

Statistics and Data Science 265

Introductory Machine Learning

Thursday, September 2

Yale

Outline

- Overview of course
- Perspectives on ML (and AI)
- Syllabus and logistics

Course objectives

Gain understanding of and experience with basic machine learning methodology

Course objectives

- Gain some new perspective
- Appreciate some of the power and limitations of ML
- Have fun
- Want to learn more

Related course

- This course introduced for Certificate in Data Science
- Intended to be accessible intro to ML for wide range of students
- S&DS 365 (grad number 665) will become a new course “Intermediate Machine Learning” this spring. Can be taken as a follow up course; more technical and in-depth
- Happy to chat if unclear this course is right for you

Common questions

“What’s the difference between AI and Machine Learning?”

“Is Deep Learning the same as Machine Learning?”

“What’s the difference between Statistics and Machine Learning?”

August 31, 1955

John McCarthy, Marvin L. Minsky, Nathaniel Rochester,
and Claude E. Shannon

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug. 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

SUNDAY, OCTOBER 19, 1958

MACHINE TO COPY BRAIN'S METHODS

Huge Computer in London
to 'Think' Like a Person
for Study of Learning

Special to The New York Times.

LONDON, Oct. 15—Investigators in neurology at University College here are building a massive automatic computer for the principal purpose of testing theories about the learning capacity of the brain.

The machine will "think"; that is, it will scan shapes such as the letters of the alphabet and simple words and after analyzing and absorbing this visual information it will "say" (through a loudspeaker) what it has seen at precisely the same rate as that of a fairly intelligent human subject.

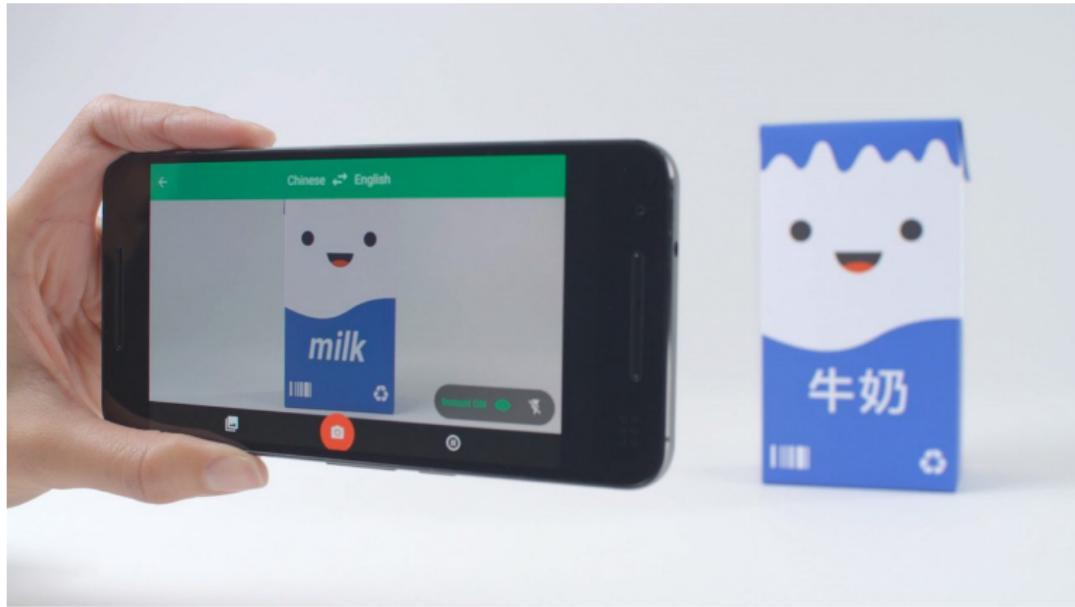
This is being achieved by



Today: Home assistants



Translation



<http://www.sciencemag.org/>

Pricing and recommending homes

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

\$1 for 2 months

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate 



CIO JOURNAL



Zillow Develops Neural Network to 'See' Like a House Hunter

Granite or stainless steel countertops? Zillow's visual recognition effort can recognize the difference

By **SARA CASTELLANOS**

Nov 11, 2016 3:29 pm ET

Data scientists at Zillow Group are developing complex computer programs that detect specific attributes in photographs of homes, which could aid in estimating their value. Advances in deep learning, big data and cloud computing have converged to allow the online real estate database firm and others to develop technology that mimics how the human brain [...]

