CIS 490 Machine Learning

Lecture 1

Instructor: (Julia) Hua Fang

1

Lecture 1: Contents

- Vaccination requirement and personal responsibility
- Overview of Course Logistics
  - Instructor and TA: hrs, location, and contact info
  - Course description and policy: Syllabus, Tentative Schedules, Reference books, Grading Rubric, Late and attendance policy, etc.
  - Announcements: MyCourses.
- Overview of Machine Learning:
  - Course Organization
  - Tentative Schedules
- Online Background Survey

2

Testing & Vaccination Requirement

- COVID19 testing is required for students who wish to live, learn, or physically come to campus before the start of the Spring 2022 semester. Further details on this requirement are provided here: https://www.umassd.edu/covid/test/
- UMass Dartmouth requires COVID-19 vaccinations and vaccination boosters for faculty, staff, and all students who wish to live, learn, or physically come to campus for the Spring 2022 semester. Further details on this requirement are provided here: https://www.umassd.edu/covid/vaccine/
- Students are required to upload their proof of vaccination to the Health Services Portal. Students claiming an exemption to the requirement must also submit either a written request for exemption upon religious grounds or medical exemption documentation from a healthcare provider. Follow these links on how to upload your COVID vaccine documentation guide to add your information to the Health Services portal.

3

Personal Responsibility

Every member of the UMassD community must do their part to protect one another, including:

- Using face coverings in all public places on campus unless a medical exemption is obtained.
- Complying with the MA face covering requirement, that is, students and faculty must wear a face covering at all times in classrooms and labs.
- Practicing good personal hygiene including handwashing per CDC guidelines.
- Avoiding eating and drinking in classrooms and labs.
- Staying home or isolated if they feel sick or exhibit known COVID-19 symptoms.

4

Zoom meets in unexpected situations

- Will notify all for zoom meets in unexpected situations.

5

Instructor

- Instructor: Dr. Hua Fang (Julia), hfang2@Umassd.edu;

  Tweets: @DrJuliaHuaFang; Office: DION 317A.

  Class Hours: Tues, Thurs, 2:00am -3:15pm

  Class Location: Science & Engr 113;

  Office Hours (subject to change):

  Tuesday Thursday: 11:00-12: 30pm (in person/zoom)

  Friday: 10:00-11:00am (zoom)

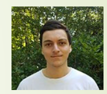
  By appointment: send emails
- CSDS Lab: https://www.umassmed.edu/fanglab/

6

TA/Grader

- **Teaching Assistants:**
  Hieu X Ngo (Henry) hngo1@umassd.edu;
  Office hours (subject to change): Zoom
  Tuesday & Thursday: 3:00-5:00pm, or on demand
  By appointment: send email

  Stefan Bruendl (Stefan) sbruendl@umassd.edu;
  Office hours (subject to change): Zoom
  Monday & Wednesday: 2:00-4:00pm, or on demand
  By appointment: send email



  Salvador Balkus (Sal) sbalkus@umassd.edu;
  Office hours (subject to change): Zoom
  Monday & Friday: 2:00-4:00pm, or on demand
  By appointment: send email

Rule: Please contact TA by email for appointments; Make your appointments with one TA; Don't make "Double/Triple" appointments. TAs work as a group.

- "Lab" location: DION 311, available 8AM-6PM, Mon-Fri, except CIS lab time
  Lab Technician: "Paul Naylor" pnaylor@umassd.edu.

7

Suggestions for Appointments

- Try your best to schedule your meets in regular business hours and days.
- Midnights and Weekends are discouraged. Last minute appointments before deadlines are discouraged.
- The best practice is to review the assignments on the day they are assigned and if you have questions, please make appointments in advance.

8

Syllabus: CIS490 Prerequisites

- Probability and Statistics: e.g., Distributions, densities, marginalization, moments.
- Calculus and Linear Algebra: e.g., Matrix multiplication, eigenvalues, positive semi-definiteness, multivariate derivates.
- Algorithms: e.g. CIS 360 - Algorithms and Data Structures
- Programming: R for HWs and Sectional Projects. Your language of choice for **final** project

  -- no coding, compilation, or debugging support.
- Ability to deal with abstract mathematical concepts.

