Fundamentos de Data Science

Prof Alberto Paccanaro
Lic. Aldo Galeano, Lic. Santiago Noto (monitores)

Sobre mim

- Background

  MSc Computer Science (Un. Milan) 

  PhD in Machine Learning (Un. Toronto) 

  Computational Biology/Bioinformatics (Yale Un.) 

- Pesquisa

  www.paccanarolab.org

- Disciplinas em FGV

  Técnicas e Algoritmos em Ciência de Dados (UG)

  Fundamentos de Ciência de Dados (MSc)

  Ciência de Redes (PhD)

Sobre o curso

- **Aulas:**
  - Terças-feiras 14:20-16:00
  - Quintas-feiras 14:20-16:00

  https://fgv-br.zoom.us/j/96697557455?pwd=ZEtBaHgyWE14OGJKcXRDQitWQ0F0dz09

- **Horário de atendimento:**
  - Terças-feiras, das 17h às 18h
  - Quintas-feiras das 17h às 18h

  Entrar em contato comigo por e-mail, e irei fornecer um link de zoom para a reunião.

- **Site do curso**: on eClass

  Aqui você encontrará os slides das aulas, tarefas, soluções e qualquer outra informação sobre o curso

Objetivos da disciplina

Fornecer uma visão geral das principais **ideias de aprendizado de máquina**.

Fornecer uma **compreensão aprofundada de algoritmos importantes** na aprendizagem supervisionada (e possivelmente alguns não supervisionados)

Ementa

- Introduction to Data Science and AI; a taxonomy of AI and Machine Learning;
- key concepts: classification, regression, training, regularization, curse of dimensionality; Bayes theorem; decision theory;
- parametric and nonparametric density estimation; K-nearest neighbour; kernel density estimation; Naïve Bayes; the EM algorithm;
- Regression by linear combination of basis functions; maximum likelihood and least squares; regularized least square: weight decay, the LASSO method, elastic net;
- Linear Models for Classification; Fisher's linear discriminant; the perceptron; probabilistic generative models for classification; logistic regression; iterative reweighted least squares;
- Neural networks for classification and regression; stochastic gradient descent; the backpropagation algorithm;
- Decision trees for classification and regression: learning, pruning, Gini index, entropy; bias and variance;
- the Bootstrap; Bagging; Random Forests; Boosting for regression; Adaboost;
- Unsupervised learning: general concepts; Principal Component Analysis; K-means clustering; Mixture of Gaussians; the ClusterONE algorithm;
- Embedding techniques: multidimensional scaling, locally linear embedding, T-SNE.

Objetivos da disciplina

- **Fundamentos matemáticos dos algoritmos** – queremos entender por que eles funcionam e como

- Para entendê-los, vamos implementá-los!

- Vamos aplicá-los aos conjuntos de dados do mundo real.

Procedimentos de ensino

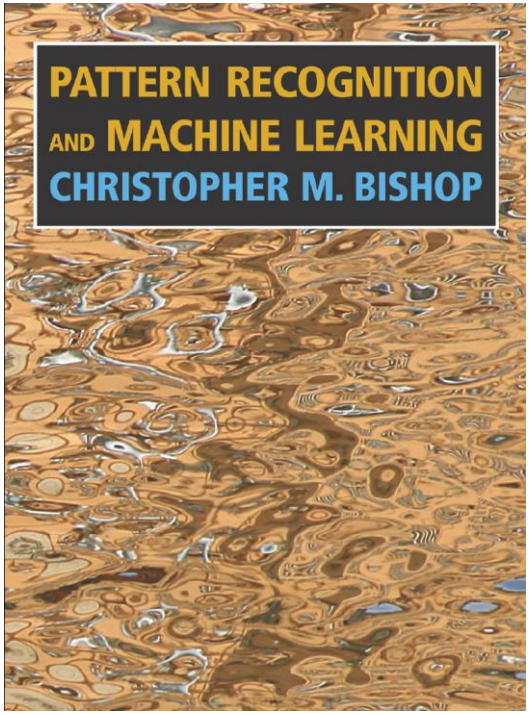
- A cada semana
  - **primeira aula**: algoritmos + fundamentos teóricos e matemáticos.
  - **segunda aula**: laboratório para programar esses algoritmos e aplicá-los (Matlab)

- Todo o material e a gravação de cada aula --> na eClass.

