Statistical Machine Learning
Theory and Applications

Jun Zhu

dcszj@mail.tsinghua.edu.cn

http://ml.cs.Tsinghua.edu.cn/~jun

State Key Lab of Intelligent Technology & Systems

Tsinghua University

February 27, 2018

A bit about the Instructor

- Jun Zhu, Associate Professor, Depart. of Computer Science. I received Ph.D. in 2009. My research interest includes machine learning, Bayesian methods, 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 100+ 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".

- IEEE AI's 10 to Watch; MIT TR35 China (pioneers)

- Homepage: http://ml.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
    - Better to make an appointment in advance

Teaching Assistants

◆ Jiaxin Shi (Head TA)

- ❑ E-mail: ishijiaxin@126.com
- ❑ Phone: 62795869, 18810690095
- ❑ Bayesian methods, Deep learning
- ❑ Publish at VAST, NIPS, ICLR

◆ Yucen Luo (Head TA)

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

Teaching Assistants

◆ Jialian Li
  ❑ E-mail: lijialian7@163.com
  ❑ Phone: 18510243737
  ❑ Reinforcement learning

◆ Kun Xu
  ❑ E-mail: vofhqn@gmail.com
  ❑ Phone: 15701006589
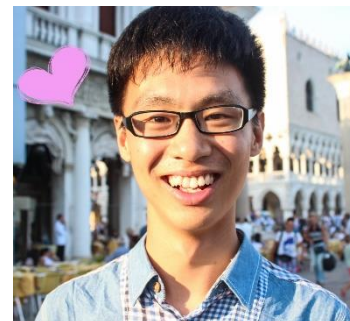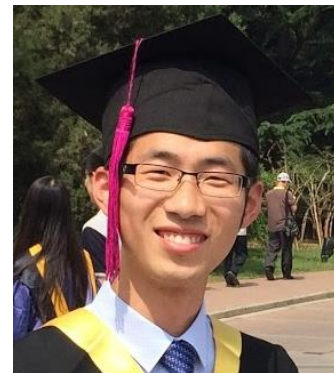 ❑ Bayesian methods, Bandits

◆ Yichi Zhou
  ❑ E-mail: vofhqn@gmail.com
  ❑ Phone: 15701006589
  ❑ Bayesian methods, Bandits

◆ Haosheng Zou
  ❑ E-mail: zouhs16@mails.tsinghua.edu.cn
  ❑ Phone: 18800120568
  ❑ Reinforcement learning

◆ TA office hours: Wed afternoon 3:00pm-5:00pm
◆ Office: Rm 1-508/509, FIT Building; 62795869

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
- Sparse learning
- Deep learning
- Reinforcement Learning

Units
3 units
6 units
6 units
3 units
6 units
3 units
3 units
6 units
6 units

HW1 out

HW1 due
HW2 out

HW2 due
HW3 out

HW3 due
HW4 out

HW4 due
June 7

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)
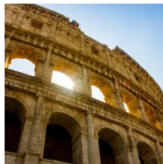
Some example Kaggle tasks
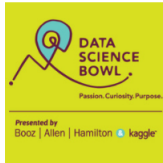
2018 Data Science Bowl
Find the nuclei in divergent images to advance medical discovery
Featured · 2 months to go · 🏷 biology

$100,000
1,743 teams

Google Landmark Recognition Challenge
Label famous (and not-so-famous) landmarks in images
Research · 3 months to go · 🏷 image data

$2,500
81 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

Other Projects

- Self-defined topics
  - Need to propose as early as possible to filter out improper ones

- Other candidates
  - Chinese handwritten characters generation and recognition
  - Adversarial attacks and defense of deep learning
  - Reinforcement learning
  - More to come

◆ If the end date is later than June 12, report the position in the leaderboard;

◆ Otherwise, ask TAs to define a train/test split and compare your methods with 1 or 2 baselines.

About final report

◆ We expect to see

- ❑ Problems (what?)
- ❑ Motivations (why?)
- ❑ Techniques (how?)
- ❑ Results & Analysis (did you verify what you claimed above?)
- ❑ Conclusions

◆ The final report should look like a NIPS technical paper

Questions?