Download a copy of CIS490 Syllabus at myCourses for details

9

Syllabus: CIS490 Objectives

- Differentiate between supervised and unsupervised learning;
- Illustrate the concepts and algorithms of common machine learning methods;
- Apply common machine learning techniques in one applied area, analyze and interpret data;
- Describe and discuss the concepts of machine learning techniques and their utility in applied areas.

10

Syllabus: Textbook Options

No textbook is required*. *Lecture slides, Handouts, or freely available online course readings will be posted on "mycourses" .

Reference Books:
Entry level:

- An Introduction To Statistical Learning, (2013) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2013, ISBN: 9781461471387 (online) and 9781461471370.
- Pattern Recognition and Machine Learning, (2006) Christopher Bishop. Springer, New York. (online) ISBN-13: 9780387310732

Intermediate-advanced level:

- The Elements of Statistical Learning, (2009) Hastie, Tibshirani, and Friedman. Springer. (online) ISBN 978-0-387-84858-7
- Machine Learning: A Probabilistic Perspective. (2012) Kevin P. Murphy. The MIT press. Cambridge, Massachusetts. ISBN 978-0-262-01802-9

11

Syllabus: **Grading Rubric**

- 5% Learning Activities (~ 5 **LA**, Total score: 100)
- 25% Homework Assignments (**HW**) and Sectional Projects(**SP**) (approximate every 2 weeks)
    - 10% HW (Total score of 3 written assignments: 100)
    - 15% SP (Total score of 2 hands-on projects: 100): submit source codes, URL and demo slides for presentation
- 15% Exam I (Total score: 100)
- 20% Exam II (Total score: 100)
- 30% Final Project (Total Score: 100)
    - 20% Final Project: A complete application; submit source codes, URL and demo slides for presentation
    - 10% Final Project Written Report
- 5% Class participation/attendance

Example: Exam 1 (100points, 15%)	Exam 2 (100points, 20%)	Learning Activities (100Points, 5%)	Homework Assignment (100 Points, 10%)	Sectional Projects (100 points, 15%)	Final Project: (100 points, 30%)	Class Attendance (100 points; 5%)	TOTAL	GRADE
95	95	89	92	99	90	100	93.75	A

the UMass Dartmouth grading system can be found here:
https://catalog.umassd.edu/content.php?catoid=69&navoid=5637&hl=grading&returnto=search#Grades_and_Grading_System

12

Reading and Practice expectation

- Reading time: at least 5 hrs a week
- "Lab" time: at least 5 hrs a week

13

Syllabus: Assignment Late Policy

- Deadlines:
  - 11:59pm EST for assignments
- You have 1 late day (1 x 24hour-chunk, Incl. Weekends) for assignments: Free
  - After 24 Hours (Incl. Weekends): 0
  - Not for **Exam 1**, **Exam 2**, and the sectional/final **project presentation session**

Note: Due to Covid, in the event that an illness requires a student to miss more than the equivalent of two weeks of classes (i.e., four classes for a class meeting twice a week for 75 minutes), it is important for the student to communicate the circumstances surrounding their absences to the instructor immediately to be given instructions on how to make up missed material.

14

Syllabus: Exam Policy & Academic integrity Policy

- Exam Policy: No Cheating
- Student Academic Integrity Policy

(see details in CIS490 Syllabus)

15

CIS 490: Announcements

- Announcements:
  - Announce via MyCourses and you will receive via email.
  - Click on "What's new" at MyCourses to check on all announcements.
- Learning activities (LA), Sectional and Final Projects, Exams: announced in class, on slides and posted at myCourses

16

Tentative Schedule

Download a copy of "Tentative Schedule" at MyCourses

17

Course Organization

18

Course Organization:

First two weeks: Course logistics, Get to Know R, Review Probability Theory. Learning activity and Homework 1

