

**Data Science
Academy**

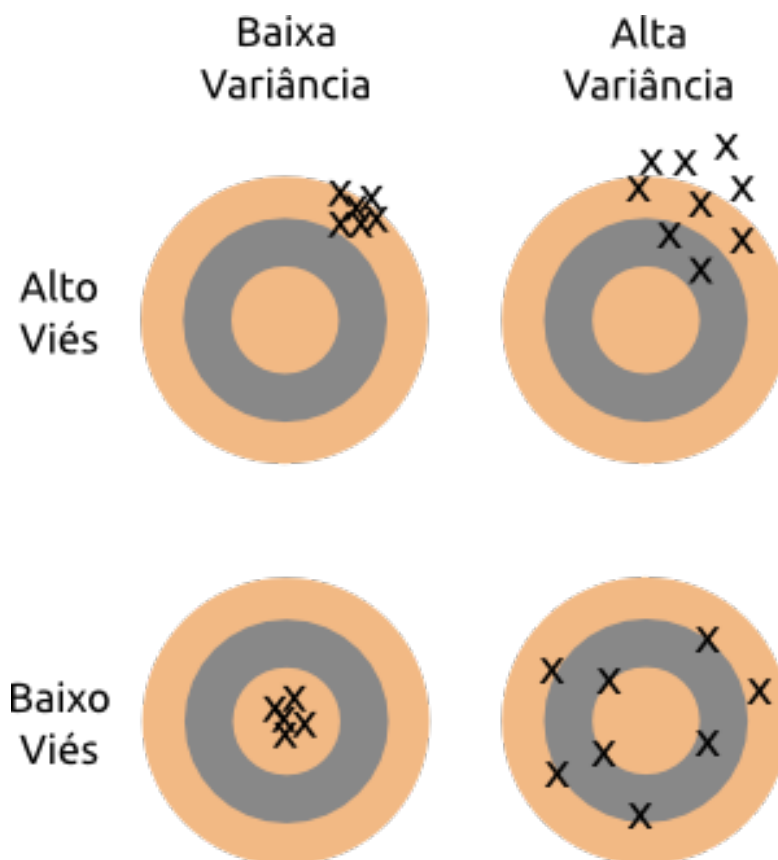
www.datascienceacademy.com.br

Machine Learning

Viés e Variância

Você já deve ter percebido que construir modelos de Machine Learning eficientes não é tarefa simples. Como se não bastassem todas as técnicas de pré-processamento, mais todas as opções de configuração e otimização do algoritmo, temos ainda os erros e problemas gerados durante a fase de aprendizagem e isso sem falar nas questões de armazenamento dos dados. E o viés e variância são mais 2 desses problemas com os quais precisamos lidar.

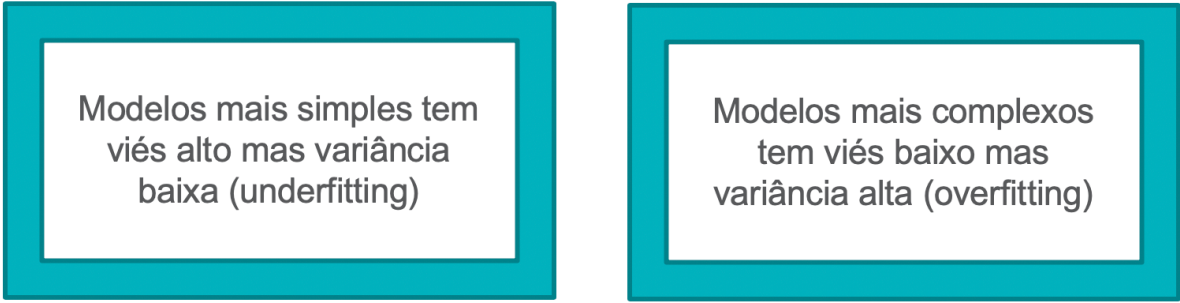
Quando construímos um modelo que generaliza incorretamente particularidades dos dados, temos o problema conhecido como overfitting. Quando um modelo tem 100% de acerto no dataset de treino e 50% no dataset de teste, enquanto ele deveria ter 75% de acerto em ambos os conjuntos, dizemos que este modelo sofre de overfitting. Uma forma ilustrativa de se compreender tal fenômeno, é separando o erro de generalização em viés (bias) e variância (variance).



Viés é a tendência de o modelo aprender consistentemente uma generalização incorreta (por exemplo, pela simplicidade do modelo). O viés é a distância entre a média do conjunto de estimativas e o único parâmetro a ser estimado.

Variância é a tendência de se aprender fatos aleatórios independentemente do sinal real. A variância é simplesmente o valor esperado dos desvios quadrados de amostragem. Ele é usado para indicar quão distante, em média, o conjunto de estimativas está do valor esperado das estimativas.

Um modelo muito complexo, tem alta variância por ser capaz de aprender padrões que possam não ser reais. Isso faz com que, de forma contra intuitiva, os modelos mais complexos não sejam sempre a melhor alternativa e isso traz o desafio adicional de buscar o modelo que não seja tão complexo, mas ainda assim generalizável.



Modelos mais simples tem
viés alto mas variância
baixa (underfitting)

Modelos mais complexos
tem viés baixo mas
variância alta (overfitting)

Normalmente, modelos mais simples tem viés alto mas variância baixa (underfitting), enquanto que modelos mais complexos tem viés baixo mas variância alta (overfitting). Mas o que significa um modelo ser mais complexo? Em geral, a complexidade de um modelo aumenta conforme o número de variáveis preditoras aumenta e conforme a capacidade do modelo de captar relações não lineares e interações entre as variáveis preditoras aumenta.

No caso de inúmeras variáveis preditoras, até mesmo um modelo de regressão linear pode se tornar complexo demais, exigindo técnicas de regularização como LASSO ou Regressão Ridge, para diminuir a possibilidade de overfitting. Como veremos, no próximo capítulo!