CIS 490 Machine Learning

Lecture 2

Instructor: (Julia) Hua Fang

1

Last time

- 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

3

Last time: 2 tasks assigned

Reminder:

- Task 1: Study groups
- Task 2: Install R and RStudio
- Task 3: Background Survey due this Friday, 1/21.

3

4

Lecture 2: Outline

- Breaking News in ML field since 2017 CIS602/490 machine learning class: Latest Machine Learning (ML) application/advancement in bio-medicine and game industry.
- ML databases: e.g. UCI Machine Learning Repository
- Review: Probability Theory (I)
- R basics: See the posted Word file called "CIS490_RIntro_final.docx" under Lecture 2 link at MyCourses. In class demo.

4

5

When CIS602/CIS 490 took place since 2017, how fast is AI development?

See breaking news in application areas:

- 2017Spring: AlphaGo, deep reinforcement learning
- 2018 Fall: Andrew Ng' Chexnet (in later lectures)
- 2019 Spring: Deepmind & Blizzard open StarCraft II as an AI research environment: deep reinforcement learning, etc.
- 2020 Spring: Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks, Univ. of Michigan Ann Arbor
- 2021 Spring: DALL*E: Creating Images from text

5

ML Advancement in Game Industry, Spring 2019

6

Jan22, 2019

Tweets	Following	Followers	Likes
750	111	207K	216

Tweets   Tweets & replies   Media

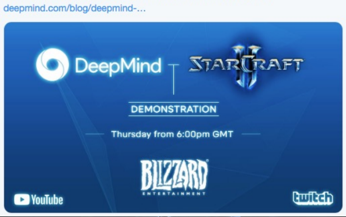
Pinned Tweet

DeepMind @DeepMindAI · Jan 22

Join us and @Blizzard_Ent this Thursday at 6:00pm GMT for an exciting #StarCraft demonstration, hosted by @Artosis and @RotterdaM08!

Livestream on YouTube: youtube.com/c/deepmind

Read more about #StarCraft2 as an environment for AI research: deepmind.com/blog/deepmind-…



DeepMind and Blizzard open StarCraft II as an AI research environment

DeepMind's scientific mission is to push the boundaries of AI by developing systems that can learn to solve complex problems. To do this, we design agents and test their ability in a wide range of environments from the purpose-built DeepMind Lab to established games, such as Atari and Go.

Testing our agents in games that are not specifically designed for AI research, and where humans play well, is crucial to benchmark agent performance. That is why we, along with our partner Blizzard Entertainment, are excited to announce the release of SC2LE, a set of tools that we hope will accelerate AI research in the real-time strategy game StarCraft II. The SC2LE release includes:

- A Machine Learning API developed by Blizzard that gives researchers and developers hooks into the game. This includes the release of tools for Linux for the first time.
- A dataset of anonymised game replays, which will increase from 65k to more than half a million in the coming weeks.
- An open source version of DeepMind's toolset, PySC2, to allow researchers to easily use Blizzard's feature-layer API with their agents.
- A series of simple RL mini-games to allow researchers to test the performance of agents on specific tasks.
- A joint paper that outlines the environment, and reports initial baseline results on the mini-games, supervised learning from replays, and the full 1v1 ladder game against the built-in AI.

https://deepmind.com/blog/deepmind-and-blizzard-open-starcraft-ii-ai-research-environment/

https://deepmind.com/research/publications/deepmind-lab

Read for fun: https://arxiv.org/pdf/1612.03801.pdf

6

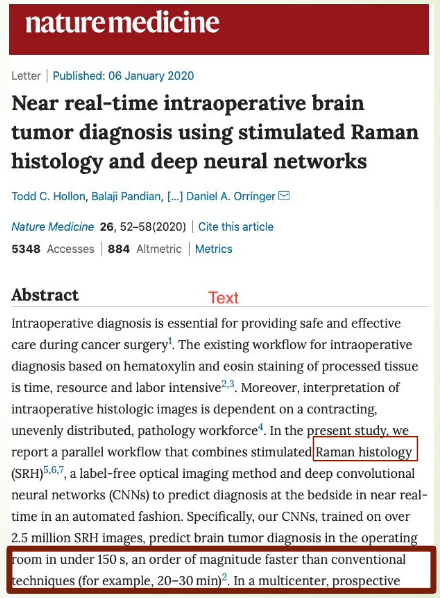
ML Advancement in Biomedicine, Spring 2020

7

Jan 6, 2020: Deep Convolutional Neural Networks

"CNN-based diagnosis of SRH images was noninferior to pathologist-based interpretation of conventional histologic images (overall accuracy, 94.6% versus 93.9%)."

