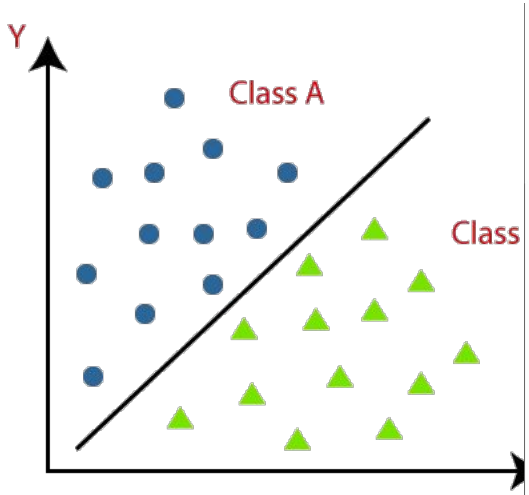
Conceitos Básicos de Machine Learning com Python

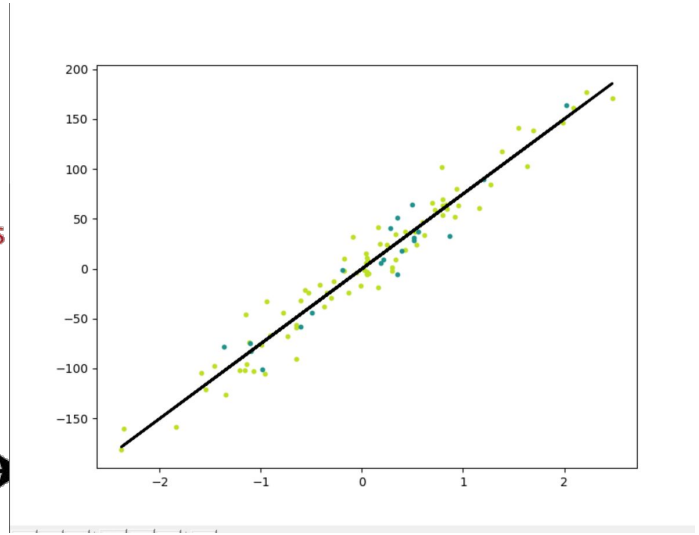
python

PET CC

O que os algoritmos explorados nesse curso propõem?

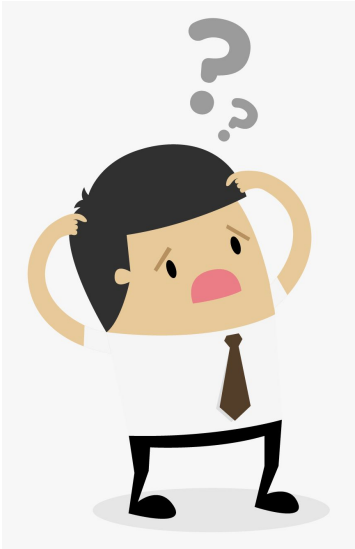
- Classificação
- Regressão





Principais desafios do Machine Learning

- Problemática dos Dados
- Underfitting X Overfitting
- Bias X Variance



Problemática dos Dados

Insuficiência de Dados

Imagine que você quer explicar para uma criança o que é um carro. Você aponta para 2 ou 3 carros e avisa a criança que aquilo é um carro. Simples assim.

Essa facilidade é dada pois o cérebro humano é incrivelmente complexo e consegue aprender facilmente.

Problemática dos Dados

Insuficiência de Dados

Com machine learning não é tão simples assim. Até os problemas mais simples exigem milhares de exemplos para aprendizado. Os mais complexos, então, demandam milhões de exemplos.




