Final "Project" Guidelines

Stats 231/CS229T: Statistical Learning Theory

May 29, 2017

1 Introduction

In this class, we have periodically mentioned the idea of a final project. While doing a final project with a serious research component is of course a great desiderata, we are aware that many of you are not actively doing research in the theoretical aspects of statistical learning, and that doing a project for a class in which one is not already actively doing research is not always the greatest learning experience.

With this in mind, for the final part of the class, we provide three options, each of which (we believe) requires non-trivial work, and each of which will be worth precisely the same amount of credit for the class. The three options, which we describe more carefully below, are as follows.

1. Do a research project related to the class.
2. Write a non-trivial homework problem, along with its solution.
3. Do a final homework, which we will release on or before May 30.

2 Final Project

The idea of the project is to contribue to the body of theoretical work in statistical learning, keeping in mind that we view statistical learning quite broadly: it is a combination of

(1) Statistics

(2) Optimization

(3) Computing

With that in mind, any project that touches on some convex combination of these three, while at least touching some type of plausible process that could provide data, is within the purview of the course.

As some broad guidelines, we present the following basic outline of what we expect. A common project type will involve an innovation in modeling (in the sense used in optimization and statistics), i.e., the way a practical problem is formulated as a statistics or optimization problem. In such a project, you develop a new approach to some practical problem, possibly in simplified form. Such a project will likely have the following components:

- Background and problem in general terms. You must clearly describe the engineering problem, giving the background, describing how the problem is solved now, and whats lacking or inadequate in how its done now. For example, current design practice might be ad hoc, ignore a number of constraints, etc. This description can be vague, as in "The goal is to estimate the original uncorrupted signal, without excessive sensitivity to noise." Dont mix in anything here that is related to how youre going to solve the problem.

- Your new formulation. Explain your formulation of the problem (or some part of it, or some simplified variation) as a learning problem. Be clear about what the variables and constraints are, and whether the problem is high-dimensional, based on generlization, convex, nonconvex, stochastic, infinite dimensional, etc. How accurate is your modeling (formulation)?
- Solution. To solve the problem you formulate you can use the techniques we have developed in class, gain insights via simulation, provide a number of plausible heuristic derivations, or whatever makes most sense given your problem and formulation. What performance guarantees can you give on your solutions to your problem? Did you develop new algorithms? Did you develop new concentration bounds?

Deliverables The final project submission should consist of a paper, written using the JMLR style of latex (see `http://jmlr.org/author-info.html#Links` and `http://www.jmlr.org/format/jmlr2e.sty`) of at least 4 pages, but whose length may be anything longer as needed. You should try to be as concise as possible, but feel free to use whatever space needed to convey the important ideas.

Due date: The due date is June 13, 23:59, with no extensions given because we must grade all work before graduation.

3 Writing a Homework Question

The second option for completing the course is to write a homework question. As you have likely noticed in the class, the homework questions tend to be somewhat long and—often—are non-trivial. Generally, we base the homework questions off of (part) of a recent or more classical paper in the statistics, machine learning, or optimization literature. With this in mind, the homework question you provide should essentially be a synthesis of some piece of work that you have read and understood.

The basic goal should be this: your question should be less than one page, but in that page it shold

(1) Define the problem in a way that is understandable to the average student in this class, including any extra notation or additional concepts that someone would need to solve it

(2) Motivate the problem

(3) Present a clear question (or set of questions) that student might answer; someone who solves the question should reasonably expect to understand the paper off of which it is based very well

You should also provide a solution to your question, which should be written at least as well as any of the solutions we provide, with clear references to any literature you use. Your solutions should read clearly—this is a pedagogical exercise—and help *you* understand the problem as well. For references, you may use any of the solutions we have provided in class.

Deliverables We will provide a style file for you to use, which we *expect* you to use, to write the problem. The style file is available on the course website at `http://web.stanford.edu/class/cs229t/solutions/exercises.sty` with an example solution set at `http://web.stanford.edu/class/cs229t/solutions/solutions-2.tex`. The deliverable for this project is a `.tex` file, along with a compiled `.pdf`, of your question and the solution to the question.

Due date: The due date is June 13, 23:59, with no extensions given because we must grade all work before graduation. You should email your submission to the course staff list.

4 Final Homework

This one is pretty self-explanatory. You do the last homework. It should be turned in either in person in the turn-in box or by email to the course list.

Deliverables