[70240413 Statistical Machine Learning, Spring, 2015]

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

March 3, 2015

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)
- 2012: visiting professor at CMU
- Have published more than 50 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; 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
  - Phone: 62772322, 18810502646
  - Office hours: Thursday afternoon 3:00pm-5:00pm

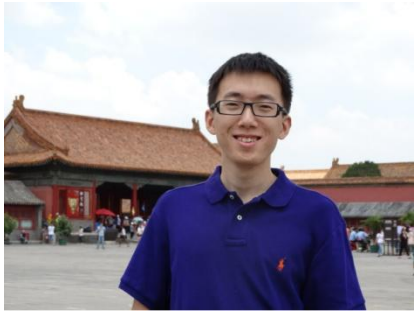
Teaching Assistants

- Minjie Xu (Head TA)
  - Office: Rm 4-506, FIT Building
  - E-mail: chokkyvista06@gmail.com
  - Phone: 62795869, 15901038918
  - Latent variable models, Bayesian nonparametrics, distributed Bayesian inference
  - 2013-2014: visit Oxford University
  - Publish several papers at ICML, NIPS, etc
  - http://bigml.cs.tsinghua.edu.cn/~minjie/

Teaching Assistants

- Wenbo Hu
  - Office: Rm 4-506, FIT Building
  - E-mail: hwb13@mails.tsinghua.edu.cn
  - Phone: 62795869, 13164210311
  - Scalable learning algorithms

- Tian Tian
  - Office: Rm 4-506, FIT Building
  - E-mail: rossowhite@163.com
  - Phone: 62795869, 15210588652
  - Learning from crowds; Bayesian methods



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

Resources

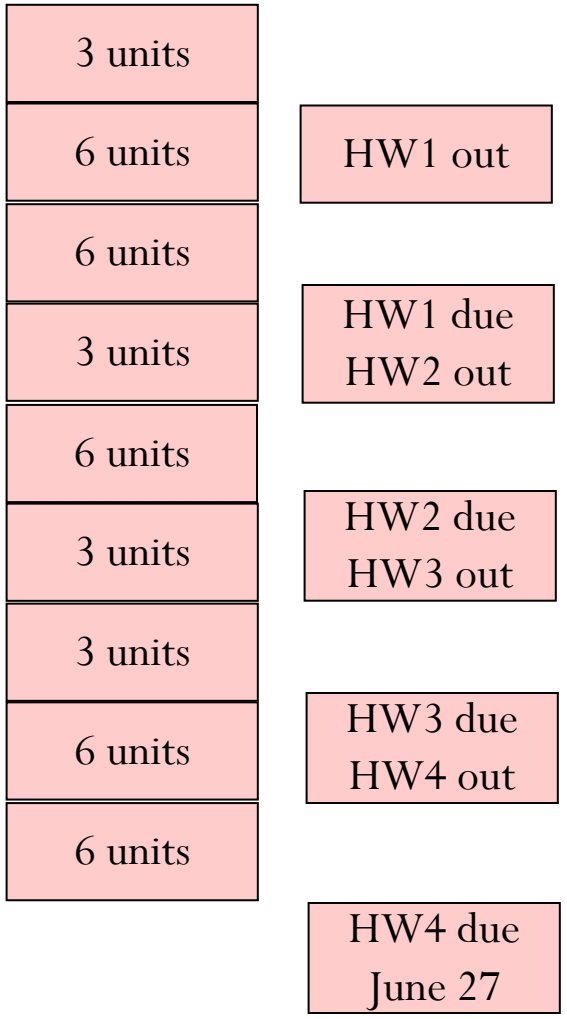
- Mainly class slides/notes
- Recommended text books
  - Christopher M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2007.
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning*. 2nd Edition, Springer, 2009.
- Further readings:
  - Conferences:
    - 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:
    - 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



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 (5th week), a mid-term report (9th week), a final report (16th week), and the implementation code (16th week)
  - All reports should be in NIPS format, written in English: (http://nips.cc/Conferences/2014/PaperInformation/StyleFiles)
  - Poster presentation

Some example Kaggle tasks

- Bag of words meets bags of popcorn **(end June 30)**
  - **Sentiment analysis**
- Diabetic Retinopathy Detection **(end July 27)**
  - **Image processing**
- Microsoft Malware Classification Challenge (BIG 2015) **(end April 17)**
- Large Scale Hierarchical Text Classification
- Higgs Boson Machine Learning Challenge
  - **UCI dataset: https://archive.ics.uci.edu/ml/datasets/HIGGS**

- If the end date is later then June 20, report the position in the leaderboard;
- Otherwise, TAs will define a train/test split and compare your methods with 1 or 2 baselines.

Questions?