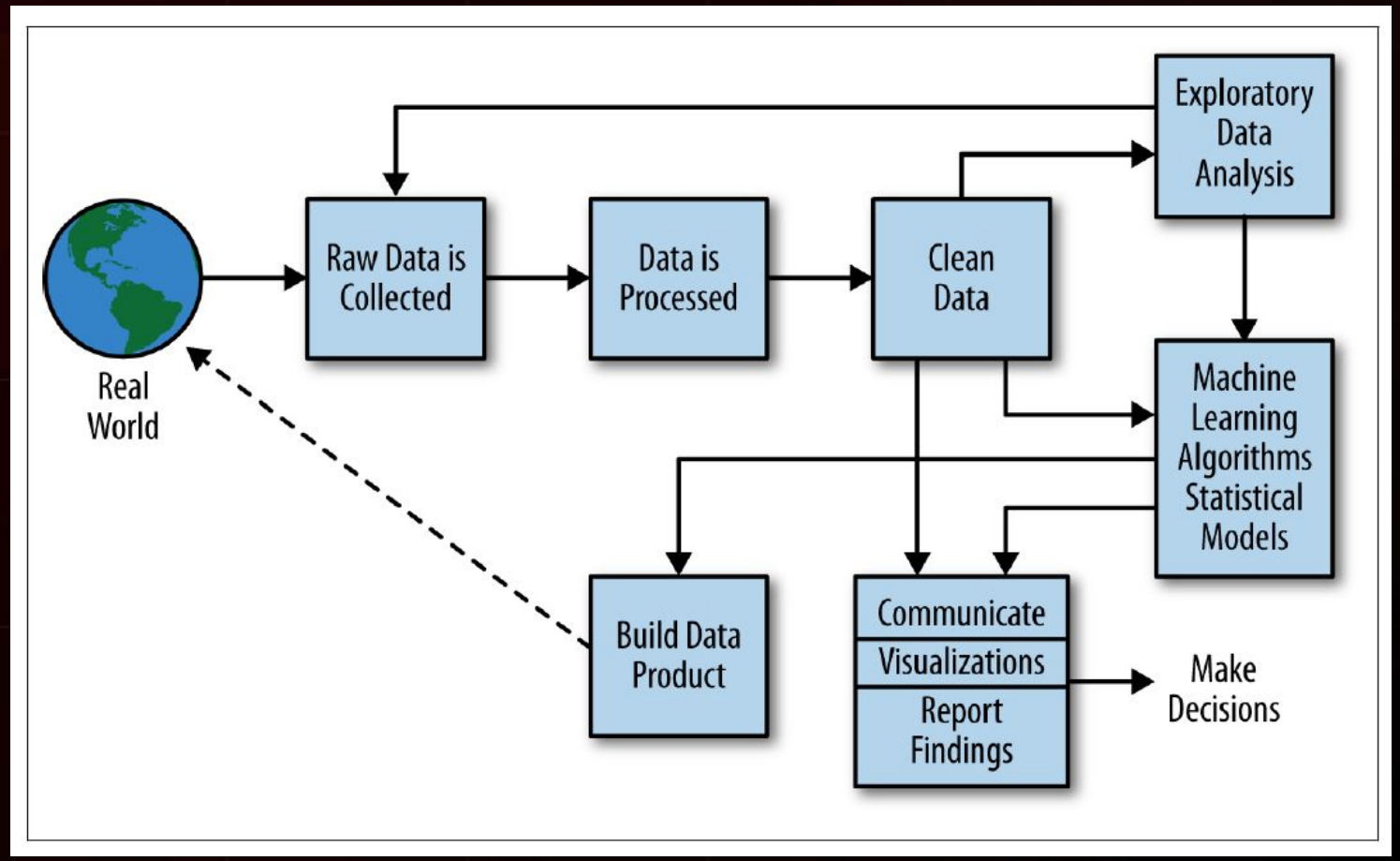
INTRODUÇÃO À CIÊNCIA DE DADOS

Introdução à Machine Learning

MACHINE LEARNING (ML)



Fonte: Shutt, R. and O'Neil, C.; Doing Data Science, 2014

COMPUTAÇÃO TRADICIONAL E ML

- Na computação tradicional são as pessoas (programadores) que criam programas (rotinas) para que os computadores possam executá-las.

- Em ML os computadores criam seus próprios procedimentos a partir da observação sobre dados e seus resultados, ou sobre a estrutura e organização de um conjunto de dados.

