Machine Learning no Mundo Imperfeito

Marina Fouto

Mike Townsend
@Mikettownsend

Follow

Machine learning is like highschool sex. Everyone says they do it, nobody really does, and no one knows what it actually is.

2:04 PM - 26 Sep 2016

376 452

O que é Machine Learning?

Algoritmos capazes de “aprender” a partir de dados empíricos e fazer predições sobre estes

Conceitos Básicos I

Classes

Features (Características)

Datasets

Conceitos Básicos II

Aprendizado Supervisionado

- Classificação: divide os dados em classes, precisa de treinamento
- Regressão: infere novos dados a partir de observações passadas

Aprendizado Não-Supervisionado

- Clusterização: agrupamento de dados similares
- Aproximação: “pessoas que compraram X também compraram Y”

Machine Learning… Por Onde Começar?

Qual é o problema que você quer resolver?

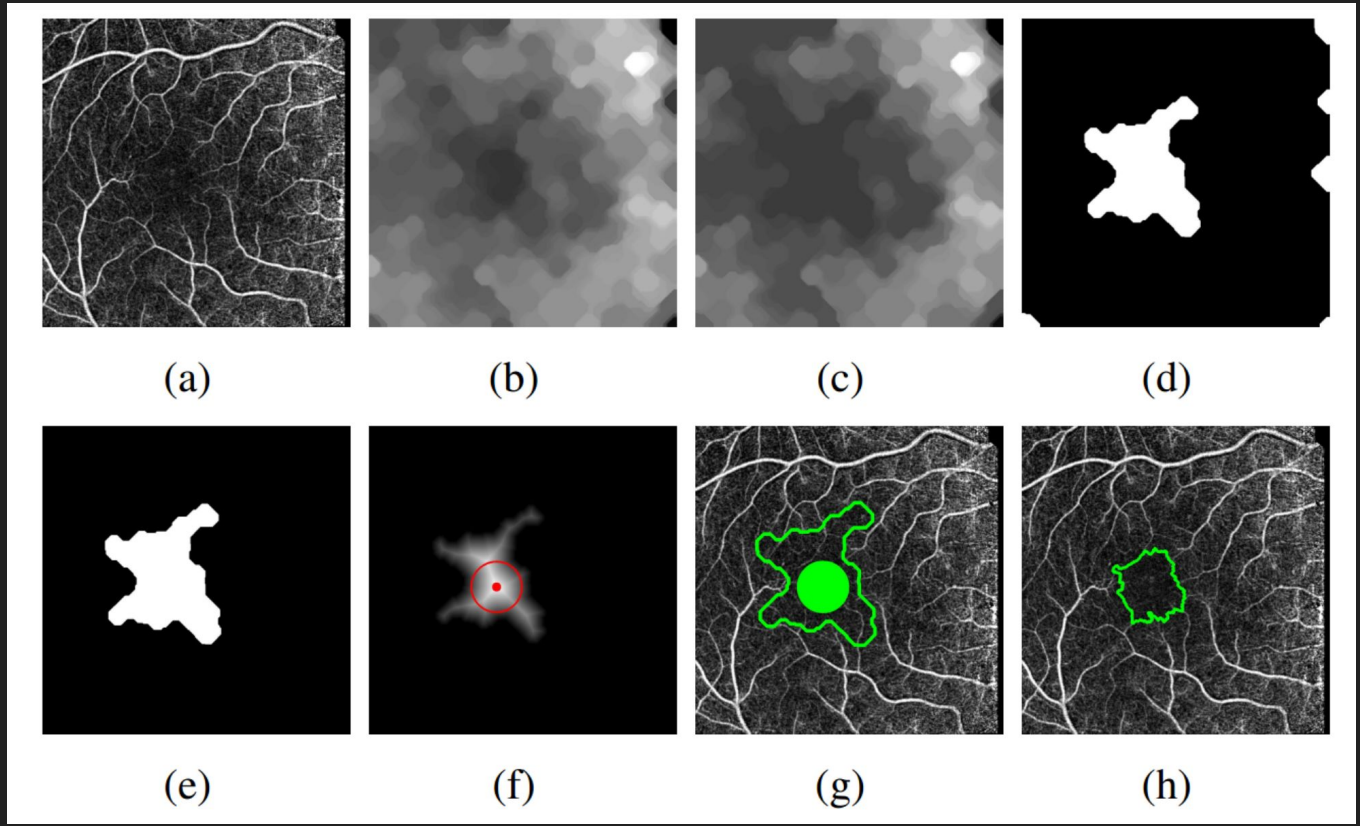
Você sabe o que são os seus dados?