Recommended Videos

1. Film Clip: Pirates of the Caribbean: Dead Men Tell No Tales'



2. What to do in your 40s to retire a millionaire



<https://blogs.wsj.com/cio/2016/11/11/zillow-develops-neural-network-to-see-like-a-home-buyer/>

Email suggestions

The screenshot shows a user interface for composing an email. A dark gray rectangular box, likely a suggestion or placeholder, contains the text "Taco Tuesday". Above this box, there is some very faint, illegible text. In the top right corner of the main area, there is a red square icon with the letters "ET" in white.

Jacqueline Bruzek x

Taco Tuesday

Hey Jacqueline,

Haven't seen you in a while and I hope you're doing well.

SUBSCRIBE

<https://www.youtube.com/watch?v=nZ-C8I-8BZw&t=0m16s>

YouTube



Amazing ways YouTube uses ML and AI: <https://www.forbes.com/sites/bernardmarr/2019/08/23/the-amazing-ways-youtube-uses-artificial-intelligence-and-machine-learning>

YouTube

- Each month: 1.9 billion users
- Each day: 1 billion hours of video watched
- Each minute: 300 hours of video uploaded
- ML: Automatically remove objectionable content
- ML: “Up Next” feature

Amazing ways YouTube uses ML and AI: <https://www.forbes.com/sites/bernardmarr/2019/08/23/the-amazing-ways-youtube-uses-artificial-intelligence-and-machine-learning>

Translation

HOME > SPORTS

A Belarusian Olympian who complained about her coaches used Google Translate to relay her plea for help to Japanese police

Lauren Fries Aug 5, 2021, 6:46 PM



Belarusian Olympic sprinter Krystsina Tsimanouskaya said she was taken to the airport against her wishes and would not return home. Reuters

With a zero trust strategy, C you're in the driver's seat

See how →

IBM

What is Machine Learning?

The study of algorithms and statistical models to develop computer programs that improve with experience.

What is Machine Learning?

Machine Learning is closely aligned with Statistics, but with a focus on computation, scalability, prediction, representation, and complex problems

- Speech recognition
- Machine translation
- Object recognition and scene classification
- Autonomous driving...

Subproblems of these and other complex problems are concrete, statistical estimation and inference problems that can be studied in isolation.

AI vs. ML

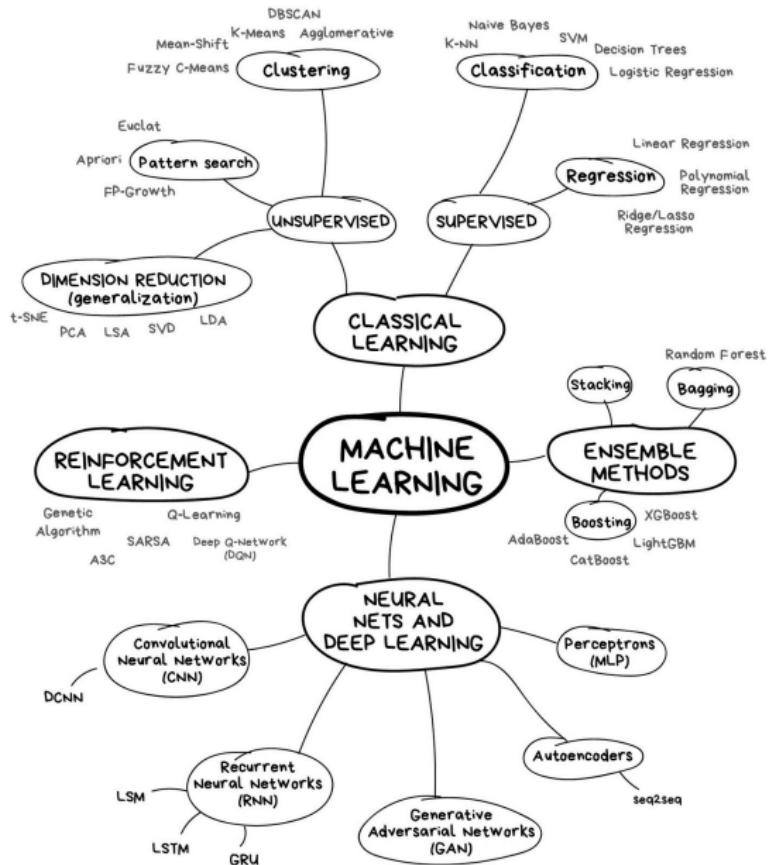
Machine learning focuses on making predictions and inferences from data.