MACHINE LEARNING

- Machine Learning (ML) ou Aprendizado de Máquina, é o protagonista do que alguns chamam "Nova Inteligência Artificial".
- O uso de dados como evidência e para tomada de decisão e o contexto de Big Data são os principais impulsionadores do ML.
- ML é essencialmente a capacidade de computadores aprenderem a realizar uma tarefa ao invés de serem programados para tal tarefa.

MACHINE LEARNING

- Na inteligência artificial e em ML, deve ocorrer o desenvolvimento de métricas, a partir das quais as máquinas devem ser capazes de criar hipóteses e, assim, resolvê-las.

- Exemplo: baseado no perfil, na navegação e/ou na compra de um produto em um e-commerce, o sistema é capaz de recomendar um novo produto para o cliente (com máxima capacidade de acerto que ele também vai adquiri-lo).

MACHINE LEARNING

- A esse processo de indução de uma hipótese (ou aproximação de função) a partir da experiência passada, dá-se o nome de ML.

- Os algoritmos de ML aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido.

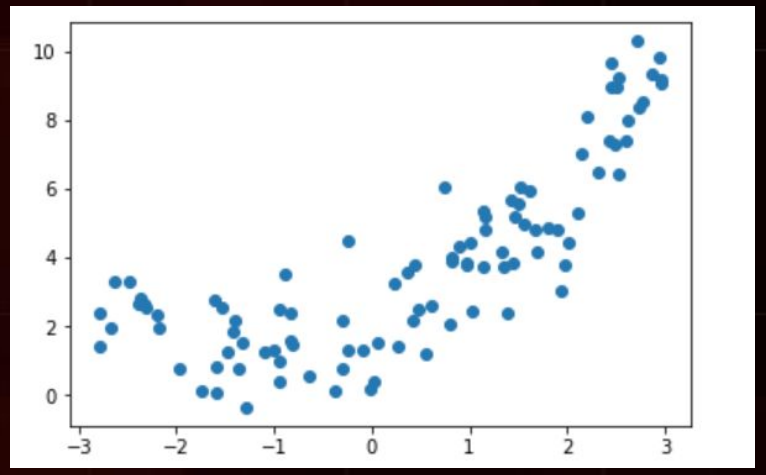
MACHINE LEARNING

- Conforme Amaral (2016), ML é a aplicação de técnicas computacionais com o objetivo de encontrar padrões ocultos em dados. Segundo o autor, esses padrões ocultos são aquelas características que não podem ser observadas tão claramente nos dados.

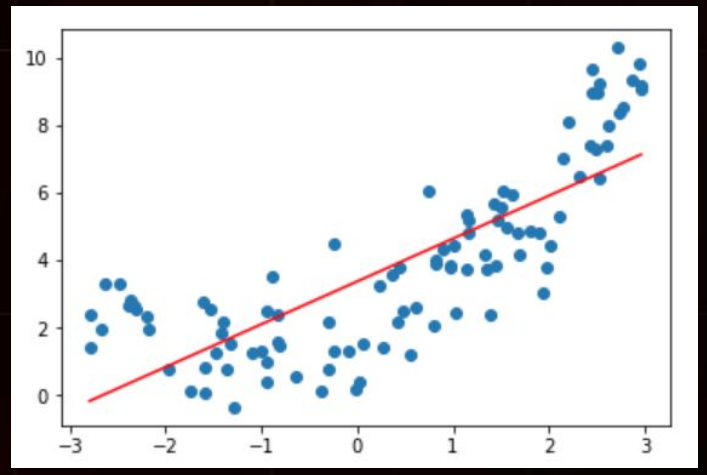
- ML é a capacidade de computadores criarem um modelo (induzirem uma função ou hipótese) que aprenderam com os dados.

MODELO E MODELAGEM

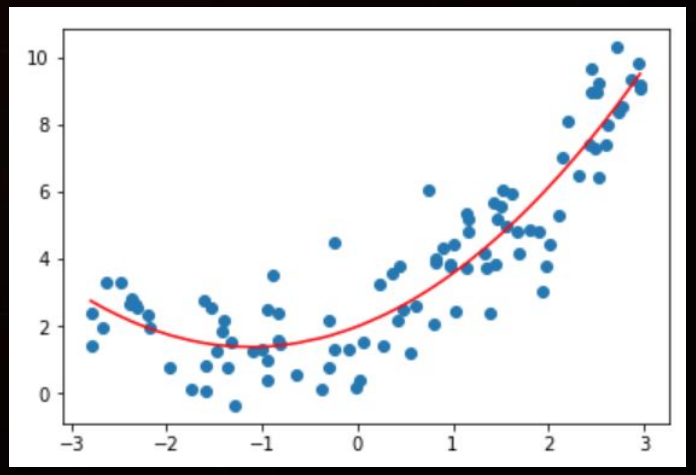
- Um modelo é uma especificação de uma relação matemática (ou probabilística) entre diferentes variáveis.



