
ESCOLA DE MATEMÁTICA APLICADA

DISCIPLINA: Machine Learning 1 (**Aprendizagem por Máquinas 1**)

PROFESSOR: Alberto Paccanaro

CARGA HORÁRIA: 60h

PERÍODO: 04/01/24 a 01/03/24

HORÁRIO: Terças, quartas quintas e sextas-feiras 14h20min às 16h20min.

PLANO DE ENSINO

1. Ementa

1. Introduction: AI, machine learning, data science
 - a. Logic based AI
 - b. Machine learning based AI
 - c. Data Science
 - d. taxonomy of Machine Learning techniques
2. Key concepts
 - a. classification problems
 - b. regression problems
 - c. training and regularization
 - d. the curse of dimensionality
 - e. Bayes theorem
 - f. decision theory
3. Assessing generalization performance
 - a. confusion matrices
 - b. precision recall trade-off
 - c. F1 measure, ROC analysis
 - d. train, validation and test sets
 - e. crossvalidation
4. Parametric and nonparametric density estimation
 - a. the mixture of gaussians model
 - b. the EM algorithm
 - c. K-nearest neighbour
 - d. kernel density estimation
 - e. Naïve Bayes
5. Linear Regression
 - a. regression by linear combination of basis functions
 - b. maximum likelihood and least squares
 - c. the Moore-Penrose pseudo-inverse
 - d. stochastic gradient descent
 - e. regularized least square: weight decay, the LASSO method, elastic net
6. Linear Classification
 - a. The geometry of linear discriminants

- b. Fisher's linear discriminant
 - c. the perceptron learning algorithm
 - d. probabilistic generative models for classification
 - e. logistic regression
 - f. iterative reweighted least squares
7. Neural networks
- a. combining basis functions into networks
 - b. error functions and activation functions
 - c. neural networks for classification
 - d. neural networks for regression
 - e. stochastic gradient descent
 - f. the backpropagation algorithm
 - g. regularization in neural networks
8. Decision trees
- a. Non metric methods
 - b. Learning decision trees for regression
 - c. Cost complexity pruning
 - d. Learning decision trees for classification
 - e. Gini index, entropy; information gain, Gini gain
9. Ensemble Methods
- a. Bootstrap
 - b. Bagging
 - c. Random Forests
 - d. Boosting for regression
 - e. Adaboost
10. Unsupervised learning
- a. general concepts
 - b. Principal Component Analysis
 - c. K-means clustering
 - d. mixture of Gaussians
 - e. the ClusterONE algorithm
 - f. embedding techniques: multidimensional scaling, locally linear embedding, T-SNE.

Durante o curso, o aluno usará o Matlab para entender os algoritmos, implementando-os e aplicando-os a conjuntos de dados do mundo real. As horas de laboratório nas primeiras duas semanas do curso serão dedicadas a um minicurso que apresentará o Matlab como uma linguagem para prototipagem rápida de algoritmos de Machine Learning.

2. Objetivos da disciplina

Esta disciplina tem como objetivo geral fornecer uma visão geral de algumas das principais ideias em Machine Learning, juntamente com técnicas importantes na aprendizagem supervisionada e não supervisionada. Ele irá abranger os fundamentos matemáticos dos algoritmos, bem como sua implementação prática e aplicação a conjuntos de dados do mundo real.

3. Procedimentos de ensino (metodologia)

O curso tem quatro aulas por semana.

A cada semana, durante a primeira e terceira aula, irei abordar novos algoritmos e seus fundamentos teóricos e matemáticos. Fornecerei aos alunos os slides que usarei em sala de aula e os indicarei os capítulos de livros relevantes e o material onde o conteúdo pode ser encontrado.

A segunda e quarta aula de cada semana serão aulas de laboratório, onde os alunos aprenderão a programar esses algoritmos e aplicá-los a conjuntos de dados do mundo real. Aqui, os alunos receberão dados, código de exemplo no Matlab e problemas para eles resolverem. O objetivo é consolidar os conhecimentos apresentados na aula teórica anterior.

As horas de laboratório nas primeiras duas semanas do curso serão dedicadas a um minicurso que apresentará o Matlab como uma linguagem para prototipagem rápida de algoritmos de Machine Learning.

Todo o material mencionado acima serão disponibilizados na plataforma eClass.

4. Procedimentos de avaliação

O conhecimento dos fundamentos teóricos e matemáticos dos algoritmos de aprendizagem de máquina será avaliado através de uma prova que será realizada ao final do curso.

A capacidade dos alunos de implementar e aplicar os algoritmos em problemas do mundo real será avaliada por meio de dois trabalhos individuais que exigirão que os alunos enviem código Matlab que implementa todo o fluxo de trabalho de aprendizado de máquina (carregamento de dados, pré-processamento, treinamento, avaliação, etc.). Cada trabalho individual valerá 30% da nota final. Estas são as datas principais para os trabalhos individuais e a prova:

	Entrega a alunos	Prazo	Feedback fornecido em	Valor
trabalho individual 1	19/1	26/1	2/2	30%
trabalho individual 2	9/2	16/2	23/2	30%
prova	<i>Depois do carnaval</i>			40%

6. Bibliografia Obrigatória

Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani An Introduction to Statistical Learning, Springer, 2013

Bibliografia Complementar

Richard Duda, Peter Hart, David Stork. Pattern classification, Wiley, New York, 2 edition, 2001.

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

7. Minicurrículo do Professor

Alberto Paccanaro www.paccanarolab.org - Sou Professor Titular da Escola de Matemática Aplicada (EMAp) da FGV do Rio de Janeiro, onde ingressei em 2020. Obtive meu doutorado em Ciência da Computação em 2002 pela Universidade de Toronto, com especialização em Aprendizado de Máquina sob orientação de Geoffrey Hinton. Entre 2002 e 2006, fiz pós-doutorado em Biologia Computacional, primeiro no laboratório de Mansoor Saqi na Queen Mary University of London, e depois no laboratório de Mark Gerstein na Yale University. Tornei-me PI em 2006, obtendo o cargo de professor na Royal Holloway University of London. Em 2014 tornei-me Professor Titular de Aprendizagem de Máquina e Biologia Computacional e Diretor do Centro de Sistemas e Biologia Sintética, da mesma Universidade. Sou professor visitante da Universidade Católica de Assunção (Paraguai), onde conduz um posto avançado de meu laboratório. Também fui professor / membro visitante em Cornell, Yale e na Universidade de Veneza. Sou responsável por várias colaborações internacionais na área de Aprendizagem de Máquina aplicada à Biologia e Medicina. Eu co-dirijo bolsas de pesquisa junto com acadêmicos da Yale University, Cornell University, University of Tennessee e da Catholic University of Asuncion. Vários de meus algoritmos de aprendizado de máquina foram publicados em revistas como Nature, Nature Methods, Nature Communications, Cell, PNAS.

8. Link para o Currículo Lattes

Aberto Paccanaro - <http://lattes.cnpq.br/9819989502690120>