Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 14 = Review for Final Exam

Agenda:

1) Go over solutions for the last 2 homeworks
2) Discuss Final Exam
3) A few Final Exam sample questions

Homework Assignment:

Chapter 5 Homework Part 2 and Chapter 8 Homework is due Tuesday 8/14 at 9AM.

Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

SCPD students may use email or fax or mail.

The assignment is posted at

http://www.stats202.com/homework.html

Important: If using email, please submit only a single file (word or pdf) with your name and chapters in the file name. Also, include your name on the first page. Finally, please put your name and the homework # in the subject of the email.

Homework Solutions

- As of 9AM Tuesday, August 14, solutions to all homework assignments will be posted at

http://www.stats202.com/solutions.html

- Review these for the exam

- Note that even if you had a perfect score, you may still have missed some parts, so check your answers against these solutions carefully

Final Exam

I have obtained permission to have the final exam from 9 AM to 12 noon on Thursday 8/16 in the classroom (Terman 156)

I will assume the same people will take it off campus as with the midterm so please let me know if

1) You are SCPD and took the midterm on campus but need to take the final off campus

or

2) You are SCPD and took the midterm off campus but want to take the final on campus

Final Exam

The exam will cover all the material from the course, but 75% of the weight will be on new material

The exam is worth 200 points, which is 40% of your final grade

As you did for the midterm, bring a pocket calculator

You may bring one 8.5” by 11” sheet of paper (front and back) containing notes, just as we did for the midterm

If you oversleep or get caught in traffic, come in anyway even if you are very late

There will be some multiple choice questions, but most of the questions will require you to solve problems or explain concepts – see examples on next slides

Sample Final Question #1:

Which of the following describes bagging as discussed in class?

A) Bagging builds different classifiers by training on repeated samples (with replacement) from the data

B) Bagging combines simple base classifiers by upweighting data points which are classified incorrectly

C) Bagging usually gives zero training error, but rarely overfits which is very curious

D) All of these

Sample Final Question #2:

Using the ten observations below having two categorical attributes, construct the optimal 2-node decision tree according to the Gini index.

(the exam would have actual data but I did not include it here)

Sample Final Question #3:

The following R code is meant to compute the training error and test error for a classifier c(x,y). What is wrong with this code?

(the exam would have actual code with a major mistake but I did not include it here)

Sample Final Question #4:

Give a general explanation of how AdaBoost works.