Machine learning lecture slides

COMS 4771 Fall 2020

Overview

Questions

- Please use Piazza Live Q&A

Outline

- A "bird's eye view" of machine learning
- About COMS 4771

Figure 1: Predict the bird species depicted in a given image.

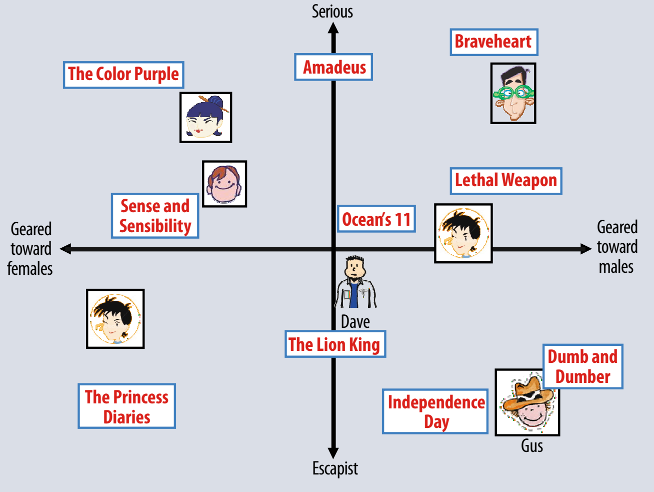

Figure 2: Predict how a given user would rate a given movie.

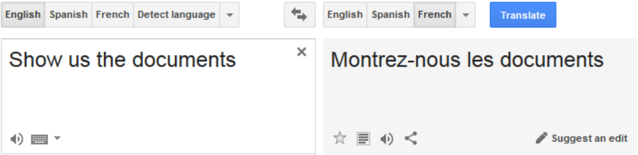

Figure 3: Predict the French translation of a given English sentence.

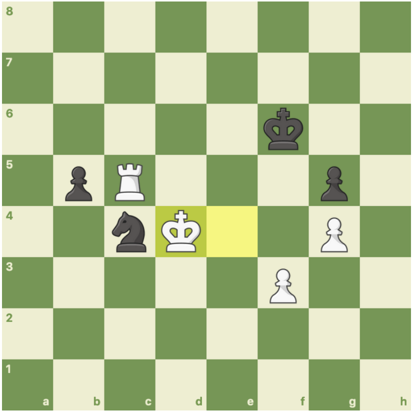

Figure 4: Predict the “win probability” of a given move in a given game state.

How to “solve” problems without ML?

- Image classification:
  - Recruit a “bird expert” to teach you about different birds features (e.g., beak shape, feather color, typical environment)
  - Recognize these features in a given image, and then come up with a best guess of the bird species
- Recommender system:
  - Ask user to self-report specific movie genres of interest (e.g., horror, sci-fi)
  - Ask movie suppliers to categorize movies into the same genres
  - Predict a high rating for any movie in a user’s genre-of-interest; low rating for all other movies
- Machine translation: . . .
- Chess: . . .

Work in ML

- Applied ML
  - Collect/prepare data, build/train models, analyze performance/errors/etc
- ML developer
  - Implement ML algorithms and infrastructure
- ML research
  - Design/analyze models and algorithms

Note: These roles are not mutually exclusive!

Mathematical and computational prerequisites

- Math
  - Linear algebra, probability, multivariable calculus
  - Understand and reason about the concepts (not just calculations)
- Software/programming
  - Much ML work is implemented in python with libraries such as numpy and pytorch

Basic setting: supervised learning

- *Training data*: dataset comprised of *labeled examples*
  - *Labeled example*: a pair of the form (input, label)
  - *Input*: what you see before you make a prediction (a.k.a. context, side-information, features, etc.)
  - *Label*: output value (a.k.a. output, response, target, etc.)
- Goal: learn *predictor* (i.e., prediction function) to predict label from input for new examples

if age ≥ 40 **then**
    if genre $=$ western **then**
        return 4.3
    else if release date > 1998 **then**
        return 2.5
    else
        $\vdots$
    end if
else if $\cdots$ **then**
    $\vdots$
end if

Figure 5: Decision tree

if $0.335 \cdot x_1 \; + \; 2.5 \cdot x_2 \; + \; \cdots \; + \; 6.35 \cdot x_{10^6} \; > \; 4.3$ **then**
    return spam
else
    return not spam
end if

Figure 6: Linear classifier (“Perceptron”)

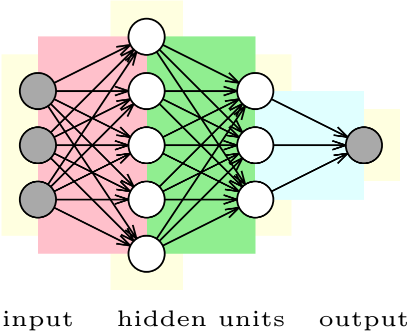

Figure 7: Neural network

Types of prediction problems

- Binary classification
  - Given an email, is it spam or not?
  - (What's the probability that it is spam?)
- Multi-class classification
  - Given an image, what animal is depicted?
  - (Or which animals are depicted?)
- Regression
  - Given clincal measurements, what is level of tumor antigens?
  - (In absolute level? Log-scale?)
- Structured output prediction
  - Given a sentence, what is its grammatical parse tree?
  - (Or dependency tree?)
- ...