Conseguindo Dados I

Seus dados podem ser:

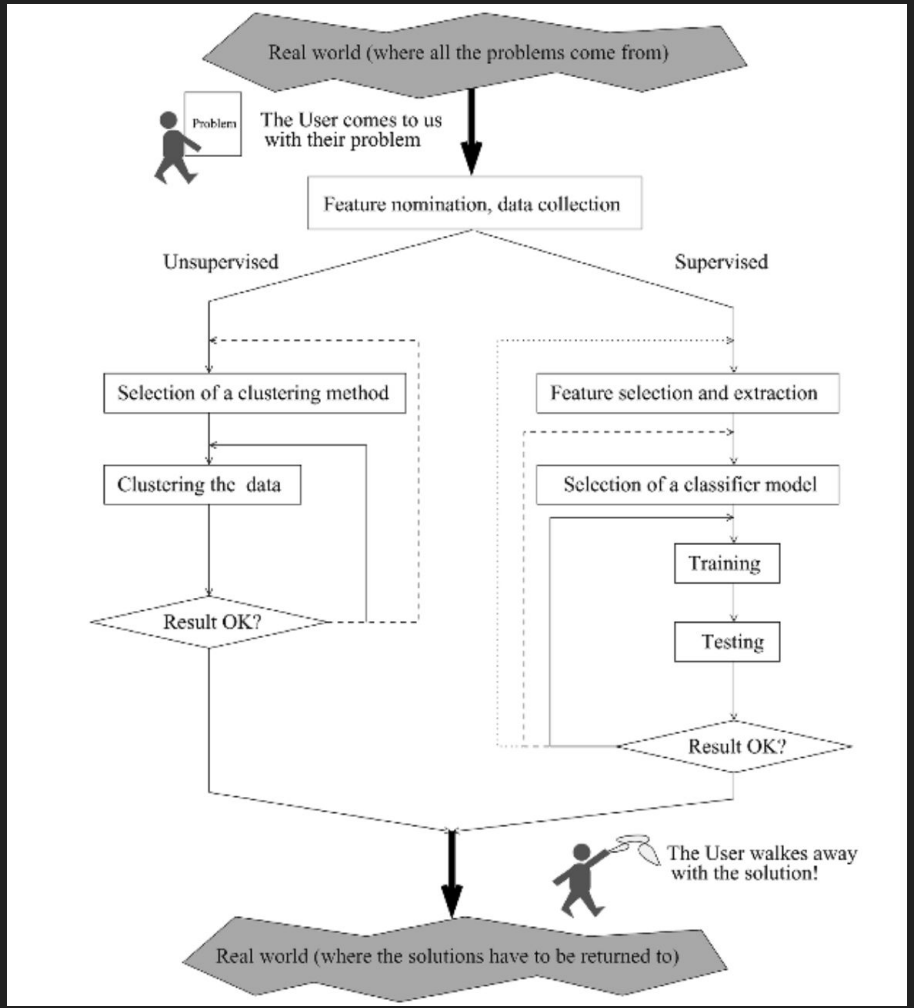
- Organizados e lindos como o Iris dataset
- Cheios de observações faltantes
- Você pode ter que extrair eles de algo

Conseguindo Dados II



19 características extraídas baseadas no formato da região obtida pela segmentação

7 datasets construídos a partir da combinação das características



Kuncheva, 2004

Aprendizado Supervisionado - Classificação

Support Vector Machines (SVM)

K-Nearest Neighbors (KNN)

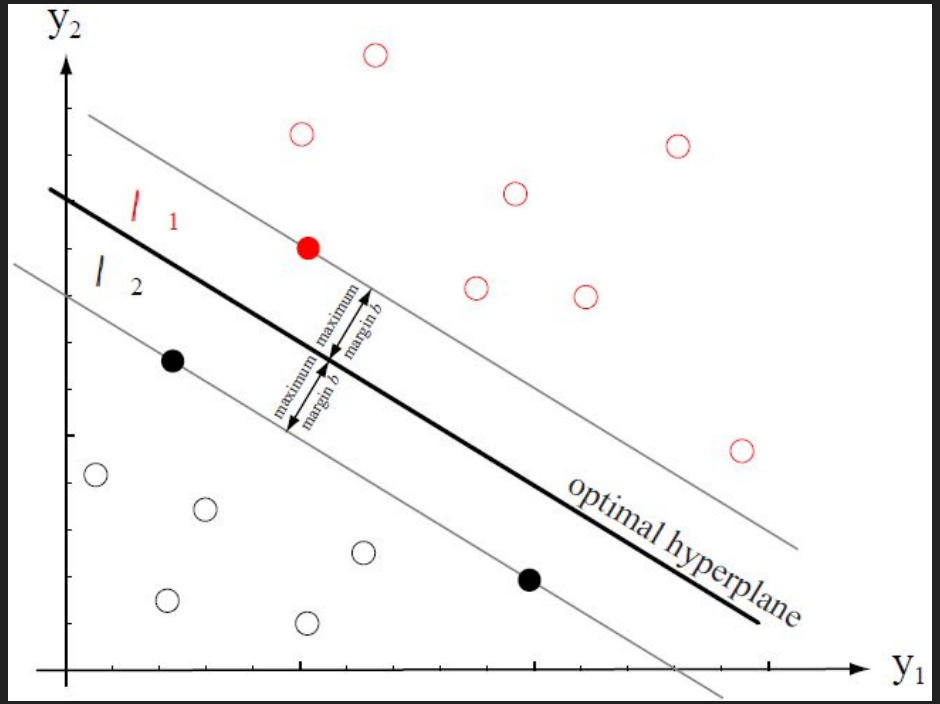
Naive Bayes

Árvores de Decisão

Ensemble Methods

Redes Neurais

Support Vector Machines (SVM)