Steamlined AI based diagnosis vs. conventional techniques in lab:
150s vs 20-30 min.


naturemedicine

Letter | Published: 06 January 2020

Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks

Todd C. Hollon, Balaji Pandian, [...] Daniel A. Orringer

Nature Medicine **26**, 52–58(2020) | Cite this article

5348 Accesses | **884** Altmetric | Metrics

Abstract

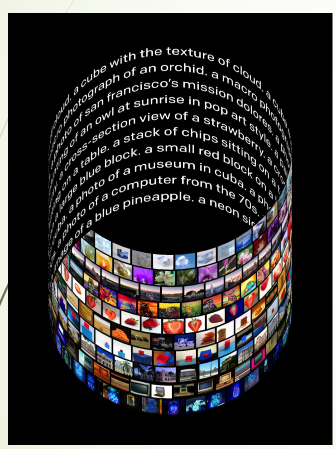
Text

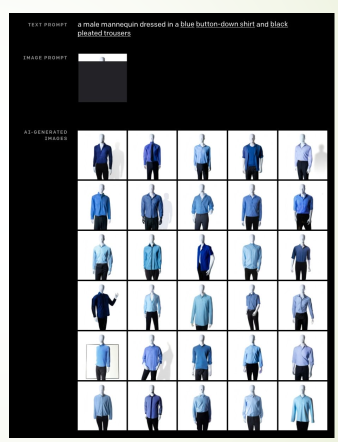
Intraoperative diagnosis is essential for providing safe and effective care during cancer surgery[1]. The existing workflow for intraoperative diagnosis based on hematoxylin and eosin staining of processed tissue is time, resource and labor intensive[2,3]. Moreover, interpretation of intraoperative histologic images is dependent on a contracting, unevenly distributed, pathology workforce[4]. In the present study, we report a parallel workflow that combines stimulated Raman histology (SRH)[5,6,7], a label-free optical imaging method and deep convolutional neural networks (CNNs) to predict diagnosis at the bedside in near real-time in an automated fashion. Specifically, our CNNs, trained on over 2.5 million SRH images, predict brain tumor diagnosis in the operating room in under 150 s, an order of magnitude faster than conventional techniques (for example, 20–30 min)[2]. In a multicenter, prospective

7

8

ML Advancement in Text-Image pair, Spring 2021





citation: https://openai.com/blog/dall-e/ & by courtesy, Andrew Ng's tweeter

8

9

ML databases:

- UCI Machine Learning Repository
- Kaggle https://www.kaggle.com/datasets

Other databases:

- Image and videos: e.g.
  http://www.image-net.org/
  https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/
  http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm
- Biosensor: e.g.
  https://biosensordb.ucsd.edu/
  https://catalog.data.gov/dataset/biosensors-for-exploration-medical-system
- Speech/NLP:e.g.
  https://lionbridge.ai/datasets/the-best-25-datasets-for-natural-language-processing/
  https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38

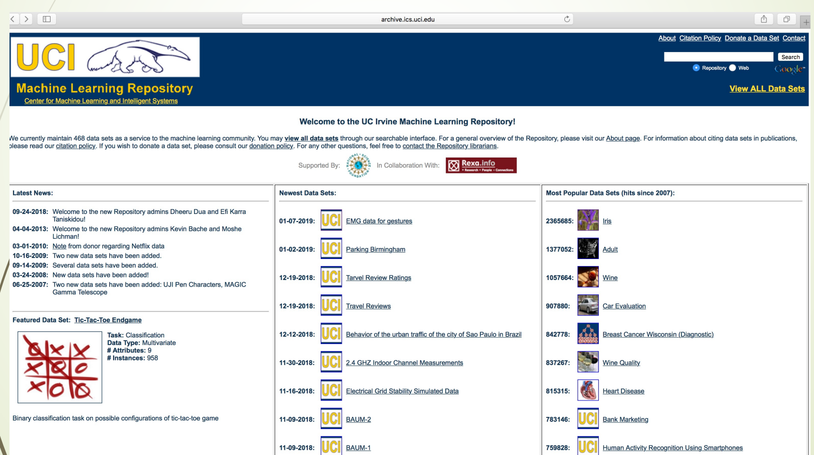
9

10

Real Data Resources:

- UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/index.php



10

11

Real Data Example: CIS490 2017 Fall Survey Data

	ID	Y_outcome	Q1_Major	Q1_advisor
1	1	1	2	7
2	2	0	0	0
3	3	1	0	0
4	4	0	0	0
5	5	1	0	.
6	6	0	0	8
7	7	1	0	0

11

12

Review: Probability Theory (I)

- Brief review of Probability Theory
- Random Variable

--Types of Random Variables

--Types of Probability Distribution

Q1: what are pdf, pmf, and cdf?