Procedimentos de avaliação

	ENTREGA A ALUNOS	PRAZO	FEEDBACK FORNECIDO EM	VALOR
trabalho individual 1	13/7	20/7	27/7	**30%**
trabalho individual 2	17/8	24/8	31/8	**30%**
PROVA	3/09			**40%**

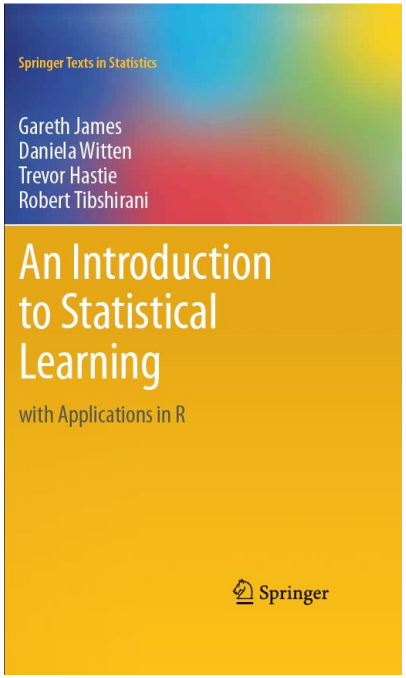
Bibliografia Obrigatória



Christopher Bishop
Pattern Recognition and Machine Learning
Springer, 2006

Disponível em: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

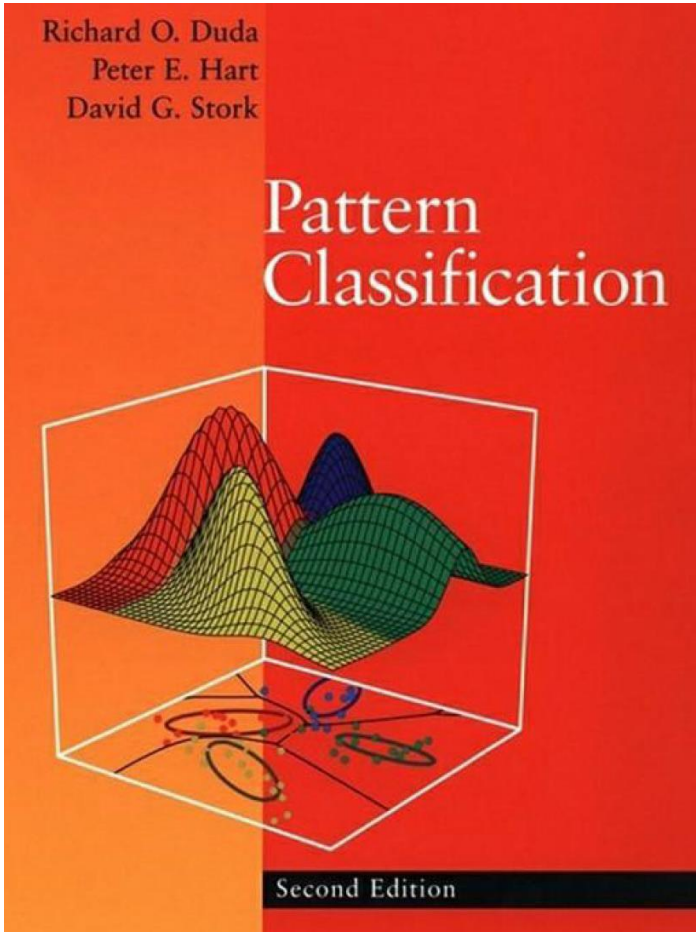
Bibliografia Obrigatoria



James, G., Witten, D., Hastie, T., Tibshirani, R
An Introduction to Statistical Learning
Springer, 2013

