

[70240413 Statistical Machine Learning, Spring, 2016]

# **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 23, 2016

# 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
- ◆ Have published more than 70 research papers on the top-tier ML conferences and journals, including JMLR, IEEE. Trans. PAMI, ICML, NIPS, etc.
- ◆ Served as Area Chair for ICML, NIPS, UAI, AAI, IJCAI; Associate Editor for PAMI
- ◆ 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

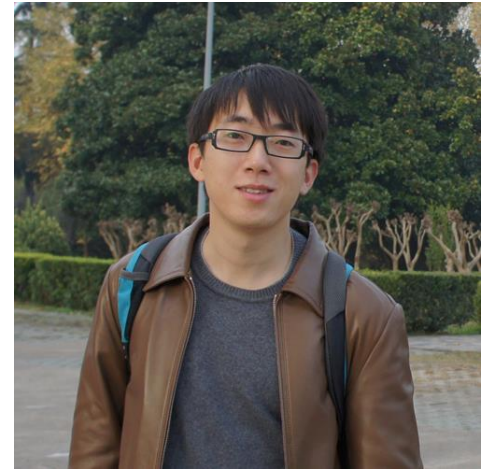
## ◆ Jun Zhu

- 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](mailto:dcszj@tsinghua.edu.cn)
- Phone: 62772322, 18810502646
- Office hours: Thursday afternoon 3:00pm-5:00pm

# Teaching Assistants

## ◆ Tian Tian (Head TA)

- Office: Rm 4-504, FIT Building
- E-mail: [rossowhite@163.com](mailto:rossowhite@163.com)
- Phone: 62795869, 15210588652
- Learning from crowds, Latent variable models, Bayesian inference
- Publish at NIPS, WWW, AAI, etc.
- 2015清华大学研究生特奖、西贝尔奖学金等获得者
- <http://bigml.cs.tsinghua.edu.cn/~tian>



# Teaching Assistants

## ◆ Chongxuan Li

- E-mail: [chongxuanli1991@gmail.com](mailto:chongxuanli1991@gmail.com)
- Phone: 62795869, 15201523592
- Deep learning
- Publish at NIPS

## ◆ Jianfei Chen

- E-mail: [chris.jianfei.chen@gmail.com](mailto:chris.jianfei.chen@gmail.com)
- Phone: 62795869, 18518316949
- Large-scale Machine Learning
- Publish at NIPS, ICML, AAI, WWW, etc.



## ◆ Jingwei Zhuo

- E-mail: [chris.jianfei.chen@gmail.com](mailto:chris.jianfei.chen@gmail.com)
- Phone: 62795869, 15201519430
- Deep learning, Bayesian methods
- Publish at IJCAI



◆ TA office hours: Wednesday afternoon 3:00pm-5:00pm

◆ Office: Rm 4-506, 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*. 2<sup>nd</sup> 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:
    - Derivatives, integrals of multivariate functions
  - Linear Algebra
    - Matrix inversions, eigendecomposition, ...
  - 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




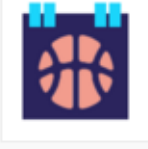



3 units	
6 units	HW1 out
6 units	
3 units	HW1 due HW2 out
6 units	
3 units	HW2 due HW3 out
3 units	
6 units	HW3 due HW4 out
6 units	
	HW4 due June 7



# Grading

- ◆ Participation (10%)
  - 2 random quiz (5 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 (5<sup>th</sup> week), a mid-term report (9<sup>th</sup> week), a final report (16<sup>th</sup> week), and the implementation code (16<sup>th</sup> week)
  - All reports should be in NIPS format, written in English:  
(<http://nips.cc/Conferences/2014/PaperInformation/StyleFiles>)
  - Poster presentation

S

Competition Name	Reward	Teams	Deadline
 <b>Second Annual Data Science Bowl</b> Transforming How We Diagnose Heart Disease	\$200,000	574	21 days
 <b>Home Depot Product Search Relevance</b> Predict the relevance of search results on homedepot.com	\$40,000	1049	2 months
 <b>BNP Paribas Cardif Claims Management</b> Can you accelerate BNP Paribas Cardif's claims management process?	\$30,000	1145	56 days
 <b>March Machine Learning Mania 2016</b> Predict the 2016 NCAA Basketball Tournament	\$25,000	280	19 days
 <b>Telstra Network Disruptions</b> <b>Data Scientist at Telstra</b> Sydney, Melbourne, and other cities in Australia	Jobs	846	7.7 days
 <b>San Francisco Crime Classification</b> Predict the category of crimes that occurred in the city by the bay	Knowledge	1303	3 months
 <b>Digit Recognizer</b> Classify handwritten digits using the famous MNIST data	Knowledge	857	10 months

- ◆ If the end date is later than June 7, report the position in the leaderboard;
- ◆ Otherwise, TAs will define a train/test split and compare your methods with 1 or 2 baselines.

# Petuum

- ◆ You are encouraged to build your own ML experiments / algorithms on Petuum
- ◆ Analyzing Petuum or Developing on Petuum can be a good course project



Petuum is a distributed machine learning framework. It aims to provide a generic algorithmic and systems interface to large scale machine learning, and takes care of difficult systems "plumbing work" and algorithmic acceleration, while simplifying the distributed implementation of ML programs - allowing you to focus on model perfection and Big Data Analytics. Petuum runs efficiently at scale on research clusters and cloud compute like Amazon EC2 and Google GCE.

[View on GitHub](#)

**Questions?**