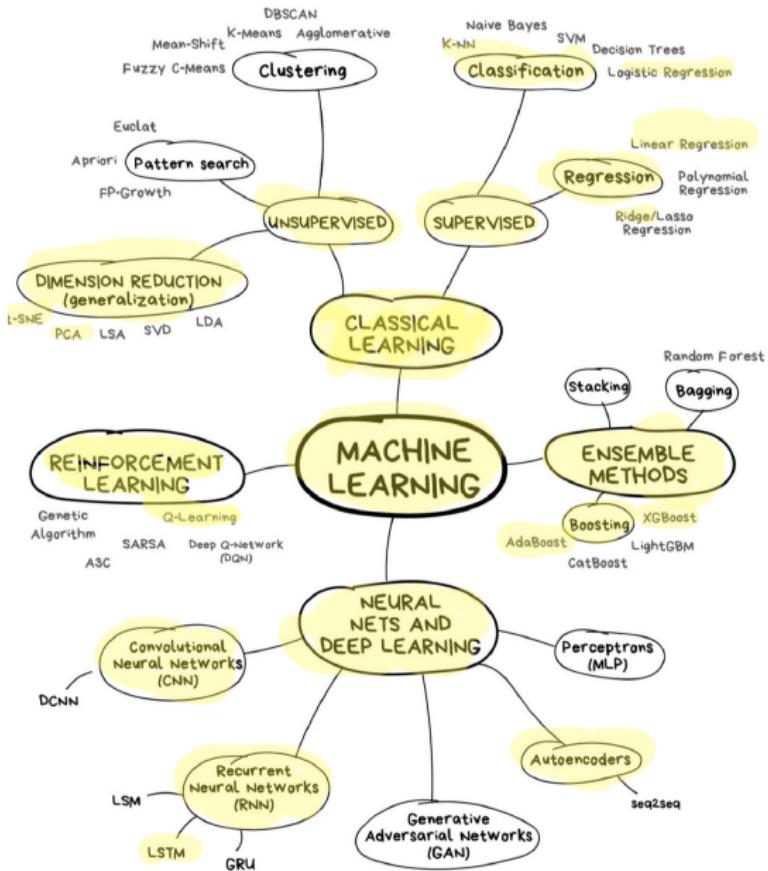
AI combines machine learning components into a larger system that includes a decision making component.

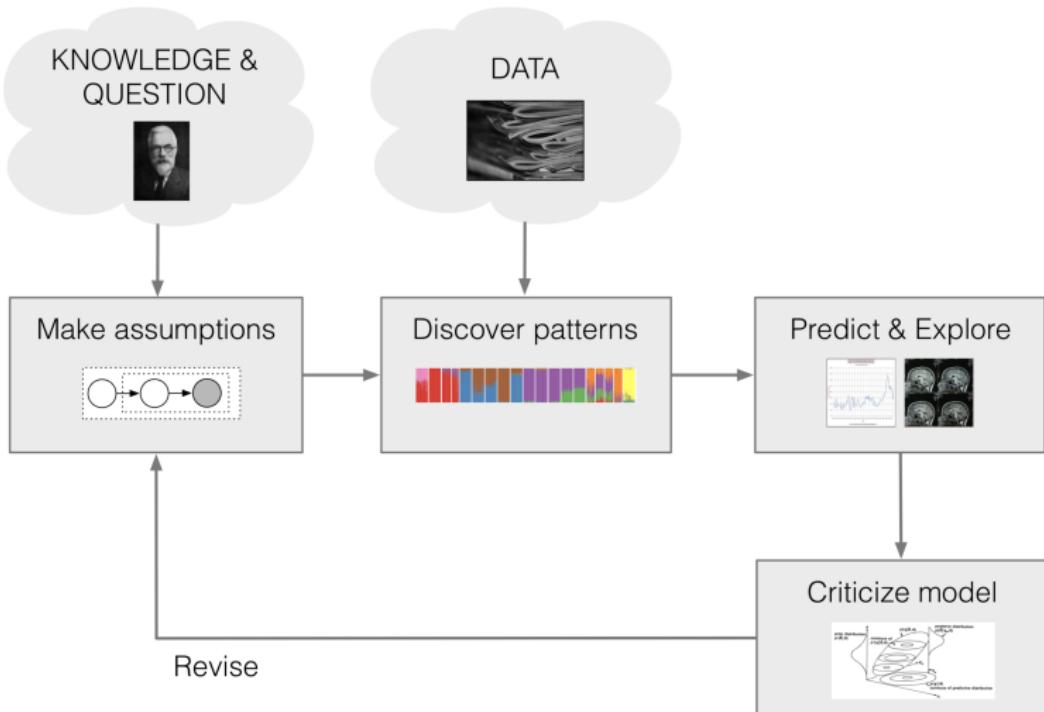
An AI system exhibits a behavior, resulting from the collective decisions that are made.

