[70240413 Statistical Machine Learning, Spring, 2017]

Statistical Machine Learning
Theory and Applications

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.Tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

February 21, 2017

A bit about the Instructor

- Jun Zhu, Associate Professor, Depart. of Computer Science & Technology. I received my Ph.D. in DCST of Tsinghua University in 2009. My research interests include statistical machine learning, Bayesian nonparametrics, and data mining
- I did post-doc at the Machine Learning Department in CMU with Prof. Eric P. Xing. Before that I was invited to visit CMU for twice. I was also invited to visit Stanford for joint research (with Prof. Li Fei-Fei)
- 2015: Adjunct Associate Professor at CMU

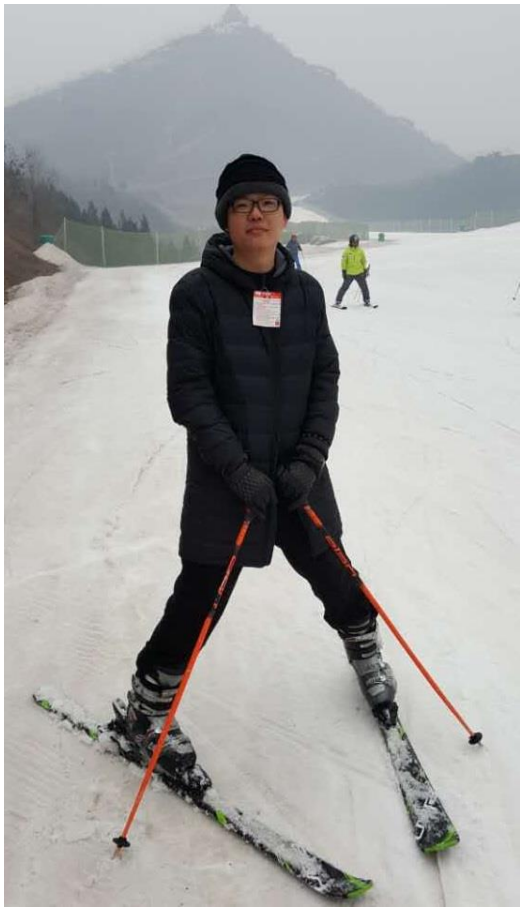
- Published 80+ research papers on the top-tier ML conferences and journals, including JMLR, TPAMI, ICML, NIPS, etc.
- Served as Area Chairs for ICML, NIPS, UAI, AAAI, IJCAI; Associate Editor for PAMI, AI Journal
- Research is supported by National 973, NSFC, "Tsinghua 221 Basic Research Plan for Young Talents".
- Homepage: http://bigml.cs.tsinghua.edu.cn/~jun

Contact Information

- Jun Zhu
  - State Key Lab of Intelligent Technology and Systems, Department of Computer Science, Tsinghua U.
  - Office: Rm 4-513, FIT Building
  - E-mail: dcszj@tsinghua.edu.cn
  - Phone: 62772322, 18810502646
  - Office hours: Thursday afternoon 3:00pm-5:00pm

Teaching Assistants

- Chongxuan Li (Head TA)
  - Office: Rm 1-509, FIT Building
  - E-mail: chongxuanli1991@gmail.com
  - Phone: 62795869, 15201523592
  - Deep Learning, Latent variable models, Bayesian inference
  - Publish at NIPS, ICML, etc.
  - http://bigml.cs.tsinghua.edu.cn/~chongxuan/



Teaching Assistants

- Jiaxin Shi
  - E-mail: ishijiaxin@126.com
  - Phone: 62795869, 18810690095
  - Deep learning
  - Publish at VAST, NIPS

- Yucen Luo
  - E-mail: luoyucencen@163.com
  - Phone: 62795869, 18810301080
  - Deep learning, Latent variable models
  - Publish at ICML.