Problemática dos Dados

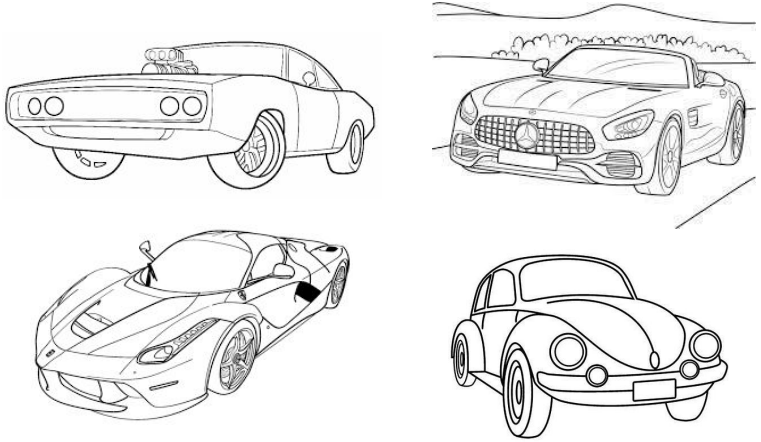
Insuficiência de Dados

- Infelizmente, é notável grande dificuldade em conseguir grandes quantias de dados, fato que dificulta muito sistemas de Machine Learning

Problemática dos Dados

Dados não Representativos

Outro grande problema são os dados não representativos. Quando os dados utilizados para treinar o sistema não representam os dados aos quais queremos generalizar, o algoritmo não irá performar bem.



Problemática dos Dados

Dados sem qualidade

Quando a base de dados está cheia de erros e valores inconsistentes, o sistema encontrará dificuldade em compreender relações entre os dados

$ 5 000

$ 500 000

Problemática dos Dados

Dados Irrelevantes

Dados irrelevantes sobre as instâncias podem levar a interpretações erradas do algoritmo

Ex: Essa casa possui 3 vasos de flor na entrada. Isso claramente não tem relevância alguma para o preço dela, porém o algoritmo pode interpretar essas features erroneamente

Overfitting X Underfitting

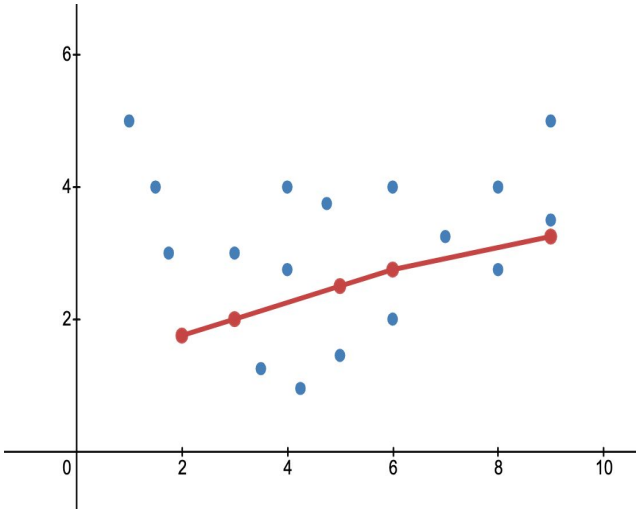
Já abordados os problemas nos dados, vamos aos problemas nos algoritmos. Overfitting e Underfitting são termos diários na vida de alguém que trabalha com machine learning, e as maneiras de evitar esses problemas são pautas até hoje

Underfitting

Ocorre quando o modelo é simples demais para “entender” a estrutura dos dados.

O que fazer?

- Selecionar modelos de ML mais poderosos
- Alimentar o modelo com dados que tem mais qualidade
- Reduzir as regularizações (limites) do modelo (hyperparameters)