Machine learning frameworks

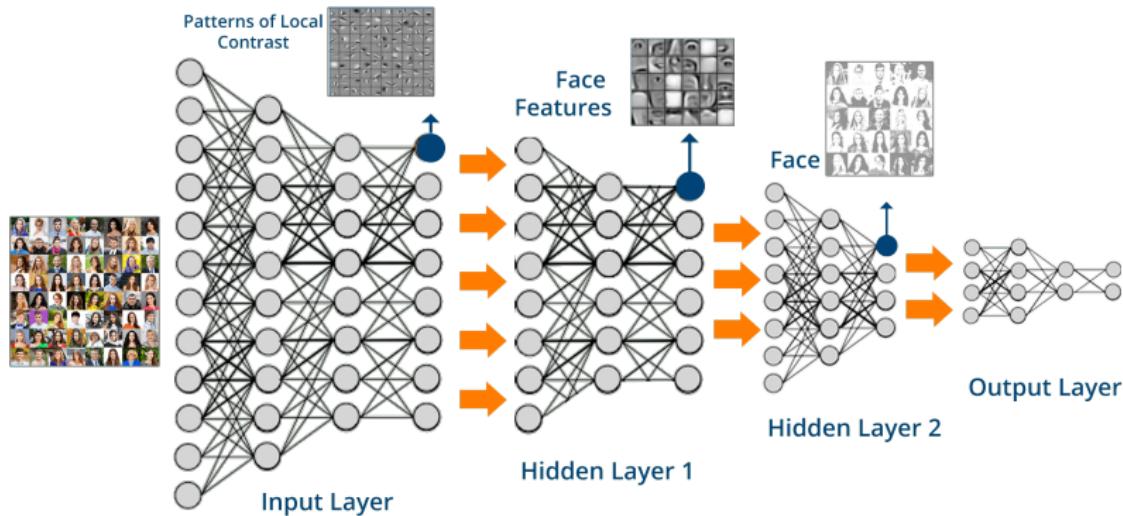
- Supervised, unsupervised, semi-supervised
- Reinforcement learning
- Generative vs. discriminative models
- Representation learning







Deep learning is a type of machine learning



- Heuristics motivated from simplified view of the brain
- A particular form of nonlinear classification/regression
- Not well-suited to latent variables

Culture of Code

- Great deal of current AI/ML work is purely engineering based
- Informal input/output reasoning

*“that program gave this output...
maybe this program will give that output”*

- Deep learning software engineers develop sophisticated intuitions
- The code is the product

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.



