CAS MA751 -- Statistical Machine Learning

Spring 2022

Final Project Guidelines

(Due Monday, May 2)

The final project for this class should allow you to expand your understanding/knowledge of the topics in this course, either in the direction of learning new methodologies and approaches, or better understanding approaches we have studied along with and their implementations. The grade for the project will contribute 25% to your overall grade. In the project you and your team will be asked to select an advanced statistical method (or methods) of interest to you. You should research it in the literature, explore it thoroughly in a numerical fashion (this is a primary purpose of the project), and summarize what you have learned. This can be a method above and beyond what we have covered in detail (i.e., not covered in depth in lectures or problem sets, though possibly covered in the text). Alternatively, you may also take a topic we have learned/studied carefully, and advance your level of understanding/detail of this topic (from what was covered in the class/problems) and explore/implement it numerically.

I will encourage you to self-organize into teams of size two or three for the work on final projects. Each team should submit a project proposal to be modified/approved on the schedule mentioned below. The writeups of the projects will be in the format of jointly written technical reports/papers following the format guidelines of a typical journal, i.e., with an abstract, introduction, paper contents, conclusion, references, etc. There will be a section at the end (see below) specifying project authors' contributions.

Please note that the number of topics we have studied in detail in class or in the problems is much smaller than the number of topics covered in your text, and even smaller relative to the large number of statistical machine learning methodologies. Thus there are a large number of 'new' topics to choose from. In addition, as mentioned, a deeper analysis of a topic we have covered, together with thorough numerical study and implementation of this method, is also a good option.

The project proposal should contain both a theoretical and applied component, i.e. theory and a simulation of some type. There can be exceptions, but please check with me on those.

One reasonable idea for some proposals is to emulate what already exists in a paper that you have found (that is, to reproduce it partially or fully, or extend it if that is sensible). This typically is best done with 'fairly' recent papers in 'good' journals, i.e. ones with high impact factors. These would include journals like Nature, Science, Machine Learning, Neural Networks, Journal of Machine Learning Research, or Neurocomputing. They

would also include articles that occur in good conference proceedings like NIPS and ICML

Here is a brief outline of what is sought in your project.

1. If you wish to choose a topic we have not covered in detail, these include a number of reproducing kernel Hilbert space methods, wavelet methods, neural networks, unsupervised (e.g. clustering) methods, etc. To give you an idea of what we plan to cover beyond March, these topics include primarily text sections 8.1-8.3, 8.7, 8.8, 9.1, 9.2, 10.1-10.5, 10.7-10.9, 11.1-11.8, 12.1-12.4, 13.1-13.3, 15.1-15.3, 14.1, 14.3-14.5, 14.7, 14.10, 18.1-18.7. It may be a good idea to choose a topic related to your chosen field of research, as long as this expands rather than just reflects the current state of your work.

2. The end product of your work should be a final report (roughly 15-20 pages, including figures and/or tables). The structure of the report should be relatively formal (you can consult a journal style you choose as a guideline organizing your report, if needed). The report should include

(i) A one-paragraph abstract, providing an overview of what you have done.

(ii) An introductory section with background on the topic geared toward a general reader. In particular please do not assume that your reader (i.e., the instructor) is very familiar with the area.

(iii) A section describing the methodology you have chosen to work with.

(iv) A section presenting and discussing the numerical examples you will use in exploring your methods. These could include a well-designed simulation study as well as an application to real data, with any appropriate comparisons to reasonable alternatives that we may have covered in class.

(v) Your conclusions.

(vi) An appendix containing annotated and commented code, figures, and tables.

