Relatório de Regressão Polinomial

Instituto Federal do Ceará, Campus Maracanaú

Disciplina: Reconhecimento de Padrões
Professor: Hericson Araújo
Aluno: Francisco Aldenor Silva Neto
Matrícula: 20221045050117

Introdução

Este relatório descreve a aplicação de modelos de regressão polinomial para prever os preços das casas em Boston utilizando o conjunto de dados **boston.csv**. O objetivo é treinar modelos de regressão polinomial variando a ordem dos polinômios de 1 a 11 e avaliar o desempenho desses modelos com e sem regularização L2 (Ridge Regression).

A regressão polinomial é uma extensão da regressão linear que busca modelar relações não lineares entre as variáveis independentes e dependentes, ajustando um polinômio de grau (d). Para evitar o sobreajuste (overfitting), utilizamos a regularização L2, que adiciona uma penalidade aos coeficientes do modelo.

Fórmula da Regressão Polinomial

g A regressão polinomial busca minimizar o erro quadrático entre os valores previstos e os reais. A função objetivo é definida como:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta d x^d$$

onde:

- x são as variáveis independentes,
- $\beta_0, \beta_1, \ldots, \beta d$ são os coeficientes do modelo,
- d é o grau do polinômio.

A função de custo (erro quadrático médio) é dada por:

$$J(\beta) = (1/n) * \Sigma_i (y_i - \hat{y}_i)^2$$

Regularização L2

Para evitar que o modelo ajuste demais os dados de treino (overfitting), a regularização L2 adiciona um termo de penalização baseado na soma dos quadrados dos coeficientes:

$$J_{Ridge}(\beta) = (1/n) * \Sigma_i (y_i - \hat{y}_i)^2 + \lambda * \Sigma_j \beta_j^2$$

onde:

- λ é o parâmetro de regularização que controla a intensidade da penalização,
- $\sum_j \beta_j^2$ é o termo de regularização que penaliza coeficientes grandes.

Com $\lambda = 0$, a regularização não é aplicada, resultando na regressão polinomial padrão. Para $\lambda > 0$, a regularização L2 reduz a magnitude dos coeficientes, limitando a complexidade do modelo.

Conjunto de Dados

O conjunto de dados **boston.csv** contém 14 colunas, sendo 13 atributos (independentes) e 1 variável de saída (dependente), que corresponde ao preço das casas em Boston na década de 1970.

Metodologia

1. Divisão dos Dados

Os dados foram divididos aleatoriamente em dois conjuntos:

- **Treinamento (80%)**
- **Teste (20%)**

2. Normalização dos Dados

Os dados de entrada foram normalizados utilizando a faixa entre o menor e o maior valor dos dados de treinamento.

3. Modelos de Regressão Polinomial

Foram treinados modelos de regressão polinomial de grau 1 a 11, tanto com quanto sem regularização L2. O processo de treinamento foi realizado utilizando a biblioteca **scikit-learn** para manipulação de dados e modelos.

4. Regularização L2

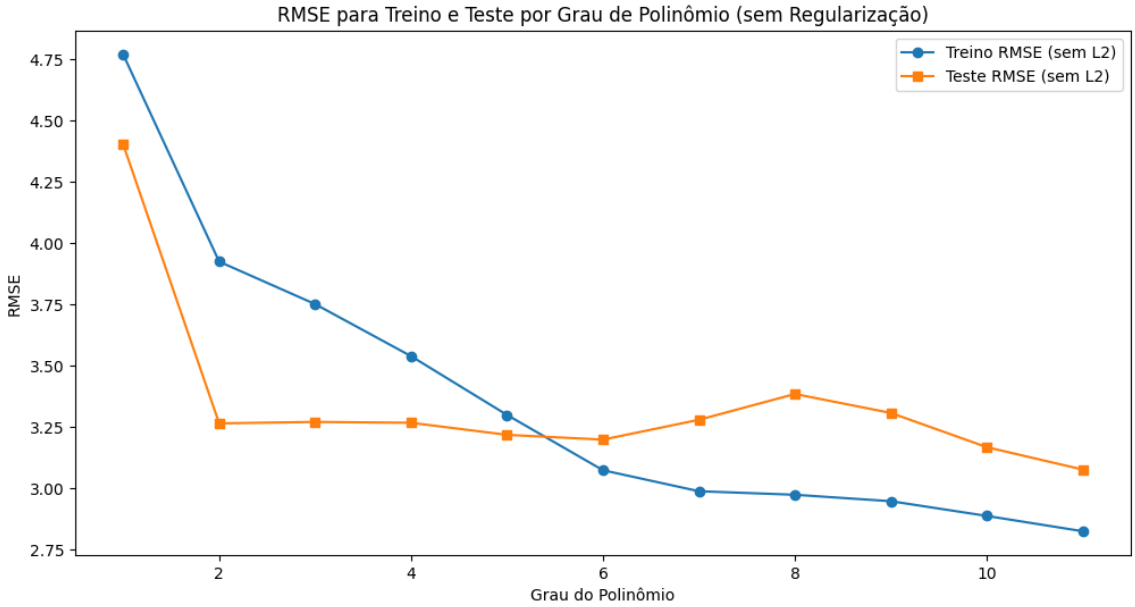
A regularização L2 foi aplicada utilizando o modelo de **Ridge Regression**, com um valor de **lambda** igual a 0,01.

