Machine Learning for Big Data (CSE 599)

Statistics for Big Data (STAT 592)

(Or how to do really kickass research in the age of big data)
Course Staff

Instructors:
• Emily Fox (Stat)
• Carlos Guestrin (CSE)

TAs:
• Jay Gu (CSE)
• Linda Li (Stat)
CONTENT

What is the course about?
Course Structure

• 4 “case studies”
  – Estimating Click Probabilities
  – Document Retrieval
  – fMRI Prediction
  – Collaborative Filtering

• Not comprehensive, but a sample of tasks and associated solution methods

• Methods broadly applicable beyond these case studies
1. Estimating Click Probabilities

- **Goal:** Predict whether a person clicks on an ad
- **Basic method:** logistic regression, online learning
1. Estimating Click Probabilities

- **Challenge 1:** Overfitting, high-dimensional feature space
- **Advanced method:** L2 regularization, hashing
1. Estimating Click Probabilities

- **Challenge II**: Dimension of feature space changes
  - New word, new user attribute, etc.
- **Advanced method**: sketching, hashing
2. Document Retrieval

• **Goal**: Retrieve documents of interest

• **Methods**: fast K-NN, k-means, mixture models, spectral clustering, Hadoop
2. Document Retrieval

• **Challenge:** Document may belong to multiple clusters

• **Methods:** mixed membership models (e.g., LDA)
3. fMRI Prediction

- **Goal:** Predict word probability from fMRI image
- **Challenge:** $p >> n$ (feature dimension $>>$ sample size)
- **Methods:** L1 regularization (LASSO), parallel learning

[Diagram of brain scan with model and parameters]
3. fMRI Prediction

- **Goal:** Predict fMRI image for given stimulus
- **Challenge:** zero shot learning (generalization)
- **Methods:** features of words, Mechanical Turk, graphical LASSO
4. Collaborative Filtering

- **Goal:** Find movies of interest to a user based on movies watched by the user and others
- **Methods:** matrix factorization, GraphLab
What do I recommend???
4. Collaborative Filtering

- **Challenge:** Cold-start problem (new movie or user)
- **Methods:** use features of movie/user
Scalability

• Throughout case studies, introduce notions of parallel learning and distributed computations
Assumed Background

Comfortable with:
- Linear regression
- Basic optimization (e.g., gradient descent)
- EM algorithm
- Java

Have seen:
- Graphical models (as a representational tool)
- Gibbs sampling

Computational and mathematical maturity
LOGISTICS

How is the course going to operate?
Website and Google Group

• Course website:

• Google Group:
  – Used for all discussions
  – Post all questions there (unless personal)
  – See website for sign-up details
Reading

• No req’d textbook, but background reading in:
  “Machine Learning: A Probabilistic Perspective”
  Kevin P. Murphy

• Readings will be from papers linked to on course website

• Please do reading before lecture on topic
Homework

• 4 HWs, one for each case study
• Collaboration allowed, but write-ups and coding must be done individually
• Submitted at beginning of class
• Allowed 2 “late days” for entire quarter
• 3rd assignment must be completed individually
Project

• Individual, or teams of two
• New work, but can be connected to research
• Schedule:
  – Proposal (1 page) – January 31
  – Progress report (3 pages) – February 21
  – Poster presentation – March 14
  – Final report (8 pages, NIPS format) – March 19
Grading

• HWs 1, 2, 4 (15% each)
• HW 3 (20%) – midterm exam
• Final project (35%)
Support/Resources

• Office Hours
  – TAs: MW 4-5pm in CSE 216
    T 3-5pm in CSE 220
  – Emily: Th 12:45-1:45pm in Padelford B-305
  – Carlos: F 1:30-2:30pm in CSE 568

• Recitations
  – Optional tutorial/example-based sections will be held weekly on Thursdays from 5:30-7pm
  – MUE 153, to be confirmed
Conclusion

• I like Big Data and I cannot lie

[INSERT SONG HERE]

Or, let’s just carry on with the first lecture...