Before Exam I: Supervised Learning
(Learning activity, Homework 2 and 3, Sectional Project 1, Exam I)

After Exam I: Unsupervised Learning
(Learning activity, Sectional Project 2)

Last two weeks: Final project presentation, written report and Exam II

19

Overview of Machine Learning

20

Definition of Machine Learning

- Machine learning is defined as a set of methods that can **automatically** detect patterns in data, and then use the uncovered patterns to predict future data, or perform other kinds of decision making under uncertainty. (Murphy, 2012)

General agreement:

-Machine learning is a branch of artificial intelligence that focuses on **automation** and **algorithms.**

-**Math & statistics** are the basis for machine learning.

Machine Learning: A Probabilistic Perspective. (2012) Kevin P. Murphy. The MIT press. Cambridge, Massachusetts. ISBN 978-0-262-01802-9

21

Machine Learning **Overlap** with Statistical Learning

- Statistical Learning refers to

"a set of tools for modeling and understanding complex datasets. Recently developed are in statistics and blends with parallel developments in computer science and, in particular, machine learning." "These tools can be classified as supervised or unsupervised."(Preface & Chapter 2.1, James, Witten, Hastie, Tibshirani, 2013)

(Note: classical statistics focus more on population inference and hypothesis testing using samples)

An Introduction To Statistical Learning, (2013) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Online ISBN : 9781461471370

22

Key types of machine learning

- Supervised Learning
- Unsupervised Learning

(*Semi-supervised Learning*)

- *Reinforcement Learning (advanced; not covered; suppl.)*

23

Supervised Learning: aka. predictive approach

- Fact: "Most machine learning – about 70 percent – is supervised learning"(Wayne Thompson, SAS)
- Goal: to build a function (***f***) that maps inputs (**X**) to outputs (**Y**), given a "training" dataset with a number of dimensions (e.g. columns/variables) and a number of objects (e.g., rows/people), then use this function to predict the label (outcome) for new unseen objects.

X: aka. predictors, independent variables, covariates, features, attributes, which are measured, observed (Note: in applied areas, **X** may be defined more specific, eg. confounding variables)

f: aka. learner or predictive model such as linear regression

Y: aka. Label, response, outcome, dependent variable, which are measured, or observed, or known.

24

Supervised Learning: based on James, Witten, Hastie, Tibshirani, 2013

- "Supervised statistical learning involves building a statistical model for predicting, or estimating, an *output* based on one or more *inputs*." (James, Witten, Hastie, Tibshirani, 2013)
- Typically rely on historical, or observed data
- It is called "**supervised**" because of the presence of **Y**, the (known) label/outcome/dependent variable to guide the learning process. (Hastie, Tibshirani, Friedman, 2009)

25

Unsupervised Learning: aka. Descriptive approach

- Fact: "About 10 to 20 percent of machine learning is **unsupervised learning**, although this area is growing rapidly. " (Wayne Thompson, SAS)

  (http://www.sas.com/en_us/whitepapers/statistics-machine-learning-at-scale-107284.html)

- Goal: to discover / describe "interesting structure" (clusters/latent patterns) from **X**. (sometimes, called knowledge discovery)

  -- Observe only **X** (e.g. attributes/features) and have no measurements of **Y**, ie., no associated response/outcome, or unknown label.

- It is referred to as *unsupervised* because we lack **Y**, a response variable, that can supervise our analysis. (James, Witten, Hastie, Tibshirani, 2013)

So, our task is rather to describe how the data are organized or clustered.