5. Avaliação dos Modelos

A performance dos modelos foi avaliada utilizando o **RMSE** (Raiz Quadrada do Erro Quadrático Médio) para as previsões tanto no conjunto de treino quanto no conjunto de teste.

Resultados

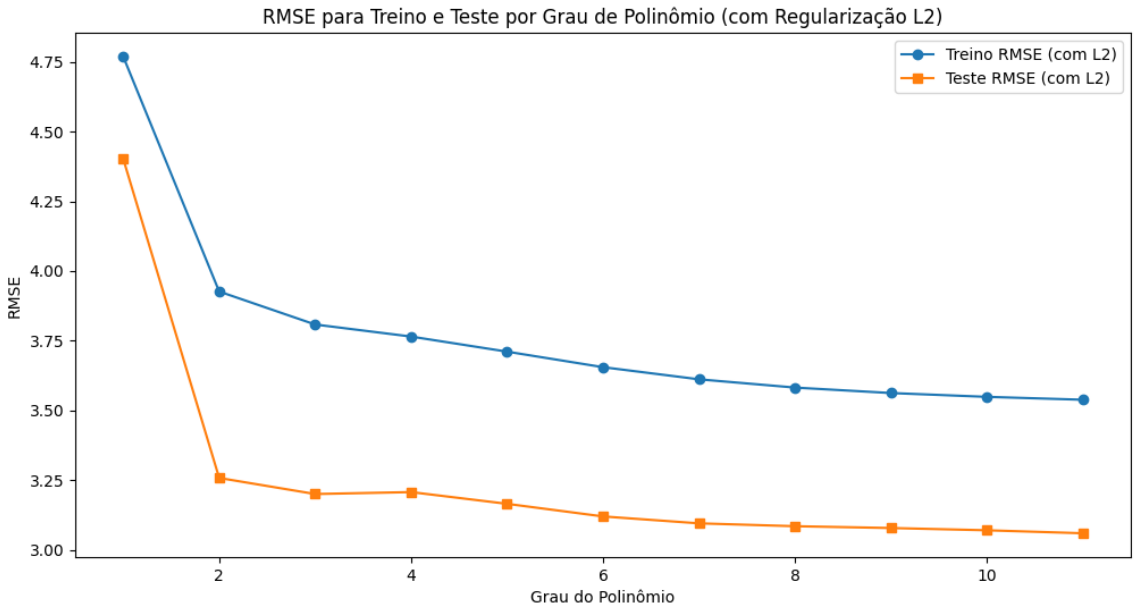
RMSE sem Regularização L2



Grau do Polinômio	RMSE Treinamento	RMSE Teste
1	4.7689	4.4022
2	3.9248	3.2657
3	3.7518	3.2716
4	3.5395	3.2681
5	3.3000	3.2189
6	3.0747	3.1995
7	2.9891	3.2803
8	2.9747	3.3855
9	2.9485	3.3082
10	2.8885	3.1686
11	2.8260	3.0775

Observa-se que o overfitting começa a se tornar evidente a partir do grau 3. A partir desse ponto, o RMSE no conjunto de treino diminui de forma expressiva, enquanto o RMSE no conjunto de teste permanece estável ou aumenta levemente, indicando que o modelo está ajustando ruídos ao invés de padrões reais.