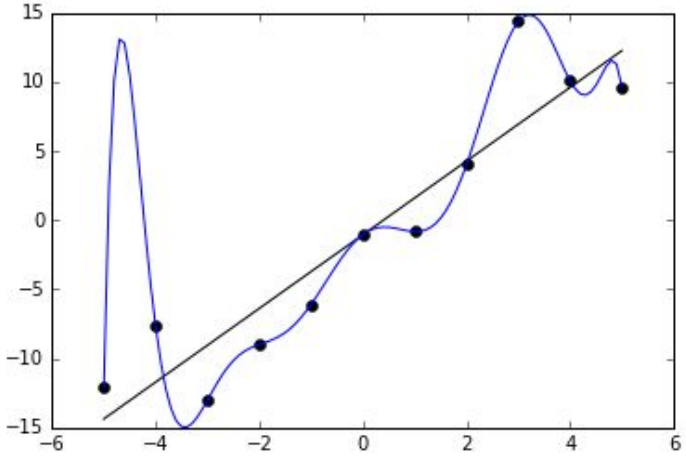


Overfitting

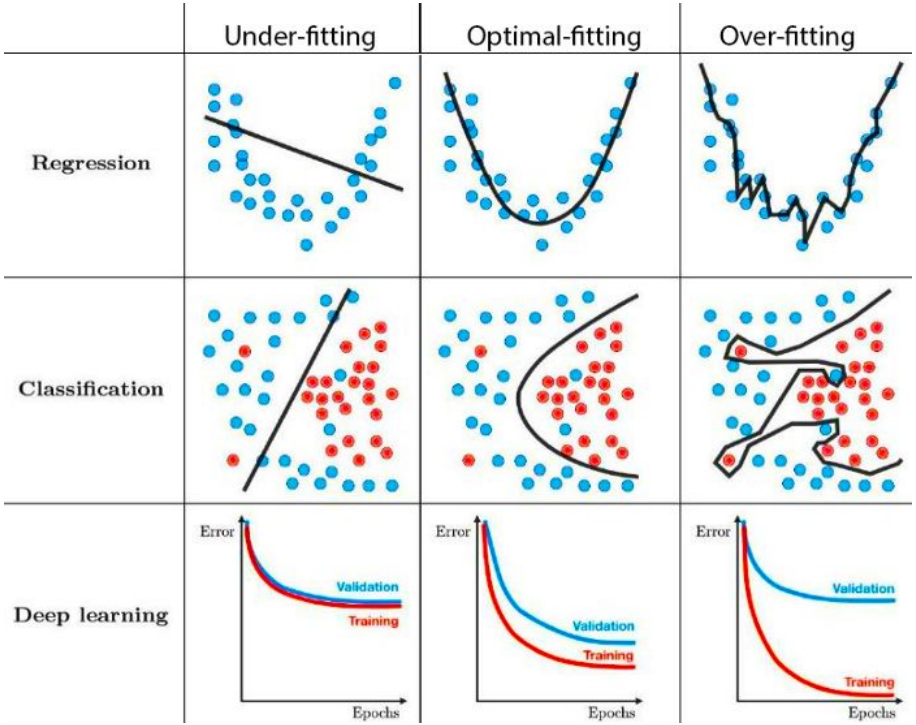
É o contrário de underfitting. O modelo é muito complexo para o problema, então ele se adapta “perfeitamente” aos dados do treino, mas assim que são adicionados novos dados, sua precisão cai, pois o modelo não generaliza.

O que fazer?

- Simplificar modelo
- Reunir mais dados
- Reduzir imprecisões nos dados



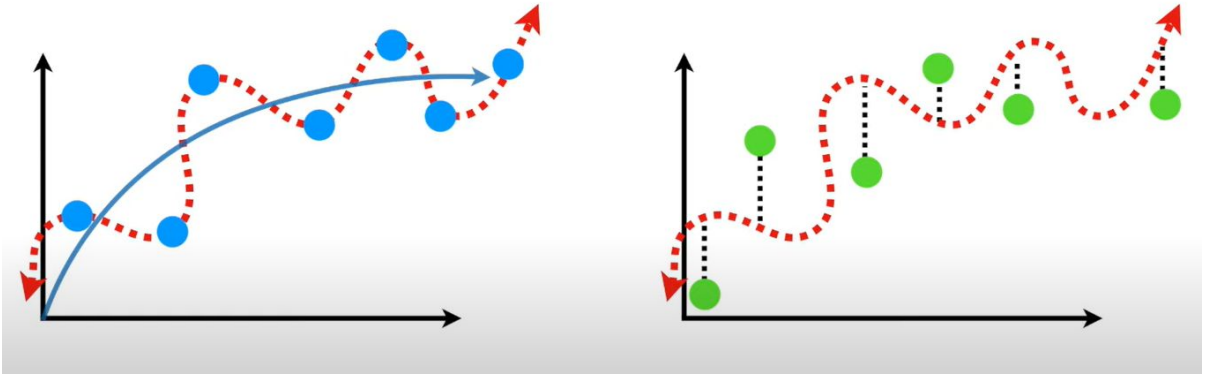
Overfitting e Underfitting em diferentes situações