Template of supervised learning pipeline

- Get data
- Determine representation of and predictive model for data
- Train the predictor (a.k.a. model fitting, parameter estimation)
- Evaluate predictor (test the “goodness of fit”)
- Deploy predictor in application

Questions

- What is the core prediction problem?
- What *features* (i.e., predictor variables) are available?
- Will these features be available at time of prediction?
- Is there enough information (e.g., training data, features) to learn the relationship between the input and output?
- What are the modeling assumptions?
- Is high-accuracy prediction a useful goal for the application?

Where do assumptions / domain expertise come in?

- Form of the prediction function
- Choice of features
- Choice of training data
- Choice of learning algorithm
    - Choice of objective function and contraints

Challenges

- Might not have the “right” data
- Might not have the “right” model
- Might under-fit the data
- Might over-fit the data
- Data might be corrupted, noisy, . . .

Key statistical/algorithmic ideas in ML

- Plug-in principle
- Inductive bias
- Linearity
- Mathematical optimization

About COMS 4771

- Basic principles and methods of supervised machine learning
  1. Appetizer: nearest neighbor rules (a "non-parametric" method)
  2. Statistical model for prediction
  3. Regression
     - Why? Clean, simple, and illustrates important concepts (linearity, inductive bias, regularization, kernels)
  4. Classification
  5. Optimization methods for machine learning
     - Convex optimization, neural networks
  6. Maybe one other topic if time permits ...
- This is not a course about how to use sklearn, tensorflow, etc.
- Also not about latest nonsense on arXiv
- Good stuff beyond COMS 4771:
  - COMS 4252, 4773: Mathematical theory of learning
  - COMS 4774: Unsupervised learning
  - COMS 4775: Causal inference
  - ...

About me

- Professor Daniel Hsu
  - Okay to call me "Daniel"!
  - "Professor Hsu" also okay
    - "Professor Daniel" is a little weird
  - At Columbia since 2013
  - Previously at Microsoft Research, Rutgers, UPenn, UC San Diego, UC Berkeley, . . .
  - Research interests: algorithms, statistics, & combining the two
  - Good at: LaTeX-hacking
  - Bad at: making slides

About you

- I assume you have fluency in
  - multivariable calculus,
  - linear algebra, and
  - elementary probability (no measure theory needed)
- I also assume you can read and write programs in Python
  - (and read online documentation to learn, e.g., how to do I/O with CSV files)
  - See Courseworks for a "Python basics" Jupyter notebook to brush up on Python, Numpy, etc.
- Let me know why you are interested in ML!
  - Part of HW 1.

Administrative stuff

- Website: `https://www.cs.columbia.edu/~djhsu/ML`
  - Schedule for office hours/lectures/homework/quizzes/exam
  - Syllabus
- Course format:
  - Lecture/recitation: online over Zoom
  - "On Campus" people: check email about in-person lectures
- Course assistants (CAs):
  - Andy, Andrea, Wonjun, Serena
  - Links for online office hours will be posted on Courseworks
- Technology:
  - Piazza: communicate with course staff (live Q&A and offline)
  - Courseworks: retrieve assignments, quizzes, data files, etc.
  - Gradescope: submit homework write-ups, code
  - Slack: discussion with fellow classmates
- Disability services:
  - Please make arrangements with disability services ASAP

Academic rules of conduct

- See syllabus
- **Cheating**: don't do it
  - If unsure about something, ask!
  - Consequence is automatic fail
- **Cheating out of desperation** is also cheating
  - Instead: get help early
  - We are here to help
- Okay to work on homework in groups of ≤ 3
  - No collaboration across groups
  - No diffusion of responsibility
- No collaboration at all on quizzes or exams

Reading assignments

- There are some required reading assignments (mostly from handouts posted on website)
- Unfortunately, most textbooks on ML are not appropriate for this course
  - Closest is "A Course in Machine Learning" by Daumé
  - I have selected some optional reading assignments from a few books that may be used to supplement the lectures
  - All books available online

Questions?