

# Comportamento do Overfitting

Rodrigo Azevedo Santos  
Trabalho II de Inteligência Computacional II  
2017.2  
Email: rodrigo4zevedo@gmail.com

Instituto Alberto Luiz Coimbra de Pós-Graduação  
e Pesquisa de Engenharia  
Universidade Federal do Rio de Janeiro  
Rio de Janeiro, Brasil

## I. INTRODUÇÃO E OBJETIVO

Overfitting é um caso muito comum no aprendizado de máquinas. Ocorre quando o ajuste aos dados não indica um bom resultado fora-da-amostra, ou seja, no mundo real. Em poucas palavras, overfitting significa “ajustar os dados mais que o necessário”. Quando o modelo é mais complexo que o necessário para representar uma função alvo, o overfitting ocorre [1], levando suas previsões a resultados inesperados fora-da-amostra, porém com um bom resultado dentro-da-amostra. Este trabalho tem como objetivo estudar quando o overfitting ocorre e demonstrar que o uso de graus de liberdade adicionais (modelos mais complexos) acabam sendo usados para se “ajustar demais” aos dados e incorporando por exemplo, ruídos ao aprendizado. Portanto, o seguinte experimento visa estudar o comportamento do aprendizado diante do impacto de três parâmetros: nível  $\sigma^2$  do ruído adicionado aos dados, a complexidade  $Q_f$  da função alvo (grau- $Q_f$  do polinômio) e a quantidade de dados  $N$  disponível ( $N$  pontos no conjunto de dados). Para a realização do experimento, a diferença entre duas hipóteses finais de diferentes complexidades  $g_2 \in \mathcal{H}_2$  e  $g_{10} \in \mathcal{H}_{10}$ , um modelo menos complexo e um mais complexo respectivamente, é utilizado para a determinar a ocorrência de overfitting.

## II. METODOLOGIA

Como o intuito do experimento é compreender o comportamento do Overfitting em relação à complexidade das funções alvo, nível de ruídos apresentados aos dados e tamanho do conjunto de dados, as funções alvo foram geradas de forma aleatória para cada cenário e iteração. De modo que a média dos experimentos apresentasse uma noção do impacto de tais fatores ao Overfitting. As funções alvos utilizadas para o aprendizado foram geradas no espaço  $\chi = [-1, 1]$ , e sendo polinômios de grau  $Q_f$ , escrita da forma

$$f(x) = \sum_{q=0}^{Q_f} a_q L_q(x) \quad (1)$$

Onde  $L_q(x)$  são polinômios Legendre e  $a_q$  são coeficientes gerados aleatoriamente de uma distribuição Normal. Os coeficientes foram então normalizados de modo que  $E[f^2] = 1$ , multiplicando-os por uma constante de normalização  $\theta$  obtido da seguinte forma

$$\theta = \frac{1}{\sqrt{\frac{\int_{-1}^1 f^2 dx}{\int_{-1}^1 1 dx}}} \quad (2)$$

Devido a propriedade de ortogonalidade dos polinômios Legendre com respeito à norma Euclidiana no intervalo  $[-1, 1]$ , temos que

$$\int_{-1}^1 P_m(x) P_n(x) dx = \frac{2}{2n+1} \delta_{mn} \quad (3)$$

Com  $\delta_{mn}$  sendo o Delta de Kronecker, descrito da seguinte forma

$$\delta_{mn} = \begin{cases} 1, & \text{se } i = j \\ 0, & \text{se } i \neq j \end{cases}$$

Aplicando a equação (1) do polinômio Legendre à constante de normalização definida em (2) e utilizando a propriedade de ortogonalidade apresentada em (3) temos que

$$\theta = \frac{1}{\sqrt{\sum_{q=0}^{Q_f} \frac{a_q^2}{2q+1}}} \quad (4)$$

Portanto a função alvo  $f$  é normalizada pela constante  $\theta f$ , que garante  $E[(\theta f)^2] = 1$ , restringindo o impacto do ruído ao nível do sinal. [1]

Para a geração do conjunto de dados,  $N$  pontos foram gerados aleatoriamente, selecionando  $x_1, \dots, x_N$  a partir de uma distribuição uniforme no intervalo  $[-1, 1]$ , ou seja, com uma função de densidade de probabilidade  $P(x) = \frac{1}{2}$ . Os ruídos artificiais foram adicionados a cada ponto  $x_n$ , independente dos outros valores, utilizando uma distribuição Normal com média  $\mu = 0$  e variação  $\sigma^2$  apresentada para cada caso do experimento. Portanto, o conjunto de dados é um conjunto  $\mathcal{D} = (x_1, y_1), \dots, (x_N, y_N)$ , onde

$$y_n = f(x_n) + \sigma \epsilon_n \quad (5)$$

e  $\epsilon_n$  são variáveis aleatórias da distribuição Normal descrita.

