Seminar in Computational Biomedicine

Place & Time: Smith Hall, 201, Tuesdays & Thursdays, 12:30AM – 1:45PM

Instructor: Hagit Shatkay, shatkay@udel.edu, Smith Hall 416
WWW: http://www.eecis.udel.edu/~shatkay/Course/CISC844S2020
Office hours: Tuesdays, 2-3pm and by appointment
Prerequisites: Basic course-level background in probability and/or statistics. Background in CS.
Other background useful – but note required: AI, Machine Learning, Biomed-related knowledge

General Information

Computational Biomedicine is a seminar-based course, in which we will survey mostly well-established – but also some newer – approaches in machine learning, algorithms and computational theory, along with their recent applications in biomedical informatics. The past 15 years have seen tremendous growth in the amount of biomedical data; much recent research in biomedical computing/informatics is focused on its interpretation. The ultimate goal is to understand and predict normal function of organisms and, more importantly, the mechanisms underlying disease. Given the wealth of data – interpretation cannot be done manually. It requires advanced algorithms and computational tools, mimicking some aspects of the manual interpretation process, but expediting it several folds.

Machine Learning is concerned with the formulation and automatic acquisition of models from data, as well as with using such models for automatic inference and prediction. Both modeling and inference/prediction are inherent to biomedical data-analysis problems; as such machine learning methods are well-suited for this domain and are indeed applied to a wide variety of biological and medical problems.

The course will begin with a broad survey of machine learning methods and algorithms, and will also provide a high-level introduction to fundamental biological concepts. It will then proceed as a sequence of themes, where each theme starts with a seminar discussing the relevant algorithms and machine learning methods, followed by seminars presenting their application central biological or medical problems. We shall cover both supervised and unsupervised machine learning methods, ranging from probabilistic clustering to kernel methods and various types of network and of models (some probabilistic, some not). In terms of applications, these include: malignancy identification in clinical data, characterizing types of anomalies in heartbeats, modeling protein function/locations or mining electronic health records. We shall also discuss probabilistic language models and generative models, which play a role in text, image and data integration. Several guest lectures will showcase concrete biological and clinical contexts in which computational methods are employed.

Course Mechanics and Grading

Two main components: *seminars* (70% of the total grade) and a *project* (30%).

Seminars

Basics: *Presentation & leading discussion: 50% (of total grade).*
Reading (including at least ***12 hand-in notes***), *presence in classes,* participation and filling in evaluation forms: 20% (of total grade).

- Both the speakers and the attendees are responsible for the seminar's success.
- The speakers are in charge of presenting and leading a discussion.
- The rest of the class is responsible for reading the papers and contributing to the discussion.
- ***The seminar presentation is of the TOPIC*** discussed in the paper(s) as a whole (see guidelines about what this entails). The presentation should cover the ***topic*** that is covered in the paper(s), in a way that is clear and self-contained (including: ***Understanding and Explaining the Background, the Motivation For and the Impact Of the work***, ***and all fundamental concepts***).
Speakers should plan on 50-55 minutes talk and 20-25 minutes for discussing the papers pros/cons.
At the beginning of each class all students will share what they thought about the paper(s), what they got out of it, and bring up 1-2 points for discussion.
The presentation follows + discussion at the end.
The discussion will be organized as mock-decision-session, where the points brought up for discussion are covered and the value of the paper is considered; We as a class decide if to accept/reject the paper. (See guidelines/suggestions in handout #4).
- ***Working in pairs / reading partners:*** *Each person in class should have a reading-partner to consult with about his/her presentation material*. The roles will reverse when it's the reading-partner's turn to give a seminar. The reading partner is there to discuss problematic points in the paper with – ***well ahead*** of the seminar, and help with suggestions about what to present. The speaker may also want to consult with the reading partner about the contents/clarity of the presentation.
- Signing-up for seminar presentations will happen during the first two weeks of classes.

- ***Reading – Hand-in notes for AT LEAST 12 seminars*** *of your choice*: Often each seminar has 2 papers in the ***required*** reading list (and some helpful/background papers that are not required reading).
For ***Each of the Required-Reading Papers***, make a bulleted list consisting of :
2 (or more) points for discussion; ***2 positive aspects; and 2 negative aspects*** that you find while reading it. The notes should be handed in at the beginning of each seminar.

    Positive aspects are points that you find significant or interesting (*e.g.* an interesting application of algorithm X to problem Y – and what makes it interesting, a clear explanation of a particularly tricky concept, an effective utilization of a very small training set etc.).
    Negative aspects are those points that you found problematic or hard (*e.g.* a specific concept that was not well-explained, insufficient testing, lack of improvement over earlier methods, etc.).
    The comments *must be specific*. (That is, a comment of the form "A really cool method! I liked it!" is ***not*** an acceptable note…)

    Each aspect that you write about should take NO more than 2 lines of writing, totaling ~20 lines per seminar. ***Reading notes should NEVER Exceed a single page. (See more on the next page)***

    Hand-in notes about a topic are due at the beginning of the class discussing it.
    To receive full credit, at least Twelve (12) hand-in notes should be submitted during the semester.
    (Note that each hand-in note typically covers 2 papers.)

