

Introduction

CSC 461: Machine Learning

Fall 2021

Prof. Marco Alvarez
University of Rhode Island

Course Logistics

Welcome

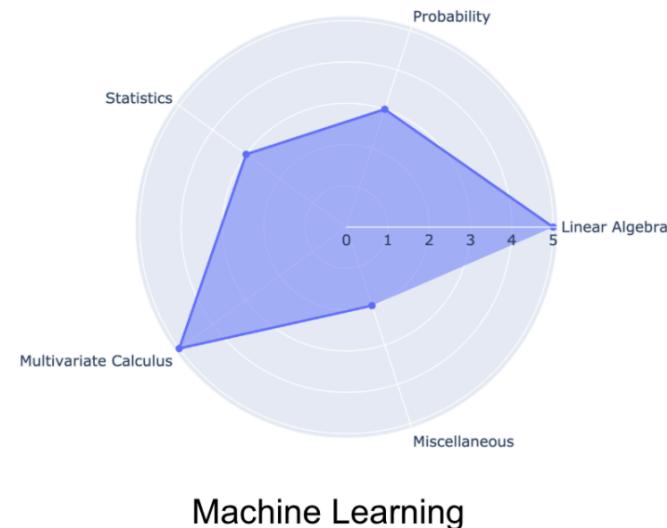
- Instructor
 - ✓ Prof. Marco Alvarez
- TAs
 - ✓ Mikel Gjergji
 - ✓ Derek Jacobs
- Lectures
 - ✓ MW 3 - 4:15p
- Office Hours
 - ✓ TBA
- Course Website
 - ✓ <https://homepage.cs.uri.edu/~malvarez/teaching/csc-461/>

Course goals

- Understand how ML algorithms work
 - ✓ the **learning** problem and limitations
 - ✓ theoretical foundations of major techniques
- Be able to develop ML applications
 - ✓ problem design, algorithm/platform choice
- Be able to read ML papers

Should I take this class?

- Requires more math than traditional CS courses
- Programming experience is required
 - ✓ consider taking this course at a later time if necessary
- Little emphasis on ‘how to use this library’
 - ✓ focus on implementing major algorithms
- High grades require high effort
 - ✓ long and challenging assignments



<https://www.analyticsvidhya.com/blog/2019/10/mathematics-behind-machine-learning/>

Math resources

- Linear Algebra Review and Reference
 - ✓ <http://cs229.stanford.edu/summer2019/cs229-linalg.pdf>
- Review of Probability Theory
 - ✓ <http://cs229.stanford.edu/summer2019/cs229-prob.pdf>
- Computational Linear Algebra for Coders
 - ✓ <https://github.com/fastai/numerical-linear-algebra>
- Mathematics for Machine Learning
 - ✓ <https://gwthomas.github.io/docs/math4ml.pdf>

Python/Numpy resources

- Google's Python Class
 - ✓ <https://developers.google.com/edu/python/>
- From Python to Numpy
 - ✓ <https://www.labri.fr/perso/nrougier/from-python-to-numpy/>
- Python Numpy Tutorial
 - ✓ <http://cs231n.github.io/python-numpy-tutorial/>

Tentative topics (order not relevant)

Introduction, Supervised Learning	Bagging, Boosting
Linear Regression, Linear Classifiers, Loss Functions	Maximum Likelihood Estimation, Naive Bayes
Logistic Regression, Gradient Descent	k-Means, Hierarchical Clustering
Generalization, Bias/Variance, Overfitting/Underfitting, Model Selection	PCA, SVD, Matrix Factorization
Linear SVM, Empirical Risk Minimization, Regularization, Kernels	Reinforcement Learning
K-NN, Curse of Dimensionality	Neural Networks
Decision Trees, Regression Trees	Convolutional Networks

Free textbooks



More textbooks



Grading

- ▶ Assignments (35%)
 - ✓ ~6 Homework Assignments
- ▶ Technical Presentation (25%)
 - ✓ groups of 2
- ▶ Final Project
 - ✓ progress report (10%)
 - ✓ final report (15%)
 - ✓ presentation (15%)

Final Project

- Find some new data/problem or find new approach to old data/problem
 - ✓ 2-3 students per group
- Deliverables
 - ✓ progress report (mid October)
 - ✓ final report (end of semester)
 - ✓ presentation (end of semester)

Logistics



Your job

- Attend lectures (**synchronous**)
- Participate
 - ✓ in-class and also on Ed
- Work hard
 - ✓ read textbooks and papers (schedule is ambitious)
 - ✓ work on your assignments (focus on **excellence** rather than just “getting a good grade”)
 - ✓ this is about developing highly-sought skills and competencies

The badges game

Badges data

- » COLT conference 1994
 - ✓ attendees received badges labeled as **positive** or **negative**
- » The author (Haym Hirsh) knew the function that generated the labels
- » Challenge: look at the names and find the hidden function
- » <https://www.seas.upenn.edu/~cis519/fall2019/assets/lectures/lecture-0/game.html>

