STAT/CSE 416: Intro to Machine Learning

Welcome

Emily Fox
University of Washington
March 27, 2018

Machine learning is changing the world
Old view of ML

Data → ML Algorithm → My curve is better than your curve → Write a paper

Disruptive companies differentiated by INTELLIGENT APPLICATIONS using Machine Learning

©2018 Emily Fox
What is machine learning?

Generically...

Study of algorithms that improve their performance at some task with experience.
The machine learning pipeline

Data → ML Method → Intelligence

ML case studies
Case Study 1: 
Predicting house prices

Data → Regression → Intelligence

Price ($) = ??

What is regression?
From features to predictions

Data → Regression → Intelligence

Input x: features derived from data

Learn $x \rightarrow y$
relationship

Predict y: continuous “output” or “response” to input
Salary after STAT/CSE 416

- How much will your salary be? \( y = \$\$ \)
- Depends on \( x = \) performance in courses, quality of programming assignments, \# of discussion responses, ...

Stock prediction

- Predict the price of a stock \( y \)
- Depends on \( x = \)
  - Recent history of stock price
  - News events
  - Related commodities
Tweet popularity

• How many people will retweet your tweet?
• Depends on # followers, # of followers of followers, features of text tweeted, popularity of hashtag, # of past retweets,...

Reading your mind

Inputs x are brain region intensities
Case Study 2: Sentiment analysis

Sushi was awesome, the food was awesome but the service was awful.

All reviews:

Score(x) > 0

Score(x) < 0

Sushi was awesome, the food was awesome, but the service was awful.

What is classification?

From features to predictions

Input x: features derived from data

Learn $x \rightarrow y$ relationship

Predict y: categorical “output”, class or label
Spam filtering

Input: $x$  
Output: $y$

Text of email, sender, IP,...

Not spam  
Spam

Multiclass classifier

Output $y$ has more than 2 categories

Input: $x$

Webpage

Output: $y$

Education  
Finance  
Technology
Image classification

Input: \( x \) Image pixels
Output: \( y \) Predicted object

Personalized medical diagnosis

Input: \( x \)
Output: \( y \)

Disease Classifier MODEL

Healthy
Cold
Flu
Pneumonia
...
Reading your mind

Output $y$

“Hammer”

“House”

Inputs $x$ are brain region intensities

Case Study 3:
Document retrieval

Data $\rightarrow$ Nearest neighbor $\rightarrow$ Intelligence
What is retrieval?

Search for related items

Input $x, \{x'\}$: features for query point + features of all other datapoints

Compute distances to other $x'$

Output $x^{NN}$: “nearest” point or set of points to query

Retrieve “nearest neighbor” article

Space of all articles, organized by similarity of text
Or set of nearest neighbors

Space of all articles, organized by similarity of text

Retrieval applications

Just about everything...

Streaming content:
- Songs
- Movies
- TV shows
- ...

Images

Products

Social networks
(people you might want to connect with)
Case Study 3++:
Document structuring for retrieval

What is clustering?
Discover groups of similar inputs

Input \{x\}: features for points in dataset
Separate points into disjoint sets
Output \{z\}: cluster labels per datapoint
Clustering images

For search, group as:
- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...

Or users on websites...

Discover groups of users for better targeting of content
Embedding
Example: Embedding images to visualize data

Images with thousands or millions of pixels

Can we give each image a coordinate, such that similar images are near each other?

[Saul & Roweis ’03]

Case Study 4:
Product recommendation

Your past purchases:
+ purchase histories of all customers

Recommended items:
Recommender systems applications

Movies

Songs

Friends, apps, …

Case Study 5:
Visual product recommender
What is (supervised) deep learning?

Flexible method for performing classification or regression

Input \( \{x\} \): raw data or extracted features for points in dataset

Nonlinear feature representation

Output \( \{z\} \): label or value

ImageNet 2012 competition: 1.2M training images, 1000 categories

Top 3 teams

SuperVision

ISI

OXFORD_VGG

Huge gain

Exploited hand-coded features like SIFT
Examples of deep learning success stories

• Image classification
• Image segmentation
• Image captioning
• Object detection
• Speech recognition
• Speech synthesis
• Machine translation
• Handwriting recognition
• ...

Other ML topics we won’t cover

• Reinforcement learning
• Learning theory
• Active learning
• Multi-task and transfer learning
• Spectral methods
• ...
Will learn about the ML pipeline...

