CSE 446: Machine Learning

Welcome

Emily Fox
University of Washington
January 4, 2017

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
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

Regression
Example: Predicting house prices

Data → Regression → Intelligence

$ = ??$
What is regression?

From features to predictions

- **Input x:** features derived from data
- **Learn x → y relationship**
- **Predict y:** continuous “output” or “response” to input

**Salary after CSE 446**

- How much will your salary be? (y = $$)
- Depends on x = performance in course, quality of project, # of discussion board 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? (y)
• Depends on $x =$ # followers,
  # of followers of followers,
  features of text tweeted,
  popularity of hashtag,
  # of past retweets,...
Classification
Example: Sentiment analysis

Inputs $x$ are brain region intensities

Output $y$

very sad

very happy

Classification

Data

Intelligence

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

All reviews:

"awesome"

"awful"

Score($x$) > 0

Score($x$) < 0

Count "awesome"

Count "awful"

positive

negative
What is classification?

From features to predictions

Data -> Classifier -> Intelligence

Input x: features derived from data
Learn $x \rightarrow y$ relationship
Predict y: categorical "output", class or label

Spam filtering

Text of email, sender, IP,...

Not spam

Spam

Input: $x$  Output: $y$
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
Similarity/finding data
Example: Document retrieval

What is retrieval?
Search for related items
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...

Social networks (people you might want to connect with)

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

Products

Images

News articles

Clustering

Example: 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
- ...

©2017 Emily Fox
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?
Embedding words

Embedding words (zoom in)
Deep Learning
Example: Visual product recommender

What is (supervised) deep learning?
Flexible method for performing classification or regression
ImageNet 2012 competition: 1.2M training images, 1000 categories

Top 3 teams

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

- Recommender systems
- Reinforcement learning
- Learning theory
- Active learning
- Multi-task and transfer learning
- Spectral methods
- ...

Syllabus
Will learn about the ML pipeline...

**Detailed topics**

**Models**
- Linear regression, regularized approaches (ridge, Lasso)
- Linear classifiers: logistic regression, SVMs, perceptron
- Non-linear models: decision trees
- Nearest neighbors, clustering, mixtures of Gaussians
- PCA
- Naive Bayes, graphical models
- Deep learning

**Algorithms**
- Gradient descent, stochastic gradient descent
- Boosting
- K-means
- Expectation-maximization (EM)
- Eigen decomposition, SVD

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

- Formally:
  - Either CSE 326 or CSE 332; either STAT 390, STAT 391, or CSE 312
- Probabilities
  - Distributions, densities, marginalization...
- Basic statistics
  - Moments, typical distributions, regression...
- Algorithms
  - Dynamic programming, basic data structures, complexity...
- 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”
Tutorials

• Python
  - Thurs, Jan 5 (tomorrow!)
  - No 12:30-1:20pm section

• Linear algebra
  - Thurs, Jan 12

• Midterm review
  - Thurs, Feb 2

• Final review
  - Thurs, Mar 9

Communication channels

• Catalyst discussion board
  - For all non-personal questions
  - Answering your question will help others
  - Feel free to (and please do!) chime in

• Instructor email list: cse446-staff@cs.washington.edu
  - Only for personal issues
Course staff + office hours

Instructor:
- Emily Fox
  - Office hours: Fridays, 10:30-11:30am, Padelford B-305

TAs:
- Dae Hyun Lee
  - Office hours: Mondays, 2-3pm, CSE 021
- Sachin Mehta
  - Office hours: Wednesdays, 2-3pm, CSE 220
- David Wadden
  - Office hours: Thursdays, 2-3pm, CSE 021
- Kaiyu Zheng
  - Office hours: Tuesdays, 2-3pm, CSE 4th floor breakout

Textbooks

- Required textbook:
  - Machine Learning: A Probabilistic Perspective; Kevin Murphy

- Optional Books:
  - Pattern Recognition and Machine Learning; Chris Bishop
  - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman
  - Machine Learning; Tom Mitchell
  - Information Theory, Inference, and Learning Algorithms; David MacKay
Homeworks

Homeworks are hard, start early 😊

Submission procedure and late policy:
- Use Catalyst to submit homeworks
- Due before the start of class
- 33% subtracted per late day
- All homeworks must be handed in, even for zero credit

Collaboration policy:
- You may discuss the questions
- Each student writes their own answers
- Write on your homework anyone with whom you collaborate
- Each student must write their own code for the programming part
- Please don’t search for answers on the web, Google, previous years’ homeworks, etc.
  - please ask us if you are not sure if you can use a particular reference

Exams

- Midterm
  - In-class
  - Weds, Feb 8

- Final
  - Finals week
  - Date/time/location TBD
  - (Likely Weds, Mar 15, 8:30-10:20am)
Project

- Projects can be selected from a list of ideas or proposed based on your interests
- Make sure you have data available and a clear roadmap. *Quarter is short!*
- Can work as an individual or teams of 2

Project proposals
- Mon., Feb 6 by 9:30am

Project milestone
- Fri., Feb 24 by 9:30am

Poster session
- Fri., Mar 10, 2-4pm in CSE Atrium

Project report
- Mon., Mar 13 by 9:30am

Grading

- 4 homeworks (40%)
  - First homework due January 20\textsuperscript{th}

- Final project (20%)

- Midterm (15%)

- Final (25%)
Course overload process

Majors:
• Within **60 minutes** of this class ending, fill out the course overload form here: [http://tinyurl.com/hz9sxd](http://tinyurl.com/hz9sxd)
• Need code word: `bison`

Non-majors:
• Unfortunately, the course has a huge number of majors waiting and there is almost no chance of getting in
• Please talk to the CSE ugrad advisors in a week or two

Other exciting ML opportunities

• Next year: new non-majors ML course (CSE/STAT 416)

• STAT 435 – Introduction to Statistical Machine Learning
• INFX 574 – Data Science II: Machine Learning and Econometrics
• BIOST 546 – Machine Learning for Biomedical and Public Health Big Data
• DATA 558 – Statistical Machine Learning for Data Scientists
You’ll be able to do amazing things…