Para cada experimento com valores específicos para  $N$ ,  $Q_f$  e  $\sigma^2$ , uma função alvo foi gerada aleatoriamente como descrito anteriormente e duas hipóteses finais foram obtidas,  $g_2$  e  $g_{10}$ , respectivamente do conjunto de hipóteses  $\mathcal{H}_2$  e  $\mathcal{H}_{10}$ . Onde  $\mathcal{H}_2$  é o conjunto de um modelo mais simples (polinômios de grau 2) e  $\mathcal{H}_{10}$  um modelo mais complexo (polinômios grau 10). As hipóteses finais  $g_2$  e  $g_{10}$  foram ajustadas ao conjunto de dados de modo a obter o erro mínimo quadrático, resultando em polinômios Legendre com

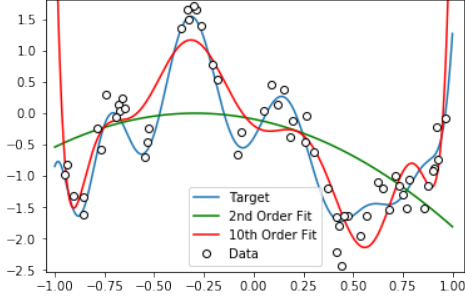


Figura 1. Exemplo de função alvo gerada com complexidade  $Q_f = 15$ ,  $N = 60$  pontos gerados com nível  $\sigma^2 = 0.1$  de ruído e seus respectivos modelos  $g_2$  e  $g_{10}$  ajustados conforme os pontos. Exemplo do experimento de ruído determinístico.

seus respectivos graus. As hipóteses foram ajustadas na forma de polinômios Legendre para minimizar o custo computacional no cálculo dos erros fora-da-amostra comparados à função alvo. A Figura 1 exemplifica uma simulação com sua função alvo e modelos ajustados. Os erros fora-da-amostra  $E_{out}(g_2)$  e  $E_{out}(g_{10})$  foram calculados através do erro quadrático médio de forma analítica, para todos os pontos no intervalo  $[-1, 1]$ , apresentada pela fórmula seguinte

$$E_{out}(g) = \frac{\int_{-1}^1 (f - g)^2 dx}{\int_{-1}^1 1 dx} \quad (6)$$

Utilizando a propriedade de ortogonalidade dos polinômios Legendre, de forma análoga a equação (4), temos que

$$E_{out}(g) = \sum_{q=0}^{Q_f} \frac{(a_q^g - a_q^f)^2}{2q + 1} \quad (7)$$

onde  $a_q^g$  e  $a_q^f$  são respectivamente os coeficientes dos polinômios Legendre  $g$  e  $f$ . A medida de overfitting é obtida com a diferença entre o erro fora-da-amostra do modelo mais complexo e o modelo mais simples  $E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)$ . A média dessas diferenças em um alto número de simulações apresenta uma boa expectativa da esperança do impacto de  $Q_f$ ,  $N$  e  $\sigma^2$  no overfitting.

### III. RESULTADOS

Os seguintes resultados foram obtidos através de experimentos realizados em diversas máquinas com ambiente Anaconda e Python 3.6. As simulações foram divididas em grupos de 1000, armazenando suas respectivas médias para  $E_{out}(\mathcal{H}_{10})$  e  $E_{out}(\mathcal{H}_2)$  em arquivos de texto, de modo que todo o processo pudesse ser interrompido e reiniciado a qualquer momento sem perda crítica dos cálculos. Para obter um resultado final, foi realizado a média dos grupos. Tem-se o comportamento do ruído estocástico gerado a partir do intervalo  $[60, 61, \dots, 130]$  para  $N$  pontos da amostra de dados e o intervalo  $[0.0, 0.05, \dots, 2.5]$  para a variância  $\sigma^2$  do ruído. O valor de  $Q_f$  que diz a complexidade da função alvo mantém-se constante em 20. Foram necessários cerca de 750 grupos de 1000 simulações, totalizando 750000 simulações, para obter o *heat map* exibido na Figura 2. O comportamento do ruído determinístico, como

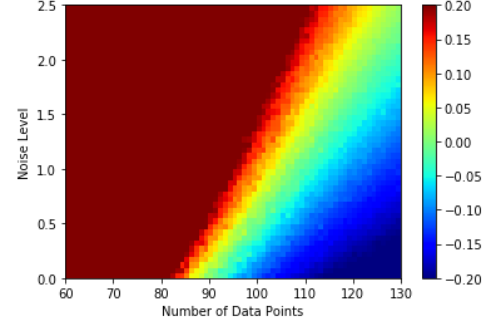


Figura 2. Resultados para o impacto do ruído  $\sigma^2$ , número de pontos  $N$  e complexidade da função alvo  $Q_f = 20$ . O espectro de cores mostra a medida de overfit  $E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)$ . Pelo espectro percebe-se o quanto o overfitting depende do nível de ruído  $\sigma^2$  e do tamanho do conjunto de dados  $N$ . Resultado obtido com cerca de 100 mil simulações.