The Bias and Variance Tradeoff

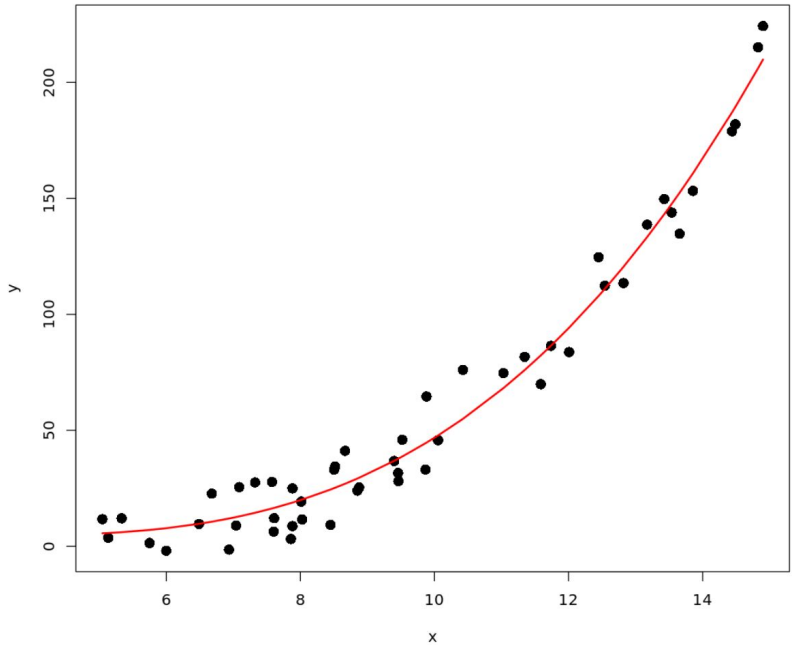
Bias - Incapacidade do algoritmo entender a relação entre as features, ligado a Underfitting

Variance - Diferença entre conjuntos de dados, onde um algoritmo demasiadamente complexo performa muito bem em um conjunto, mas é falho no outro



The Bias and Variance Tradeoff

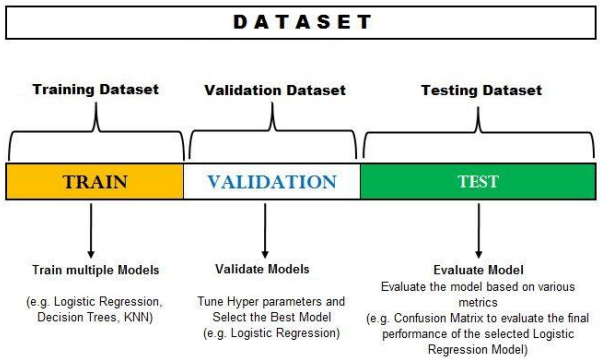
A redução do bias tende a aumentar os problemas com a variância. O grande desafio é achar um ponto de equilíbrio entre os dois



Teste e Validação de Modelos

A única forma de saber se um algoritmo generaliza para novos dados é testando ele em novos dados.

Para isso, sempre que estamos treinando um sistema de ML, dividimos os dados em dois. O *Training set* e o *Testing set* .



Teste e Validação de Modelos

Fazemos o algoritmo aprender no Training set, e para checar a precisão e os erros dele, utilizamos o Testing set.