xkcd.com/1838

Latent variables: The elephants in the room



Example of representation learning: Word embeddings

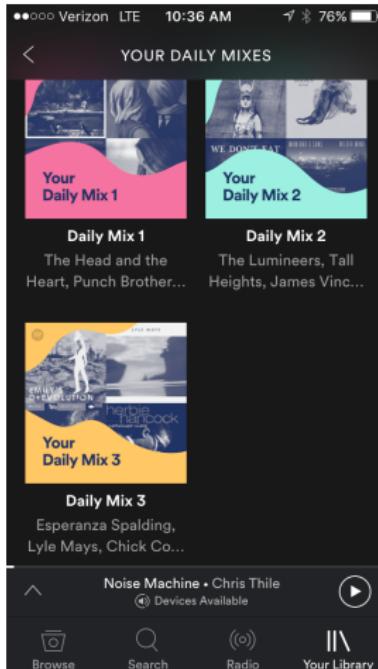
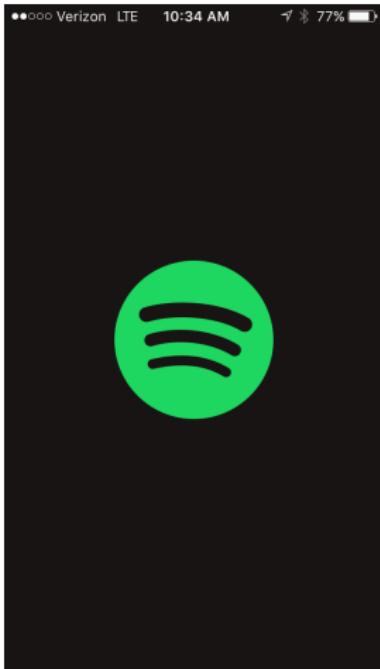
- Each word in vocab is mapped to 100 or 500 dimensional vector
- Based solely on co-occurrence statistics in corpus of text

Example of representation learning: Word embeddings

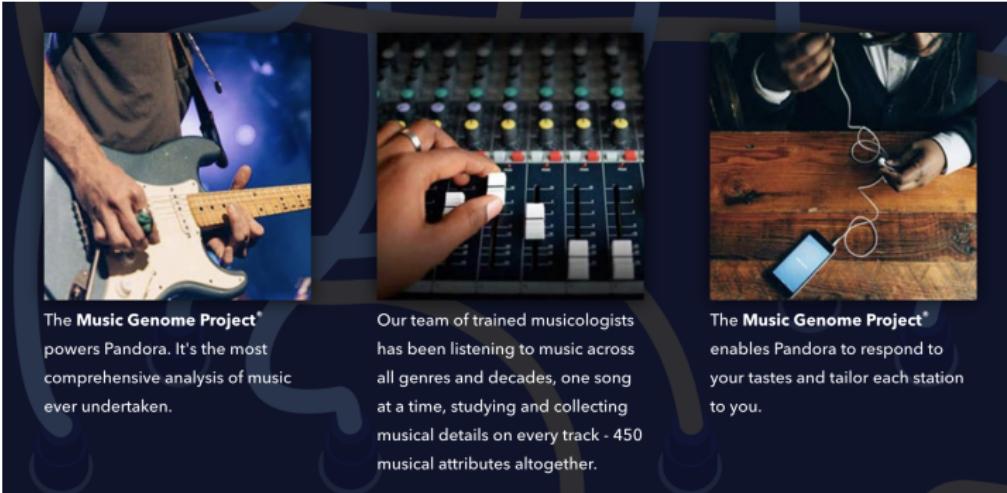
Yale:

```
[ 0.78310001, 0.51717001, -0.38207 , -0.23722 , -0.31615999, 0.30805001, 0.76389998, 0.064106 , -0.74913001,  
 0.60585999, -0.23871 , -0.16876 , -0.25634 , 1.07270002, -0.29967999, 0.020095 , 0.54500997, -0.17847 , -0.26675999,  
 -0.11798 , -0.48692 , 0.22712 , 0.017473 , -0.4747 , 0.44861001, -0.084281 , -0.30412999, -1.13510001, -0.14869 , -0.11182 ,  
 -0.32530001, 1.0029 , -0.35742 , 0.35148999, -1.10679996, -0.064142 , -0.72284001, 0.14114 , -0.41247001, -0.16184001,  
 -0.54576999, -0.12958001, -0.88356 , -0.089722 , 0.10555 , -0.12288 , 0.92851001, 0.50032002, 0.1349 , 0.21457 ,  
 0.35073999, -0.73132998, 0.39633 , -0.43239999, -0.38815999, -1.34669995, 0.37463999, -0.79386002, 0.11185 , 0.18007 ,  
 -0.75142998, 0.24975 , -0.094948 , -0.36341 , 0.24869999, -0.22667 , 0.32289001, 1.29489994, 0.42658001, 1.29120004,  
 -0.13954 , 0.68976003, 0.21586999, 0.13715 , -1.00919998, 0.028827 , 0.11011 , -0.1912 , -0.073198 , -0.52449 , 0.49199 ,  
 0.14463 , -0.18844 , -0.75536001, -0.28704 , 0.019113 , 0.30349001, -0.74425 , -0.072221 , -0.40647 , 0.26899001, -0.28318  
, 0.72409999, 0.50796002, -0.37845999, -0.13008 , -0.13808 , 0.098928 , 0.16215999, 0.16293 ]
```