- Yong Ren
  - E-mail: reny11@foxmail.com
  - Phone: 62795869, 17601648338
  - Deep learning, Bayesian methods, Optimization
  - Publish at NIPS, PAMI

- TA office hours: Wednesday afternoon 3:00pm-5:00pm
- Office: Rm 1-508/509, FIT Building

Resources

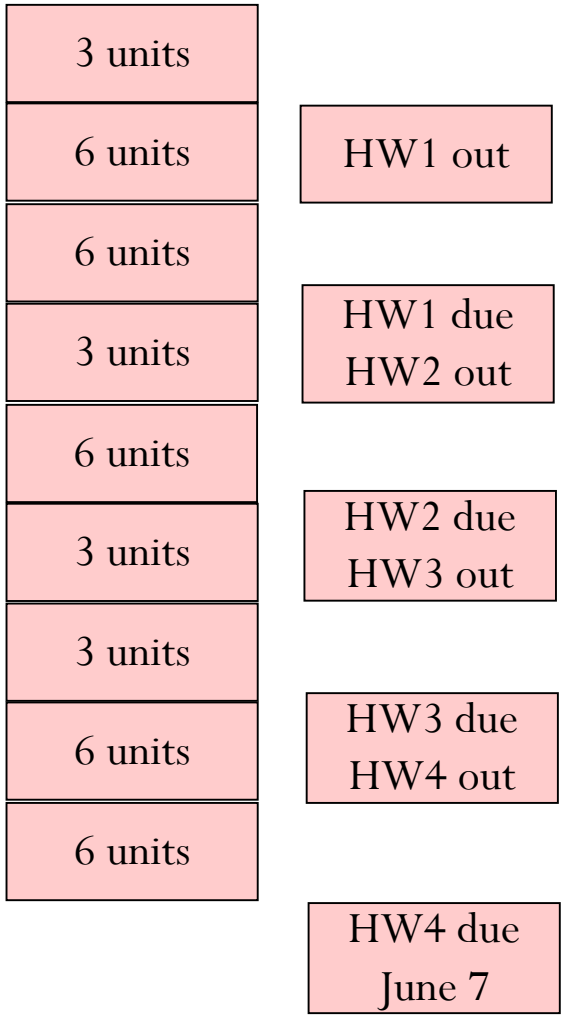
- Mainly class slides/notes
- Recommended text books
  - Christopher M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2007.
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning*. 2nd Edition, Springer, 2009.
- Further readings:
  - Conferences:
    - Theory: ICML, NIPS, UAI, COLT, AISTATS, AAAI, IJCAI
    - App: KDD, SIGIR, WWW, ACL
  - Journals:
    - JMLR, PAMI, MLJ

Prerequisites

- Knowledge of probability, linear algebra, statistics and algorithms
  - Calculus:
    - Derivative, integral of multivariate functions
  - Linear Algebra
    - Matrix inversion, eigen-decomposition, …
  - Basic Probability and Statistics
    - Probability distributions, Mean, Variance, Conditional probabilities, Bayes rule, …
- Knowledge of programming languages, e.g., C/C++, Java, matlab, Python
- **Homework 0**: take the Self-Evaluation
  - Minimum & modest background tests (available at course webpage)

Overview of Class

- Introduction
- Unsupervised learning
- Supervised learning
- Learning theory
- Probabilistic graphical models
- Bayesian methods
- Online learning
- Sparse learning
- Deep learning



Grading

- Participation (10%)
  - 1 mid-term quiz (10 points each time)
- Homeworks (40%)
  - 4 homeworks (10 points each time)
- Project (50%)
  - 2~4 students to form a team
  - Apply machine learning to solve a real problem
    - Choose one task at Kaggle (http://www.kaggle.com/competitions)
  - Submit materials:
    - a proposal (6th week), a mid-term report (9th week), a final report (18th week), and the implementation code (18th week)
  - All reports should be in NIPS format, written in English: (http://nips.cc/Conferences/2014/PaperInformation/StyleFiles)
  - Poster presentation (16th or 17th week)

Some example Kaggle tasks

Data Science Bowl 2017
Can you improve lung cancer detection?
Featured · 2 months to go · 511 kernels
$1,000,000
1,149 teams

The Nature Conservancy Fisheries Monitoring
Can you detect and classify species of fish?
Featured · 2 months to go · 269 kernels
$150,000
1,547 teams

Google Cloud & YouTube-8M Video Understanding Challenge
Can you produce the best video tag predictions?
Featured · 3 months to go · 14 kernels
$100,000
95 teams

Digit Recognizer
Classify handwritten digits using the famous MNIST data
Getting Started · 3 years to go · 2,361 kernels
1,422 teams

Titanic: Machine Learning from Disaster
Predict survival on the Titanic using Excel, Python, R & Random Forests
Getting Started · 3 years to go · 6,074 kernels
5,864 teams

- If the end date is later than June 5, report the position in the leaderboard;
- Otherwise, TAs will define a train/test split and compare your methods with 1 or 2 baselines.

Questions?