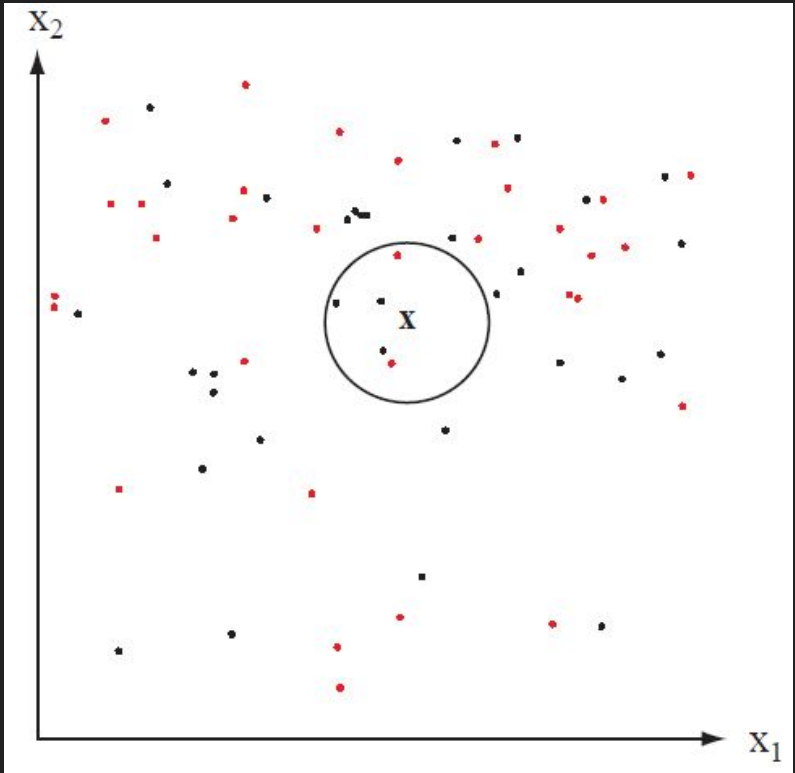


Classificador binário

Versátil, pode utilizar várias funções como separador

Útil em espaços de alta dimensionalidade e eficaz na classificação quando a quantidade de dimensões é maior que a de amostras

K-Nearest Neighbors (KNN)