Embeddings for music recommendations



Experts vs. Data: The case of Pandora vs. Spotify



The Music Genome Project* powers Pandora. It's the most comprehensive analysis of music ever undertaken.

Our team of trained musicologists has been listening to music across all genres and decades, one song at a time, studying and collecting musical details on every track - 450 musical attributes altogether.

The Music Genome Project* enables Pandora to respond to your tastes and tailor each station to you.

- Pandora's "Music genome": Over 450 musical attributes
- Melody, harmony, rhythm, form, composition, lyrics...

<https://arstechnica.com/tech-policy/2011/01/digging-into-pandoras-music-genome-with-musicologist-nolan-gasser/>

Experts vs. Data: The case of Pandora vs. Spotify

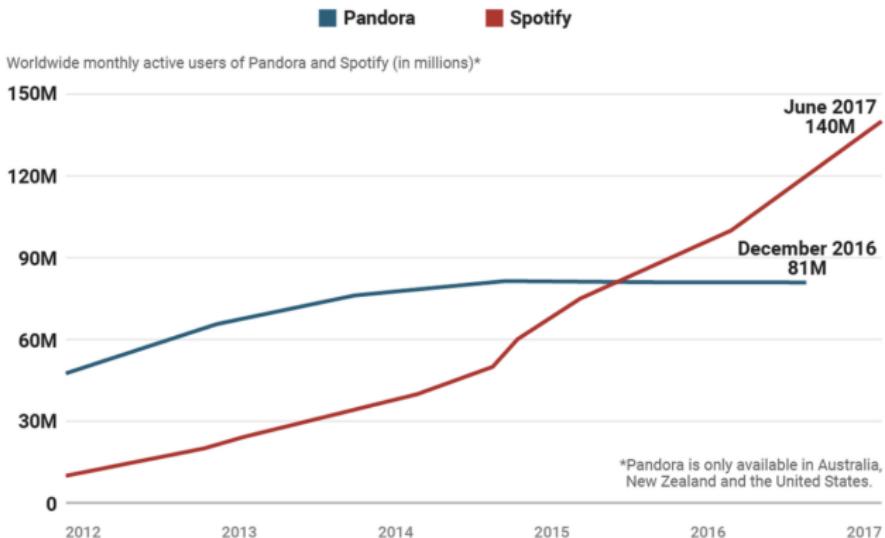


Spotify: Word embeddings trained from playlists

Experts vs. Data: The case of Pandora vs. Spotify

TECH ■ CHART OF THE DAY

PANDORA'S GROWTH STALLS AS SPOTIFY PULLS AHEAD



SOURCE: Company filings/announcements

BUSINESS INSIDER

Hacking ML Systems

[SUBSCRIBE](#)[SIGN IN ▾](#)

TESLA AUTOPILOT —

Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/1/2019, 8:50 PM

Keen Security Lab



Machine learning at a large Internet company

- Typical project lifetime: 6 months to 1 year
- Ads projects involve thousands of software engineers
- Often adding new “feature” to existing black box model
- No single person understands entire model
- Not interpretable
- Users are hashed; employees don’t have access to friends’ data

Reasons for optimism

- Increasingly part of academic research across disciplines
- Engaging a broad community
- We're still in very early stages

Questions or discussion?

