

de 19 a 28
de outubro de 2022

SNCT_{na}
UERJ



Estudo e decomposição de viés e variância em algoritmos de Machine Learning^a

Vitor Bueno Entringe de Souza^b Cristiane Oliveira de Faria^c

Resumo

Nos dias de hoje existe uma facilidade para armazenar e processar dados como nunca houve antes na história. Esse fato abre espaço para novos tipos de análises, fazendo com que surjam oportunidades completamente novas em várias áreas diferentes. Uma destas novas áreas é a Ciência de Dados, que permite adotar estratégias diferentes de entender o mundo através da análise de dados [4]. Ferramentas estatísticas tradicionais se aliaram à computação moderna permitindo o surgimento de modelos robustos e eficazes. Dentre estes métodos, temos o de *Machine Learning* (ou Aprendizagem de Máquina) que encontram soluções para problemas de regressão e de classificação. Neste trabalho serão abordados os fundamentos de alguns algoritmos de Machine Learning como: Árvore de decisão, Árvore de decisão com Bagging [1], Floresta aleatória e Regressão logística [3].

Também é realizado um estudo da decomposição do Erro Quadrático Médio em viés e variância para problemas de regressão e também é apresentado um paralelo para a sua aplicação em problemas de classificação, ambos tem o objetivo de avaliar a eficiência do modelo. Em relação a interpretação dessas grandezas no contexto de *Machine Learning*, de forma simplificada, pode se dizer que a variância é a

^aSessão de painéis de Pós-Graduação, XV Semana do IME.

^bDoutorando no PPG em Ciências Computacionais, IME, UERJ e vitor.bueno@pos.ime.uerj.br

^cIME, UERJ, Rio de Janeiro, RJ e cofaria@ime.uerj.br



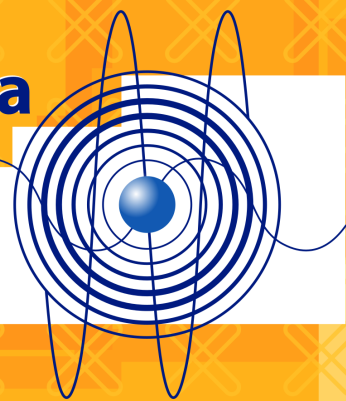
19ª SEMANA
NACIONAL DE
CIÊNCIA E
TECNOLOGIA

BICENTENÁRIO DA
INDEPENDÊNCIA
200 ANOS DE CIÊNCIA, TECNOLOGIA E INOVAÇÃO NO BRASIL



de 19 a 28
de outubro de 2022

SNCT^{na}
UERJ



variabilidade das funções construídas pelo modelo em relação a media de todas as funções, em outras palavras, ela diz respeito a capacidade do modelo de entregar uma função que se adapte ao conjunto de dados usado em sua construção. Enquanto que o Viés pode ser interpretado como a diferença entre o valor da observação e a predição média do modelo, um modelo com viés alto tende a produzir uma função que trata o problema de forma demasiadamente simplificada. Através desse estudo e da interpretação dessas grandezas é possível detectar quando a função gerada pelo modelo pode estar sofrendo de *overfitting* ou o *underfitting*, isto é, de forma simplificada, quando a função gerada pode estar captando variações aleatórias dos dados usados em sua construção ou quando o modelo utilizado não é capaz de captar a complexidade do problema em questão [2].

Como validação da eficiência destes métodos, eles serão aplicados em um conjunto de dados do naufrágio do Titanic, disponibilizado pela plataforma de competição em ciência de dados *Kaggle*, com o objetivo de fazer a classificação dos passageiros em sobreviventes ou não. Para a avaliação dos algoritmos é feita a decomposição de viés e variância em cada modelo construído de modo a analisar as suas características e fazer melhoras nos pontos possíveis, visando aumentar o acerto dos modelos com dados novos. Por fim é feita uma comparação dos resultados obtidos e suas diferenças, onde conclui-se que o modelo que obteve o maior sucesso com dados novos foi o modelo construído pelo algoritmo de Floresta aleatória.

Palavras-chave: Aprendizado de máquina. Algoritmos. Variância. Viés. Decomposição.

Referências

- [1] L. BREIMAN, *Bagging predictors*, Machine learning, 24 (1996), pp. 123–140.
- [2] T. G. DIETTERICH AND E. B. KONG, *Machine Learning bias, statistical bias, and statistical variance of decision tree algorithms*, tech. rep., Citeseer, 1995.
- [3] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The elements of statistical learning. Springer series in statistics*, Springer, 2001.
- [4] L. IGUAL AND S. SEGUÍ, *Introduction to Data Science*, Springer, 2017.

Realização e Apoios



de 19 a 28
de outubro de 2022

SNCT_{na}
UERJ



- [5] G. STRANG, *Linear algebra and learning from data*, Wellesley-Cambridge Press Cambridge, 2019.



Apoio:

Secretaria de
Ciência e Tecnologia

GOVERNO DO
DISTRITO FEDERAL

FNDCT

Finep

CNPq

Realização:

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES



19ª SEMANA
NACIONAL DE
CIÊNCIA E
TECNOLOGIA

BICENTENÁRIO DA
INDEPENDÊNCIA
200 ANOS DE CIÊNCIA, TECNOLOGIA E INOVAÇÃO NO BRASIL