KNN classifica a amostra a partir de voto majoritário

Escolhe-se vizinhança ímpar para evitar empates

Naive Bayes

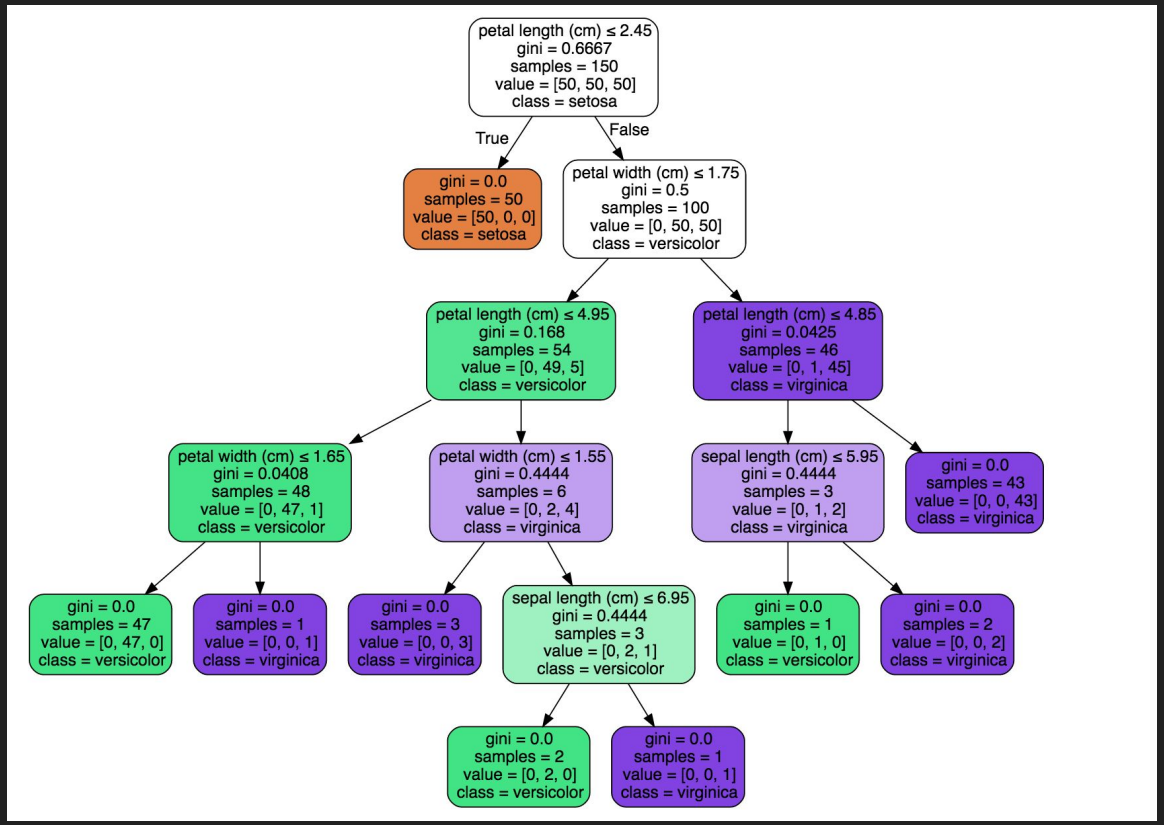
Naive = Ingênuo

Assume que os atributos são condicionalmente independentes em uma mesma classe

Muito utilizado em detecção de spam e é um bom classificador quando existem poucos dados a serem analisados

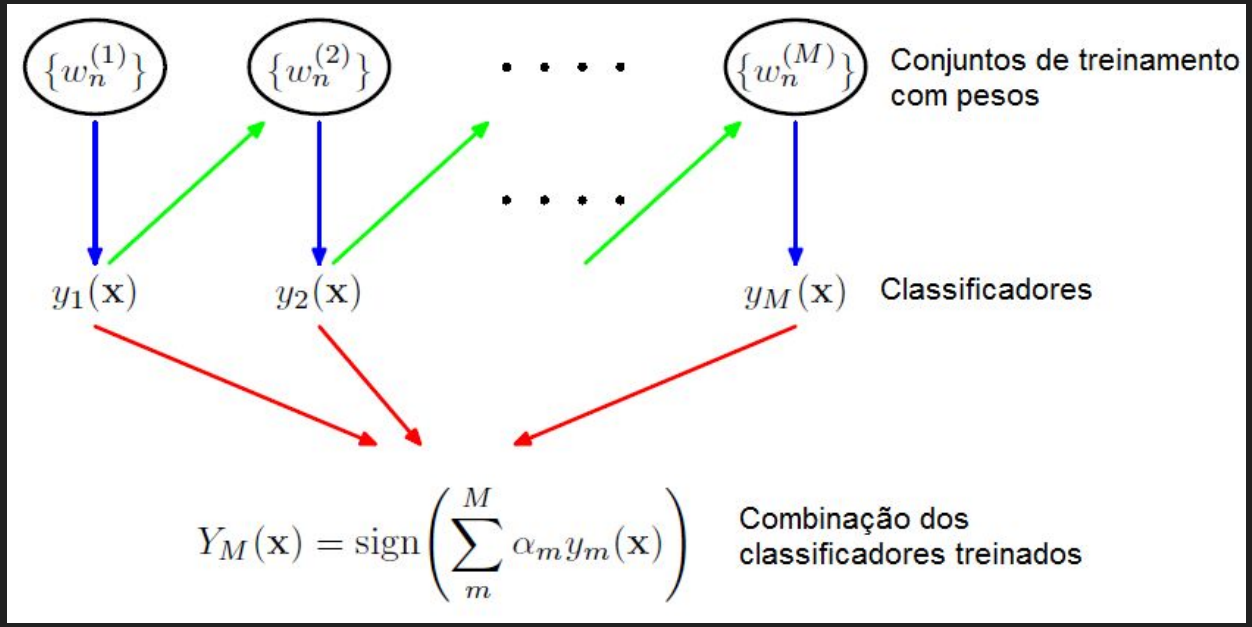
(http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf)

Árvores de Decisão



- Fáceis de entender e visualizar
- Multi-classe
- Podem gerar árvores muito complexas e levar a overfitting
- Pode favorecer uma classe, caso ela seja maioria no dataset