Dados



Modelo 1



Modelo 2

ML - TIPOS

- Existem alguns tipos de aprendizado de máquina, dentre os quais podemos citar:
  - Supervisionado: normalmente preditivo, traz um objetivo estabelecido e pode ser dividido entre problemas de regressão e de classificação.
  - Não supervisionado: normalmente descritivo, quando o objetivo não está bem definido e temos o intuito de compreender melhor os dados para realizar o agrupamento.
  - Por reforço: quando as saídas não estão bem definidas e as respostas só podem ser aferidas após algumas execuções.

HIERARQUIA ML

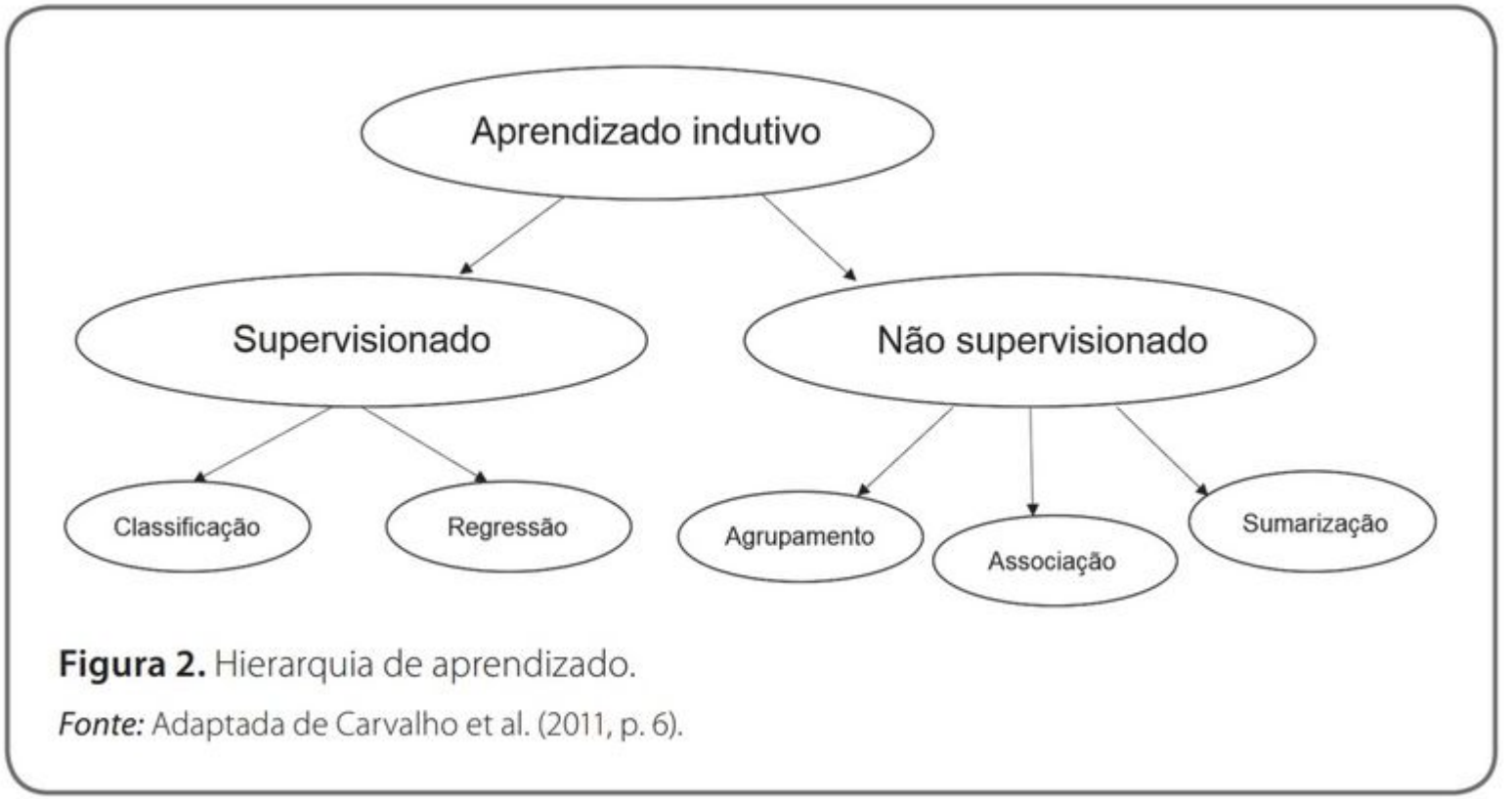


Figura 2. Hierarquia de aprendizado.