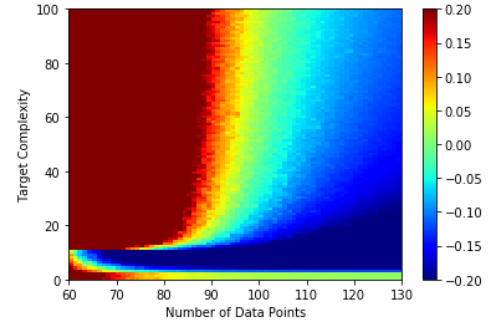


Figura 3. Resultados para o impacto da complexidade da função alvo  $Q_f$ , número de pontos  $N$  e ruído  $\sigma^2 = 0.1$ . O espectro de cores mostra a medida de overfit  $E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)$ . Vemos na figura o quanto o overfitting depende da complexidade  $Q_f$  da função alvo e do tamanho do conjunto de dados  $N$ . Resultado obtido com cerca de 100 mil simulações.

visto na Figura 3, foi gerado para um intervalo  $[60, 61, \dots, 130]$  para  $N$  pontos da amostra de dados e o intervalo  $[0, 1, \dots, 100]$  para a complexidade  $Q$  da função alvo. A variância  $\sigma^2$  do ruído foi mantida constante em 0.1 e o experimento foi realizado com um total de 650000 simulações.

### IV. DISCUSSÃO

Nos resultados foi possível perceber claramente a influência dos parâmetros  $N$ ,  $\sigma^2$  e  $Q_f$  no problema do aprendizado. No ruído estocástico, apresentado na Figura 2, percebeu-se como o esperado que menos overfitting ocorre quando há uma menor influência do nível do ruído  $\sigma^2$  ou uma quantidade maior de dados  $N$  utilizados para o aprendizado. Isso ocorre devido a deformação apresentada aos dados pelo ruído, que muda o “formato” da função alvo enquanto o algoritmo com modelo mais complexo aproveita seus graus de liberdade adicionais para “aprender” o ruído. Tal comportamento resulta num erro dentro-da-amostra  $E_{in}$  menor, porém um maior erro fora-da-amostra  $E_{out}$ , representado pelas regiões vermelhas mais escuras no mapa. O algoritmo de aprendizado apenas tem conhecimento dos dados e não da função alvo, seu aprendizado não deveria tentar se ajustar a curva, entretanto, não é possível distinguir o ruído de um sinal verdadeiro do problema [1].

Na Figura 3, denominado como ruído determinístico, percebe-se que a complexidade  $Q_f$  da função alvo afeta o

comportamento do overfitting de uma maneira semelhante ao ruído  $\sigma^2$ . O overfitting ocorre com um maior impacto a medida que a complexidade da função alvo aumenta, entretanto diminui conforme o número de dados  $N$  disponíveis aumenta. É possível perceber também uma mudança drástica no padrão do comportamento do overfitting após a complexidade  $Q_f$  da função alvo no valor 10, devido a hipótese mais complexa determinada para o experimento ser da ordem 10.

O resultado obtido foi bem semelhante ao obtido pelo experimento realizado pelo Abu-Mostafa, demonstrado em seu livro e sua vídeo-aula [1][2]. Apesar de utilizarmos uma resolução menor e um número menor de simulações, o impacto dos parâmetros ao overfitting ficou claro. Foi possível mostrar que a crença de que melhores resultados são obtidos agregando o máximo possível de informação sobre a função alvo não é verdadeira [1], uma vez que o modelo menos complexo pode ser mais “estável” ao generalizar bem a função alvo enquanto o mais complexo é mais suscetível a ruídos.

#### REFERÊNCIAS

- [1] Yaser S. Abu-Mostafa; Malik Magdon-Ismail and Hsuan-Tien Lin, *Learning from Data*. 2012.
- [2] Yaser S. Abu-Mostafa, *Learning from Data*, Lecture 11 Overfitting. Hameetman Auditorium at Caltech, Pasadena, CA, USA, 2012.