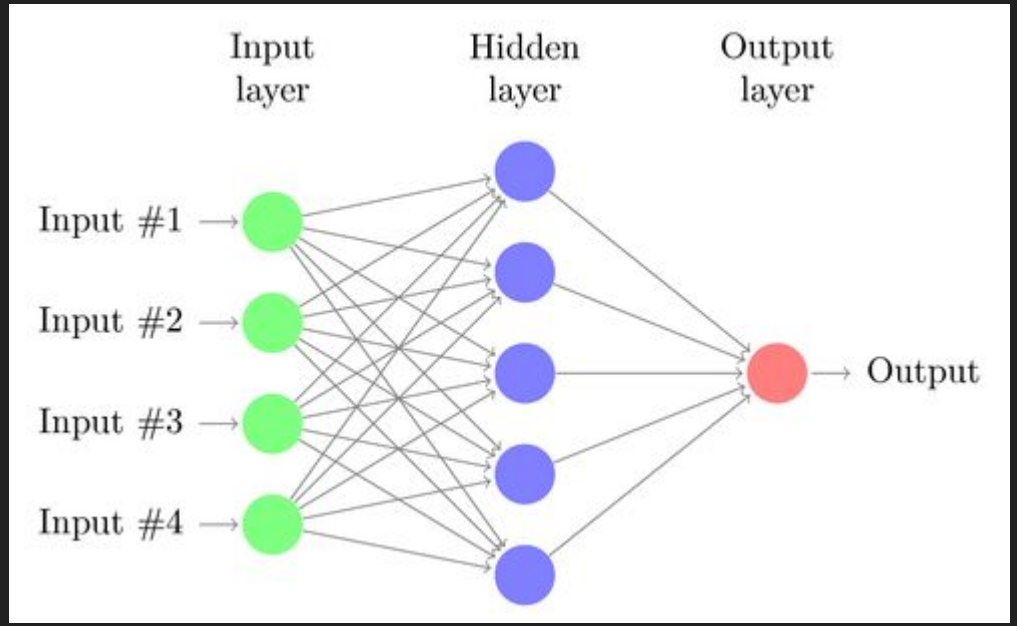
Ensemble Methods



Métodos de média: a média dos classificadores aplicados independentemente sobre os dados é tomada como resultado

Métodos de boosting: um classificador é construído após o outro, baseado no resultado anterior

Redes Neurais



Multi-layer Perceptron (MLP)

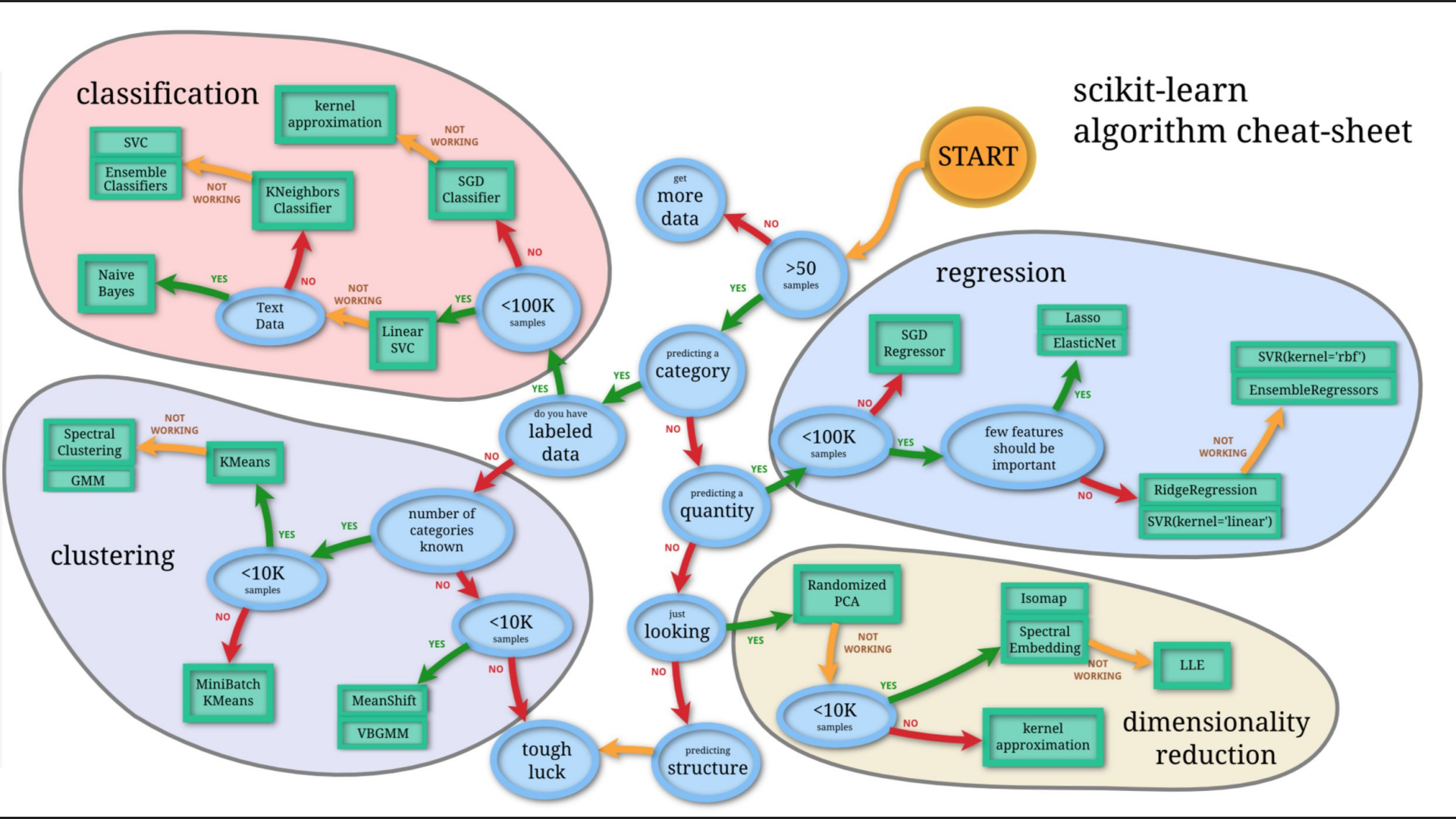
Redes Neurais Convolucionais (CNN)