- Data
- ML Method
- Intelligence
- Task
- Model & parameters
- Optimization algorithm
- Evaluation

Use pre-specified (black box)
Level of the course

Motto:  
*tough concepts made intuitive and applicable*

minimize prereq knowledge  
maximize ability to develop and deploy  
learn concepts through case studies

Detailed topics

**Models**
- Linear regression, regularized approaches (ridge, Lasso)
- Linear classifiers: logistic regression
- Non-linear models: decision trees
- Nearest neighbors, clustering
- Recommender systems
- Deep learning

**Algorithms**
- Gradient descent
- Boosting
- K-means

**Concepts**
- Point estimation, MLE
- Loss functions, bias-variance tradeoff, cross-validation
- Sparsity, overfitting, model selection
- Decision boundaries
Prerequisites

- Formally:
  - Either CSE 143 or CSE 160; either STAT 311 or STAT/MATH 390 or STAT 391
- Basic Probability + Statistics
  - Distributions, densities, independence, marginalization, conditioning, expectation, variance...
- Programming
  - Python will be very useful, but we’ll help you get started

- We provide some background, but the class will be fast paced!
- Ability to deal with “abstract mathematical concepts”
Computing needs

• Everything will be on JupyterHub
  – Just need to log in
  – No need to install and run Python locally
  – Email sent with username/password

iPython notebooks are the thing!!!
(Real tool people use)
JupyterHub will make things seamless

Course staff + office hours

Instructor:
• Emily Fox
  – Office hours: Thursdays, 11:30am – 12:30pm, CSE 568

TAs:
• Devin Didericksen
  – Office hours: Tuesday 3:30 – 5:00pm, 3rd floor CSE breakout

• Varun Mahadevan
  – Office hours: Wednesdays, 12:30 – 2pm, 5th floor CSE breakout

• John Kaltenbach
  – Office hours: TBA

• Hunter Schafer
  – Office hours: Mondays, 12:30 – 2pm; Tuesdays 12:30 – 1:30pm, CSE 220

• Patrick Spieker
  – Office hours: Wednesdays and Fridays, 10:30 – 11:30am, 3rd floor CSE breakout
Quiz Sections

• Important to attend weekly
• Topics:
  – Intros to and demos of running things in Python
  – Reinforcing concepts from lecture
  – Bonus material to supplement lectures

Communication channels – us to you

• Course email list
  – Announcements from us. Please check your email!

• Course website
  – https://courses.cs.washington.edu/courses/cse416/18sp/
  – Lecture slides, quiz section handouts, high-level (static) course info

• Canvas
  – Discussion board, access to concept quizzes, submissions of work, and grades

• Google calendar
  – Live updates to schedules (also via email to course mailing list)
  – Shared url to be announced...stay tuned
Communication channels – you to us/each other

- **Canvas discussion board**
  - For all non-personal questions
  - Answering your question will help others
  - Feel free to (and please do!) chime in
  - **Guidelines and expectations:**
    - Look through threads before posting a new one
    - Reflect on question before posting
    - Our goal is to respond within 24 hrs

- **Instructor email list:** cse416-staff@cs.washington.edu
  - Only for personal issues

Textbooks

- **None! Come to lectures and quiz sections**
  - Annotated slides will be posted
  - Quiz section handouts will be posted
  - Blog posts and other sources will sometimes be referenced, too

- **Optional Books:**
  - A Course in Machine Learning; Hal Duame III
    [http://ciml.info](http://ciml.info)
  - Machine Learning: A Probabilistic Perspective; Kevin Murphy
  - Pattern Recognition and Machine Learning; Chris Bishop
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman
Programming assignments

Programming assignments are hands-on experience with ML methods on real data. The assignments are hard, start early 😊

Submission procedure and late policy:
• Use Canvas to submit code and answers related to running the code
• 2 late days per quarter, and then 33% subtracted per late day
• All assignments must be handed in, even for zero credit

Collaboration policy:
• You may discuss the questions
• Each student must write their own code and submit their own answers
  – We will be using a cheating detection software
• Submit the names of anyone with whom you collaborate
• Please don’t search for answers on the web, Google, etc.
  – please ask us if you are not sure if you can use a particular reference

Exams

• Concept quizzes
  – Online!!!
  – Spread throughout the quarter
  – At least one per major topic
  – Primary purpose is to make sure you are following content
  – Must be completed 100% individually

• Final
  – Finals week
  – Monday, June 4, 10:30-12:20 in MLR 301
Grading

- Programming assignments (60%)
  - Bonus Assignment 0 to get setup with tools (0%)

- Concept quizzes (15%)
  - Bonus Concept quiz 0 to refresh prob/stat background (0%)

- Final (25%)

Getting started in CSE 416

- Concept quiz 0
  - Recall basic prob/stat topics
- Programming assignment 0
  - Intro to iPython notebooks and Turi Create tutorial

- Resources:
  - Java-to-Python guide (thanks to Hunter!)
  - Videos on Python and Turi Create fundamentals
  - Quiz section intro to running things on JupyterHub
You’ll be able to do amazing things...