RMSE com Regularização L2 (lambda = 0,01)



Grau do Polinômio	RMSE Treinamento	RMSE Teste
1	4.7689	4.4022
2	3.9264	3.2579
3	3.8081	3.2002
4	3.7649	3.2071
5	3.7108	3.1651
6	3.6551	3.1199
7	3.6116	3.0951
8	3.5820	3.0848
9	3.5624	3.0783
10	3.5487	3.0702
11	3.5383	3.0596

Com a regularização L2, o overfitting é mitigado significativamente. A regularização penaliza os coeficientes mais elevados, limitando a complexidade do modelo e promovendo um melhor equilíbrio entre os erros de treino e teste. Observa-se que, para os graus mais elevados, os valores de RMSE no conjunto de teste são menores do que na versão sem regularização, destacando a eficácia da regularização para modelos mais complexos.

Fins de Comparação: Análise dos Efeitos de Diferentes Valores de Regularização

Para avaliar os impactos da regularização L2 no comportamento dos modelos e sua capacidade de resolver o problema de overfitting, foram realizados testes com os valores de $\lambda = 0.1$ e $\lambda = 10$. O objetivo foi observar se ajustes na penalidade poderiam reduzir a discrepância entre os erros de treino e teste, garantindo uma melhor generalização.

RMSE com Regularização L2 (lambda = 0,1)

O gráfico gerado a partir desse teste está apresentado na **Figura abaixo**, e a tabela seguinte lista os RMSEs obtidos.

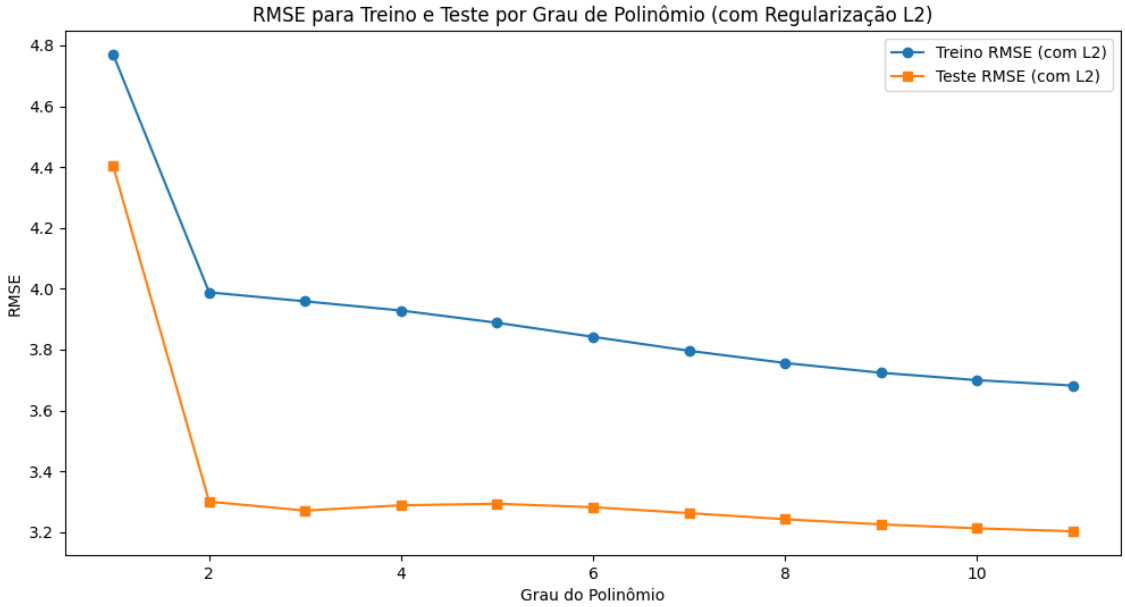


Tabela RMSE com (\lambda = 0.1):

Grau do Modelo	RMSE Treino	RMSE Teste
1	4.7701	4.4029
2	3.9883	3.3003
3	3.9588	3.2710
4	3.9286	3.2883
5	3.8883	3.2934
6	3.8421	3.2820
7	3.7963	3.2626
8	3.7562	3.2425
9	3.7240	3.2256
10	3.6998	3.2126
11	3.6820	3.2030

Resultados para (\lambda = 10):

Na sequência, foi realizado o mesmo experimento com (\lambda = 10), cujos resultados estão ilustrados na **Figura abaixo** e descritos na tabela seguinte.