Redes Neurais Recorrentes (RNN)

Redes Neurais Probabilísticas (PNN)

Aprendizado Não Supervisionado - Clusterização

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NO
Ensemble Classifiers
NOT WORKING
SVC
SGD Classifier
NOT WORKING
kernel approximation
NO
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
tough luck
NO
predicting structure
NO
just looking
NO
predicting a quantity
YES
<100K samples
YES
few features should be important
YES
Lasso
ElasticNet
NO
SGD Regressor
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
clustering

Redução de Dimensionalidade

Maldição da dimensionalidade

Serve para datasets com muitas dimensões, onde é difícil entender e avaliar os dados

Principal Component Analysis (PCA), decompõe um dataset de alta dimensionalidade em um conjunto de componentes ortogonais que explicam a maior quantidade de variância

Dados desbalanceados

Datasets médicos tendem a ter dados desbalanceados

Não confiar somente em % de acerto. 95% de acerto pode não ser útil

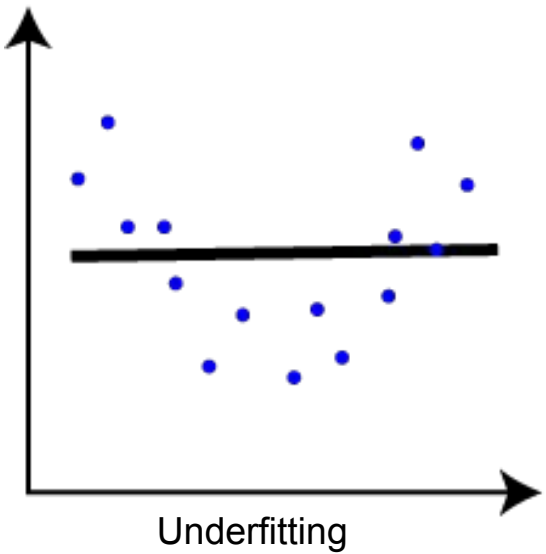
Sensitividade (Verdadeiro Positivo)

Especificidade (Verdadeiro Negativo)

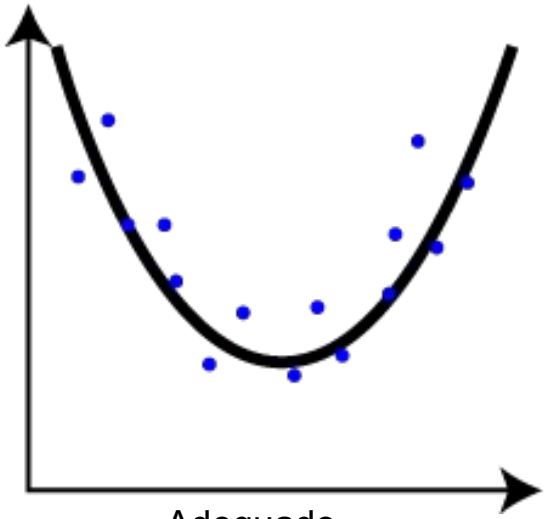
Balancear os dados com oversampling ou undersampling

Overfitting

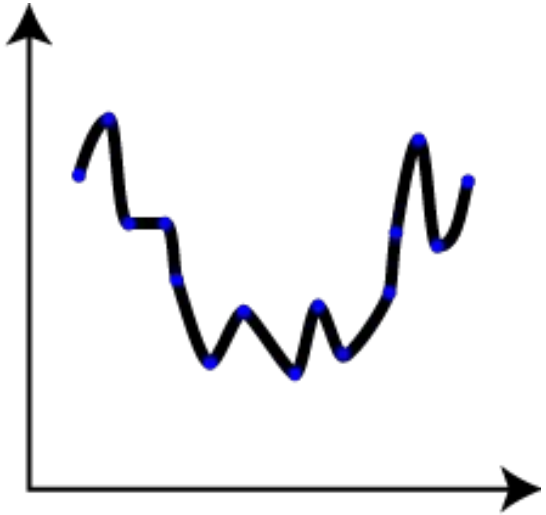
Um modelo flexível demais não generaliza!