(vii) A section on Authors' contributions. Here you indicate exactly which parts of the project were done by which authors. A typical section of this type can be found in papers in various journals. Here is a typical such section, just for style and formatting purposes (from a larger multi-author paper):

Authors' contributions

IS and SLT conceptualized the research project. DJ formulated and implemented the algorithm, generated initial simulated data and performed the analysis of simulated and real data. DH designed the extended simulation study. DH and DS performed the extended simulation study on ENVirT and other existing methods. DS developed the

graphical user interface of ENVirT and manages the software distribution. YS and SKH contributed in formulating the optimization algorithm. CYY, IS, BC and SLT processed and contributed in the analysis of real data. IS and DJ prepared the initial draft of the manuscript. DH, DS, SLT and SKH contributed in preparing the final version of the manuscript. All authors read and approved the final manuscript.

Submission: Please turn in your project electronically (by email), with a separate file including supplementary materials, such as any code you have used.

Plagiarism: Please remember that plagiarism of any type (including use of writing which in any form has appeared elsewhere) is punishable by sanctions including possible expulsion from the University. See the GRS Academic Conduct Code, posted at https://www.bu.edu/academics/policies/academic-conduct-code/.

Project Proposal:

To begin the project, I would like you/your team to send me a short project proposal by Wednesday, March 16. The proposal should be one or two paragraphs, describing your topic. This should also cite one or two references from which you will be working. I will plan to return these to you within a week, working with you to modify any details if needed. In the end you will have approximately 3 weeks to complete your project.

Machine Learning Resources:

There are a number of excellent machine learning resources in R, Matlab, and Java. Those below are only a few of them:

General:

Tensor Flow is a Google package that tries to universalize processes like machine learning as operations on large matrices and tensors (multidimensional matrices). It attempts to be a universal conduit to the different types of operations that occur in processes like machine learning.
https://www.tensorflow.org

The IBM-SPSS Modeler software allows the construction of graphical user interfaces where datasets are represented as graphical tokens, and operations on these datasets (e.g. balancing, imputation, machine learning operations) are represented as icons into which the tokens can be transported.

H2O is a package of open source software for big-data analytics to an extent coordinated with some of the statistical learning topics in your textbook, and for this reason may be worth investigating. See https://en.wikipedia.org/wiki/H2O_(software) .

In R:

General information on R software for ML:
http://cran.r-project.org/web/views/MachineLearning.html
Rweka:
http://cran.r-project.org/web/packages/RWeka/index.html
Bioconductor:
This package is a good one for bioinformatics/computational biology-type data:
Package:
http://www.bioconductor.org/
Machine learning interfaces:
http://www.bioconductor.org/packages/release/bioc/html/MLInterfaces.html
Background:
http://www.bioconductor.org/help/course-materials/2011/CSAMA/Monday/Afternoon%20Labs/MLprac2_2.pdf

In Matlab:
Matlab has a fairly comprehensive machine learning toolkit within its Statistics Toolbox
http://www.mathworks.com/products/statistics/features.html#machine-learning

In Java:
Weka is a comprehensive and very widely-used all-purpose machine learning package; Spider uses it as its back-end for some algorithms

http://www.cs.waikato.ac.nz/ml/weka/

In Python:
Scikit is a popular Python-based toolkit for machine learning and data mining
http://scikit-learn.org/stable/

Theano is a matrix/array based toolkit largely used to study deep learning.
http://deeplearning.net/software/theano/

Data Sources:
Please remember that simulated data are always a good option on which to study the performance of any algorithm, since you can then control the 'ground truth' on which your data are based, and identify your algorithm?s ability to discover it.

As for real data sets, there are plenty of sources on which to try methods you are studying.

For example, the UCI Machine Learning Repository:
(http://archive.ics.uci.edu/ml/index.php) is one.

Another is GEO, a database of gene expression datasets:
(http://www.ncbi.nlm.nih.gov/geo/).

The KDD (Knowledge Discovery and Data-mining) cup contest data:
(http://www.kdd.org/).

Amazon Web Services (AWS) datasets:
(http://aws.amazon.com/publicdatasets/)

Meta-databases: compilations of database locations:
(http://www.kdnuggets.com/datasets/)
(https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research/)
(https://github.com/caesar0301/awesome-public-datasets/)

The Hastie text data web site:
(https://web.stanford.edu/~hastie/ElemStatLearn/data.html)