Introduction

LING 572

Fei Xia

Outline

- General course information
- Course contents

General course information

Prerequisites

- CS 326 (Data Structures) or equivalent:
  - Ex: hash table, array, tree, …
- Stat 391 (Prob. and Stats for CS) or equivalent: Basic concepts in probability and statistics
  - Ex: random variables, chain rule, Bayes' rule
- Programming in C/C++, Java, Perl, Python, or Ruby
- Basic unix/linux commands (e.g., ls, cd, ln, sort, head): tutorials on unix
- LING570
- **If you don't meet the prerequisites, you should wait and take ling572 later.**

Topics covered in Ling570

- FSA, FST
- LM and smoothing
- HMM and POS tagging
- Classification task and Mallet
- Chunking, NE tagging
- Information extraction

Grades for LING572

- No midterm or final exams.
- Programming Assignments (9): 90%
- Reading assignments (4-5): 10%
- Remove the lowest score to calculate average.
- The average is then mapped to the final grade.

Tentative mapping from the class average to the final grade

98-100	4.0	77-79	3.3
95-97	3.9	74-76	3.2
92-94	3.8	71-73	3.1
89-91	3.7	68-70	3.0
86-88	3.6	65-67	2.9
83-85	3.5	62-64	2.8
80-82	3.4	59-61	2.7

Office hours

- If you have not, please fill out the doodle.

- Fei:
  - Email: fxia@uw.edu
  - Office hours: TBA

TA office hours

- Leanne Rolston:
  - Email: rolston@uw.edu
  - Office hour: TBA
- David Inman: (shared with ling571)
  - Email: davinman@uw.edu
  - No additional office hour

Questions about grades

- Comments and Grades are available at CollectIt and GradeBook resp.

- If you have any questions about hw grades, please email the TAs and cc me within a week after the grade is posted.

Slides

- The slides will be online before class.
- The revised slides (if any) will be uploaded on the same day of the class.
- “Additional slides” are not required and not covered in class.

Attending class remotely

- Url for Adobe: http://uweoconnect.extn.washington.edu/ling572/
- TA will monitor the chat window during class. If you have any questions, please type in your questions in the chat window.
- When you ask questions,
  - provide as complete info as possible
  - respond to our replies promptly
- Adobe Meeting Room is far from perfect. Please be patient.

Recording

- Lectures will be recorded and the urls are posted to GoPost after class.

- Remind me to
  - record the class
  - repeat the questions (if needed)

- Try to speak louder as the mic might not be very sensitive.

Attending live

- Sometimes, recordings do not work well: e.g., no sound.
- Recordings should be used as a supplement only. They are not meant to replace class time.
- If you cannot attend class live most of the time, I highly recommend you NOT to take the course now.

Course website and email

- Course url: **http://courses.washington.edu/ling572**
  - Syllabus (incl. slides, assignments, and papers):
  - GoPost, CollectIt, GradeBook

- Email:
  - Please use your UW email address when emailing us.
  - You should use emails only for confidential subjects (e.g., grades, extended absences, other problems) and short clarification questions.
  - For other questions, please ask in class or during office hours.
  - We will send course-related announcements to the mailing list; so please check your uw emails every day.
  - We try to reply to emails promptly, but if you do not hear back from us soon, please talk to us after class or during office hours.

GoPost

- Recording urls will be posted to GoPost. Other than that, GoPost serves as a discussion board for students ONLY.
- We (the TAs and I) will not check questions on, or post answers to, GoPost.
- If your questions on GoPost are not answered by others, please raise the questions in class or during office hours.

Using GoPost

- You need to decide what's the best way to use GoPost:
  - Pros: By posting to GoPost, you may get questions answered faster by your peers.
  - Cons: Going through posts can be time consuming, and some posts could be misinterpreted if you are not at the same "spot". So there is no need to read and try to understand every post.
  - Rule of thumb: do not spend more than 4 hours per week on GoPost.
  - For complex questions, it is much better to ask them in class or during the office hours.

The structure of GoPost

- Main discussion areas:
  - General information
  - Recordings
  - Hw1, Hw2, …
- A discussion area can have multiple threads, and each thread can have multiple posts.
- Please start a new thread when the subject changes.
- Each thread should have a clear title: e.g., “Q1: …”

Reading assignments

- You will answer some questions about the papers that will be discussed in next class.
- Your answer to each question should be concise and no more than a few lines.
- Your answers are due at 11am. Submit it to CollectIt before class.
- If you make an effort to answer those questions, you will get full credits.

Programming assignments

- Due date: every Wed at 11pm unless specified otherwise.
- The submission area is closed two days after the due date.
- There is 1% penalty for every hour after the due date.

Programming assignments

- Programming languages: C, C++, Java, Perl, or Python
- Write a simple condor submit script.
- For information about how to use patas and condor submit, please see the tutorial at http://staff.washington.edu/brodbd/orientation.pdf
- Your code MUST run on Patas with condor submission.