Team

- Instructor: John Lafferty (Prof, DS2 and CS)
- Teaching Fellows
 - Wendy Luo (PhD student in Biomedical Engineering)
 - Chris Xu (PhD student in Statistics and Data Science)
 - Yiyang Wang (MS student in Statistics and Data Science)
- ULAs
 - Fred Zhang
 - Nina Huang
 - Ben Christensen
 - Gandhar Mahadeshwar
 - Will Sun
 - Nanyan Wu
 - Aman Heyer
 - Alden Tan

Course materials

Materials posted to <https://iml.ydata123.org>; sometimes to Canvas

Please use Ed Discussion for any questions about lectures, homework, etc. first, before email!

For email

- For logistical issues (e.g. adding to Canvas): E-mail Yiyang, copy me, Chris, Wendy
- All other: E-mail me, copy TAs

Syllabus

Introductory Machine Learning covers the key ideas and techniques in machine learning without the use of advanced mathematics. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods and a more formal understanding of how and why they work. Assignments give students hands-on experience with the methods on different types of data.

Syllabus

Topics include linear regression and classification, tree-based methods, topic models, word embeddings, recurrent neural networks, deep learning and reinforcement learning. Examples come from a variety of sources including political speeches, archives of scientific articles, real estate listings, natural images, and several others. Programming is central to the course, and is based on the Python programming language.

Prerequisites

- At least two of the following courses: S&DS 230, 238, 240, 241 and 242
- Previous programming experience (e.g., R, Matlab, Python, C++), Python preferred. The course will make extensive use of Python programming, using Jupyter notebooks.

Installing Jupyter

- A beginner's guide to installing Python and Jupyter on your computer is here: <https://bit.ly/22KVCfsV>
- See installation guide on course Canvas site: Files > Getting Started
- Use Python 3.x version

Course objectives

Gain understanding of and experience with basic machine learning methodology

Course objectives

- Gain some new perspective
- Appreciate some of the power and limitations of ML
- Have fun
- Want to learn more

Evaluation

- Seven assignments (50%)
- Mid-semester exam (20%)
- Quizzes (10%)
- Final exam: 20%

Lowest assignment score will be dropped. Late assignments not accepted

Assignments

- Roughly every 1.5 weeks
- Due at 11:59pm on the day
- Submitted using Gradescope
- Mix of problem solving and data analysis
- Prepared using Python notebooks

Collaboration

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had any discussions concerning the problem. You may *not* share written work or code—after discussing a problem with others, the solution must be written by yourself.

Week	Dates	Topics	Lecture Materials	Assignments & Exams	Readings
1	Sept 2	Course overview	Slides		
2	Sept 7, 9	Python and background concepts	Example notebook  Open in Colab Slides	Thu: Quiz 0	Data8 Chapters 3, 4, 5
3	Sept 14, 16	Linear regression and classification	Classification examples  Open in Colab	Tue: Assn 1 out	
4	Sept 21, 23	Stochastic gradient descent	SGD examples  Open in Colab	Tue: Quiz 1 Thu: Assn 1 in; Assn 2 out	
5	Sept 28, 30	Bias and variance, cross-validation	Bias-variance tradeoff  Open in Colab		
6	Oct 5, 7	Tree-based methods	Trees and forests  Open in Colab	Tue: Assn 2 in; Assn 3 out	
7	Oct 12, 14	PCA and dimension reduction	PCA examples  Open in Colab	Tue: Quiz 2 Thu: Assn 3 in; Assn 4 out	
8	Oct 19	Midterm exam (in class)			

9	Oct 26, 28	Language models, word embeddings	Word embeddings  Open in Colab		
10	Nov 2, 4	Bayesian inference, topic models	Bayesian inference  Open in Colab Topic models  Open in Colab	Tue: Assn 4 in; Assn 5 out	
11	Nov 9, 11	Introduction to neural networks	Minimal neural network  Open in Colab	Thu: Assn 4 in; Assn 6 out	
12	Nov 16, 18	Deep neural networks	Autoencoder examples  Open in Colab	Tue: Quiz 3	
13	Nov 19-28	No class, Thanksgiving break			
14	Nov 30, Dec 2	Reinforcement learning		Tue: Assn 6 in; Assn 7 out	
15	Dec 7, 9	Societal issues for machine learning		Tue: Quiz 4 Thu: Assn 7 in	

Auditing

- Auditors are welcome!
- Full access to Canvas
- Just expected to regularly attend class

Questions?