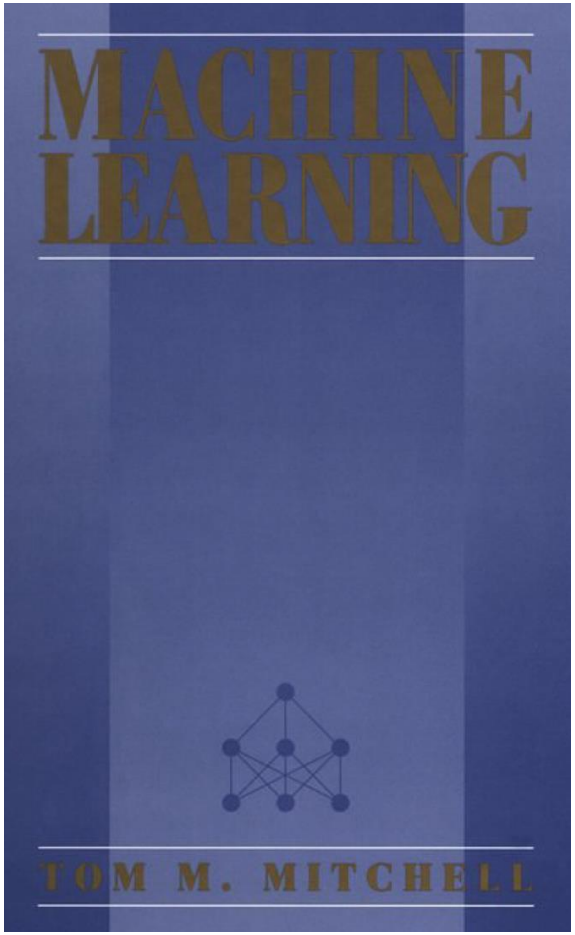
Disponível em: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Bibliografia Complementar



DHS 2006

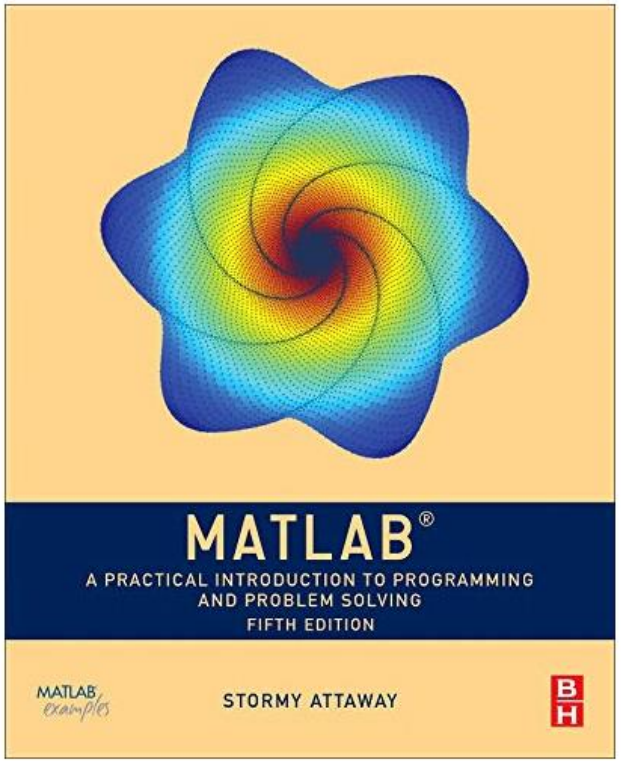
Excellent reference book



1996

A classic, very "CS" textbook

Bibliografia Complementar



Stormy Attaway
MATLAB: A Practical Introduction to Programming and Problem Solving
5th Edition, Butterworth-Heinemann, 2018

- **Trevor Hastie, Robert Tibshirani, Jerome Friedman** *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York: Springer, 2009
- **Jeff Phillips**, Mathematical Foundations for Data Analysis, Springer, 2021

Conselhos sobre o curso

Don’t catch yourself behind

PROGRAM !!! PROGRAM !!!

PROGRAM !!!

- **COMO EU ENSINO**
  - Eu sigo o livro
  - Eu sigo os slides

- **REGRAS DA CLASSE**
  - 5 minutos de pausa após 40 minutos
  - Pergunte!
  - Não chegue tarde!
  - **LIGUE O VÍDEO, POR FAVOR ☺**

IMPORTANT

In the next few classes, on Thursday, I will teach you the basics of Matlab, in case you don't know it.

Make sure you have a version of Matlab that you can access. WE START THIS THURSDAY.