Fonte: Adaptada de Carvalho et al. (2011, p. 6).

ML - SUPERVISIONADO (PREDITIVO)

- No ML supervisionado, os métodos recebem como entrada dados rotulados e usam esses dados e seus atributos para determinar um novo conjunto de dados desconhecidos.

- Os algoritmos do ML supervisionado passam por uma etapa denominada treinamento, na qual o classificador vai aprender um determinado padrão de acordo com os dados utilizados para treinar o sistema.

ML - SUPERVISIONADO (PREDITIVO)

- O ML supervisionado resolve problemas de:
  - Regressão: mapeiam um exemplo em um valor real. Um exemplo de regressão é prever o tempo de internação de um paciente em um hospital.
  - Classificação: associa a descrição de um objeto a uma classe. Um exemplo de classificação é determinar a doença de um paciente pelos seus sintomas.

ML - NÃO SUPERVISIONADO (DESCRITIVO)

- No ML não supervisionado ou descritivo, agrupam-se objetos de acordo com suas características criando associações entre eles. O ML não supervisionado resolve problemas de:
  - Agrupamento (clustering): os dados são agrupados de acordo com sua similaridade.
  - Sumarização: busca encontrar uma descrição simples e compacta para um conjunto de dados.
  - Associação: consiste em encontrar padrões frequentes de associações entre os atributos de um conjunto de dados.

ML - TIPOS E ALGORITMOS

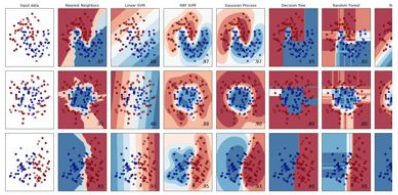
- Dentro de cada tipo (regressão, classificação, agrupamento) temos os mais diversos algoritmos, com suas diversas características.
- Na terceira semana do curso nós vimos a biblioteca Scikit-Learn, que disponibiliza os mais variados algoritmos, já prontos, para cada modelo que seja necessário. Com certeza ela vai ser muito importante em seus projetos!
- Não é tarefa do cientista de dados desenvolver novos algoritmos, mas sim saber aplicá-los.
- *claro que há quem queira desenvolver novos algoritmos!

SCIKIT-LEARN

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

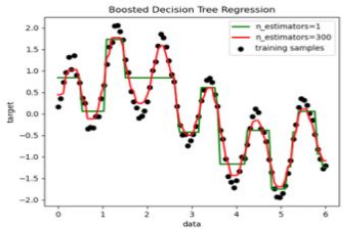


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

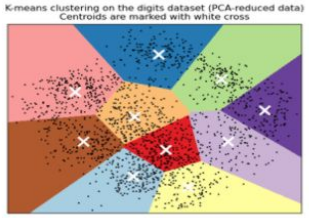


Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

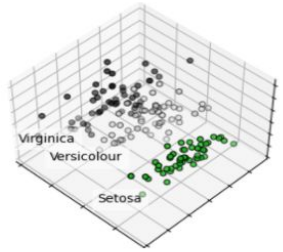


Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...

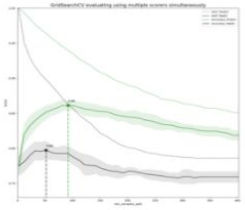


Examples

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, metrics, and more...

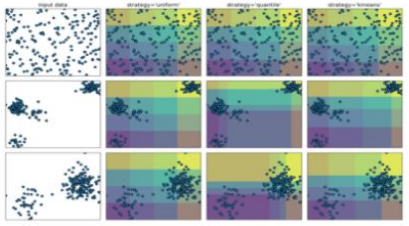


Examples

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...

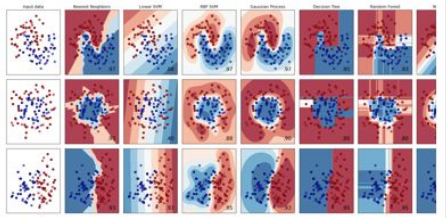


Examples

Classification

Identifying which category an object belongs to.