12

13 Brief review of Probability Theory

- A ***probability model*** is a mathematical representation of a random phenomenon. It is defined by its ***sample space***, ***events*** within the sample space, and ***probabilities*** associated with each event.
  The ***sample space S*** for a probability model is the set of all possible outcomes.

  -- An ***event E*** is a subset of the sample space **S**.
- A ***probability*** is a numerical value assigned to a given event *E*. The probability of an event is written ***P(E)***, and describes the long-run relative frequency of the event.

The first two basic rules of probability are the following:

Rule 1: *Any probability P(E) is a number between 0 and 1 ($0 \leq P(E) \leq 1$).*

Rule 2: *The probability of the sample space S is equal to 1 ($P(S) = 1$).*

13

14 Random Variable (rv.)

- A ***random variable***, usually written *X*, is a variable whose possible values are numerical outcomes of a random phenomenon.

  Classical examples:

  Can you give an example?
- Two types of Random Variables (rv.)

  -- Discrete

  -- Continuous

14

15

Random Variable (rv.)

- A ***random variable***, usually written *X*, is a variable whose possible values are numerical outcomes of a random phenomenon.

  Classical examples:

  toss a coin {heads, tails};

  roll an unbiased six- sided die {1, 2, 3, 4, 5, 6}.

- Two types of Random Variables (rv.)

  -- Discrete

  -- Continuous

15

16

Random Variable (RV.): Discrete

- A ***discrete rv.*** is one which may take on only a countable number of distinct values such as 0,1,2,3,4,........

-- Discrete random variables are usually, but not necessarily, counts.

-- If a random variable can take only a finite number of distinct values, then it must be discrete.

Can you give an example?

16

17

Random Variable (RV.): Discrete

- The ***probability distribution*** of a **discrete** rv. is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the **probability mass function (pmf)**.

- Suppose a random variable X may take k different values, with the probability that $X = x_i$ defined to be $P(X = x_i) = p_i$. The probabilities p_i must satisfy the following:

1: $0 \leq p_i \leq 1$ *for each i*

2: $p_1 + p_2 + \ldots + p_k = 1$.

In-class exercise next

http://www.stat.yale.edu/Courses/1997-98/101/ranvar.htm

17

18

Discrete Random Variable (RV.): In-class exercise/review

- Example:

Suppose a random variable X can take the values 1, 2, 3, or 4.
The probabilities associated with each outcome are described by the following table:

Outcome	**1**	**2**	**3**	**4**
Probability	0.1	0.3	0.4	0.2

Q1: What is the probability that X is equal to 2 or 3 ?

Q2: what is the probability that X is greater than 1?

18

19

Common Discrete Distributions

- Bernoulli: the probability distribution of a rv. which takes the value 1 with success probability of p and the value 0 with failure probability of q=1-p.

E.g. toss a coin, 1 = head; 0= tail.

--- one of the simplest yet most important random processes in probability.

--- Suppl.: *See proof and examples at*

http://www.randomservices.org/random/dist/Discrete.html

- Binomial: The probability distribution of the number of successes in a sequence of **n** independent yes/no trials, each of which yields success with probability **p**.

E.g. toss the coin multiple times.

http://www.stat.yale.edu/Courses/1997-98/101/binom.htm

---Bernoulli is a special case of Binomial when the number of trial n =1.

19

20

Common Discrete Distributions

- Multinomial

-- a generalization of the Binomial distribution. i.e., when more than 2 categories.

e.g., Disease risk (outcome random variable) has three levels: high, low, none risk.

- Poisson

-- popular for modelling the number of times an event occurs in an interval of time or space

-- used for count data (e.g, in queue theory)

e.g., the number of patients arriving in an emergency room between 11 and 12 pm

https://en.wikipedia.org/wiki/Multinomial_distribution
https://en.wikipedia.org/wiki/Poisson_distribution

20

21

Random Variables: Continuous

A ***continuous random variable*** is one which takes an infinite number of possible values.

- A continuous random variable is not defined at specific values. Instead, it is defined over an *interval* of values, and is represented by the ***area under a curve*** (ie. an ***integral***).

Can you give an example?

21

22

Common Continuous Distributions

- The probability of distribution of continuous random variables is described by **probability density function (pdf)**

Note: both discrete and continuous random variable X, evaluated at x, have **cumulative distribution function, cdf**. It refers to the probability that X will take a value <= x .