Subset of the original dataset

+ Naoki Abe	- Myriam Abramson	+ David W. Aha
+ Kamal M. Ali	+ Eric Allender	+ Dana Angluin
+ Siddhant Apte	+ Minow Asadi	+ Lars Asker
+ David B. Beck	+ Michael Athanasiou	+ Jordi Alberca
+ Timothy P. Barber	+ Michael W. Barley	+ Cristina Baroglio
+ Peter Bartlett	- Eric Baum	+ Welton Becket
- Shai Ben-David	+ George Berg	+ Neil Berkman
+ Malini Bhandaru	+ Bir Bhanu	+ Reinhard Blasig
- Avrim Blum	+ Anselm Blumer	+ Justin Boyan
+ Carl B. Brodsky	+ Barbara Webb	+ Wayne Brattine
- Andrey Burago	+ Tom Bylander	+ Bill Byrne
- Claire Cardie	+ Richard A. Caruana	+ John Case
+ Jason Catlett	+ Nicolo Cesa-Bianchi	- Philip Chan
+ Mark Changizi	+ Pang-Chieh Chen	- Zhixiang Chen
+ Wan-Pi Chiang	- Steve A. Chien	+ Jeffrey Clouse
+ William Cohen	+ David Colling	- Clarence Bates Congdon
+ Antonio Correa-Joia	+ Kirk W. Craven	+ Robert E. Daley
+ Lindley Darroch	- Chris Darken	- Bhaskar Dasgupta
- Brian D. Davidson	+ Michael De la Maza	- Olivier De Vel
- Scott E. Decatur	+ Gerald F. DeJong	+ Kan Deng
+ Thomas G. Dietterich	+ Michael J. Donahue	+ George A. Drastal
+ Harris Drucker	- Chris Drummond	+ Hal Duncan
- Bob Elman	+ Tapio Elomaa	+ Susan L. Epstein
+ Usama Fayyad	+ Aaron Feigenson	+ Tom Farber
+ David Finton	+ John Fischer	+ Nicolas Fischer
+ Seth Flanders	+ Lance Fortnow	+ Paul Fischer
+ Judy A. Franklin	+ Yoav Freund	- Ameer Foued
+ Merrick L. Furst	+ Jean-Gabriel Ganascia	+ Johannes Furnkranz
+ Michael J. Gavara	+ Michael J. Gervasio	+ William Gasarch
- David Gilman	- Attilio Giordano	+ Volker Goltz
+ Paul W. Goldberg	+ Sally Goldman	+ Karl Goelz
+ Geoffrey Gordon	+ Jonathan Gratch	+ Diana Gordon
+ William A. Greene	+ Russell Greiner	+ Leslie Grate
+ Tal Grossman	+ Margo Guertin	+ Marko Grobelnik
+ Earl S. Harris Jr.	+ David Haussler	+ Tom Hancock
+ Lisa H. Hartenstein	+ David Held	+ Matthias Heger
+ Haym Hirsh	+ Jonathan Hodgson	+ Daniel Heyen
+ Jiarong Hong	- Chun-Man Hsu	+ Robert C. Holte
+ Masayuki Inaba	- Drago Indic	+ Kazushi Ikeda
+ Jeff Jackson	+ Sanjay Jain	+ Nitin Indurkhya
- Klaus P. Jantke	+ Nathalie Japkowicz	+ Wolfgang Jankow
+ Randolph Jones	+ Michael J. Jordan	+ George H. John
+ Bradley L. Kandararam	- Michael J. Kammerer	+ Leslie Pack Kaelbling
+ Michael Kearns	+ Neela Khan	+ Georgia Karrasikolas
+ Dennis F. Kibler	+ Jorg-Uwe Kietz	+ Ron Khader
- Jyrki Kivinen	- Emanuel Knill	- Elin Kinber
+ Ron Kohavi	+ Pascal Koiran	- Craig Knoblock
+ Daniel Koenenkamp	+ Matvez Kovacic	+ Mosh Koppel
+ Michael P. Laski	+ Martin Kummer	- Stefan Kramer
- Stephen R铸ek	+ Mai Lan	- Boyd Krahlevitz
- Steffen Lange	+ Pat Langley	+ Ken Lang
+ Wee Sun Lee	+ Moshe Leshno	+ Mary Soon Lee
		+ Long-Ji Lin

Lets play ...

- » Analyze the training data ...
- » Lets calculate your accuracy on these names ...

Brian Tester
Lyle H. Ungar
Paul Vitanyi
Gary Weiss
Bradley L. Whitehall
Janusz Wnek
Holly Yanco
Jean-Daniel Zucker

Chen K. Tham
Paul Utgoff
Xuemei Wang
Sholom Weiss
Alma Whitten
Kenji Yamanishi
John M. Zelle
Darko Zupanic

Can we use a computer?

- » We could extract features (attributes/characteristics) then use ML
 - ✓ length
 - ✓ number of vowels
 - ✓ number of consonants
 - ✓ number of dots
 - ✓ number of words
 - ✓ number of whitespaces
 - ✓ vowel/consonant ratio
 - ✓ length even/odd?
 - ✓ starts with a vowel?
 - ✓ is second letter a vowel?
 - ✓ ...

Features and Data

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

<https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b>