes de um modelo de regressão, usando *bootstrap* dos resíduos.

O método:

1. Ajusta o modelo nos dados originais.
2. Calcula os resíduos.
3. Gera novos conjuntos de respostas y com base na predição + amostra com reposição dos resíduos.
4. Reajusta o modelo com os dados (X,y∗)
5. Repete várias vezes e calcula a variabilidade dos coeficientes.

Os resultados estão no E2_Bonus.

Referência

The Elements of Statistical Learning (Hastie, Tibshirani e Friedman)