Examples of Good comments (*Do's*):

<u>Positive</u>

"I liked the idea of methodically integrating seemingly unrelated cues to obtain a reliable classification of proteins according to their function".

"I found Figures 2a and 2b very useful as they showed how increasing the data dimensionality, K, increases the accuracy to some extent, but may also reduce it. I did not understand this idea at first, and the image helped a lot."

<u>Negative</u>

"In Section 3.5 the authors say that their approach effectively uses homology to improve localization prediction. However, they never actually demonstrate that their method improve upon any previous results.".

"Figure 5 is supposed to show accuracy as the function of the number of iterations. However, the X and the Y axes are not marked, and the plot seems to be sloping the wrong way. I don't see how the figure demonstrates the point."

Examples of Bad comments (*Don'ts*):

<u>Positive</u>

"I liked the idea presented in the paper". (Too general – does not show reading or understanding).
"Figures 2a and 2b very useful. I did not understand this idea at first, and the image helped a lot."

<u>Negative</u>

"The paper was boring and too long. Went right over my head." (One can say this about any paper, regardless of content…)
"I think the basic idea presented in the paper is bad." (Gives no information about what the idea was and why it is bad…)

<u>Participation</u>

* Reading the material even when you don't present it
* Bringing in discussion points (see reading notes)
*Attendance: ON-TIME arrival (Coming in late is disruptive to the presenters and to the whole class!)
* Filling in evaluation forms at the end of the course
are all necessary, but are not sufficient…

The course (and your grade ☺) relies on your *presence, active participation, debate, discussion, asking questions and offering answers.*

Course Mechanics and Grading *(cont.)*

Projects

Basics:
Proposal: 1 page write-up + additional pages for references if needed, 5%
Project presentation: 10%
Project report: 15%

Detail:
- Each participant will choose a topic ***relevant to the course material (computational aspects of biological or medical data),*** and carry out a small research project. It can be an implementation of an existing method, a survey of a specific domain, or the development of a new idea. Your seminar presentation topic is a good starting point. If you need help choosing a project topic, please come see me well in advance.
  Projects may be done in pairs.
  A *1 page* proposal should be submitted on April 21 by email. The write up should describe the problem or the topic you will address along with the work plan for the project.
  In addition to the 1-page write up you can submit separate references pages. You will receive feedback on the proposal within a few days past the submission.
- Final project presentations will be on ***May 18 (Most Likely)***. Each presentation will take **15-20 minutes,** and will shortly review the problem/challenge, the methods applied, and the results. Please let me know ahead of time if you plan a demonstration of a system, or have any special AV needs.
- Project reports of about 10 pages (no more than 10 + extra pages for references if needed), should be formatted as research papers, introducing the problem, providing background information and references to the relevant literature, and describing methods, results, conclusions, future directions and bibliography. Proper use of cross-references into the bibliography is expected.
- Project reports should be submitted by email to *shatkay@udel.edu* as PDF or word files, by ***May 21*** at *midnight*.

  If the project includes an implementation, the *clearly-documented* code, as well as a user's manual with installation and execution details should be provided in addition to the project report (the written project report may be shorter in this case – please talk to me well ahead of time about it).

  All presentations and reports will be graded based on the quality of the scientific contents, as well as on adhering to good writing style and clear presentation in your writing.

Reading Material

Most of the course material are papers (see Handout #3), available online from the course home-page. Older papers and some other selected reading will be provided as handouts.
 NOTE: Due to copyright issues some of the papers may be password protected or otherwise limited in accessibility. To avoid problems accessing the papers, please do not wait for the last minute and verify you have the paper several days before the seminar.

The following books, which contain relevant background and extended discussions, are placed on reserve at the Morris library:

- *Machine Learning*. T.M. Mitchell. 1997. [Q325.5 M58]. (Chapters 3-6).
- *Pattern Classification and Scene Analysis*. R.O. Duda and P.E. Hart. 1973.[Q327 .D83]. (Chapter 6 in particular).
- *Artificial Intelligence: A Modern Approach*. S. Russell and P. Norvig.
  Note: Selected relevant chapters are distributed as handouts
- *Bioinformatics: The Machine Learning Approach*. P. Baldi and S. Brunak. [QH 506 .B35] 2nd Ed. 2001.
- *Molecular Biology of the Cell*. B. Alberts et al.

Links to some useful web sites are also placed on the course home page.

Additional relevant journals:
Bioinformatics
Journal of Computational Biology (JCB)
BMC Bioinformatics
Journal of the American Medical Informatics Association (JAMIA)
Journal of Bioinformatics and Computational Biology
IEEE/ACM Transactions on Computational Biology and Bioinformatics
PLoS Computational Biology, PLoS One
Science (and its various specific journals)
Nature (and its various specific journals)

Additional relevant conference proceedings include:
The International Conference on Intelligent Systems for Molecular Biology (ISMB)
Research in Computational Biology (RECOMB)
Pacific Symposium on Biocomputing (PSB)
The European Conference on Computational Biology (ECCB)
IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
The Workshop on Algorithms in Bioinformatics (WABI)
Symposium of the American Medical Informatics Association (AMIA)
AMIA Summit on Translational Bioinformatics