A divisão dos dados geralmente é feita das seguintes formas:

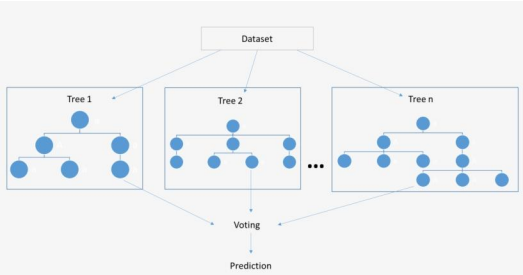
- 80% Training - 20% Testing
- 70% Training - 30% Testing

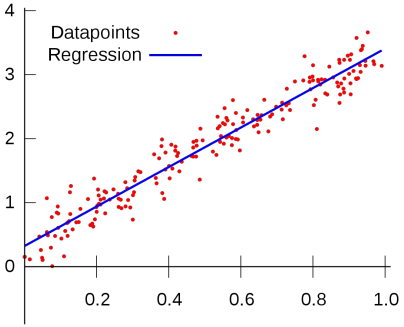
Hyperparameter Tuning

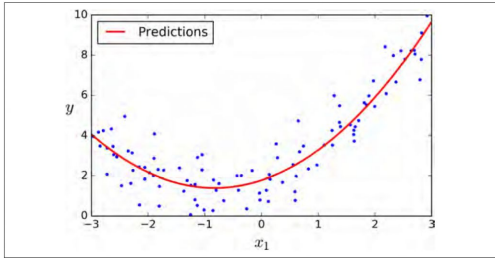
- Hyperparameters são parâmetros que regularizam todos os algoritmos de Machine Learning. Por serem termos mais complexos, que exigem vasto conhecimento sobre os algoritmos em específico, não iremos tratá-los no curso. O importante é saber que existem, e que mínimas mudanças neles podem alterar o sistema inteiro. Eles servem muito para controlar Overfitting e Underfitting e para adicionar tendências no sistema

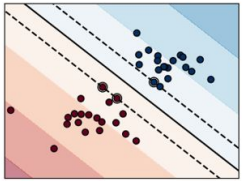
Model Selection

Muitos algoritmos de ML tem as mesmas aplicações, mas enquanto uns performam bem em certos casos, outros podem performar mal. Devemos testar diversos algoritmos em um problema e escolher o melhor, que nem sempre é o mesmo.







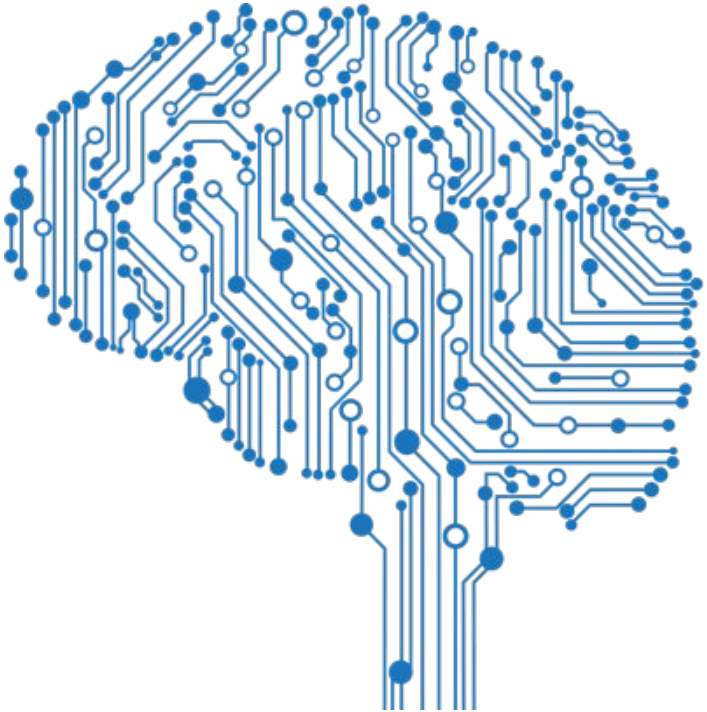


No Free Lunch Theorem

- Vários modelos para um mesmo problema. Temos que testar os que assumimos que irão performar melhor
- Tradeoffs em sistemas ML

Por hoje é isso!

Até amanhã!



Canal do youtube recomendado

StatQuest with Josh Starmer

766 mil inscritos • 218 vídeos

Statistics, Machine Learning and Data Science can sometimes seem like very scary topics, but since each technique is really just ...

INSCRITO

Livros Recomendados

Python Machine Learning: A Practical Beginner’s Guide (Brandon Railey)
Python For Data Analysis (Samuel Burns)
Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow (Aurélien Géron)