Homework Submission

- For the assignments, submit the tar file
  - E.g., tar –cvf hw1.tar hw1_dir

- Each submission includes
  - a note file: hw1-notes.(txt | pdf) for hw1.
    - If your code does not work, explain in the note file what you have implemented so far.
  - a set of shell scripts: e.g., kNN.sh
  - source code: e.g., kNN.C
  - binary code (for C/C++/Java): kNN.out
  - data files if any.
  - The TA will NOT compile or debug your code.

Patas

- If you need to have a patas account, you need to email linghelp@uw.edu right away to get an account.

- The directory for LING572: ~/dropbox/16-17/572/
  - hw1/, hw2/, ….: Assignments and solutions
  - misc/: Solution to exams and misc slides that are not on the course url.

- To run the job, use the cluster submission commands.

Summary of assignments

	Assignments (hw)	Reading assignments
Num	9	4
Distribution	Download from the course url	
Discussion	Allowed	
Submission	CollectIt	
Due date	11pm every Wed	11am on Tues or Thurs
Extension	1% penalty per hour	Disallowed
Estimate of hours	10-15 hours	2-3 hours
Solution files	On Patas	Discussed in class

Workload

- On average, students will spend around
  - 10-20 hours on each assignment
  - 3 hours on lecture time
  - 2 hours on GoPost
  - 2-3 hours on each reading assignment

  ➔ 15-25 hours per week; about 20 hrs/week

- You need to be realistic about how much time you have for 572. If you cannot spend that amount of time on 572, you should take 572 later when you can.

- If you often spend more than 25 hours per week on 572, please let me know. We can discuss what can be done to reduce time.

Programming assignments

- Try to reuse code from previous assignments

- Results:
  - No need to get exactly the same results: e.g., if the gold standard is 83.8, getting 83.1 is fine.

  ➔ Spend time on high-level ideas, not on debugging.

- Teamwork allowed for some assignments:
  - Discuss pseudo code together, but only one person has to type in the code and debug.

Extension and incomplete

- Extension and incomplete are given only under extremely unusual circumstances (e.g., health issues, family emergency).

- The following are NOT acceptable reasons for extension:
  - My code does not quite work.
  - I have a deadline at work.
  - I am going to be out of town for a few days.
  - …

Course Content

Textbook

- No textbook
- Readings are at the course website.
- Reference / Background:
  - Jurafsky and Martin, *Speech and Language Processing: An Introduction to NLP, CL, and Speech Recognition,* 2nd edition, 2008.
  - Manning and Schutze, *Foundations of Statistical NLP*

Types of ML problems

- Classification problem
- Estimation problem
- Clustering
- Discovery
- …

➔A learning method can be applied to one or more types of ML problems.

➔We will focus on the classification problem.

Course objectives

- Covering basic statistical methods that produce state-of-the-art results
- Focusing on classification and sequence labeling problems
- Some ML algorithms are complex. We will focus on basic ideas, not theoretical proofs.

Main units

- Basic classification algorithms (2 weeks)
  - kNN
  - Decision tree
  - Naïve Bayes
- Advanced classification algorithms (5 weeks)
  - MaxEnt
  - CRF
  - SVM
  - Introduction to neural network

Main units (cont)

- Other learning methods (1-2 weeks)
  - TBL
  - Introduction to semi-supervised learning (??)

- Misc topics (2 weeks)
  - Introduction
  - Two packages: Mallet and libSVM
  - Feature selection
  - Converting Multi-class to binary classification problem
  - Review and summary

Questions for each ML method

- Learning methods:
  - kNN and SVM
  - DT and TBL
  - NB and MaxEnt
  - Perceptron

- Modeling:
  - What is the model?
  - What kind of assumption is made by the model?
  - How many types of model parameters?
  - How many “internal” (or non-model) parameters?
  - …

Questions for each method (cont)

- Training: how to estimate parameters?
- Decoding: how to find the “best” solution?
- Weaknesses and strengths:
  - Is the algorithm
    - robust? (e.g., handling outliers)
    - scalable?
    - prone to overfitting?
    - efficient in training time? Test time?
  - How much data is needed?
    - Labeled data
    - Unlabeled data

Coming up

- Email me by 6pm tomorrow if
  - you have any question about the course, or
  - you have not taken 570

- Office hours: please fill out the doodle by 11pm today.
  http://doodle.com/poll/9zgthfikwu8ggx2y

- Hw1 is due at 11pm on 1/11.

Please go over self-study slides

- All are on the ling572 website.

- All have been covered in LING570
  - Probability Theory
  - Overview of Classification Task
  - Using Mallet
  - Patas and Condor

Final words

- It is better not to take the course if any of the following is true:
  - You do not meet the prerequisites.
  - You cannot attend class live most of the time.
  - You cannot spend 15-20 hours/week on the course.
  - You know most of the course material already.

- The deadline for dropping the course without a fee is this Friday.

- If you are unsure, please come to see me.