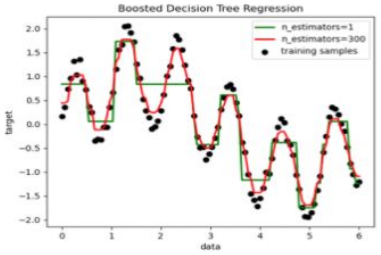
Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...



Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...



SCIKIT-LEARN - REGRESSION

1. Supervised learning

1.1. Linear Models

- 1.1.1. Ordinary Least Squares
- 1.1.2. Ridge regression and classification
- 1.1.3. Lasso
- 1.1.4. Multi-task Lasso
- 1.1.5. Elastic-Net
- 1.1.6. Multi-task Elastic-Net
- 1.1.7. Least Angle Regression
- 1.1.8. LARS Lasso
- 1.1.9. Orthogonal Matching Pursuit (OMP)
- 1.1.10. Bayesian Regression
- 1.1.11. Logistic regression
- 1.1.12. Generalized Linear Regression
- 1.1.13. Stochastic Gradient Descent - SGD
- 1.1.14. Perceptron
- 1.1.15. Passive Aggressive Algorithms
- 1.1.16. Robustness regression: outliers and modeling errors
- 1.1.17. Polynomial regression: extending linear models with basis functions

1.2. Linear and Quadratic Discriminant Analysis

- 1.2.1. Dimensionality reduction using Linear Discriminant Analysis
- 1.2.2. Mathematical formulation of the LDA and QDA classifiers
- 1.2.3. Mathematical formulation of LDA dimensionality reduction
- 1.2.4. Shrinkage and Covariance Estimator
- 1.2.5. Estimation algorithms

1.3. Kernel ridge regression

1.4. Support Vector Machines

- 1.4.1. Classification
- 1.4.2. Regression
- 1.4.3. Density estimation, novelty detection
- 1.4.4. Complexity
- 1.4.5. Tips on Practical Use
- 1.4.6. Kernel functions
- 1.4.7. Mathematical formulation
- 1.4.8. Implementation details

1.5. Stochastic Gradient Descent

- 1.5.1. Classification
- 1.5.2. Regression
- 1.5.3. Stochastic Gradient Descent for sparse data
- 1.5.4. Complexity
- 1.5.5. Stopping criterion
- 1.5.6. Tips on Practical Use
- 1.5.7. Mathematical formulation
- 1.5.8. Implementation details

1.6. Nearest Neighbors

- 1.6.1. Unsupervised Nearest Neighbors
- 1.6.2. Nearest Neighbors Classification
- 1.6.3. Nearest Neighbors Regression
- 1.6.4. Nearest Neighbor Algorithms
- 1.6.5. Nearest Centroid Classifier
- 1.6.6. Nearest Neighbors Transformer
- 1.6.7. Neighborhood Components Analysis

1.7. Gaussian Processes

- 1.7.1. Gaussian Process Regression (GPR)
- 1.7.2. GPR examples

ESCOLHER ALGORITMO

- Não é uma tarefa fácil escolher o algoritmo adequado, inclusive há estudos que colocam algoritmos para testar os melhores algoritmos para uma determinada situação!

TREINO E TESTE

- Depois dos dados já terem passado pelo pré-processamento, inicia-se uma divisão dos dados para que o algoritmo possa ser treinado.
- Geralmente divide-se o conjunto de dados em torno de 8/2, 7/3 entre treino e teste, alguns profissionais também gostam de usar um pouco dos dados para validação.
- Os dados de treino são usados para criar o modelo (treinar) e os dados de teste para verificar se o modelo realmente atende às expectativas, esperando que sim!!
  - Essa frase resume nossa disciplina por completo!

FINALIZANDO

- Iniciamos a compreensão de que ML é um algoritmo que busca gerar um modelo (induzir uma função ou hipótese) a partir de dados.
- Há ML supervisionado, quando os dados têm resultados e também ML não supervisionado, para o qual é necessário descrever os dados.
- A biblioteca scikit-learn será nossa aliada em nossos processos.

REFERÊNCIAS

- Conteúdo dessa aula é baseado nos livros:

CARVALHO, André Carlos Ponce de Leon Ferreira et al. *Inteligência Artificial* - Uma Abordagem de Aprendizado de Máquina. Disponível em: Minha Biblioteca, (2nd edição). Grupo GEN, 2021.

Grus, Joel. Data Science do Zero. Disponível em: Minha Biblioteca, (2nd edição). Editora Alta Books, 2021.

INTRODUÇÃO À CIÊNCIA DE DADOS

Introdução à Machine Learning