22

23

Common Continuous Distributions

- Uniform
- Gaussian/Normal
- Student t
- Chi-squared
- Gamma
- Beta
- Pareto

(Q: Which ones belong to exponential family distributions?)

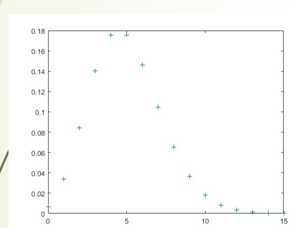
Note: Review online if you forget these common Continuous Distributions

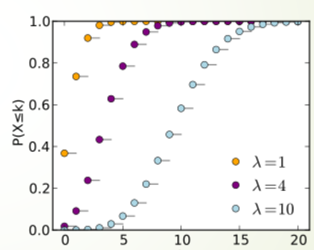
23

24

Quick Quiz: 1

- Which statistical distribution do these two graphs show?
- Are they continuous or discrete distributions?
- Which one shows the pmf of this distribution? Which one shows cdf?





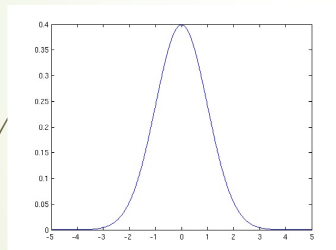
https://en.wikipedia.org/wiki/Poisson_distribution

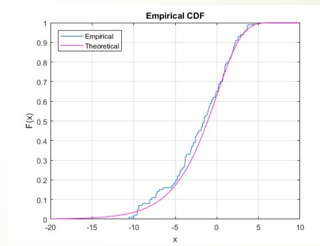
24

25

Class exercise: 2

- Which statistical distribution do these two graphs show?
- Are they continuous or discrete distributions?
- Which one shows the pdf of this distribution? Which one shows cdf?





Refer to: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html

25

26

Suppl: Types of probability spaces

Define | *Ω* | = number of possible outcomes

- Discrete space | *Ω* | is finite
  - Analysis involves *summations* (∑)
- Continuous space | *Ω* | is infinite
  - Analysis involves *integrals* (∫)

26

27

R: Basics

- See the posted Word file called "CIS490_RIntro_final.docx" under Lecture 2 link at MyCourses. Also feel free to check at http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R1_GettingStarted/R1_GettingStarted8.html

In-class demo

Get started!

In two weeks, going to work on supervised learning…

27

Learning Activity 1(LA1): 20 points, Due: Jan 26.

28

Part I: a. Review Lecture 1 Slides(LS1) and read *Chapter 1.1* "***Machine Learning: A Probabilistic Perspective***". (2012) Kevin P. Murphy (simply, **Murphy Chpt 1.1 later**), describe the definition of Machine Learning (ML); read *Preface* & *Chapter* 2.1 of ***An Introduction To Statistical Learning***, (2013) (Simply, **ITSL Preface & Chpt 2.1 later**), describe statistical learning and its overlap with ML.

b. Review Lecture 2 Slides (LS2), read Murphy Chpt 2, or refer to online materials e.g., Wikipedia (https://en.wikipedia.org/wiki/Poisson_distribution), and previous probability textbooks you can find:

- Write pmf and cdf, of these discrete distributions: (1)Bernoulli, (2) binomial, (3)Poisson. Note: give an example for each of these 3 distributions, eg. The Poisson distribution is useful for modeling such an event: the number of covid patients arriving in ICU between 9am -5pm.
- Write pdf and cdf of : (1) Uniform, (2) Gaussian/Normal, (3) Student t, (4) Chi-squared, (5) Gamma, (6) Beta and (7) Pareto; give an example for each of these 7 distributions.
- What is RV? What are the two types of RV?

Note: Don't copy and paste contents from Lecture Slides. Please use your own language and write your answers in a Word document (*.docx)

Part II: Refer to the instruction in "CIS490_RIntro_final.docx" posted under LA1 to install R and RStudio.

- Download the IRIS data from UCI repository (https://archive.ics.uci.edu/ml/datasets/Iris), then use R to import and export this dataset: **copy and paste your R code into the same Word document (*.docx) you used for PART I;**

Submission: Put Part I and Part II work into one Word file (*docx) and only submit this one file under LA1 link at MyCourses.

Note: Don't attach data (we have it!). Don't submit a zipped file.

The late policy does NOT apply to Learning Activity (LA) Assignments. LAs are not group assignment. To receive your score, each individual must submit your Complete work on Time.

28