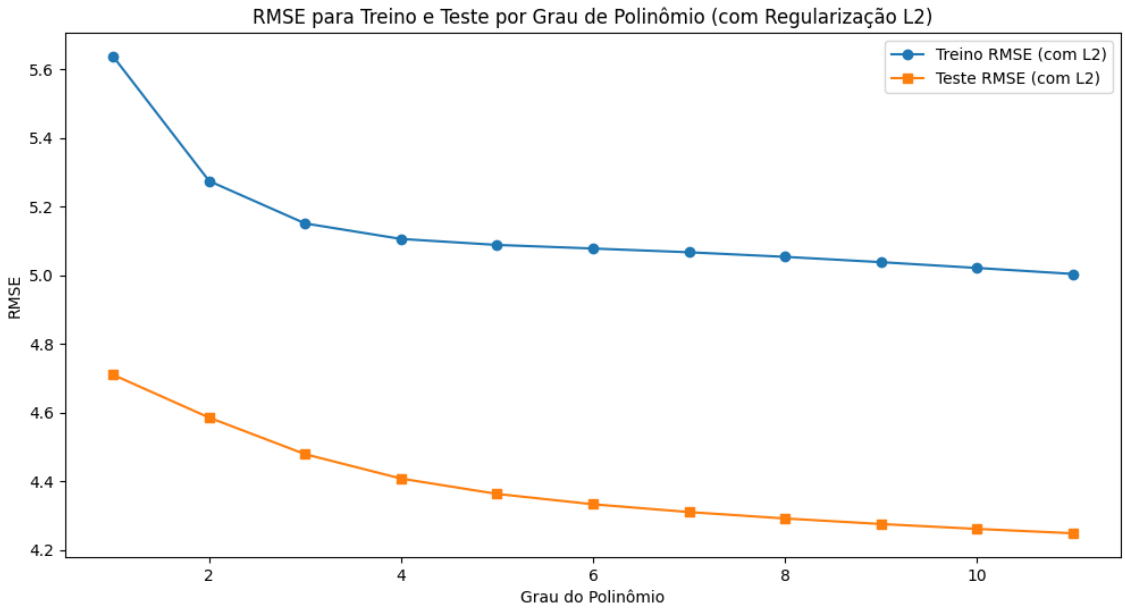


Tabela RMSE com (\lambda = 10):

Grau do Modelo	RMSE Treino	RMSE Teste
1	5.6372	4.7098
2	5.2745	4.5851
3	5.1509	4.4786
4	5.1057	4.4075
5	5.0883	4.3628
6	5.0779	4.3327
7	5.0669	4.3100
8	5.0536	4.2913
9	5.0381	4.2752
10	5.0213	4.2609
11	5.0038	4.2482

Discussão dos Resultados:

Ao aumentar (\lambda) de 0.1 para 10, observou-se uma diminuição do efeito de overfitting, especialmente em modelos de grau elevado. A regularização mais intensa suprime coeficientes polinomiais excessivos, melhorando a estabilidade dos resultados no conjunto de teste, mas com aumento no erro no conjunto de treino. Essa troca reflete o compromisso entre bias e variância, típico em problemas de regressão com regularização.

Conclusão

Os resultados demonstram que a regularização L2 ajuda a controlar o sobreajuste, especialmente para polinômios de grau mais elevado, ao limitar o ajuste excessivo aos dados de treino. No entanto, o efeito de overfitting ainda é observado para graus maiores que 4, indicando que, apesar da regularização, esses modelos ainda capturam ruídos específicos do conjunto de treino.

Sem regularização, o modelo apresenta um ajuste quase perfeito no treino para graus elevados, mas com um aumento drástico do erro no conjunto de teste, evidenciando um sobreajuste severo. A regularização L2 reduz o impacto desse comportamento, estabilizando os erros de teste, mas não eliminando completamente o efeito.

Observações

- Para graus mais baixos (1 e 2), o desempenho dos modelos com e sem regularização é semelhante, pois a complexidade do modelo é gerenciável.
- A regularização L2 limita o ajuste excessivo em graus mais altos, melhorando a estabilidade dos modelos. No entanto, o efeito de overfitting ainda persiste para modelos mais complexos (graus acima de 4).

Repositório no GitHub

O código fonte deste trabalho está disponível no seguinte repositório:
Regressão Polinomial - RMSE