26

Unsupervised Learning: Two types

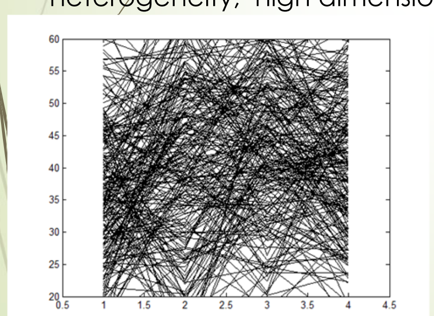
- **Dimension reduction**: decrease the number of variables (D), ie. the columns. E.g. Principle Component Analysis (PCA)
- **Clustering**: combine the objects (e.g. patients) into latent groups, or "decrease" the number of rows (N). E.g. Kmeans

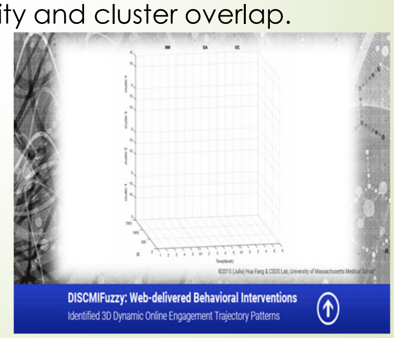
27

Unsupervised Learning: aka. Descriptive approach (example)

MIFuzzy: National Institutes of Health (NIH) funded computational project.
https://www.umassmed.edu/fanglab/research-projects/

- Behavior trajectory pattern recognition : unsupervised learning; latent variable model approach;
- Handle small and big longitudinal data with missing values, heterogeneity, high dimensionality and cluster overlap.





28

Supplement (Suppl): Semisupervised Learning

A semi-supervised learning problem:

- Suppose that we have a set of *n* observations. For *m* of the observations, where $m < n$, we have both **X** predictor measurements and a **Y** response measurement. For the remaining $n - m$ observations, we have **X** predictor measurements but no **Y** response measurement.

Such a scenario can arise if

the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect.

(http://www.sas.com/en_us/whitepapers/statistics-machine-learning-at-scale-107284.html)

29

Suppl: Reinforcement Learning

- Useful to learn how to act or behave when given occasional rewards or punishment signals. (Murphy, 2012)
- Discovers for itself which actions yield the greatest rewards through trial and error. Reinforcement learning has three primary components:

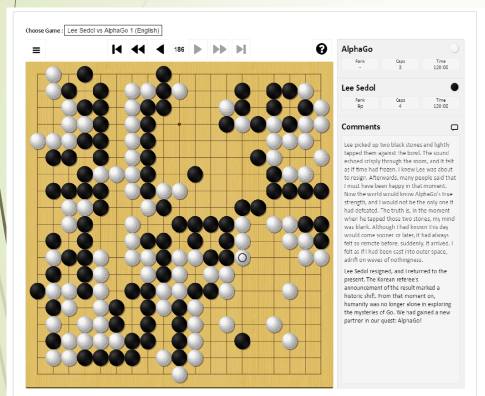
1. **The agent** – the learner,
2. **The environment** – everything the agent interacts with,
3. **Actions** – what the agent can do.

Markov decision processes (MDPs) are popular models used in reinforcement learning.

30

Application Examples : AlphaGo

- [Alphago: Machine vs. Human.]



- Deep (reinforcement) Learning
(evolving fast; material provided in Chapter Deep Learning "may soon be outdated", Murphy, 2012)

- Start from the common machine learning methods

https://www.nature.com/articles/nature16961
https://www.natureasia.com/en/research/highlight/12229

31

To-Do Tasks

32

Task 1: Study Groups

- By Jan 25, First Round: Find your study group members; select your group leader who will be the contact person and coordinator for your group.

  Submit the *.docx file to MyCourses, in the folder called "Course Logistics".

  -- Title: "CIS490 Study Group Members"

  contents: include Full names, Student IDs, Emails

  Mark Group leader (contact person/coordinator);

  -- Each team: no more than 3. **If you prefer to work on your own, please do submit your intention here as well.**
- ´After Jan 25, Second Round: **If you don't want to work on your own** but haven't found a partner, we will randomly assign you to a group that can accept a member.
- By Jan 28, Final group list announced at myCourses, ie., You agreed to work on your own or on a team

33

Task 2: Software Installation: R

Install R:

https://www.r-project.org/about.html

https://stats.idre.ucla.edu/r/

34

Task 3: **Background Survey**

Due Jan 21,

Posted in the "Course Logistics" folder

Please submit your survey at MyCourses

35