Machine Learning for Big Data
(CSE 547 / STAT 548)

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

Course Staff

Instructor:
• Emily Fox

TAs:
• Marco Ribeiro
• Alden Timme
CONTENT

What is the course about?

Course Structure

• 5 “case studies”
  – Estimating Click Probabilities
  – Document Retrieval
  – fMRI Prediction
  – Collaborative Filtering
  – Document Mixed Membership Modeling

• 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 I**: 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, Hadoop
3. fMRI Prediction

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

---

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, latent factor models, GraphLab

What do I recommend???
4. Collaborative Filtering

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

5. Document Mixed Membership

- **Challenge**: Document may belong to multiple clusters
- **Methods**: mixed membership models (e.g., LDA), distributed Gibbs, stochastic variational inference
Scalability

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

Assumed Background

Official Prereq (strict): CSE 546 or STAT 535

Specific topics:
• Linear and logistic regression, ridge regression, LASSO
• Basic optimization (e.g., gradient descent, SGD)
• Perceptron algorithm
• K-NN, k-means, EM algorithm

Comfortable with:
• Java or Python
• Probabilistic and statistical reasoning

Computational and mathematical maturity
LOGISTICS

How is the course going to operate?

Website and Catalyst

- Course website: http://www.cs.washington.edu/education/courses/cse547/15sp/

- Catalyst:
  - Used for all discussions
  - Post all questions there (unless personal)
  - Homework collection
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, approx one for each case study
• Collaboration allowed, but write-ups and coding must be done individually
• On due date, due at beginning of class time
• Allowed 2 “late days” for entire quarter
• 3rd assignment must be completed individually
  → “Midterm”
Project

- Individual, or teams of two
- New work, but can be connected to research
- Schedule:
  - Proposal (1 page) – April 21
  - Progress report (3 pages) – May 14
  - Poster presentation – *Friday*, June 5, 4:00-6:00pm (??)
  - Final report (8 pages, NIPS format) – June 9

Grading

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

• Office Hours
  – TAs: M 10-12, T 1:30-3:30 (CSE 218)
  – Emily: Th 11-12 in CSE 346
• Blog Posts
• Discussion Board

Conclusion

• I like Big Data and I cannot lie

[INSERT SONG HERE]

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