Underfitting



Adequado



Overfitting

Cross Validation

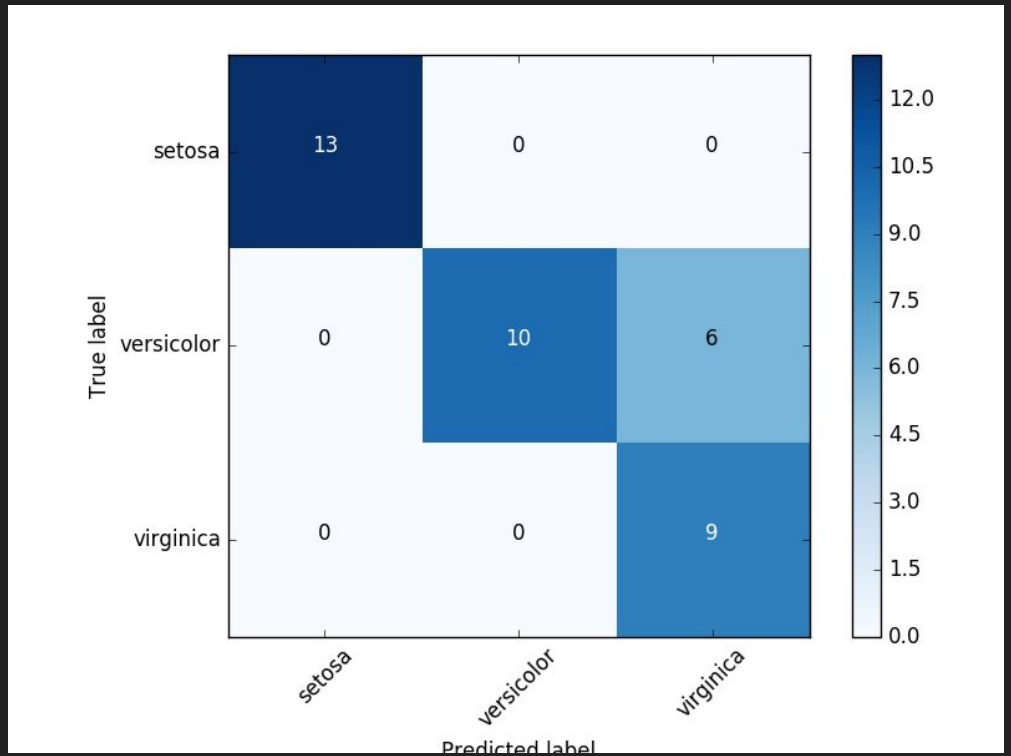
Evita overfitting

Divide o conjunto de treinamento em treinamento e validação

2-fold (50% treinamento, 50% validação)

K-fold (treina com k-1 partes, valida com 1)

Matriz de Confusão



	P' (Predicted)	N' (Predicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

Type I e Type II

Um erro Type I acontece quando um valor negativo é classificado como positivo

Um erro Type II é quando um valor positivo é classificado como negativo

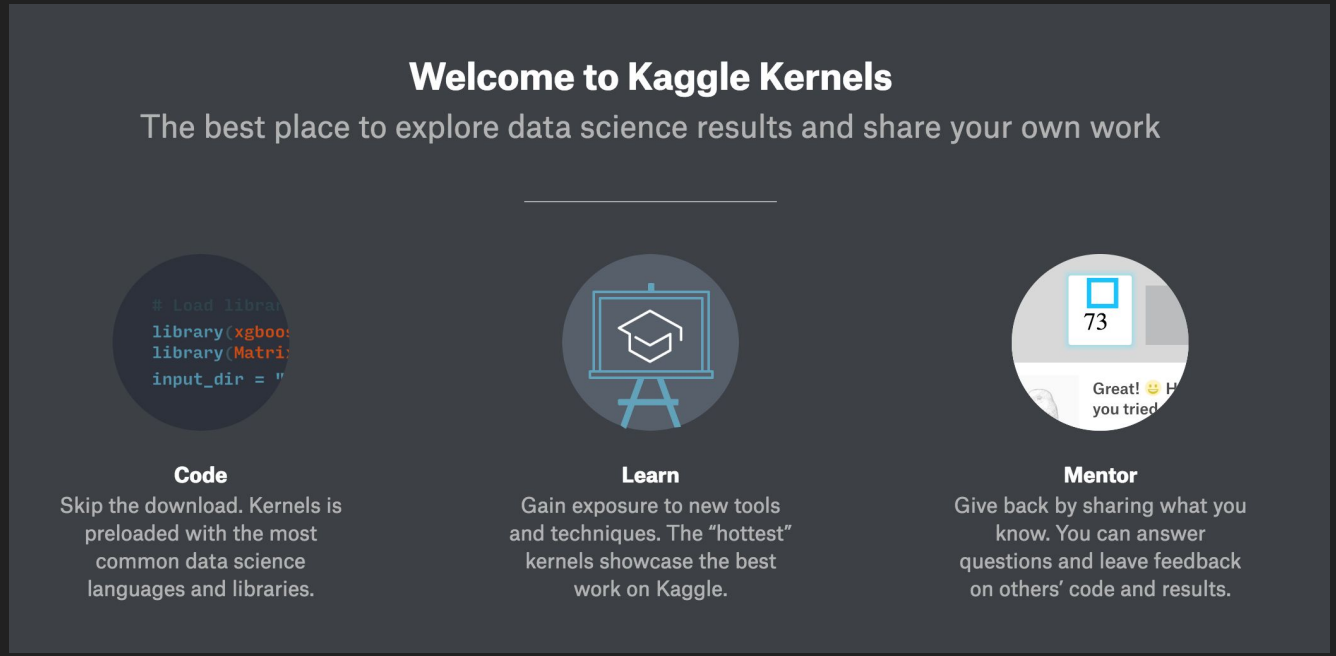
TPOT

https://github.com/rhiever/tpot

Biblioteca escrita em Python para otimizar pipeline de machine learning utilizando algoritmos genéticos

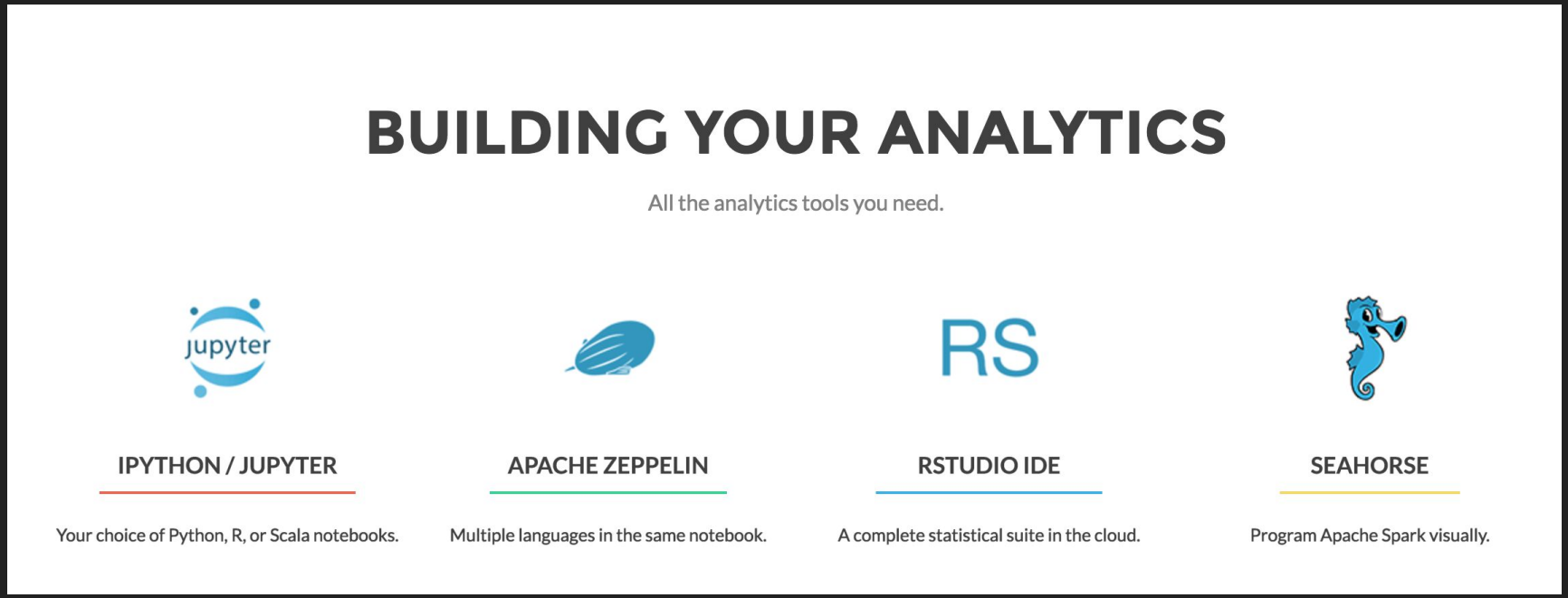
Para Saber Mais

Kaggle! https://www.kaggle.com/kernels



Para Saber Mais

Data Scientist Workbench! https://datascientistworkbench.com/



Para Saber Mais

Introdução a Machine Learning no Big Data University:
https://bigdatauniversity.com/courses/introduction-to-machine-learning/

Nanodegree em Machine Learning no Udacity

https://br.udacity.com/course/machine-learning-engineer-nanodegree--nd009/

edX: https://www.edx.org/course?search_query=machine+learning

Machine Learning no Mundo Imperfeito

Marina Fouto